Facial Expression Translation Using Landmark Guided GANs

Hao Tang and Nicu Sebe

Abstract—We propose a simple yet powerful Landmark guided Generative Adversarial Network (LandmarkGAN) for the facial expression-to-expression translation using a single image, which is an important and challenging task in computer vision since the expression-to-expression translation is a non-linear and non-aligned problem. Moreover, it requires a high-level semantic understanding between the input and output images since the objects in images can have arbitrary poses, sizes, locations, backgrounds, and self-occlusions. To tackle this problem, we propose utilizing facial landmark information explicitly. Since it is a challenging problem, we split it into two sub-tasks, (i) category-guided landmark generation, and (ii) landmark-guided expression-to-expression translation. Two sub-tasks are trained in an end-to-end fashion that aims to enjoy the mutually improved benefits from the generated landmarks and expressions. Compared with current keypoint-guided approaches, the proposed LandmarkGAN only needs a single facial image to generate various expressions. Extensive experimental results on four public datasets demonstrate that the proposed LandmarkGAN achieves better results compared with state-of-the-art approaches only using a single image.

Index Terms—GANs, facial landmark, facial expression generation, image-to-image translation

1 INTRODUCTION

In this paper, we mainly study how to enable machines to perform the facial expression-to-expression translation task, which is a classic task in computer vision. This task has many applications such as human-computer interactions, entertainment, and virtual reality. Another important benefit of this task is that they can augment training data by generating images with given input images, which could be employed to improve the expression recognition accuracy. However, this task is difficult since (i) it needs to handle complex backgrounds with different illumination conditions, objects, and occlusions. (ii) it needs a high-level semantic understanding of the mapping between the input image and the output image since the objects in the inputs images can have arbitrary poses, sizes, locations, backgrounds, and self-occlusions.

Recently, Generative Adversarial Networks (GANs) [1] have shown the potential to solve this challenging task, and it can be utilized, for example, to convert a face with a neutral expression into different expressions. GANs have demonstrated promising results in many generative tasks such as photo-realistic image generation [2], [3], [4], [5], [6], [7]. Moreover, impressive image-to-image translation results have been obtained by using Conditional GAN (CGAN) [8], in which a conditional image is taken as input and the model outputs another image with a different style. Given $m$ image domains, Pix2pix [8] needs to train $m(m-1)$ models, which is inefficient and ineffective. Recently, ComboGAN [9] and StarGAN [10] are proposed to solve multi-domain image-to-image translation problem. ComboGAN [9] requires $m$ models and StarGAN [10] only needs to train one model. However, these methods cannot handle some specific image translation tasks such as pose generation [11], [12] and gesture generation [13] since person/gesture can have arbitrary poses, sizes, appearances, and locations in the wild, leading to infinity image domains.

To address these limitations, several works have been proposed to generate images based on object keypoint or human skeleton. For instance, Reed et al. [14] proposed the Generative Adversarial What-Where Network (GAWWN), which generates birds conditioned on both text descriptions and object location. Reed et al. [15] presented an extension of Pixel Convolutional Neural Networks (PixelCNN) to generate images part keypoints and text descriptions. Ma et al. [11] proposed a two-stage reconstruction pipeline that generates novel person images. Korshunova et al. [16] used facial keypoints to define the affine transformations of the alignment and realignment steps for face swap.

Unlike these methods, we focus on facial expression-to-expression tasks. Wei et al. [17] proposed a Conditional Multi-Mode Network (CMM-Net) for landmark-guided smile generation. Di et al. [18] proposed the Gender Preserving Generative Adversarial Network (GPGAN) to synthesize faces based on facial landmarks. Qiao et al. [19] presented the Geometry-Contrastive Generative Adversarial Network (GCGAN) to generate facial expression conditioned on geometry information of facial landmarks, which explicitly employs the facial landmark to control the appearances and locations of the facial action units. In the task, the facial landmark is especially useful because different expressions

---

*Hao Tang is with the Department of Information Technology and Electrical Engineering, ETH Zurich, 8092 Zurich, Switzerland. E-mail: haotang@vision.ee.ethz.ch.*

*Nicu Sebe is with the Department of Information Engineering and Computer Science (DISI), University of Trento, 38123 Trento, Italy. E-mail: sebe@disi.unitn.it.*

Manuscript received 18 April 2022; revised 5 September 2022; accepted 5 September 2022. Date of publication 15 September 2022; date of current version 15 November 2022.

This work was supported in part by the Italy-China Collaboration Project TALENT and in part by the EU H2020 Project AH4Media under Grant 951911.

(Corresponding author: Hao Tang.)

Digital Object Identifier no. 10.1109/TAFFC.2022.3207007
have different facial action units that have different shapes of mouths, lips, eyes, and face contours. The expressions are mainly encoded by the landmarks. In other words, different expressions produce different landmark shapes w.r.t. the eyes, eyebrows, and lips. Therefore, we explicitly employ landmarks as the guidance for expression translation which can embed the locations of the eyes, eyebrows, and lips into the conditional image.

But during the testing period the aforementioned methods employ a keypoint detector to extract the landmarks of the target object, which introduces an extra module and then reduces the flexibility of the proposed system. To overcome these difficulties, we propose a novel LandmarkGAN in this paper which can generate landmarks and images without any keypoint detectors during the testing stage, which thus can be used in practical applications with less network overhead.

Specifically, we propose a novel LandmarkGAN framework, as shown in Fig. 1. LandmarkGAN comprises two generation stages, (i) category-guided landmark generation and (ii) landmark-guided expression-to-expression translation. In the first stage, we employ a CGAN model to generate the targeted object landmark conditioned on the object label and the input image. We can also generate the targeted object landmark conditioned on the object label and the landmark of the input image. However, this requires an additional object landmark detector to detect the landmarks of the input image. Therefore, instead of using the detected landmarks of the input image, we use the input image directly as the conditioning prior. To further generate more accurate landmarks, we use the Mean Squared Error (MSE) loss to calculate the errors on the coordinates of the landmarks.

In the second stage, the generated object landmarks and the conditional image are concatenated together as the input to a U-Net [8] to generate the targeted image with the corresponding facial expression. Note that in the first stage, facial expressions are represented by one-hot vectors in which there is only one element set to 1 to represent the corresponding expression while the others are set to 0. First, the one-hot vector is encoded by the fully connected layers and then the encoded label representations are concatenated with the image representations to generate the target facial landmarks. Two stages are trained jointly by an end-to-end fashion, in this way, the generated landmark can help generate more photo-realistic faces, and high-quality faces then improve the generation of landmarks.

Compared with state-of-the-art keypoint-guided methods such as C2GAN [12] which need image and keypoint as input during the testing stage, the proposed LandmarkGAN only needs a single face image to generate diverse landmarks and faces. Extensive experiments on several public datasets demonstrate that LandmarkGAN can generate better results compared with state-of-the-art methods. Thus, these generated images by the proposed method can augment the training data and improve the performance of the facial expression classifier.

Overall, the contributions of this paper are summarized as follows:

- We propose a novel LandmarkGAN model for facial expression-to-expression translation tasks. LandmarkGAN is comprised of two sub-modules, one for category-guided landmark generation and the other one for landmark-guided expression generation. Compared with current methods that need images and landmarks as input during the testing stage, the proposed LandmarkGAN only needs a single image to generate faces with different novel expressions.
- The proposed LandmarkGAN is trained in an end-to-end fashion that aims to enjoy the mutually improved benefits from the generated landmarks and expressions. To the best of our knowledge, we are the first to be able to make one GAN framework work on both facial expression image generation and facial landmark generation tasks.
- Experimental results on four challenging datasets demonstrate the superiority of LandmarkGAN over
the state-of-the-art models on the unconstrained facial expression-to-expression task. Moreover, the generated images have high-quality and preserve the identity information, and thus they can be used to boost the performance of facial expression classifiers. The code will be released upon the acceptance of the paper at https://github.com/Ha0Tang/LandmarkGAN

2 RELATED WORK

Generative Adversarial Networks (GANs) [1] have shown the capability of generating high-quality images [2], [3], [4], [5], [6], [7], [20], [21], [22]. GANs have two important components, i.e., generator G and discriminator D. The goal of G is to generate photo-realistic images from a noise vector, while D tries to distinguish between the real image and the image generated by G. Although it is successful in many tasks, it also has challenges such as how to generate images of high visual fidelity and how to control the position and the shape of generated objects.

To generate some specific kind of images given the input for users’ goal, Conditional GAN (CGAN) [23] is proposed. A CGAN always combines a basic GAN and an external information, such as labels [24], [25], [26], text descriptions [14], [27], [28], segmentation maps [29], [30], [31], [32], [33], [34], [35], [36], [37], [38] and images [8], [39]. For example, GANmut [24] introduces a novel GAN-based framework that learns an expressive and interpretable conditional space to generate a gamut of emotions, using only the categorical emotion labels.

Image-to-Image Translation frameworks use input-output data to learn a parametric mapping between inputs and outputs, e.g., Isola et al. [8] proposed Pix2pix, which uses a CGAN to learn a translation function from input to output image domains. Zhu et al. [39] introduced CycleGAN, which achieves unpaired image-to-image translation using the cycle-consistency loss. However, existing image-to-image translation models (e.g., CycleGAN [39], DiscoGAN [40], DualGAN [41], Pix2pix [8], ComboGAN [9], StarGAN [10], AttentionGAN [42]) are inefficient and ineffective. These approaches cannot handle some specific image generation tasks such as pose [11] and gesture [13] generation, which could have infinity image domains since person/gesture can have arbitrary poses, sizes, appearances and locations in the wild.

Keypoint-Guided Image-to-Image Translation. To address these limitations, several works have been proposed to generate images based on object keypoint [43], [44], [45], [46], [47], [48], [49], [50], [51], [52]. For instance, Di et al. [18] proposed GPGAN to synthesize faces based on facial landmarks. Reed et al. [14] proposed GAWWN, which generates birds conditioned on both text descriptions and object location. Ma et al. proposed PG2 [11], which achieves person image translation using a conditional image and a target pose image. Sun et al. [53] proposed a two-stage framework to perform head inpainting conditioned on the generated facial landmark in the first stage. Korshunova et al. [16] used facial keypoints to define the affine transformations of the alignment and realignment steps for face swap. Wei et al. [17] proposed a Conditional MultiMode Network (CMM-Net) for landmark-guided smile generation. Qiao et al. [19] presented GCGAN to generate facial expression conditioned on geometry information of facial landmarks. Song et al. [54] proposed G2GAN for facial expression synthesis guided by fiducial points. These methods employ object keypoints to guide the image generation process since the object keypoints provide four types of information for generation at the same time, i.e., category, location, scale, and orientation of objects.

However, existing approaches such as [11], [12], [19], [54] employ a keypoint detector to generate keypoints during the testing stage, which reduces the flexibility of the proposed system or method. In this paper, we propose the LandmarkGAN which can generate facial landmarks and facial expressions without any keypoint detectors, which thus can be used in practical applications with less network overhead. Moreover, our method can generate both facial landmarks and images (see Fig. 8), which is not investigated in both existing methods [11], [12], [19], [54].

3 LANDMARKGAN FORMULATION

The goal of GANs is to learn a mapping from random noise z to output image y, G: z → y [1]. The generator G is trained to generate images that “fools” the discriminator D. The discriminator D, which is optimized to distinguish real from fake images. The objective of the original GAN can be formalized as follows:

\[ L_{GAN}(G, D) = E_{y}(\log D(y)) + E_{z}(\log (1-D(G(z)))) \]  

(1)

While, for conditional GANs, which learn the mapping with a conditional image x, i.e., G: [x, z] → y. To utilize x during training phase, the objection can be updated to conditional version:

\[ L_{cGAN}(G, D) = E_{x,y}(\log D(x, y)) + E_{x,z}(\log (1-D(x, G(x, z)))) \]  

(2)

where D tries to maximize this objective while G tries to minimize it. Thus, the solution is \( G^* = \arg \min_{G} \max_{D} L_{cGAN}(G, D) \). For our LandmarkGAN, which comprises two stags, i.e., facial landmark and facial expression generation (see Fig. 1).

3.1 Stage I: Facial Landmark Generation

During the first stage, the translation mapping can be formulated as \( G^*_{l} : [x, z_1, l] → l \), where \( x \) denotes the input image, \( z_1 \) represents a random noise, \( l \) denotes expression label, and \( l = \sum_{i=1}^{68}(p_i, q_i) \) denotes the target landmark. For generating training data, we employ OpenFace [55] as facial landmark detector to extract 68 facial landmarks, and then encode these points as a color image. Specifically, each pixel in a radius of \( S_r = 4 \) around the corresponding facial landmark is filled with the color in the point of the corresponding image \( x \) and 255 (white color) elsewhere. This embedding strategy ensures that the model to learn both locations and colors information of facial landmarks. In this way, the facial landmarks not only provide the clues regarding facial shapes to the generator but also provide the color information of different facial parts to the generator. Then, the objective of first stage can be formulated as follows,
\[ \mathcal{L}_{\text{GAN}}(G_l, D_l) = E_x[\log D_l(x)] + E_{x, z_1}[\log (1 - D_l(G_l(x, z_1, l_e)))] , \]  

(3)

where \( l_e \) is the one-hot vector of expressions and \( l_e \) is connected with the image embedding at the bottleneck fully connected layer.

In addition, some prior works have verified it favorable to combine the GAN loss with a more traditional reconstruction loss, such as \( L_1 \) loss [8] and \( L_2 \) loss [56]. Both \( L_1 \) loss [8] and \( L_2 \) loss [56] are calculated based on images, while for improving the accuracy of the generated landmarks, we propose the landmark reconstruction loss. Unlike the general MSE loss that calculates on the pixels of the image, the landmark reconstruction loss calculates on the coordinates of the landmarks, which can be expressed as,

\[ \mathcal{L}_{\text{Landmark}}(G_l) = E_{x, l} \left[ \frac{1}{68} \sum_{i=1}^{68} \sqrt{(p_i - \tilde{p}_i)^2 + (q_i - \tilde{q}_i)^2} \right] , \]  

(4)

where \( \tilde{l} = G_l(x, z_1, l_e) = \sum_{i=1}^{68} (\tilde{p}_i, \tilde{q}_i) \), \( p \) and \( q \) represents the \( x \) and \( y \) coordinate of the landmark \( i \), respectively.

Thus, the final loss of the first stage is,

\[ \mathcal{L}_1 = \mathcal{L}_{\text{GAN}}(G_l, D_l) + \lambda_1 \mathcal{L}_{\text{Landmark}}(G_l) , \]  

(5)

where the parameter \( \lambda_1 \) controls the relative importance of both \( \mathcal{L}_{\text{GAN}}(G_l, D_l) \) and \( \mathcal{L}_{\text{Landmark}}(G_l) \).

### 3.2 Stage II: Facial Expression Generation

Facial expression generation is conditioned on the landmark image \( G_l(x, z_1, l_e) \) generated in the first stage, a conditional image \( x \), and expression label \( l_e \). We first define \( x, G_l(x, z_1, l_e) = \tilde{x} \), thus the adversarial objective of the second stage can be expressed as,

\[ \mathcal{L}_{\text{GAN}}(G_e, D_e) = E_{x, y}[\log D_e(x, y)] + E_{x, z_2}[\log (1 - D_e(x, G_e(\tilde{x}, z_2)))] , \]  

(6)

in this way, the input image and the output image can be better aligned. Our model only needs to pay attention to the areas related to the expression change, and not to the difference unrelated to the expression change, thereby reducing the difficulty of learning.

Next, the face reconstruction loss of generator \( G_e \), measured by \( L_1 \) and \( L_2 \) distances can be defined as \( \mathcal{L}_{\text{L1}}(G_e) = E_{x, y, z_2}[||y - G_e(\tilde{x}, z_2)||_1] \) and \( \mathcal{L}_{\text{L2}}(G_e) = E_{x, y, z_2}[||y - G_e(\tilde{x}, z_2)||_2] \), respectively.

In this paper, we also explore an alternative option and propose a robust version loss in Fig. 2, i.e., \( \mathcal{L}_{\text{L1,2}} \), which takes both \( L_1 \) and \( L_2 \) losses into consideration simultaneously. The proposed \( \mathcal{L}_{\text{L1,2}} \) loss is less sensitive to outliers than the \( L_2 \) loss, and in some cases it could prevent exploding gradients [57]. Moreover, we observe that the \( \mathcal{L}_{\text{L1,2}} \) loss can provide more image details than the \( L_1 \) loss in the preliminary experiments. Thus, the definition of the proposed \( \mathcal{L}_{\text{L1,2}} \) is:

\[ \mathcal{L}_{\text{L1,2}}(G_e) = \begin{cases} 0.5 \mathcal{L}_{\text{L2}}(G_e) & \text{if } ||y - G_e(\tilde{x}, z_2)||_1 < 1, \\ \mathcal{L}_{\text{L1}}(G_e) - 0.5 & \text{if } ||y - G_e(\tilde{x}, z_2)||_2 \geq 1. \end{cases} \]  

(7)

Note that the proposed \( L_{1,2} \) loss is similar to the Huber loss, but we are the first to use this loss in the generation task.

Moreover, to preserve the face identity while expression synthesis, we propose the face identity distance loss, which can be expressed as follows,

\[ \mathcal{L}_{\text{identity}}(G_e) = E_{x, y, z_2}[||F(y) - F(G_e(\tilde{x}, z_2))||_1] , \]  

(8)

where \( F \) is a face feature extractor. We employ [58] to compare the identity distance between the target face \( y \) and the generated face \( G_e(\tilde{x}, z_2) \), and try to minimize the distance \( F(y) - F(G_e(\tilde{x}, z_2)) \) using \( L_1 \) distance. The model proposed in [58] is pre-trained as a classifier to distinguish between tens of thousands of identities, so it has the ability to capture the most prominent feature for face identity discrimination.

Therefore, our overall objective of the second stage is:

\[ \mathcal{L}_2 = \mathcal{L}_{\text{L1,2}}(G_e) + \lambda_2 \mathcal{L}_{\text{L1,2}}(G_e) + \lambda_3 \mathcal{L}_{\text{identity}}(G_e) , \]  

(9)

where the parameter \( \lambda_2 \) and \( \lambda_3 \) control the relative importance of \( \mathcal{L}_{\text{L1,2}}(G_e) \) and \( \mathcal{L}_{\text{identity}}(G_e) \). Thus, the full objective of the proposed LandmarkGAN is \( \mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2 \).

During the training stage, the full objective is optimized by an end-to-end fashion that aims to enjoy the mutually improved benefits from both stages.

In addition, in the first stage, the discriminator \( D_l \) tries to distinguish landmark \( l \) from \( G_l(x, z_1, l_e) \). In the second stage, the discriminator \( D_e \) learns to classify fake \( (G_e(\tilde{x}, z_2), x) \) and real \( (y, x) \) pairs, as shown in Fig. 1. Thus, the loss functions for the discriminator \( D_l \) and \( D_e \) are:

\[ \mathcal{L}(D_l) = \frac{1}{2} [\mathcal{L}_{\text{BCE}}(D_l(x, l_e, 1)) + \mathcal{L}_{\text{BCE}}(D_l(G_l(x, z_1, l_e), 0))], \]  

and

\[ \mathcal{L}(D_e) = \frac{1}{2} [\mathcal{L}_{\text{BCE}}(D_e(x, y, 1)) + \mathcal{L}_{\text{BCE}}(D_e(G_e(\tilde{x}, z_2), x), 0)] , \]  

respectively, where \( \mathcal{L}_{\text{BCE}} \) represents the binary cross entropy loss function between the target and the output. We also divide the objective of discriminators by 2 to slow down the rate of discriminators \( D_l \) and \( D_e \) relative to generators \( G_l \) and \( G_e \). Moreover, Isola et al. [8] have found that the noise \( z_1 \) and \( z_2 \) are not necessary in Pix2pix, thus in our LandmarkGAN, we follow Pix2pix [8] and provide noise only through dropout operation in generators.

### 3.3 Implementation Details

**LandmarkGAN Architecture.** We employ U-Net [8] for generators \( G_l \) and \( G_e \). U-Net actually is a network with skip connections between encoder and decoder, as shown in Fig. 1. Batch normalization [59] is employed except in the first
layer of the encoders. Leaky ReLU is used in the encoders with a slope of 0.2. While all ReLUs in the decoders are not leaky. And the last layer of the decoders is the Tanh activation function. Moreover, the first three blocks of the decoders use the Dropout layer with a dropout rate of 50%. For discriminators $D_l$ and $D_r$, we adopt PatchGAN proposed in [8], PatchGAN tries to classify if each $70 \times 70$ patch in an image is real or fake. The final layer of discriminators employs a Sigmoid activation function.

**LandmarkGAN Optimization.** We follow the same optimization method in [1] to optimize the proposed LandmarkGAN, i.e., one gradient descent step on discriminator and generator alternately. We first train $G_l$ and $D_l$ with $G_r$ and $D_r$ fixed, and then train $G_r$ and $D_r$ with $G_l$ and $D_l$ fixed. The proposed LandmarkGAN is trained by an end-to-end fashion. During the training stage, we employ OpenFace [55] to extract facial landmarks as training data. We noticed that as the landmark detection accuracy decreases, the performance of the proposed LandmarkGAN decreases very slowly. This is because our model is also able to generate the correct landmark under the guidance of the expression label in the first stage. At the same time, in the second stage, our model can also generate the target image under the guidance of expression label. The most important thing is that we have the ground truth image, so it can correctly guide the model to learn by using the proposed face reconstruction loss, face identity distance loss, and adversarial loss.

While during the testing stage, we use the trained generator $G_1$ as a facial landmark detector to generate the target landmark, which means LandmarkGAN does not need extra modules, while PG2 [11] and C2GAN [12] need to employ extra module to produce landmark in order to generate variety expressions. All images are scaled to $256 \times 256$. Each model is trained 200 epochs with batch size 1. We employ the Adam [60] with momentum terms $\beta_1=0.5$ and $\beta_2=0.999$ as our solver. The initial learning rate for Adam is 0.0002.

We set $S_l=4$, $\lambda_1=2$, $\lambda_2=100$ and $\lambda_3=0.1$ according to our ablation experiments. For better training the landmark generator, we adopt the pre-trained PDM decoder used in [53]. We also employ dual discriminators as in Nguyen et al. [61] and GestureGAN [13], which have demonstrated that they improve the ability of discriminator to generate more photo-realistic images. The proposed LandmarkGAN is implemented by deep learning framework PyTorch [62].

## 4 EXPERIMENTS

### 4.1 Experimental Setups

**Datasets.** To validate the effectiveness of the proposed LandmarkGAN, we conduct experiments of unconstrained facial expression-to-expression translation on four public datasets (i.e., JAFFE [63], 3DFE [64], FERG [65] and RaFD [66]) for four different purposes.

1. We evaluate on the Japanese Female Facial Expression (JAFFE) dataset [63] to see if our model still works with limited training samples, which will inspire few-shot image generation tasks. The JAFFE dataset [63] contains 271 images of 6 basic facial expressions and 1 neutral posed by 10 Japanese females, and each subject has 3 images except one has 4 images. We randomly select 2 subjects as testing data and the remaining 8 people as training data.

2. The evaluation on the 3D Facial Expression (3DFE) dataset proves that our method has a good ability to control the intensity of expression. The 3DFE dataset [64] includes 100 subjects with 7 expressions and each expression has 4 levels of intensity, thus this dataset has 100 subjects x 7 expressions x 4 images = 2,800 images in total. For each expression, we randomly choose 10 subjects as testing data and the remaining 360 images as training samples.

3. Our evaluation on the Facial Expression Research Group (FERG) dataset shows that our method is also effective on cartoon images. The FERG dataset [65] contains about 50,000 images modeled using the MAYA software. We conduct our experiments on 1,100 images out of 1,200 images as testing data and the rest of the 1,100 images as training samples.

4. Our evaluation on the Radboud Faces (RaFD) dataset proves that our method is more effective than the existing methods such as GANimation [67] and C2GAN [12] since both use RaFD as a standard dataset. The RaFD dataset [66], which has 8,000 color face images collected from 67 subjects with eight different emotional expressions, i.e., anger, fear, disgust, sadness, happiness, surprise, neutral and contempt. Each emotion has 1,005 images captured from five cameras with different angles and each subject is asked to show three different gaze directions. For each emotion, we randomly select 67% images as training data and the rest 33% images as testing data. Similar to C2GAN [12], all the images are rescaled to $256 \times 256 \times 3$ without any preprocessing in all experiments since we aim to conduct unconstrained facial expression-to-expression translation task.

These datasets contain faces with different races and styles, and they have different illuminations, occlusions, pose conditions, and backgrounds.

**Evaluation Metrics.** We employ Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) [68], Inception Score (IS) [69], and LPIPS [70] to measure the quality of generated images. More specifically, PSNR measures the difference of two images from pixel level, while SSIM evaluates the similarity of two images from global level. Both IS and LPIPS estimate generated images from a higher semantic level. For all metrics except LPIPS, higher is better.

### 4.2 Experimental Results

**Ablation Study.** We first conduct an ablation study of the proposed LandmarkGAN and show the results without the Double discriminators strategy (D), face Identity distance loss (I), and Landmark Reconstruction loss (LR), respectively. We use the scheme of training a dual-discriminator instead of one discriminator as a more stable way to improve the capacity of discriminators similar to [71]. To be more specific, the dual-discriminator architecture can better approximate the optimal discriminator. If one of the discriminators is trained to be far superior over the generators, the generators can still receive instructive gradients from the other one. Table 1 shows the results using different components of LandmarkGAN on the JAFFE, 3DFE, and FERG datasets.
datasets. We observe that removing any terms from I, FR, LR, and D can degrade the generation performance. Thus, we conclude that these items are important for generating photo-realistic images.

Moreover, we compare the proposed $L_{1,2}$ with $L_1$ and $L_2$ losses in Table 1. We observe that the proposed $L_{1,2}$ achieves better results than $L_1$ and $L_2$ losses.

Also, in the proposed LandmarkGAN framework, the expression label is provided in the two generators, i.e., facial landmark generator $G_l$ and facial expression generator $G_e$. To evaluate the influence of the expression label, we test three different variants in Table 1, i.e., the expression label $l_e$ is only provided to the facial landmark generator (i.e., $l_e \rightarrow G_l$), the expression label $l_e$ is only provided to the facial expression generator (i.e., $l_e \rightarrow G_e$), the expression label $l_e$ is provided to both generators (i.e., $l_e \rightarrow G_l$ & $l_e \rightarrow G_e$). As we can see in Table 1, both $l_e \rightarrow G_l$ and $l_e \rightarrow G_l & l_e \rightarrow G_e$ achieve almost the same performance, which proves that the emotion information mainly lies in the generated landmarks in the first stage, further validating the effectiveness of the proposed method. Moreover, compared with $l_e \rightarrow G_l$ & $l_e \rightarrow G_e$, the method $l_e \rightarrow G_e$ achieves much worse results. The reason for this is that the expression label is not provided in the first stage, resulting in unsatisfactory landmarks generated in the first stage, which in turn leads to unsatisfactory images produced in the second stage.

Lastly, the landmark based features consists in “soft-crops” of the image around each landmark location. Therefore, we also provide an ablation study in Table 1 to show that this method (i.e., “w/ soft-crops”) achieves better results than simply giving image and landmarks to the network (i.e., “w/o soft-crops”).

### Table 1

Quantitative Results of Different Components on JAFFE, 3DFE, and FERG

| Setting                          | PSNR ↑ | SSIM ↑ | IS ↑ |
|---------------------------------|--------|--------|------|
| LandmarkGAN - D - I - LR        | 16.8467| 22.3147| 32.9560 |
| LandmarkGAN - D - I             | 17.2185| 22.8473| 36.4789 |
| LandmarkGAN - LR                | 17.1502| 22.8186| 34.1283 |
| LandmarkGAN - I                 | 17.2613| 23.1719| 37.7746 |
| LandmarkGAN - D                 | 17.2852| 23.8509| 38.8294 |
| LandmarkGAN ($l_{1,2}, l_e \rightarrow G_l & l_e \rightarrow G_e$, w/ soft-crops) | 17.2911| 23.8754| 38.9132 |
| LandmarkGAN ($L_1, l_e \rightarrow G_l & l_e \rightarrow G_e$, w/ soft-crops) | 17.1892| 23.4216| 38.3760 |
| LandmarkGAN ($L_2, l_e \rightarrow G_l & l_e \rightarrow G_e$, w/ soft-crops) | 17.2343| 23.6743| 38.7598 |
| LandmarkGAN ($L_{1,3}, l_e \rightarrow G_l$, w/ soft-crops) | 17.2807| 23.8532| 38.8750 |
| LandmarkGAN ($L_{1,3}, l_e \rightarrow G_l$, w/ soft-crops) | 17.0156| 23.6424| 37.5791 |
| LandmarkGAN ($L_{1,3}, l_e \rightarrow G_l$, w/o soft-crops) | 17.0726| 23.6752| 38.6709 |

Feature Visualization. We use t-SNE [72] to visualize the feature learned from a pre-trained VGG-19 model [73], which trained on a subset of the ImageNet dataset. The feature is extracted after “fc7” fully connected layer, which is a 4,096-dim vector. The results on the JAFFE, 3DFE, and FERG datasets are shown in Fig. 3. Note that most generated samples are well separated, which demonstrates our LandmarkGAN can preserve identity information when converting facial expressions.

Moreover, we conduct face verification experiments on the RaFD dataset. The results of CMM-Net [17], GANimation [67], GANmut [24], AttentionGAN [42], C2GAN [12], and the proposed LandmarkGAN are 0.902, 0.944, 0.948, 0.952, 0.968, and 0.981, respectively. These results show that the proposed method can better preserve the identity during translation.

Data Augmentation. To show the generated images are useful for improving the performance of the facial expression recognition task, we employ the generated images as extra data for data augmentation purpose. The intuition is that if the generated images are realistic, (i) the classifiers trained on both the real images and the generated images will be able to boost the accuracy of the real images (in this situation, the generated images work as augmented data.)

![Fig. 3. Identity feature space of JAFFE (left), 3DFE (middle), and FERG (right). Each color represents a subject.](image-url)
and (ii) the classifiers trained on real images will also be able to classify the synthesized image correctly. The results are listed in Table 2. Note that “Real+Nor./Real” means that the images produced by normal augmentation methods like adding noises, random rotation and crop for a fair comparison. In this way the total number of images in training set should be same with the setting of “Real+Syn./Real”. We can see that the recognition performance is boosted by adding the generated images by our method on all datasets.

Control of Expression Intensity. Fig. 4 shows the results with different intensity levels of expressions on the 3DFE dataset. For example, in the last column of Fig. 4, we can judge the intensity levels of the generated images by the size of the mouth opening. We present four levels of expression intensity from weak to strong. To generate expressions with different intensities, Eq. (3) needs to be updated to,

\[
\mathcal{L}_{GAN}(G_l, D_l) = \mathbb{E}_l[\log D_l(l)] + \mathbb{E}_{x,z_i}[\log (1 - D_l(G_l(x, z, l, l_i)))],
\]

where \(l_i\) is the one-hot label of expression intensity. Moreover, \(\bar{x}\) in Eq. (6) needs to be updated to \(\bar{x} = [x, G_l(x, z, l, l_i)]\).

Comparison With State-of-the-Art Methods. We compare the proposed LandmarkGAN with the most related facial expression generation models, i.e., CMM-Net [17], C2GAN [12], GANimation [67], AttentionGAN [42], and GANmut [24]. Note that C2GAN needs input a image and the target landmarks for generating the desired expressions. For GANimation, which adopts facial action units (AUs) to guide the image generation process, AUs are the local components of the face which show muscle deformations. AttentionGAN is an unsupervised method, which can identify the most discriminative foreground contents and minimize the change of the background. GANmut can generate a gamut of emotions by using only the categorical emotion labels. Specifically, we re-implemented the GANimation, GANmut, AttentionGAN, C2GAN on all the three JAFFE, 3DFE, and FERG datasets with the same training/inference configuration in their papers for fair comparisons.

As we can see from Table 3 that the proposed LandmarkGAN consistently achieves better performance than C2GAN, GANimation, GANmut, and AttentionGAN on all the JAFFE, 3DFE, and FERG datasets. We also show the visualization results compared to the state-of-the-art method, i.e., C2GAN,
in Fig. 5. We see that the proposed LandmarkGAN generate much better results than C2GAN on all the three datasets, which validates the effectiveness of our method. Moreover, the results of the RaFD dataset are shown in Fig. 6 and Table 4. We can see that the proposed LandmarkGAN achieves better results compared with the leading baselines, which validates...
the effectiveness of the proposed method. Finally, in Fig. 7 we show that the proposed LandmarkGAN can generate other different expressions by inputting only a anger or a disgust face. The advantage of our proposed LandmarkGAN over these existing frameworks is that our framework is two-stage, the landmark generated in the first stage will help the second stage to generate a better image, and conversely, the image generated in the second stage will in turn help the first stage to generate better landmark. These two stages are trained in an end-to-end fashion so that we can improve the generation results of the target landmark and image jointly.

**User Study.** We follow the same setting in [8], [39] and conduct an perceptual studies to evaluate the quality of the generated images. Results are shown in Table 5. We can see that the proposed LandmarkGAN consistently achieves better results compared with the leading method (i.e., C2GAN), which further validates the images generated by our method are more clear and visually plausible.

The metrics such as PSNR and LPIPS calculate the semantic distance between the real image and the generated image. The semantic distance between the images is not very large considering they are all faces. On the contrary, the AMT score in Table 5 measures the generation quality from a perceptual level. The difference on the perceptual level is more obvious than on the semantic level, i.e., the generated images with small artifacts show a minor difference on the semantic level, while are being judged with a significant difference from the real images by humans.

**Visualization of Landmark Generation.** We also compare landmark generation with [53]. Similar to [53], we employ $L_2$ distance between detected and generated landmarks as the evaluation metric. Quantitative results are provided in Table 6. Note that the generated landmarks have a lower $L_2$ distance than [53] as our multi-task learning strategy. In Fig. 8, we show some samples of the generated landmarks. We see that the generated

| Method         | JAFFE | 3DFE | FERG | RaFD |
|----------------|-------|------|------|------|
| Sun et al. [53] | 12.4  | 9.8  | 4.2  | 6.7  |
| LandmarkGAN (Ours) | **11.2** | **7.2** | **3.5** | **6.2** |

Fig. 7. More results of the proposed LandmarkGAN on RaFD. In LandmarkGAN, we only need to input an arbitrary expression face and thus LandmarkGAN will generate other expressions.
landmarks are very close to the target ones, which means LandmarkGAN can generate correct facial landmarks. Overall, the proposed Landmark can generate facial landmarks and expressions simultaneously, which is not embraced by any existing work.

Analysis of Failure Results. As can be seen from Fig. 9, our model cannot generate clear teeth, because we do not introduce any guidance in our framework to guide the generation of teeth. In future work, we will try to introduce the structural information of the teeth to generate realistic teeth. This could be very helpful for many generative tasks, such as image animation [74] and audio-to-image translation [75].

Applications. The proposed LandmarkGAN is a task-irrelevant framework, thus we also show the results on other three generative tasks, i.e., face colorization, face inpainting, and sketch-to-face translation. For face inpainting, images on 3DFE are re-scaled to $256 \times 256 \times 3$ first, then we set the mask size as $96 \times 96$ randomly to guarantee that at least one importance part of the face is missing. For sketch-to-face translation, we use a public software to convert images from the 3DFE dataset into sketches, then to learn the mapping between sketch and image. Results of the three tasks are shown in Fig. 10. We see that the proposed LandmarkGAN generates reasonable results on the three different generative tasks, validating the generalization ability of the proposed LandmarkGAN.
5 Conclusion

We propose a novel LandmarkGAN for facial expression translation. The proposed LandmarkGAN can generate faces with arbitrary expressions and expression intensities. The training of LandmarkGAN is comprised of two stages, (i) category-guided facial landmark generation and (ii) landmark-guided facial expression-to-expression translation. Moreover, two novel losses include the landmark GAN loss, and the $L_{1,2}$ loss are proposed to learn the expression-to-expression mapping. Experimental results demonstrate that the proposed LandmarkGAN is capable of generating higher quality faces with correct expression than the state-of-the-art approaches. Lastly, since the proposed LandmarkGAN is task agnostic, it can be employed to other generative tasks such as expression recognition, face colorization, face inpainting, and sketch-to-face translation.

In this article, we aim to generate basic emotions from a single image. Extending the proposed method to generate micro-expressions and dissimulated behavior is part of our further work.

REFERENCES

[1] I. Goodfellow et al., “Generative adversarial nets,” in Proc. Int. Conf. Neural Informat. Process. Syst., 2014, pp. 2672–2680.
[2] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive growing of GANs for improved quality, stability, and variation,” in Proc. Int. Conf. Learn. Representations, 2018.
[3] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, “High-resolution image synthesis and semantic manipulation with conditional GANs,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2018, pp. 8886–8895.
[4] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, “Semantic image synthesis with spatially-adaptive normalization,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 2332–2341.
[5] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, “Self-attention generative adversarial networks,” in Proc. Int. Conf. Mach. Learn., 2019, pp. 61–70.
[6] A. Brock, J. Donahue, and K. Simonyan, “Large scale GAN training for high fidelity natural image synthesis,” in Proc. Int. Conf. Learn. Representations, 2019.
[7] T. R. Shaham, T. Dekel, and T. Michaeli, “SinGAN: Learning a generative model from a single natural image,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 4569–4579.
[8] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2017, pp. 5967–5976.
[9] A. Aounaouche, E. Agustsson, R. Timofte, and L. Van Gool, “CombOGAN: Unrestricted scalability for image domain translation,” in Proc. Int. Conf. Learn. Representations, 2018, pp. 783–790.
[10] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, “StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2018, pp. 8789–8797.
[11] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool, “Pose guided person image generation,” in Proc. Int. Conf. Neural Informat. Process. Syst., 2017, pp. 405–415.
[12] H. Tang, D. Xu, G. Liu, W. Wang, N. Sebe, and Y. Yan, “Cycle in cycle generative adversarial networks for keypoint-guided image generation,” in Proc. 27th ACM Int. Conf. Multimedia, 2019, pp. 2052–2060.
[13] H. Tang, W. Wang, D. Xu, Y. Yan, and N. Sebe, “Gesturesgan for hand gesture-to-gesture translation in the wild,” in Proc. ACM Int. Conf. on Multimedia, 2018, pp. 774–782.
[14] S. E. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee, “Learning what and where to draw,” in Proc. Int. Conf. Neural Informat. Process. Syst., 2016, pp. 217–225.
[15] S. Reed, A. van roadway, N. Kalchbrunner, V. Bapst, M. Botvinick, and N. de Freitas, “Generating interpretable images with controllable structure,” in Proc. Int. Conf. Learn. Representations, 2017.
[40] T. Kim, M. Cha, H. Kim, J. Lee, and J. Kim, “Learning to discover cross-domain relations with generative adversarial networks,” in Proc. Int. Conf. Mach. Learn., 2017, pp. 1857–1865.

[41] Z. Yi, H. Zhang, P. T. Gong, and M. Gong, “DualGAN: Unsupervised dual learning for image-to-image translation,” in Proc. IEEE/CVF Int. Conf. Vis., 2017, pp. 2868–2876.

[42] H. Tang, H. Liu, D. Xu, P. H. Torr, and N. Sebe, “AttentionGAN: Unpaired image-to-image translation using attention-guided generative adversarial networks,” IEEE Trans. Neural Netw. Learn. Syst., to be published, doi: 10.1109/TNNLS.2021.3105725.

[43] Y. Ge et al., “FD-GAN: Pose-guided feature distilling GAN for robust person re-identification,” in Proc. Int. Conf. Neural Inform. Process. Syst., 2018, pp. 1230–1240.

[44] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in Proc. Int. Conf. Learn. Representations, 2015.

[45] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in Proc. IEEE/CVF Conf. Vis. Pattern Recognit., 2018, pp. 886–895.

[46] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, “Coding facial expressions with gabor wavelets,” in Proc. 3rd IEEE Int. Conf. Autom. Face Gesture Recognit., 1998, pp. 200–205.

[47] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, “A 3D facial expression database for facial behavior research,” in Proc. 7th Int. Conf. Autom. Face Gesture Recognit., 2006, pp. 211–216.

[48] H. Tang, H. Liu, D. Xu, P. H. Torr, and N. Sebe, “Unified generative adversarial networks,” in Proc. Int. Conf. Learn. Representations, 2016, pp. 1377–1388, 2010.

[49] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” IEEE Trans. Image Process., vol. 13, no. 4, pp. 600–612, Apr. 2004.

[50] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, “Improved techniques for training GANs,” in Proc. Int. Conf. Neural Inform. Process. Syst., 2016, pp. 2224–2232.

[51] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, “First order motion model for image animation,” in Proc. Int. Conf. Neural Inform. Process. Syst., 2019, pp. 7137–7147.

[52] S. Tulyakov, E. Ricci, and N. Sebe, “GANimation: Anatomically-aware facial animation from a single image,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 835–851.

[53] Hao Tang received the master’s degree from the School of Electronics and Computer Engineering, Peking University, China, and the PhD degree from the Multimedia and Human Understanding Group, University of Trento, Italy. He is currently a postdoctoral with Computer Vision Lab, ETH Zurich, Switzerland. He was a visiting scholar with the Department of Engineering Science, University of Oxford. His main research interests include deep learning, machine learning, and their applications to computer vision.

[54] Nicu Sebe is professor with the University of Trento, Italy, where he is leading the research in the areas of multimedia analysis and human behavior understanding. He was the General chair of the IEEE FG 2008 and ACM Multimedia 2013. He was a program chair of ACM Multimedia 2011 and 2007, ECCV 2016, ICCV 2017 and ICPAR 2020. He is a general chair of ACM Multimedia 2022 and a program chair of ECCV 2024. He is a fellow of IAPR.