Emotional facial expression transfer from a single image via generative adversarial nets

Article *in* Computer Animation and Virtual Worlds · May 2018

DOI: 10.1002/cav.1819

CITATIONS
12

READS
746

6 authors, including:

Fengchun Qiao
Chinese Academy of Sciences
3 PUBLICATIONS **36** CITATIONS
SEE PROFILE

Zirui Jiao
Chinese Academy of Sciences
3 PUBLICATIONS **36** CITATIONS
SEE PROFILE

Hui Chen
Chinese Academy of Sciences
50 PUBLICATIONS **392** CITATIONS
SEE PROFILE

Received: 16 March 2018 | Accepted: 21 March 2018

DOI: 10.1002/cav.1819

SPECIAL ISSUE PAPER

WILEY

Emotional facial expression transfer from a single image via generative adversarial nets

Fengchun Qiao[1,2,*] | Naiming Yao[1,2,*] | Zirui Jiao[1,2,*] | Zhihao Li[1] | Hui Chen[1,2] | Hongan Wang[1,2,3]

[1]Beijing Key Laboratory of Human–Computer Interaction, Institute of Software, Chinese Academy of Sciences, Beijing 100190, China

[2]University of Chinese Academy of Sciences, Beijing 100049, China

[3]State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences, Beijing 100190, China

Correspondence
Hui Chen, Beijing Key Laboratory of Human–Computer Interaction, Institute of Software, Chinese Academy of Sciences, Beijing 100190, China; or University of Chinese Academy of Sciences, Beijing 100049, China.
Email: chenhui@iscas.ac.cn

Funding information
National Key R&D Program of China, Grant/Award Number: 2017YFB1002805; National Natural Science Foundation of China, Grant/Award Number: 61661146002; Key Research Program of Frontier Sciences, CAS, Grant/Award Number: QYZDY-SSW-JSC041

Abstract

Facial expression transfer from a single image is a challenging task and has drawn sustained attention in the fields of computer vision and computer graphics. Recently, generative adversarial nets (GANs) have provided a new approach to facial expression transfer from a single image toward target facial expressions. However, it is still difficult to obtain a sequence of smoothly changed facial expressions. We present a novel GAN-based method for generating emotional facial expression animations given a single image and several facial landmarks for the in-between stages. In particular, landmarks of other subjects are incorporated into a GAN model to control the generated facial expression from a latent space. With the trained model, high-quality face images and a smoothly changed facial expression sequence can be effectively obtained, which are showed qualitatively and quantitatively in our experiments on the Multi-PIE and CK+ data sets.

KEYWORDS

dynamic expression sequence, facial expression transfer, generative adversarial nets

1 | INTRODUCTION

Facial expression transfer is a research field for mapping and generating desired images of specified subject and facial expression. Many methods achieved significant results for high-resolution images and are applied to a wide range of applications, such as facial animation, facial editing, and facial expression recognition. However, in the condition of single-image input, there still remain some problems. It is not difficult for a model to synthesize arbitrary facial expressions from a single image, but it is difficult to synthesize a smoothly changed image sequence. Furthermore, some methods of facial expression transfer rely on expression information of an input image, for example, expression label, landmarks, or neutral face. These two bottlenecks greatly restrict the diversity of synthetic results and the efficiency of facial expression transfer.

*These authors contributed equally to this work.

The methods of facial expression transfer from a single image can be roughly divided into two categories: traditional graphic-based methods and emerging generative methods. In the first category, target images are synthesized by a series of designed warps that imitate the changing process of human facial expressions or by a global statistical model. Both geometry-based features and appearance-based features play an irreplaceable role in facial expression recognition.[1] Geometry information can be represented as facial landmarks and extracted by some graphical algorithms, for example, active appearance models[2] and the dlib regression trees algorithm.[3] Landmarks of the same subject are used as an auxiliary for facial expression representation.

The second category of methods is based on data-driven generative models. However, traditional generative models have some bottlenecks: blurry generated images, incapability of fine control of facial expression, low-resolution outputs, etc. In recent years, generative adversarial nets (GANs)[4] have drawn wide attention and achieved considerable progress in the field of computer vision. Because of the high-quality generated images, GANs have been applied to facial expression transfer with some positive results obtained. The conditional difference adversarial autoencoder[5] preserves identity information when generating facial expressions for unseen subjects. The differential generative adversarial network[6] can synthesize smoothly changed images and control the specific parts of the face with a small amount of training data set. However, these GAN-based methods are still limited to discrete facial expression synthesis, that is, that they cannot generate a face sequence showing a smooth transition from an emotion to another. Recently, Song et al.[7] utilized landmarks and proposed the geometry-guided GAN (G2GAN) to generate smooth image sequences of facial expressions. However, this model needs a neutral face, which is human annotated or generated by the model, as an intermediate of the expression transfer. This dependence on neutral face intermediates is unnecessary and reduces the performance of facial expression transition.

In this paper, we propose a new GAN-based approach to generate high-quality facial expression image sequences from a single image with the guidance of facial geometry information. We combine identity and appearance information from an input image and target emotion's landmarks of other subjects in feature latent space to guide the facial expression transfer. These generated expressions can be smoothly changed because of the continuity of landmarks in Euclidean space. Our approach has been evaluated on the Multi-PIE[8] data set and the CK+[9] data set without any expression labels.

2 | RELATED WORK

2.1 | Facial expression transfer

In traditional graphics-based methods, many research efforts have been devoted to facial expression transfer from a single image. Some methods directly warp the input face to approach the target expression, by either 2D warps[10,11] or 3D warps.[12–14] Some construct the parametric global model, for example, the method of Mohammed et al.[15] learns a probabilistic model where existing and generated images obey the structural constraints. In order to carry more facial details, Averbuch-Elor et al.[16] added fine-scale dynamic details associated with facial expressions such as wrinkles and inner mouths. Although computer graphics methods have achieved some positive results in high-resolution and one-to-many image synthesis,[16] they are still limited to the sophisticated design and expensive computation.

Unlike computer graphics methods regarded as theory-driven approaches, generative modeling methods based on deep learning are data driven. Facial expression transfer is obtained through training an appropriate generative model according to the data set, including all information about identity, expression, viewing angle, etc. Generative modeling methods reduce the complicated design of the connection between facial textures and emotion states and encode intuitionistic facial features into parameters of data distribution. Some pioneering generative models have been applied to facial expression synthesis in prior work, for example, deep belief net (DBN)[17] and higher-order Boltzmann machine.[18] However, these models have some bottlenecks: blurry generated images, incapability of fine control of facial expression, low-resolution outputs, etc.

Recently, GAN has shown its value in computer vision and has been applied to facial expression transfer. Since the original GAN cannot generate facial images with a specific facial expression and person, some methods conditioned on expression categories have been proposed. Shu et al.[19] trained a generative adversarial network to learn a disentangled representation of intrinsic face properties and utilized the latent representations to manipulate facial appearance. In the work of Zhou and Shi,[20] a conditional difference adversarial autoencoder is proposed, aiming at synthesizing facial images of unseen subjects, with the guidance of action unit labels. ExprGAN[21] is designed for photo-realistic facial expression editing with controllable expression intensity. However, it still relies on facial expression labels and cannot generate smoothly changed face images. The most related work to our approach is G2GAN.[7] G2GAN uses geometry information based on dual adversarial networks to express face changes and synthesizes facial images. Through manipulating the landmarks,

smoothly changed images can also be generated. However, this method demands a neutral face of the target person as the intermediate of facial expression transfer. Although the expression removal network could generate a neutral expression of a specific person, this procedure brings additional artifacts and degrades the performance of expression transition.

In order to eliminate the dependence on expression labels and landmarks of an input image and on a neutral intermediate during transfer, we propose an approach guided by landmarks and manage to control facial expressions through continuous variables in landmark space. The construction of latent space relies on GAN models.

2.2 | Generative adversarial nets

GANs[4] are an emerging generative method to learn a generator approaching the data distribution via a zero-sum game. GANs consist of two parts: a generator (G) and a discriminator (D), which is trained to distinguish the generated data samples from real data. The value function of the game that D and G play is formulated as a min-max problem:

$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}\left[\log D(\boldsymbol{x})\right] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}\left[\log\left(1 - D\left(G(\boldsymbol{z})\right)\right)\right], \tag{1}$$

where $\boldsymbol{x}$ denotes the sample vector from a real data set with the distribution of p_{data}, and $\boldsymbol{z}$ denotes the noise vector sampled from a prior distribution $p_{\boldsymbol{z}}$. The min-max value function means the sum of two expectations that imply the game between D and G.

GANs take the advantages of sharp and high-quality generated images and are applied to facial expression transfer. However, the original GANs are difficult to train and prone to the mode collapse problem, that is, low diversity of generated data. Some recent works are devoted to resolving the problems. WGAN[22] leverages the Wasserstein distance to modify the value function in theory and makes progress toward stable training of GANs. However, WGAN still generates bad samples and fails to converge sometimes as a result of critic weight clipping. Then, WGAN with gradient penalty (WGAN-GP)[23] improves the algorithm and performs better than standard WGAN.

To integrate emotion semantics and geometry information, VAE-GAN[24] uses a variational autoencoder (VAE)[25] to provide GANs for controllable latent space. The generator is modified into a VAE model, consisting of an encoding subnetwork and a decoding subnetwork. The encoding subnetwork transforms input data into variables, which guides the generating of the whole GAN, in latent space. If images and landmarks are input into the model, expression features are extracted in latent space and can be controlled by inputs.

In our model, we modify VAE-GAN for using the combination of appearance features and geometry features to construct latent space and employ the methods of WGAN-GP for robust training.

3 | THE PROPOSED APPROACH

In this section, our new framework is presented for facial expression transfer based on GANs with geometry information. Emotional landmarks of the same input image or intermediate neutral face are no longer required.

3.1 | Overview

The problem of facial expression transfer is described mathematically as follows. Given a source facial image $\boldsymbol{I}_i^u$ and a target facial expression $\boldsymbol{g}_i^v$ represented by facial landmarks, our model aims to generate a new identity-preserving facial image $\mathcal{G}(\boldsymbol{I}_i^u, \mathcal{L}_{\text{enc}}(\boldsymbol{g}_i^v))$ conditioned on the target facial expression. In our method, $\boldsymbol{I}$ and $\boldsymbol{g}$ denote the gray or RGB matrix of an image and the vector concatenated from the xy-coordinates of facial landmarks, respectively. $\boldsymbol{I}_i^j$ or $\boldsymbol{g}_i^j$ means the image matrix or the landmark vector of any subject i and arbitrary facial expression j. $\mathcal{L}_{\text{enc}}$ and $\mathcal{L}_{\text{dec}}$ denote the encoding and decoding networks of landmark vectors; $\mathcal{G}_{\text{enc}}$ and $\mathcal{G}_{\text{dec}}$ denote the encoding and decoding networks of image vectors. In GAN, $\hat{\boldsymbol{I}}$ and $\hat{\boldsymbol{g}}$ are, respectively, the matrix and the vector sampled from the corresponding network.

With mathematic definitions above, we illustrate the framework in Figure 1, consisting of three components: a facial geometry embedding network $\mathcal{L}$, an image generator network $\mathcal{G}$, and an image discriminator network $\mathcal{D}$. $\mathcal{G}$ has a similar architecture with U-Net,[26] including batch normalization (BN)[27] and ReLU.[28] Different from the original U-Net, we add one skip connection between the output of the first convolution layer and the output of the penultimate deconvolution layer, which enables $\mathcal{G}$ to reuse low-level facial features related with identity information.[20] For discriminator network $\mathcal{D}$, we adopt a similar setting as in the work of Shrivastava et al.,[29] including layer normalization[30] and Leaky ReLU,[31] and

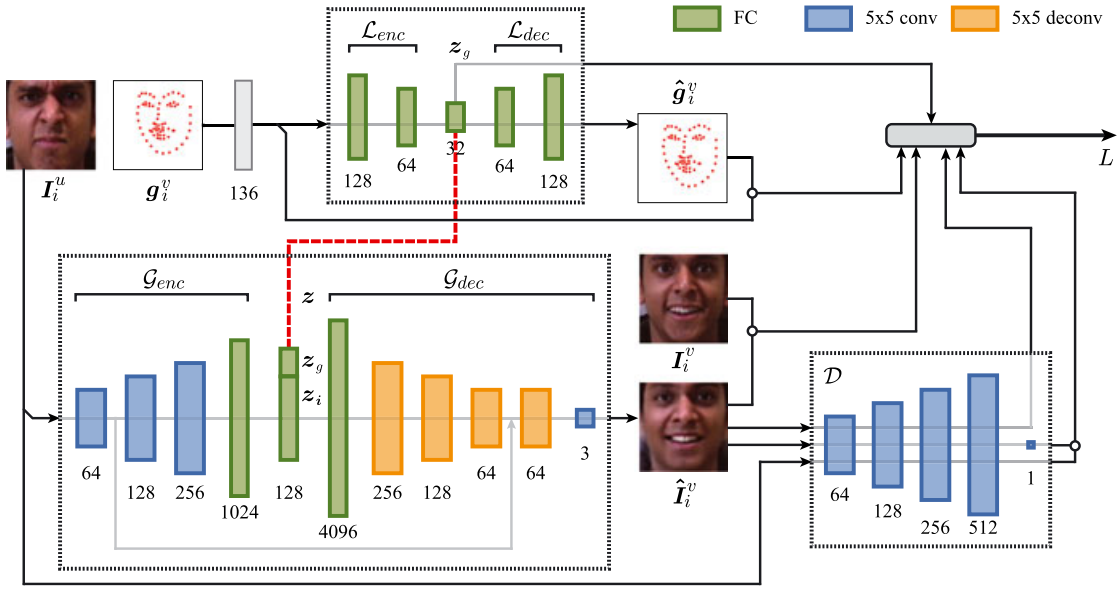

FIGURE 1 Overview of the proposed framework. In $\mathcal{G}$ and $\mathcal{L}$, input facial image $\boldsymbol{I}_i^u$ and input landmarks $\boldsymbol{g}_i^v$ are encoded by $\mathcal{G}_{\text{enc}}$ and $\mathcal{L}_{\text{enc}}$ into $\boldsymbol{z}_i$ and $\boldsymbol{z}_g$, respectively. Then, $\boldsymbol{z}_i$ and $\boldsymbol{z}_g$ are concatenated into a single vector $\boldsymbol{z}$ for $\mathcal{G}_{\text{dec}}$

the output is a 2×2 feature map. Embedding network $\mathcal{L}$ consists of five fully connected layers followed by BN and ReLU. Besides, we use $L_{(\cdot)}$ to represent the loss functions during optimization, and further definition will be showed in the following subsections.

3.2 | Geometry guidance

Facial landmarks provide the geometry information and facial expression information. Moreover, landmarks of the same emotion exhibit similar characteristics, and thus, the category information could be explored for constructing a semantic manifold of facial landmarks. First, we arrange the *xy*-coordinates of facial landmarks into a one-dimensional vector as input for the embedding network $\mathcal{L}$. Then, the concatenated vector $\boldsymbol{z}$ is the input of the main generator and guides the whole process of synthesis. In other words, given an image of arbitrary facial expression and target subject and landmarks of target facial expression, a trained $\mathcal{G}_{\text{dec}}$ is able to synthesize the desired image.

In order to enhance the robustness of our model for an unseen subject in the test set, we introduce contrastive learning. Landmark pairs $(\boldsymbol{g}_{(\cdot)}^b, \boldsymbol{g}_{(\cdot)}^c)$ of arbitrary subjects are prepared for training, in which $\boldsymbol{g}_{(\cdot)}^b$ corresponds to the target expression and $\boldsymbol{g}_{(\cdot)}^c$ corresponds to the reference expression. Our goal is to measure the similarity between $\mathcal{L}_{\text{enc}}(\boldsymbol{g}_{(\cdot)}^b)$ and $\mathcal{L}_{\text{enc}}(\boldsymbol{g}_{(\cdot)}^c)$ according to their labels of facial expressions. Thus, a contrastive loss L_{contr} is defined as follows:

$$L_{\text{contr}} = \min_{\mathcal{L}_{\text{enc}}} \left[\frac{\alpha}{2} \max\left(0, m - \left\| \mathcal{L}_{\text{enc}}\left(\boldsymbol{g}_{(\cdot)}^b\right) - \mathcal{L}_{\text{enc}}\left(\boldsymbol{g}_{(\cdot)}^c\right) \right\|_2^2\right) + \frac{1-\alpha}{2} \left\| \mathcal{L}_{\text{enc}}\left(\boldsymbol{g}_{(\cdot)}^b\right) - \mathcal{L}_{\text{enc}}\left(\boldsymbol{g}_{(\cdot)}^c\right) \right\|_2^2 \right], \tag{2}$$

where $\alpha = 0$, if the facial expression labels are the same, and $\alpha = 1$ otherwise. Moreover, m is a margin that is always greater than 0. We treat the process of optimizing contrastive loss as contrastive learning, since the model is constrained to comparing the vectors in latent space mapped from landmarks.

3.3 | Training multitask GANs

After the geometry information being integrated, the following task is to train the GAN model. The min-max game between image generator $\mathcal{G}$ and image discriminator $\mathcal{D}$ follows WGAN-GP,[23] and the adversarial loss for $\mathcal{G}$ and $\mathcal{D}$ is formulated as follows:

$$L_{\text{adver}} = \mathbb{E}_{\hat{\boldsymbol{I}} \sim p_{\hat{I}}}\left[\mathcal{D}\left(\hat{\boldsymbol{I}}_i^v\right)\right] - \mathbb{E}_{\boldsymbol{I} \sim p_I}\left[\mathcal{D}\left(\boldsymbol{I}_i^u\right)\right] + \lambda_{\text{gp}} \mathbb{E}_{\tilde{\boldsymbol{I}} \sim p_{\tilde{I}}}\left[\left(\left\|\nabla_{\tilde{\boldsymbol{I}}} \mathcal{D}(\tilde{\boldsymbol{I}})\right\|_2 - 1\right)^2\right], \tag{3}$$

where $p_{\boldsymbol{I}}$ and $p_{\hat{\boldsymbol{I}}}$ indicate the data distribution of real facial images and the generator distribution, $p_{\tilde{\boldsymbol{I}}}$ is defined to sample uniformly along straight lines between pairs of points sampled from $p_{\boldsymbol{I}}$ and $p_{\hat{\boldsymbol{I}}}$, and $\hat{\boldsymbol{I}}_i^v$ stands for $\mathcal{G}(\boldsymbol{I}_i^u, \mathcal{L}_{\text{enc}}(\boldsymbol{g}_i^v))$. Adversarial learning provides guidance for our model to learn the data distribution and generate target images.

Reconstruction loss is also introduced as a regularization to measure the reconstruction errors of faces and facial landmarks. For generated facial images $\hat{\boldsymbol{I}}_i^v$, we employ ℓ_1 distance to compute a pixel-to-pixel difference between the generated image and the target real image. For facial landmarks, we employ ℓ_2 distance to compute the difference between reconstructed landmarks and input landmarks. The reconstruction loss is formulated as follows:

$$L_{\text{recon}} = \min_{\mathcal{G},\mathcal{L}_{\text{enc}}} \left\| \boldsymbol{I}_i^v - \hat{\boldsymbol{I}}_i^v \right\|_1 + \min_{\mathcal{L}} \left\| \boldsymbol{g}_i^v - \hat{\boldsymbol{g}}_i^v \right\|_2^2. \tag{4}$$

So far, the full objective L is computed by combining Equations (2), (3), and (4), as follows:

$$L = \lambda_1 L_{\text{contr}} + \lambda_2 L_{\text{adver}} + \lambda_3 L_{\text{recon}}, \tag{5}$$

where λ_1, λ_2, and λ_3 are the weights of the losses.

4 | EXPERIMENTAL RESULTS

To evaluate our model, we conduct a series of experiments on two popular data sets: Multi-PIE and CK+.

4.1 | Data sets and settings

In Multi-PIE, we select all 337 subjects and a subset of 150,244 images with 3 poses (0°, ±15°), 20 illumination conditions, and all 6 expressions (neutral, smile, surprise, squint, disgust, and scream). Input images and target images are under the same conditions of pose and illumination. CK+ consists of 327 image sequences of 118 subjects with seven prototypical emotion categories, namely, anger, contempt, disgust, fear, happiness, sadness, and surprise. Each sequence starts with a neutral emotion and ends with a peak of the emotion. Following the widely adopted protocol introduced in the work of Jung et al.,[32] the subjects are grouped into 10 subsets by ID in Multi-PIE and CK+, of which nine subsets are used for training while the remaining subset is used for test.

In our experiment, facial landmarks are detected by *dlib** including landmark points of two eyebrows, two eyes, the nose, the lips, and the jaw. The face region of each image is detected and aligned based on inner eyes and the bottom lip through 2D affine transform. Then, facial images are cropped and resized into 64×64. Image values and the xy-coordinates of facial landmarks are both normalized into $[-1, 1]$. For the training triplets $(\boldsymbol{I}_i^u, \boldsymbol{I}_i^v, \boldsymbol{g}_i^v)$, all pairs of facial expressions per identity are used, and the reference landmarks are generated by sampling at random.

The specific parameter settings are as follows. The margin m in L_{contr} is set to 5. The coefficients λ_{gp} in L_{adver} are equal to 10. Moreover, λ_1, λ_2, and λ_3 in loss L are empirically set to 1, 10^{-3}, and 10^{-4}, respectively. All the networks $\mathcal{L}$, $\mathcal{G}$, and $\mathcal{D}$ are trained jointly in a multitask way by using the Adam optimizer[33] with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The batch size is 64, and the initial learning rate is set to 3×10^{-4}. The models are implemented using TensorFlow.[34]

4.2 | Static facial expression transfer

In order to evaluate whether the synthetic faces are generated conditioned on target facial expression, we qualitatively and quantitatively compare the generated faces with the ground truth. The qualitative results are shown in Figure 2, in which the generated faces are identity preserving and are similar to the target expressions, indicating the effectiveness of the guiding landmarks.

For quantitative measurement, we use two evaluation metrics including structural similarity index measure (SSIM)[35] and peak-signal-to-noise ratio (PSNR). The SSIM and PSNR of our approach are 0.687 and 26.731, respectively, on Multi-PIE, whereas they are 0.769 and 27.665, respectively, on CK+. G2GAN[7] is the most similar work to ours as far as we know. In their work, the quantitative results of Multi-PIE are not given, whereas SSIM and PSNR on CK+ are 0.767 and 24.420, respectively. Even though our proposed method is free from the neutral emotion of the target person, it consistently outperforms G2GAN across both measures.

*http://dlib.net/

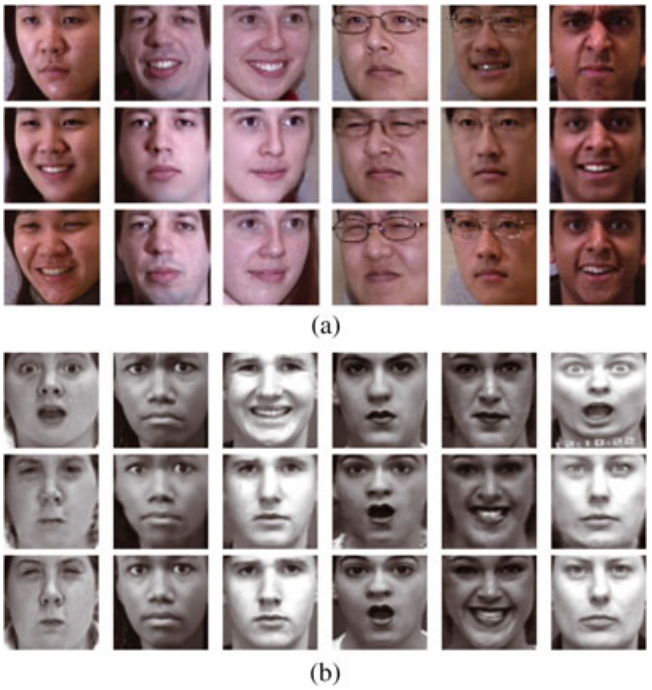

FIGURE 2 Results of facial expression synthesis on (a) Multi-PIE and (b) CK+. Three images from top to bottom in each column represent the input image, the generated image by the model, and the target image in data sets, respectively

4.3 | Smooth transition through landmarks

In order to obtain the smoothly changed synthesis images, we do the following experiments with RGB images in Multi-PIE. Given one input image, two different facial expressions of the same person, and a number N, we compute their linear interpolation of facial landmarks for N times, as follows:

$$\left[\frac{t}{N}\cdot \boldsymbol{g}_i^u + \left(1-\frac{t}{N}\right)\cdot \boldsymbol{g}_i^v\right]_{t=0}^{N}. \tag{6}$$

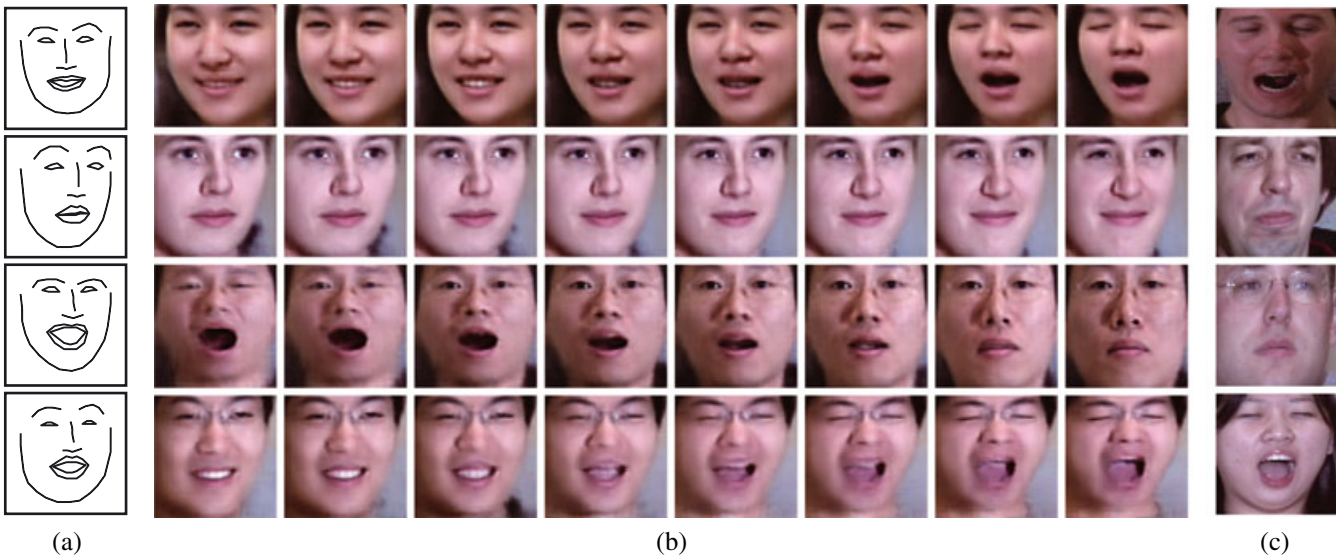

FIGURE 3 Dynamic smooth transitions of facial expressions on Multi-PIE. The images in (a) present the landmarks of source images of the start expressions, and the images in (c) show the face images of source subjects displaying the target facial expressions. The four rows in (b) show the transitions from smile to scream, from neutral to smile, from scream to neutral, and from smile to scream, respectively

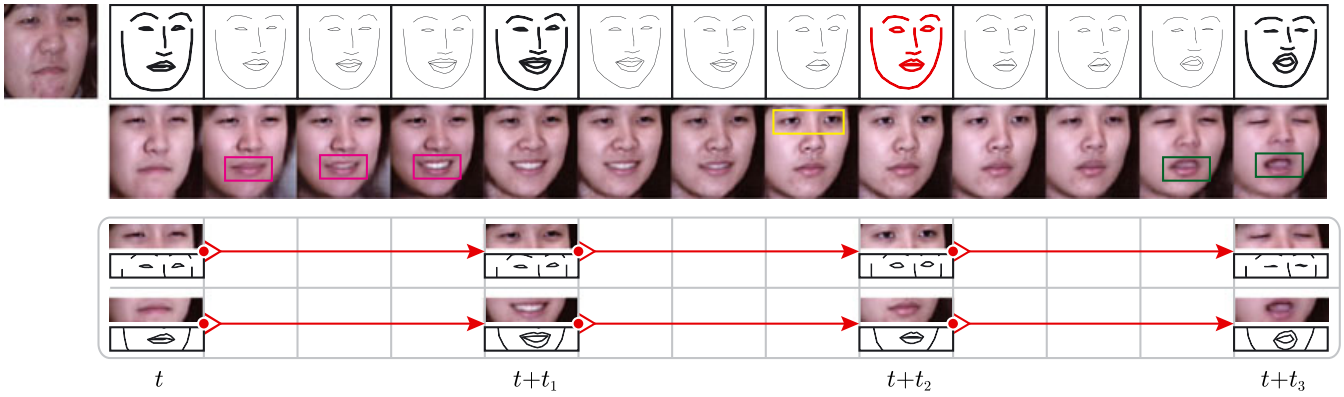

FIGURE 4 Dynamic smooth transitions of facial expressions of one subject. (Top) Input image put separately and guiding facial landmarks drawn in a wireframe. In particular, the landmarks at key frames t, $t + t_1$, $t + t_2$, and $t + t_3$ are highlighted, whereas others are computed by interpolation; the manually designed landmarks are presented in red. (Middle) Generated images and the boxes showing the generation of invisible parts. (Bottom) Transitions of key frames, in which both detailed changes around the eyes and the mouth are illustrated

Then, N groups of new landmarks are input into the model with the same input image to generate a sequence of N images in a smooth transition of facial expression.

Some generated image sequences between two arbitrary expressions for different subjects are shown in Figure 3. The results indicate that this method of interpolation is suitable for different subjects and different facial expressions.

Then, we concentrate on finely controlling the facial expressions of one person and do another experiment. The input image is fixed, that is, the identity information of generated images has been determined. We use some key frames to guide the facial expression transition. The input landmarks of key frames are obtained from some known images of specific facial expressions or manually designed by directly modifying known landmarks. Specifically, we modify the landmarks by raising or lowering the landmark points of the eyelids or moving the landmark points of the mouth corners by pixels. Between the key frames are the input landmarks calculated through interpolation, which is formulated in Equation (6). The input image, input landmarks, and the corresponding generated images are exhibited in Figure 4 with more details. The experimental results show that our model can represent both the entire facial changes and local facial changes through landmarks. That is because the combined feature $\boldsymbol{z}_g$ in the generator is capable of controlling facial expression transfer. Besides, our method can also hallucinate hidden parts of input faces such as pupils, teeth, and tongue, which are boxed in different colors in Figure 4. Consequently, dynamic smooth transitions of facial expressions are achieved primely both overall and in detail.

5 | CONCLUSION

In this paper, we have proposed a geometry guided generative adversarial network for facial expression transfer from a single image. With the combination of geometry information and appearance information, our model learns to generate an image sequence constrained by an input image and landmarks. The landmarks, uncorrelated to the input image, control the transfer via latent space without a neutral face intermediate. Some experiment results have showed the high quality of the generated static images and remarkable transfer of dynamic image sequences through the manipulation of facial landmarks. This type of face sequences shows a potential application of face animation if temporal information is given. In future work, more study about sequential expression transfer is needed.

ACKNOWLEDGEMENTS

The authors would like to thank the anonymous reviewers for their insightful comments. This work was supported by the National Key R&D Program of China (2017YFB1002805), the National Natural Science Foundation of China (61661146002), and the Key Research Program of Frontier Sciences, CAS (QYZDY-SSW-JSC041).

ORCID

Fengchun Qiao http://orcid.org/0000-0003-2714-2036

REFERENCES

1. Ghimire D, Lee J. Geometric feature-based facial expression recognition in image sequences using multi-class AdaBoost and support vector machines. Sens. 2013;13(6):7714–7734.
2. Matthews I, Baker S. Active appearance models revisited. Int J Comput Vis. 2004;60(2):135–164.
3. Kazemi V, Sullivan J. One millisecond face alignment with an ensemble of regression trees. Paper presented at: IEEE Conference on Computer Vision and Pattern Recognition; 2014 Jun 23–28; Columbus, OH.
4. Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. Paper presented at: NIPS 2014. Advances in Neural Information Processing Systems 27; 2014; Montréal, Canada. La Jolla: NIPS; 2014. p. 2672–2680.
5. Zhang Z, Song Y, Qi H. Age progression/regression by conditional adversarial autoencoder. Paper presented at: IEEE Conference on Computer Vision and Pattern Recognition; 2017; Honolulu, HI.
6. Gu G, Kim S-T, Kim K, Baddar W-J, Ro Y-M. Differential generative adversarial networks: Synthesizing non-linear facial variations with limited number of training data. 2017. arXiv preprint arXiv:1711.10267.
7. Song L, Lu Z, He R, Sun Z, Tan T. Geometry guided adversarial facial expression synthesis. 2017. arXiv preprint arXiv:1712.03474.
8. Gross R, Matthews I, Cohn J, Kanade T, Baker S. Multi-PIE. Image Vis Comput. 2010;28(5):807–813.
9. Lucey P, Cohn J-F, Kanade T, Saragih J, Ambadar Z, Matthews I. The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. Paper presented at: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops; 2010; San Francisco, CA.
10. Fried O, Shechtman E, Goldman D-B, Finkelstein A. Perspective-aware manipulation of portrait photos. ACM Trans Graph. 2016; 35(4):1–128.
11. Garrido P, Valgaerts L, Rehmsen O, Thormaehlen T, Perez P, Theobalt C. Automatic face reenactment. Paper presented at: IEEE Conference on Computer Vision and Pattern Recognition; 2016; Las Vegas, NV.
12. Blanz V, Basso C, Poggio T, Vetter T. Reanimating faces in images and video. Comput Graph Forum. 2003;22(3):641–650.
13. Cao C, Hou Q, Zhou K. Displaced dynamic expression regression for real-time facial tracking and animation. ACM Trans Graph. 2014;33(4).
14. Liu X, Mao T, Xia S, Yong Y, Wang Z. Facial animation by optimized blendshapes from motion capture data. Comput Anim Virtual Worlds. 2008;19(3–4):235–245.
15. Mohammed U, Prince SJD, Kautz J. Visio-lization: Generating novel facial images. ACM Trans Graph. 2009;28(3).
16. Averbuch-Elor H, Cohen-Or D, Kopf J, Cohen M-F. Bringing portraits to life. ACM Trans Graph. 2017;36(6):1–13.
17. Susskind J-M, Hinton G-E, Movellan J-R, Anderson A-K. Generating facial expressions with deep belief nets. In: Or J. Affective computing, focus on emotion expression, synthesis and recognition. Vienna: I-Tech; 2008. p. 421–440
18. Reed S, Sohn K, Zhang Y, Lee H. Learning to disentangle factors of variation with manifold interaction. Paper presented at: IEEE 31th International Conference on Machine Learning; 2014; Beijing, China.
19. Shu Z, Yumer E, Hadap S, Sunkavalli K, Shechtman E, Samaras D. Neural face editing with intrinsic image disentangling. Paper presented at: IEEE Conference on Computer Vision and Pattern Recognition; 2017; Honolulu, HI.
20. Zhou Y, Shi B-E. Photorealistic facial expression synthesis by the conditional difference adversarial autoencoder. 2017. arXiv preprint arXiv:1708.09126.
21. Ding H, Sricharan K, Chellappa R. ExprGAN: Facial expression editing with controllable expression intensity. 2017. arXiv preprint arXiv: 1709.03842.
22. Arjovsky M, Chintala S, Bottou L. Wasserstein GAN. 2017. arXiv preprint arXiv:1701.07875.
23. Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville A. Improved training of Wasserstein GANs. 2017. arXiv preprint arXiv: 1704.00028.
24. Larsen A-B-L, Sønderby S-K, Larochelle H, Winther O. Autoencoding beyond pixels using a learned similarity metric. Paper presented at: IEEE 33rd International Conference on Machine Learning; 2016; New York, NY.
25. Kingma D-P, Welling M. Auto-encoding variational bayes. 2013. arXiv preprint arXiv:1312.6114.
26. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. Paper presented at: International Conference on Medical Image Computing and Computer-Assisted Intervention; 2015; Granada, Spain.
27. Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. Paper presented at: IEEE 32nd International Conference on Machine Learning; 2015; Lille, France.
28. Krizhevsky A, Sutskever I, Hinton G-E. ImageNet classification with deep convolutional neural networks. Paper presented at: International Conference on Neural Information Processing Systems; 2012; Lake Tahoe, NV.
29. Shrivastava A, Pfister T, Tuzel O, Susskind J, Wang W, Webb R. Learning from simulated and unsupervised images through adversarial training. Paper presented at: IEEE Conference on Computer Vision and Pattern Recognition; 2017; Honolulu, HI.
30. Ba J-L, Kiros J-R, Hinton G-E. Layer normalization. 2016. arXiv preprint arXiv:1607.06450.

31. Maas A-L, Hannun A-Y, Ng A-Y. Rectifier nonlinearities improve neural network acoustic models. Paper presented at: IEEE 30th International Conference on Machine Learning; 2013; Atlanta, GA.
32. Jung H, Lee S, Yim J, Park S, Kim J. Joint fine-tuning in deep neural networks for facial expression recognition. Paper presented at: IEEE International Conference on Computer Vision; 2015; Los Alamitos, CA.
33. Kingma D, Ba J. Adam: A method for stochastic optimization. 2014. arXiv preprint arXiv:1412.6980.
34. Abadi M, Barham P, Chen J, Chen Z, et al. TensorFlow: A system for large-scale machine learning. Paper presented at: OSDI '16. Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation; 2016; Savannah, GA. Berkeley: USENIX Association; 2016. p. 265–283.
35. Wang Z, Bovik A-C, Sheikh H-R, Simoncelli E-P. Image quality assessment: From error visibility to structural similarity. IEEE Trans Image Process. 2004;13(4):600–612.

Fengchun Qiao is a master candidate at the Institute of Software, Chinese Academy of Sciences. He received his bachelor's degree in college of science from Beijing Forestry University in 2016. His research interest covers human-computer interaction and affective computing.

Naiming Yao is a PhD candidate at Institute of Software, Chinese Academy of Sciences, Beijing, and University of Chinese Academy of Sciences, Beijing. He received his M.S. degree in computer software and theory from Capital Normal University. His research interests include human-computer interaction, affective computing, machines learning, and computer vision.

Zirui Jiao is a master candidate at the Institute of Software, Chinese Academy of Sciences. He received his bachelor's degree of science from Peking University in 2016. His research interest covers human-computer interaction and affective computing.

Zhihao Li graduated from School of Engineering, Durham University, UK and got a master degree. He received his bachelor degree in Software Engineering from Northwest University, P.R. China. His research interest covers human-computer interaction and affective computing.

Hui Chen is a Professor at Institute of Software Chinese Academy of Sciences. She received the PhD degree in computer science from the Chinese University of Hong Kong, Hong Kong, in 2006. Her research interests include human-computer interaction, affective interaction, haptics and virtual reality.

Hongan Wang is a full professor and the director of Beijing Key Laboratory of Human-Computer Interaction, Institute of Software, Chinese Academy of Sciences, Beijing. He received his PhD degree in computer science from Institute of Software, Chinese Academy of Sciences, Beijing, in 1999. His research interests include human-computer interaction and real-time intelligence. He has published over 100 papers in human computer interaction (HCI) and real-time artificial intelligence (RTAI) fields, including ACM CHI, IJHCS, ACM IUI, ACM TIST, IEEE RTSS, etc.

How to cite this article: Qiao F, Yao N, Jiao Z, Li Z, Chen H, Wang H. Emotional facial expression transfer from a single image via generative adversarial nets. *Comput Anim Virtual Worlds*. 2018;29:e1819. https://doi.org/10.1002/cav.1819