

---

# Geometry-Contrastive GAN for Facial Expression Transfer

---

**Fengchun Qiao, Naiming Yao, Zirui Jiao, Zhihao Li, Hui Chen, Hongan Wang**  
 Institute of Software, Chinese Academy of Sciences

## Abstract

In this paper, we propose a Geometry-Contrastive Generative Adversarial Network (GC-GAN) for transferring continuous emotions across different subjects. Given an input face with certain emotion and a target facial expression from another subject, GC-GAN can generate an identity-preserving face with the target expression. Geometry information is introduced into cGANs as continuous conditions to guide the generation of facial expressions. In order to handle the misalignment across different subjects or emotions, contrastive learning is used to transform geometry manifold into an embedded semantic manifold of facial expressions. Therefore, the embedded geometry is injected into the latent space of GANs and control the emotion generation effectively. Experimental results demonstrate that our proposed method can be applied in facial expression transfer even there exist big differences in facial shapes and expressions between different subjects.

## 1 Introduction

Facial expression transfer aims to transfer facial expressions from a source subject to a target subject. The newly-synthesized expressions of the target subject are supposed to be identity-preserving and exhibit similar emotions to the source subject. A wide range of applications such as virtual reality, affective interaction, and artificial agents could be facilitated by the advances of facial expression transfer. The past few decades has witnessed numerous techniques designed for facial expression transfer. Most representative methods [5, 23, 24, 30] are conducted in a sequence-to-sequence way where they assume the availability of a video of the target subject with variants in facial expressions and head poses, which limits their applications. Averbuch *et al.* [1] proposed a warp-based method to animate a single facial image in a sequence-to-image way, capable of generating both photo-realistic and high-resolution facial expression videos, just like bringing a still portrait to life. However, there exists misalignment across different subjects without the assistance of neutral faces for alignment, leading to artifacts on the generated faces. Moreover, large manual adjustments or high computations are still required to generate realistic expressions. How to handle pixel-wise misalignments across different subjects with different emotions in an easy and automatic way is still an open problem.

Recently, Generative Adversarial Nets (GANs) [8] have received extensive attentions to generate lavish and realistic faces. Conditional GANs (cGANs) have also been widely applied in many face-related tasks, such as face pose manipulation [29], face aging [32], and facial expression transfer [3, 6, 22, 33]. Existing works relating to GAN-based facial expression transfer mainly focus on generating facial expressions with discrete and limited emotion states. Emotion states (*e.g.*, *happy*, *sad*) were usually encoded as conditions of one-hot codes to control the generation of expressions [3, 6, 33]. Facial landmarks of action units (AU) in the Facial Action Coding System (FACS) [7] were also adopted as conditions to guide the generated expressions [22]. In [22], the target face and facial landmarks of the driving face are concatenated in image space, which brings extra artifacts when there exist big differences in facial shapes and expressions between target and driving faces. Since people express emotions in a continuous and vivid way, how to inject facial geometry into GANs to generate continuous emotions of the target subject is essential in facial expression transfer.

In order to transfer continuous emotions across different subjects, a Geometry-Contrastive Generative Adversarial Network (GC-GAN) is proposed in this paper. GC-GAN consists of a facial geometry embedding network, an image generator network, and an image discriminator network. Contrastive learning from geometry information is integrated in embedding network. Its bottleneck layer, representing a semantic manifold of facial expressions, is concatenated into the latent space in GAN. Therefore, continuous facial expressions are displayed within the latent space, and the pixel-to-pixel misalignments across different subjects are resolved via embedded geometry. Our main contributions are as follows: (1) We apply contrastive learning in GAN to embed geometry information onto a semantic manifold in the latent space. (2) We inject facial geometry to guide the facial expression transfer across different subjects with Lipschitz continuity. (3) Experimental results demonstrate that our proposed method can be applied in facial expression transfer even there exist big differences in facial shapes and expressions between different subjects.

## 2 Related work

GAN-based conditional image generation has been actively studied. Mirza *et al.*[19] proposed conditional GAN (cGAN) to generate images controlled by labels or attributes. Larsen *et al.*[14] proposed VAE-GAN combining variational autoencoder and GAN to learn an embedding representation which can be used to modify high-level abstract visual features. cGANs have also been applied in facial expression transfer [3, 6, 22, 33]. Zhou *et al.*[33] proposed a conditional difference adversarial autoencoder (CDAAE) to generate faces conditioned on emotion states or AU labels. Choi *et al.*[3] proposed StarGAN to perform image-to-image translations for multiple domains using only a single model, which can also be applied in facial expression transfer. Previous studies mainly focus on generating facial images conditioned on discrete emotion states. However, human emotion is expressed in a continuous way, thus discrete states are not sufficient to describe detailed characteristics of facial expressions.

Some researchers have attempted to incorporate continuous information such as geometry into cGANs [2, 13, 17, 22, 27]. Ma *et al.*[17] proposed a pose guided person generation network (PG<sup>2</sup>), which allows to generate images of individuals in arbitrary poses. The target pose is defined by a set of 18 joint locations and encoded as heatmaps, which are concatenated with the input image in PG<sup>2</sup>. In addition, they adopted a two-stage generation method to enhance the quality of generated images. Song *et al.*[22] proposed a Geometry-Guided Generative Adversarial Network (G2GAN), which applies facial geometry information to guide the transfer of facial expressions. In G2GAN, facial landmarks are treated as an additional image channel to be concatenated with the input face directly. They use dual generators to perform the synthesis and the removal of facial expressions simultaneously. The neutral facial images are generated through the removal network and used for the subsequent facial expression transfer. This procedure brings additional artifacts and degrades the performance especially when the driving faces are collected from other subjects in different emotions.

Contrastive learning [4] minimizes a discriminative loss function that drives the similarity metric to be small for pairs of the same class, and large for pairs from different classes. Although contrastive learning has been widely used in recognition tasks [25, 26, 31], potentials of the semantic manifold in latent space haven't been fully explored, which means that inter-class transitions can be expressed in a continuous representation. In our method, we introduce contrastive learning in GANs to embed geometry information and image appearance in continuous latent space to smoothly minimize the pixel-wise misalignments between different subjects and to generate continuous emotion expressions.

## 3 Method

### 3.1 Overview

In the following, we supposed that  $I_i^e$  and  $g_i^e$  represent a facial image and a geometry of facial landmarks respectively, where the superscript  $e$  indicates an emotion state while the subscript  $i$  denotes an identity. Given a facial image  $I_j^u$  of a subject  $j$  with emotion  $u$  and a geometry variable  $g_k^v$  of a subject  $k$  with target emotion  $v$ , GC-GAN aims to generate a new identity-preserving facial image  $\hat{I}_j^v$  conditioned on the target facial expression. Our problem is described as:

$$\{I_j^u, g_k^v\} \longrightarrow \hat{I}_j^v \quad (1)$$

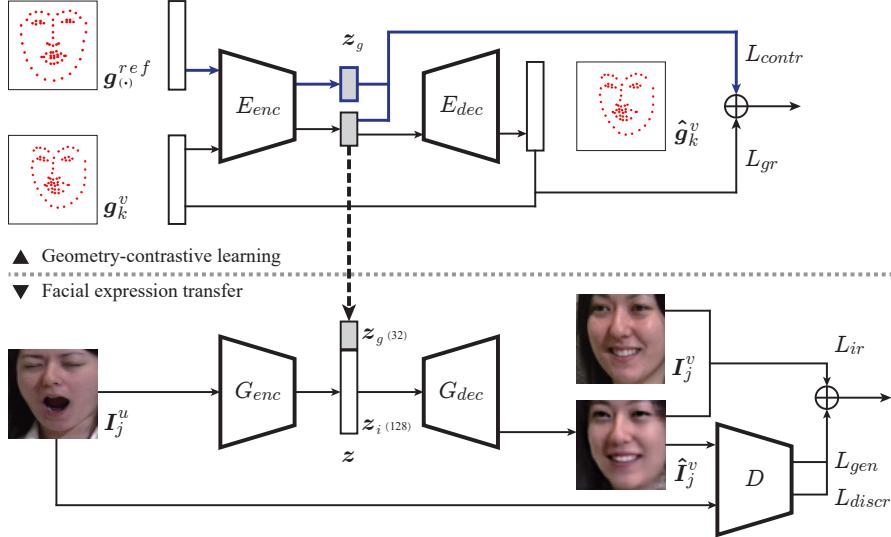


Figure 1: Overview of GC-GAN framework. GC-GAN consists of three components: facial geometry embedding network  $E$  ( $E_{enc}$ ,  $E_{dec}$ ), image generator network  $G$  ( $G_{enc}$ ,  $G_{dec}$ ), and image discriminator network  $D$ .

The overall framework of GC-GAN is shown in Figure 1. GC-GAN consists of three components: a facial geometry embedding network  $E$  ( $E_{enc}$ ,  $E_{dec}$ ), an image generator network  $G$  ( $G_{enc}$ ,  $G_{dec}$ ), and an image discriminator network  $D$ . Input facial image  $I_j^u$  and target landmarks  $g_k^v$  are encoded by  $G_{enc}$  and  $E_{enc}$  into  $z_i$  and  $z_g$  respectively. Then  $z_i$  and  $z_g$  are concatenated into a single vector  $z$  for  $G_{dec}$ .  $g_{(\cdot)}^{ref}$  is reference facial landmarks used for contrastive learning against  $g_k^v$ . Note that geometry expression features  $z_g$  and image identity features  $z_i$  are learned in a disentangled manner, thus we could modify expression and keep identity preserved.

### 3.2 Geometry-Contrastive GAN

The objective of original GANs [8] is formulated as:

$$L_{GAN} = \mathbb{E}_{\mathbf{x} \sim p_{data}} [\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z} \sim p_z} [\log(1 - D(G(\mathbf{z})))] \quad (2)$$

where  $D$  and  $G$  are the discriminator and generator respectively. In order to guide the generated samples, conditional information is input into generators as well as input prior  $\mathbf{z}$  in GANs. Classical and other cGANs mainly use a method of concatenation of prior variables and conditional variables, which is often represented as one-hot codes [19] or embedded discrete variables [21].

Now we first introduce continuous conditions into generators to provide more precise guidance. The input of generator  $\mathbf{z}$  consists of two parts,  $z_i$  and  $z_g$ , i.e.,  $\mathbf{z} = (z_i, z_g)$ .  $z_i$  and  $z_g$  obey different independent prior distributions  $p_i$  and  $p_g$ , so  $\mathbf{z} \sim p_z = p_i p_g$ , and thus  $\mathbf{z}$  is also a prior. When we use the prior to drive the generation of images, it is usually accepted that the generated images are not abruptly changed when the input  $\mathbf{z}$  is gradually changed. Referring to the LS-GAN [11], this assumption can be represented as the Lipschitz continuity of  $G(\mathbf{z})$  for  $\mathbf{z}$ , i.e.,  $\forall \mathbf{z} = (z_i, z_g), z' = (z'_i, z'_g), \exists \rho > 0$ , s.t.  $|G(\mathbf{z}) - G(\mathbf{z}')| \leq \rho \|\mathbf{z} - \mathbf{z}'\|$ . When  $z_g$  is fixed, i.e.,  $z_g = z'_g$ , we get  $|G(z_i, z_g) - G(z'_i, z_g)| \leq \rho \|z_i - z'_i\|$ , which means  $G(z_i, z_g)$  is Lipschitz continuous for  $z_i$ . Similarly, when  $z_i$  is fixed, it is concluded that  $G(z_i, z_g)$  is Lipschitz continuous for  $z_g$ :

$$|G(z_i, z_g) - G(z_i, z'_g)| \leq \rho \|z_g - z'_g\|. \quad (3)$$

We denote  $z_i \sim N(\mu, \sigma)$  and  $z_g \sim p_g$ , a certain continuous prior distribution. The Lipschitz of  $G$  for  $z_g$  indicates that the condition we introduce can continuously control the generation. It is obvious that if  $p_g$  is a discrete distribution, the model is altered into an ordinary cGAN and the Lipschitz of  $G(z_i, z_g)$  for  $z_g$  turns false, because  $z_g$  does not exist in a metric space and  $\|z_g - z'_g\|$  makes no sense anymore.

In GC-GAN, we aim to solve the problem in Eq.(1), where  $\mathbf{z}_i$  and  $\mathbf{z}_g$  represent identity information and facial expression condition of generated faces respectively. And  $\mathbf{z}_i$  and  $\mathbf{z}_g$  here are independent since the known image  $\mathbf{I}_j^u$  and geometry vector  $\mathbf{g}_k^v$  don't necessarily refer to the same subject. Compared to discrete distribution of facial expressions, if  $p_g$  is some continuous distribution derived from facial landmarks, conditions  $\mathbf{z}_g$  can guide the generation of facial expressions more precisely.

However, it increases the risk of errors to introduce geometry variables  $\mathbf{g}_k^v$  into cGANs directly *i.e.*,  $\mathbf{z}_g = \mathbf{g}_k^v$ . Original facial landmark space is badly separable for facial expressions. So when  $\mathbf{g}_k^v$  changes as conditions, the changes of generated images will get out of control, *i.e.*, the constant  $\rho$  is too large or approximately infinity. It also has bad interpretability to concatenate landmark vector and prior in the input of generator. So an embedding method is necessary to transform the original geometry information into a more effective representation. Since the objective of contrastive learning is based on the calculation of metrics, we apply a contrastive learning network  $E$  to the landmarks and define an embedding variable  $\mathbf{z}_g = E_{enc}(\mathbf{g}_k^v)$ . Through contrastive learning, landmarks are encoded into a semantic space of facial expression, which reduces the dimensions, extracts facial expression features, and provides the different features with better separability and appropriate distance. Thus, when the conditions  $\mathbf{z}_g$  in semantic space are introduced and the prior  $\mathbf{z}_i$  fixed, the generated images are Lipschitz continuous over the facial expression conditions with the identity information preserved.

### 3.3 Learning strategy

**Geometry-contrastive learning** Contrastive learning is conducted on facial geometry with reference facial landmarks, where the  $xy$ -coordinates of facial landmarks are arranged into a one-dimensional vector as input for the geometry embedding network. Landmark pairs  $(\mathbf{g}_k^v, \mathbf{g}_{(\cdot)}^{ref})$  are prepared for training, in which  $\mathbf{g}_k^v$  is the target expression while  $\mathbf{g}_{(\cdot)}^{ref}$  is the reference expression.  $(\cdot)$  represents a certain subject. After a transform function  $E_{enc}(\cdot)$ , the facial landmarks are mapped into the embedding space. Our goal is to measure the similarity between  $E_{enc}(\mathbf{g}_k^v)$  and  $E_{enc}(\mathbf{g}_{(\cdot)}^{ref})$  according to their expression labels. The contrastive loss  $L_{contr}$  is formulated as below:

$$\begin{aligned} \min_{E_{enc}} L_{contr} = & \frac{\alpha}{2} \max \left( 0, m - \left\| E_{enc}(\mathbf{g}_k^v) - E_{enc}(\mathbf{g}_{(\cdot)}^{ref}) \right\|_2^2 \right) \\ & + \frac{1-\alpha}{2} \left\| E_{enc}(\mathbf{g}_k^v) - E_{enc}(\mathbf{g}_{(\cdot)}^{ref}) \right\|_2^2, \end{aligned} \quad (4)$$

where  $\alpha = 0$ , if the expression labels are the same,  $\alpha = 1$  otherwise.  $m$  is a margin which is always greater than 0.  $L_{contr}$  enables our embedding network to focus on facial expression itself regardless of different subjects and plays an essential part for pushing facial landmarks to reside in a semantic manifold. The semantic manifold is explained in detail in Section 4.3.

**Adversarial learning** In our model, the adversarial losses for  $G$  and  $D$  are formulated as follows:

$$\min_D L_{discr} = - \mathbb{E}_{\mathbf{I} \sim p_I} D(\mathbf{I}_j^u) + \mathbb{E}_{\mathbf{I} \sim p_I, \mathbf{g} \sim p_g} D(\hat{\mathbf{I}}_j^v), \quad (5)$$

$$\min_G L_{gen} = - \mathbb{E}_{\mathbf{I} \sim p_I, \mathbf{g} \sim p_g} D(\hat{\mathbf{I}}_j^v), \quad (6)$$

$$\min_{G,D} L_{adv} = L_{discr} + L_{gen}, \quad (7)$$

where  $p_I$  and  $p_g$  indicate the distribution of real facial images and real facial landmarks, respectively.  $\hat{\mathbf{I}}_j^v$  represents the generated image computed by  $G_{dec}(G_{enc}(\mathbf{I}_j^u), E_{enc}(\mathbf{g}_j^v))$ . The adversarial losses are optimized via WGAN-GP [10].

**Reconstruction learning** The  $\ell_1$  loss and  $\ell_2$  loss are introduced for the reconstruction of faces and landmarks, respectively. For the generated facial image  $\hat{\mathbf{I}}_j^v$ , we employ  $\ell_1$  distance to compute a pixel-to-pixel difference between generated image and target real image, which is formulated as  $\min_G L_{ir} = \left\| \mathbf{I}_j^v - \hat{\mathbf{I}}_j^v \right\|_1$ . The mixture of adversarial loss and  $L_{ir}$  accelerate the convergence and enhance image visual quality. For facial landmarks,  $\ell_2$  distance is employed to compute the difference between reconstructed landmarks and input landmarks, which is formulated as  $\min_E L_{gr} = \left\| \mathbf{g}_k^v - \hat{\mathbf{g}}_k^v \right\|_2^2$ . Reconstruction learning enables latent vector  $\mathbf{z}$  to preserve enough information for the reconstruction of inputs.

**Overall learning** First, network  $E$  is pretrained on the training set. The loss function is  $\min_E L_E = \lambda_1 L_{contr} + \lambda_2 L_{gr}$ . Then network  $G$  and  $D$  are trained together while the parameters of network  $E$  are fixed. The loss function is  $\min_{G,D} L_G = \lambda_3 L_{ir} + \lambda_4 L_{adv}$ .  $\lambda_1, \lambda_2, \lambda_3$  and  $\lambda_4$  are set to 1, 1, 1 and  $10^{-3}$ , respectively. The detailed parameters of GC-GAN are provided in Appendix-A.

## 4 Experiments

### 4.1 Datasets and protocols

To evaluate the proposed GC-GAN, experiments have been conducted on four popular facial datasets: Multi-PIE [9], CK+ [16], BU-4DFE [28] and CelebA [15].

Multi-PIE consists of 754,200 images from 337 subjects with large variations in head pose, illumination, and facial expression. We select a subset of images with three head poses ( $0^\circ, \pm 15^\circ$ ), 20 illumination conditions, and all six expressions. CK+ is composed of 327 image sequences from 118 subjects with seven prototypical emotion labels. Each sequence starts with a neutral emotion and ends with the peak of a certain emotion. BU-4DFE contains 101 subjects, each one displaying six acted categorical facial expressions with moderate head pose variations. Since this dataset does not contain frame-wise facial expression, we manually select neutral and apex frames of each sequence. CelebA is a large-scale facial dataset, containing 202,599 face images of celebrities. All experiments are performed in a person-independent way, the training set and test set for each dataset are split based on subjects with proportions of 90% and 10%, respectively.

To prepare the training triplet  $(\mathbf{I}_j^u, \mathbf{g}_k^v, \mathbf{g}_{(\cdot)}^{ref})$ , we create all triplets of facial expressions per identity.  $\mathbf{g}_k^v$  is taken from  $\mathbf{I}_k^v$  and  $\mathbf{g}_{(\cdot)}^{ref}$  is generated by sampling at random. Note that  $\mathbf{g}_{(\cdot)}^{ref}$  and facial expression labels are not required in test. In our experiments, 68 facial landmark points are obtained through dlib<sup>1</sup> which implements the method proposed in [12]. Since faces in our experiments remain near-frontal, the accuracy of facial landmark detection is over 95% when dlib is used, which is sufficient for our work. The facial landmarks include points of two eyebrows, two eyes, the nose, the lips, and the jaw. The facial images are aligned according to inner eyes and bottom lip. Then, face regions are cropped and resized into  $64 \times 64$ . Pixel value of images and  $xy$ -coordinates of facial landmarks are both normalized into  $[-1, 1]$ .

### 4.2 Facial expression generation

In order to evaluate whether the synthetic faces are generated by the guidance of the target facial expression, we qualitatively and quantitatively compare the generated faces with the ground truth. The results are shown in Figure 2, in which the generated faces are identity-preserving and are similar to the target facial expressions. For quantitative comparison, SSIM (structural similarity index measure) and PSNR (peak-signal-to-noise ratio) are used as two evaluation metrics. The detailed comparison results are shown in Table 1. We conduct ablation study to evaluate contributions of three losses  $L_{ir}$ ,  $L_{gr}$ , and  $L_{contr}$ , respectively. According to Table 1, GC-GAN performs better when supervised by the three losses. We compare our method with CDAAE [33], which uses one-hot codes to represent facial expressions. As shown in Table 1, our method substantially outperforms CDAAE on both measures across all the datasets, indicating that the images generated using our method are more similar to the ground truth aided by the embedded geometry features. The experiments mentioned above demonstrate that our proposed GC-GAN effectively incorporates geometry information in conditioned facial expression generation.

### 4.3 Analysis of geometry-contrastive learning

First, we evaluate the Lipschitz continuity of generated faces  $G(\mathbf{z}_i, \mathbf{z}_g)$  over facial expression conditions  $\mathbf{z}_g$ . For interpolation of  $\mathbf{z}_g$ , we would like to seek a sequence of  $N$  images with a smooth transition between two different facial expressions of the same person (e.g.  $\mathbf{I}_j^u$  and  $\mathbf{I}_j^v$ ). In our case, we conduct equal interval sampling  $\left[ \frac{t}{N} \cdot E_{enc}(\mathbf{g}_j^u) + (1 - \frac{t}{N}) \cdot E_{enc}(\mathbf{g}_j^v) \right]_{t=0}^N$  on  $\mathbf{z}_g$ . Some generated facial expression sequences are shown in Figure 3(a). According to Eq.(3), we compute

---

<sup>1</sup><http://dlib.net/>

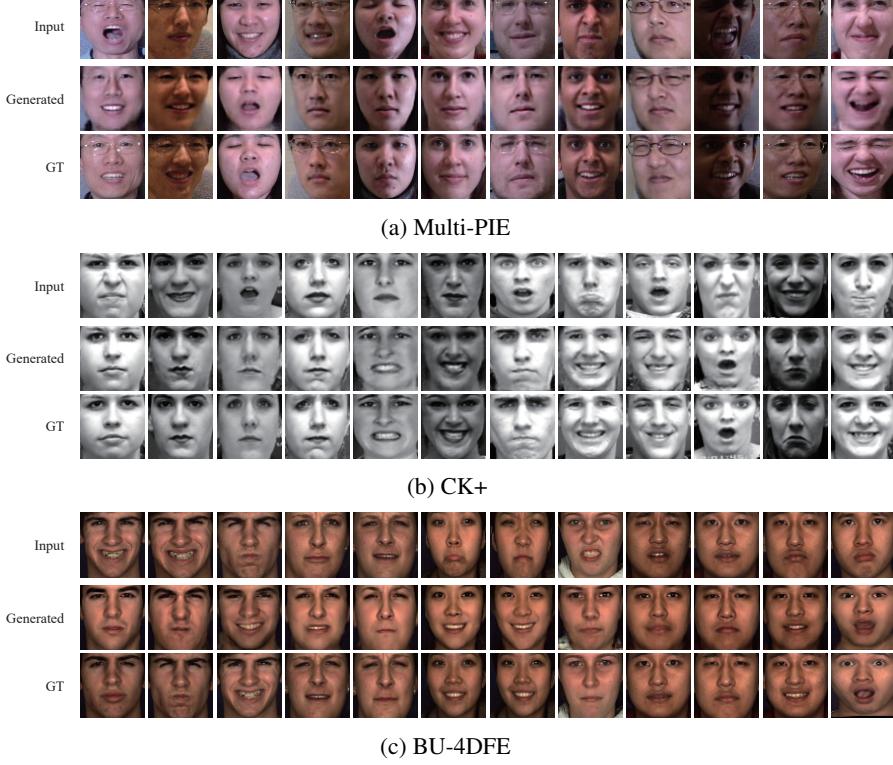


Figure 2: Results of facial expression synthesis on (a) Multi-PIE, (b) CK+, and (c) BU-4DFE. Three images from top to bottom in each column represent the input image, generated image, and ground truth, respectively.

Table 1: Comparison with other settings and method on Multi-PIE, CK+ and BU-4DFE in terms of SSIM and PSNR.

	SSIM			PSNR		
	Multi-PIE	CK+	BU-4DFE	Multi-PIE	CK+	BU-4DFE
GC-GAN (ours)	<b>0.687</b>	<b>0.769</b>	<b>0.725</b>	<b>26.731</b>	<b>27.665</b>	<b>26.915</b>
w/o $L_{ir}$	0.663	0.756	0.702	25.917	26.781	25.136
w/o $L_{gr}$	0.682	0.768	0.722	26.701	27.463	26.875
w/o $L_{contr}$	0.675	0.763	0.705	26.433	27.325	26.325
CDAAE [33]	0.669	0.765	0.710	26.371	26.973	26.122

$\rho' = \frac{|G(z_i, z_g) - G(z_i, z'_g)|}{\|z_g - z'_g\|}$  between two adjacent frames over the whole test set composed of 58,724 pairs from 33 subjects, and the violin plot is shown in Figure 3(b), where the mean at each time step is around 50 and the upper bound also exists. Therefore, the Lipschitz constant  $\rho$  in Eq.(3) is approximately found.

In order to evaluate whether the manifold of  $z_g$  is semantic-aware, we visualize the embedding manifold of facial landmarks by t-SNE [18]. For the purpose of evaluating the effectiveness of contrastive learning, we train our model without contrastive loss  $L_{contr}$ . The corresponding manifold of facial landmarks is shown in Figure 4(a), in which the embedding features cannot be clustered properly without the guidance of contrastive learning. The manifold of GC-GAN is shown in Figure 4(b) where the embedding features are clustered according to their emotion states. It can be seen that the data scatters of *disgust* and *squint* are mixed up with each other due to the intrinsic geometry similarity between the two facial expressions. In addition, we present an embedding line consisting of a sequence of points corresponding to the transition of facial expressions from *smile* to *scream* of the same person obtained by equal interval sampling on facial landmarks, which is shown in Figure

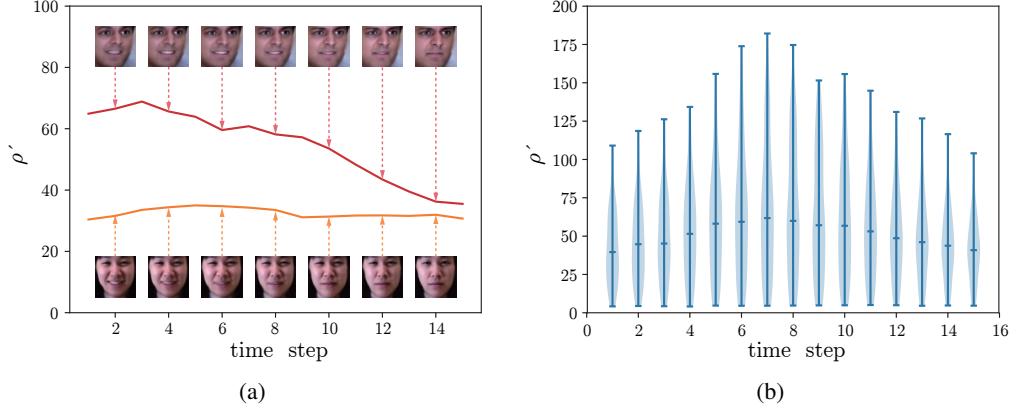


Figure 3: Evaluation of Lipschitz continuity. (a) shows two generated facial expression sequences from *smile* to *disgust* in red curve and from *smile* to *neutral* in orange curve, respectively. (b) shows the violin plot at each time step computed over the whole test set.

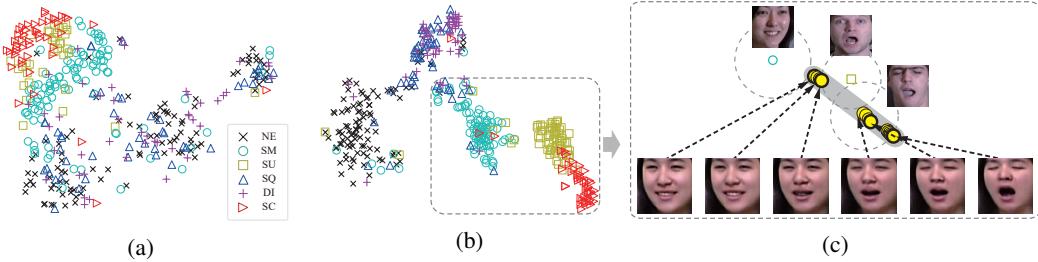


Figure 4: Visualization of the embedded manifolds. The manifold of (a) GC-GAN (w/o  $L_{contr}$ ), and the manifold of (b) GC-GAN. (c) The embedding points of a generated facial expression sequence.

4(c). It can be observed that the eyes are closing and the mouth is opening from *smile* to *scream*. So there could exist some points with open eyes and open mouth simultaneously which is exactly the characteristic of *surprise*, and could account for the reason why the embedding line comes across the region of *surprise*. The results indicate that the manifold of  $z_g$  is both continuous and semantic-aware for GC-GAN.

#### 4.4 Facial expression transfer

To further evaluate the generality of our proposed method, transferring a person’s emotion to different emotions of another person directly is also tested. Facial expression transfer between different persons are difficult due to the individual variances, especially in the absence of neutral faces for reference.

In our case, the target expressions are taken from other people of different datasets. Here, we randomly select 1000 facial images from CelebA. The facial landmarks of these images serve as guiding landmarks. Note that there are only six basic emotion states in Multi-PIE while the emotion states in CelebA are various and spontaneous. Figure 5 shows some representative samples from our experiment, in which the generated faces are identity-preserving and exhibit similar expressions with driving faces even there exist big differences in face shape between input faces and driving faces.

For comparison, a more general configuration SPG is adopted, which is similar to PG<sup>2</sup> [17] and G2GAN [22]. In detail, we treat the guiding landmarks as an additional feature map and directly concatenate it with the input face, and train our model in this configuration. The corresponding results of SPG are shown in the lowest line of Figure 5.

Since there don’t exist ground truth faces corresponding to generated faces, SSIM and PSNR are not suitable for quantitative evaluation. In order to validate whether the generated faces are identity-preserving, a pretrained face identification model VGG-Face [20] is adopted to conduct face verification on the generated faces. In detail, we extract the output of the last convolution layer for measuring the identity similarity between two faces. Two faces are considered as matched if their



Figure 5: Results of facial expression transfer on Multi-PIE from CelebA. Input faces are selected from Multi-PIE, and guiding facial geometries are taken from the corresponding driving faces in CelebA.

Table 2: Evaluation results of facial expression transfer on Multi-PIE from CelebA.

	Accuracy of face verification	Similarity of facial expression
GC-GAN (ours)	<b>0.977</b>	<b>0.896</b>
GC-GAN (w/o $L_{contr}$ )	0.942	0.880
SPG [17, 22]	0.875	0.854

cosine similarity is no less than 0.65. In order to evaluate the similarity of facial expressions between generated and driving faces, the cosine similarity of  $z_g$  of facial landmarks between generated and driving faces is adopted. The accuracy of face verification and the mean similarity of facial expressions are shown in Table 2. Compared with our method, SPG is not able to capture the target facial expression of driving faces and brings extra artifact, indicating that the model under this configuration cannot handle the misalignments between input faces and driving facial landmarks properly. Our model shows the promising capability of manipulating facial expressions while preserving identity information.

## 5 Conclusion

In this paper, we have proposed a Geometry-Contrastive Generative Adversarial Network (GC-GAN) for transferring facial expressions across different persons. Through contrastive learning, facial expressions are pushed to reside onto a continuous semantic-aware manifold. Benefited from this manifold, pixel-wise misalignments between different subjects are weakened and identity-preserving faces with more expression characteristics are generated. Experimental results demonstrate that there exists semantic consistency between the latent space of facial expressions and the image space of faces.

## References

- [1] Hadar Averbuch-Elor, Daniel Cohen-Or, Johannes Kopf, and Michael F Cohen. Bringing portraits to life. *ACM Trans. Graph*, 36(6):196:1–196:13, 2017.
- [2] Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John Guttag. Synthesizing images of humans in unseen poses. In *CVPR (To appear)*, 2018.
- [3] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR (To appear)*, 2018.

- [4] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*, pages 539–546, 2005.
- [5] Kevin Dale, Kalyan Sunkavalli, Micah K Johnson, Daniel Vlasic, Wojciech Matusik, and Hanspeter Pfister. Video face replacement. *ACM Trans. Graph.*, 30(6):130:1–130:10, 2011.
- [6] Hui Ding, Kumar Sricharan, and Rama Chellappa. Exprgan: Facial expression editing with controllable expression intensity. In *AAAI (To appear)*, 2018.
- [7] Paul Ekman and Wallace V Friesen. Facial action coding system. *Consulting Psychologists Press, Stanford University*, 1977.
- [8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014.
- [9] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-PIE. *Image and Vision Computing*, 28(5):807–813, 2010.
- [10] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. In *NIPS*, pages 5769–5779, 2017.
- [11] Qi Guojun. Loss-sensitive generative adversarial networks on lipschitz densities. *arXiv preprint arXiv:1701.06264*, 2017.
- [12] Vahid Kazemi and Sullivan Josephine. One millisecond face alignment with an ensemble of regression trees. In *CVPR*, pages 1867–1874, 2014.
- [13] Jean Kossaifi, Linh Tran, Yannis Panagakis, and Maja Pantic. GAGAN: Geometry-aware generative adversarial networks. In *CVPR (To appear)*, 2018.
- [14] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. In *ICML*, pages 1558–1566, 2016.
- [15] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, pages 3730–3738, 2015.
- [16] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *CVPRW*, pages 94–101, 2010.
- [17] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *NIPS*, pages 405–415, 2017.
- [18] Laurens van der Maaten and Geoffrey Hinton. Visualizing high-dimensional data using t-sne. *JMLR*, 9(2):2579–2605, 2008.
- [19] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.
- [20] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al. Deep face recognition. In *BMVC*, pages 41.1–41.12, 2015.
- [21] Scott E Reed, Zeynep Akata, Santosh Mohan, Samuel Tenka, Bernt Schiele, and Honglak Lee. Learning what and where to draw. In *NIPS*, pages 217–225, 2016.
- [22] Lingxiao Song, Zhihe Lu, Ran He, Zhenan Sun, and Tieniu Tan. Geometry guided adversarial facial expression synthesis. *arXiv preprint arXiv: 1712.03474*, 2017.
- [23] Justus Thies, Michael Zollhöfer, Matthias Nießner, Levi Valgaerts, Marc Stamminger, and Christian Theobalt. Real-time expression transfer for facial reenactment. *ACM Trans. Graph.*, 34(6):183–1, 2015.
- [24] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *CVPR*, pages 2387–2395, 2016.
- [25] Andreas Veit, Serge Belongie, and Theofanis Karaletsos. Conditional similarity networks. In *CVPR*, pages 1781–1789, 2017.
- [26] Jian Wang, Feng Zhou, Shilei Wen, Xiao Liu, and Yuanqing Lin. Deep metric learning with angular loss. In *CVPR*, pages 2593–2601, 2017.

- [27] Wei Wang, Xavier Alameda-Pineda, Dan Xu, Pascal Fua, Elisa Ricci, and Nicu Sebe. Every smile is unique: Landmark-guided diverse smile generation. In *CVPR (To appear)*, 2018.
- [28] Lijun Yin, Xiaochen Chen, Yi Sun, Tony Worm, and Michael Reale. A high-resolution 3d dynamic facial expression database. In *FG*, pages 1–6, 2008.
- [29] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Towards large-pose face frontalization in the wild. In *ICCV*, pages 4010–4019, 2017.
- [30] Matthew D Zeiler, Graham W Taylor, Leonid Sigal, Iain Matthews, and Rob Fergus. Facial expression transfer with input-output temporal restricted boltzmann machines. In *NIPS*, pages 1629–1637, 2011.
- [31] Kaihao Zhang, Yongzhen Huang, Yong Du, and Liang Wang. Facial expression recognition based on deep evolutional spatial-temporal networks. *IEEE Trans. Image Processing*, 26(9):4193–4203, 2017.
- [32] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *CVPR*, pages 4352–4360, 2017.
- [33] Yuqian Zhou and Bertram Emil Shi. Photorealistic facial expression synthesis by the conditional difference adversarial autoencoder. In *ACII*, pages 370–376, 2017.

## Appendix

### 6 Network architecture

The detailed architectures of GC-GAN are shown in Table 3. In detail, generator network  $G$  has a similar architecture to U-Net, but we only add one skip connection between the output of the first convolution layer and the output of the penultimate deconvolution layer, which enables  $G$  to reuse low-level facial features related with identity information. For discriminator network  $D$ , the output is a  $2 \times 2$  feature map.

In training time, the margin  $m$  in  $L_{contr}$  is set to 5. All the networks  $E$ ,  $G$ , and  $D$  are optimized by Adam optimizer with  $\beta_1 = 0.5$  and  $\beta_2 = 0.999$ . The batch size is 64 and the initial learning rate is set to  $3 \times 10^{-4}$ . We implement GC-GAN using Tensorflow.

Table 3: The detailed architecture of *GC-GAN*. In  $(De)Conv(d,k,s)$ ,  $d$ ,  $k$ , and  $s$  stand for the number of filters, the kernel size, and the stride, respectively.  $FC$  represents a fully-connected layer.  $BN$  is batch normalization and  $LN$  is layer normalization.  $LReLU$  refers to *leaky ReLU*.

<b>Embedding <math>E</math></b>	<b>Generator <math>G</math></b>	<b>Discriminator <math>D</math></b>
FC(128), BN, ReLU	Conv(64,5,2), BN, ReLU	Conv(64,5,2), LN, LReLU
FC(64), BN, ReLU	Conv(128,5,2), BN, ReLU	Conv(128,5,2), LN, LReLU
FC(32), BN, ReLU	Conv(256,5,2), BN, ReLU	Conv(256,5,2), LN, LReLU
FC(64), BN, ReLU	FC(1024), BN, ReLU	Conv(512,5,2), LN, LReLU
FC(128), BN, ReLU	FC(128), BN, ReLU	Conv(1,5,2)
FC(136), Tanh	FC(4096), BN, ReLU DeConv(256,5,2), BN, ReLU DeConv(128,5,2), BN, ReLU DeConv(64,5,2), BN, ReLU DeConv(64,5,2), BN, ReLU Conv(3,5,1), BN, Tanh	

## 7 Reviews

### Meta-Review

The paper of expression transfer in face analyses is very relevant. Although reviewers agree that including geometry in transfer with GANs and the analyses with contrastive loss and Lipschitz continuous semantic latent space are interesting, there is a generalized opinion that most of the pieces of the model are standard techniques. Overall, the paper does not achieve the minimum score required to be published at NIPS.

### Reviewer #1

1. Please provide an "overall score" for this submission.

**6:** Marginally above the acceptance threshold. I tend to vote for accepting this submission, but rejecting it would not be that bad.

2. Please provide a "confidence score" for your assessment of this submission.

4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.

3. Please provide detailed comments that explain your "overall score" and "confidence score" for this submission. You should summarize the main ideas of the submission and relate these ideas to previous work at NIPS and in other archival conferences and journals. You should then summarize the strengths and weaknesses of the submission, focusing on each of the following four criteria: quality, clarity, originality, and significance.

The paper describes a method for generating synthetic images of faces with different facial expressions. The authors propose a GAN-based architecture for the purpose, wherein, the facial identity and expression priors are disentangled and encoded in different latent spaces. The expression and identity facial latent vectors are concatenated and reconstructed by the generator to preserve the identity and change the facial expression. The facial expressions are represented by 68 facial fiducial points and encoded into a latent space. The authors use reconstruction losses for the input facial fiducial points and facial images along with adversarial training for the entire algorithm. Overall, neither the problem for synthesizing facial expressions, nor its proposed solution using GANs is novel. The paper is also specific to facial analysis and hence not directly applicable to many different problem domains. Hence the paper has decent novelty and some interesting findings, but is not earth-shattering or likely to be of broad impact.

In contrast to the prior work, the claimed superiority of this method lies in the fact that it allows to learn a Lipschitz continuous semantic latent space of facial expressions instead of a discrete one. The latter is what previous works have learnt via one-hot encoding of the facial expression classes. By comparing their approach to that of [23], the authors show empirically the value of learning a latent embedding of the facial fiducial points for continuous transitions of expressions versus using the facial fiducial points directly as inputs along with the facial image for the GAN-based approach. The latter they argue is not guaranteed to be Lipschitz continuous. This is an interesting finding.

One other finding of this work is that using the contrastive loss for encoding facial expressions helps to learn a more semantically meaningful expression latent space and hence helps to produce better expression synthesis results on moving linearly on the expression manifold. However, the authors' use of the contrastive loss is at odds with their goal of trying to learn a \*continuous\* latent expression space. The contrastive loss tends to create discontinuous clusters for discrete facial expressions classes, which is exactly what the authors observe in the figure 4(b) on visualizing their learnt latent space. This is also something that the authors set out to avoid in the first place. How then do the authors explain the helpfulness of the contrastive loss in learning a continuous latent expression space? This contradictory observation needs to be explored further and clearly addressed and explained in the paper. I would have imagined that something like a variational auto encoder with a KL divergence loss term would have helped to create a more continuously varying latent expression space.

Other than that, the paper is well written, includes sufficient experimental evidence to demonstrate the superior performance of their method versus previous work, and includes sufficient details in the paper and the supplementary document of the network and training for someone to be able to replicate it.

4. How confident are you that this submission could be reproduced by others, assuming equal access to data and resources?

3: Very confident

### Reviewer #2

1. Please provide an "overall score" for this submission.

**6:** Marginally above the acceptance threshold. I tend to vote for accepting this submission, but rejecting it would not be that bad.

2. Please provide a "confidence score" for your assessment of this submission.

**4:** You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.

3. Please provide detailed comments that explain your "overall score" and "confidence score" for this submission. You should summarize the main ideas of the submission and relate these ideas to previous work at NIPS and in other archival conferences and journals. You should then summarize the strengths and weaknesses of the submission, focusing on each of the following four criteria: quality, clarity, originality, and significance.

The paper introduces Geometry-Contrastive Generative Adversarial Network (GC-GAN), a novel deep architecture for transferring facial emotions across different subjects. The proposed architecture is based on the idea of decoupling the identity information with the motion information in the latent representation of the main network. A novel contrastive loss is employed to learn the latent motion information from landmark images. The proposed approach is evaluated on three common face image benchmarks.

Strengths:

- This paper, together with [23, 28], is one of the first attempts to inject geometry information into GAN models for facial expression transfer. This is an interesting and challenging problem.
- The idea of using contrastive learning to compute the latent representation from the landmark images is new, despite contrastive learning have been exploited in many other applications
- The experimental evaluation consider three different benchmarks where face have different characteristics. The proposed approach is shown to be more effective than previous method [23] on the proposed task.
- The paper is well written and the proposed approach clearly explained.

Weaknesses:

- In my opinion the experimental evaluation can be improved. Specifically, when analyzing the dynamics of face sequences in section 4.3 a comparison with [28] can be added. This will confirm the effectiveness of the proposed contrastive loss.
- In Table 2 it is not clear what SPG is. Which method was used exactly?
- The results from a user study could be added, as SSIM may be not the best metric for analyzing generated faces.

4. How confident are you that this submission could be reproduced by others, assuming equal access to data and resources?

2: Somewhat confident

#### Reviewer #4

1. Please provide an "overall score" for this submission.

**4:** An okay submission, but not good enough; a reject. I vote for rejecting this submission, although I would not be upset if it were accepted.

2. Please provide a "confidence score" for your assessment of this submission.

**4:** You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.

3. Please provide detailed comments that explain your "overall score" and "confidence score" for this submission. You should summarize the main ideas of the submission and relate these ideas to previous work at NIPS and in other archival conferences and journals. You should then summarize the strengths and weaknesses of the submission, focusing on each of the following four criteria: quality, clarity, originality, and significance.

This paper presented an framework - GC-GAN for facial expression transfer/generation. The proposed method consists of a facial geometry embedding network and a conditional GAN networks. The experimental on several datasets show the effectiveness of GC-GAN.

Some strong points of the paper:

1. the integration of the facial geometry embedding network is somehow novel

2. the presentation of the paper is good

Some weak points of the paper:

1. the novelty contribution of the overall framework is minor. The generator/discriminator just follow general conditional GANs. The geometry-guided idea have been explored a lot in previous works (such as G2GAN,GAGAN).

2. the experimental part is not strong: a) the model has many hyper-parameters/pretrained modules setting, making it difficult to reproduce the results. The details of those are not described clearly; b) just few baselines are compared in the results. There should be more about GANs with structured conditions.

3. The analysis of geometry-contrastive learning, i.e., the facial embedding network is not convincing. Section 4.3 can not give a convincing quantitative result. The measure in Section 4.4 only takes the overall measure but fails to show how the generated samples have better representation regarding to the continuity.

Some related and important references are missing about GANs:

a. Disentangled Representation Learning GAN for Pose-Invariant Face Recognition

b. Disentangled Person Image Generation

c. Pose Guided Person Image Generation

4. How confident are you that this submission could be reproduced by others, assuming equal access to data and resources?

2: Somewhat confident

## 8 Rebuttal

### Response to Reviewer #1

**Q:** About the novelty and potential application to other problem domains.

**A:** Yes, the problem for synthesizing facial expressions is not new, but how to integrate continuity of facial expressions in GANs hasn't been mentioned or studied carefully before, thus the detailed problem and our solution are novel to a certain extent. Facial expressions are intrinsically continuous but always annotated with discrete labels, so our method can also be applied in similar problems such as the pose of objects, the length of hair, the height of heels and so on.

**Q:** About the effect of contrastive learning for semantically continuous representation.

**A:** Contrastive learning has been widely used in face identification such as FaceNet, VGG-Face and so on. Face identification is a typical classification problem, however, these networks trained for it also generalize well on face verification where faces are unseen. Similarly, such learning mechanism enables the latent space to accommodate unseen emotion states. As seen in Figure 4(b), the data scatters of disgust and squint are not separated well with each other due to the intrinsic geometry similarity although they are different expressions. In Figure 4(c), we present an embedding line corresponding to the transition from smile to scream of the same person obtained by equal interval sampling on facial landmarks. It can be observed that the eyes are closing and the mouth is opening, which is exactly the embodiment of continuity. We try to give a general architecture of continuous latent space in cGANs and have embedded an autoencoder in current work, and we will test KL divergence loss term in the future. Thanks!

### Response to Reviewer #2

**Q:** The comparison with [28].

**A:** [28] proposes a method to generate facial image sequences of diverse smiles, which is an excellent work. We'll add this comparison in terms of analyzing the dynamics of face sequences in our future work.

**Q:** The explanation of SPG and user study.

**A:** SPG is the abbreviation of "similar to PG<sup>2</sup> and G2GAN", which means the method of concatenating geometry and images directly as in both PG<sup>2</sup> and G2GAN. PG<sup>2</sup> mainly focuses on the generation of persons of different poses and the input facial images of G2GAN should be neutral expressions. Besides, the ways of training GANs, neural network architectures and loss functions among PG<sup>2</sup>, G2GAN and GC-GAN are different. It's improper to directly compare their methods with ours considering the reasons mentioned above, so we construct such

*configuration for fair comparison. We also took user study into consideration. However, user study is somewhat subjective and may cause some trouble if other people want to compare their methods with ours. We'll conduct a user study if necessary. Thanks for your advice!*

#### **Response to Reviewer #4**

**Q:** The novelty contribution of the overall framework is minor.

**A:** *Instead of following the general conditional GAN, we concentrate on how to establish semantically continuous conditions into GANs, and formal description of Lipschitz continuity was given in Section 3.2. Despite geometry-guided ideas have been explored directly in previous works, we are the first one to embed geometry information in contrastive learning to maintain continuity into cGANs.*

**Q:** The experimental part is not strong.

**A:** *a) The detailed architecture and the key parameters are provided in supplementary materials. This work can be easily reproduced and code will be available upon publication. b) Two types of baselines have been compared. First, a standard one-hot code based cGAN CDAAE [34] and also an ablation study have been conducted in Section 4.2. Second, we compare our method with two state-of-the-art methods PG<sup>2</sup> [18] and G2GAN [23] in Section 4.4. Three different benchmarks have been tested both qualitatively and quantitatively.*

**Q:** The analysis of geometry-contrastive learning.

**A:** *Firstly, we evaluate the Lipschitz continuity of generated faces over facial expression conditions in Section 4.3, and the quantitative result is given in Figure 3. Besides, the embedded manifold is visualized qualitatively in Figure 4 for intuitive interpretation. Secondly, in Section 4.4, we mainly demonstrate how geometry-contrastive learning handles the misalignments between different persons and the generalization on the spontaneous emotion states.*

**Q:** Some related and important references are missing about GANs.

**A:** *Paper (c) has been cited as [18] in our paper. Paper (b) hasn't been officially published before the submission deadline and doesn't get much improvements compared with paper (c). Paper (a) concentrates on face frontalization and pose-invariant face recognition, which doesn't have close relationship with our work.*