Facial Expression Translation using Landmark Guided GANs

Hao Tang and Nicu Sebe

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.**

I. 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 MultiMode 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 (GC-GAN) 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 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

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

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

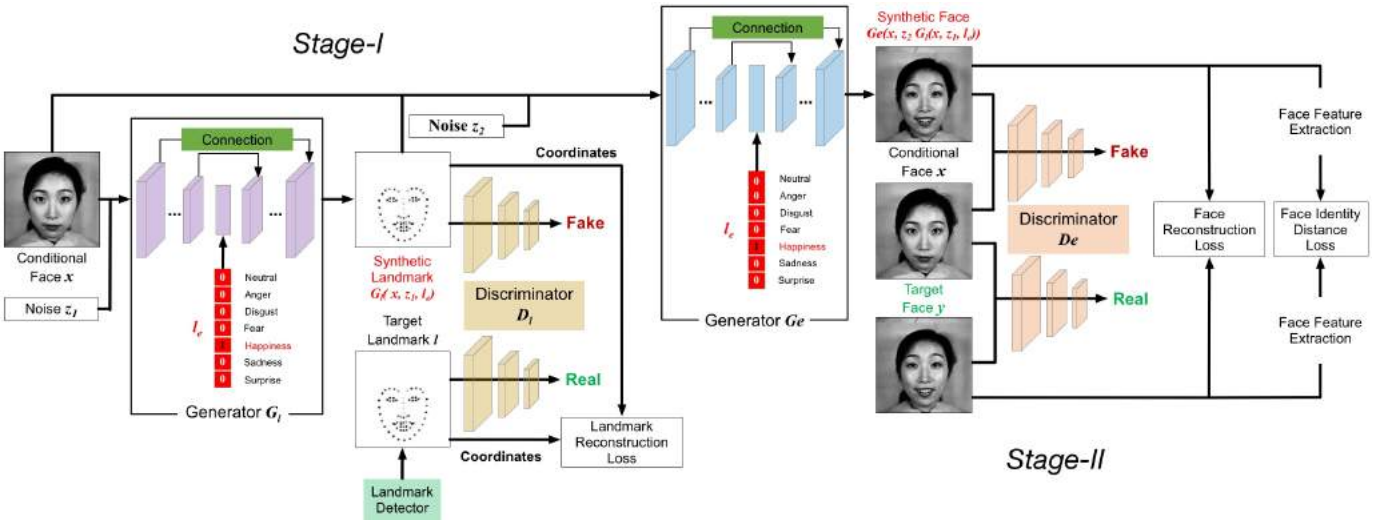

Fig. 1: The pipeline of the proposed LandmarkGAN. The input of stage-I is a face image and the expression labels, and the output is the landmark image of the target face. The inputs of stage-II are a conditional image and the synthetic landmark image generated in the stage-I, and the output is a face with the target expression but preserves identity information. The generators G_l and G_e try to generate landmark/image to fool the discriminator D_l and D_e, respectively. The discriminators D_l and D_e learn to classify fake $G_l(x, z_1, l_e)$ and real landmark l, and fake $(G_e(x, z_2, G_l(x, z_1, l_e), l_e), x)$ and real image pair (y, x), respectively.

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 Figure 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.

II. 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 trying 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 [19], [11], [12], [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 Figure 8), which is not investigated in both existing methods [19], [11], [12], [54].

III. LandmarkGAN Formulation

The goal of GANs is to learn a mapping from random noise z to output image y, $G{:}z{\mapsto}y$ [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:

$$\mathcal{L}_{GAN}(G,D){=}\mathbb{E}_y\left[\log D(y)\right]{+}\mathbb{E}_z\left[\log(1{-}D(G(z)))\right]. \quad (1)$$

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

$$\begin{aligned}\mathcal{L}_{cGAN}(G,D){=}&\mathbb{E}_{x,y}\left[\log D(x,y)\right]\\ &{+}\mathbb{E}_{x,z}\left[\log(1{-}D(x,G(x,z)))\right],\end{aligned} \quad (2)$$

where D tries to maximize this objective while G tries to minimize it. Thus, the solution is $G^*{=}\arg\min_G\max_D\mathcal{L}_{cGAN}(G,D)$. For our LandmarkGAN, which comprises two stags, i.e., facial landmark and facial expression generation (see Figure 1).

A. Stage I: Facial Landmark Generation

During the first stage, the translation mapping can be formulated as $G_l{:}[x,z_1,l_e]{\mapsto}l$, where x denotes the input image, z_1 represents a random noise, l_e 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 $S_l{=}4$ 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,

$$\begin{aligned}\mathcal{L}_{cGAN}(G_l,D_l){=}&\mathbb{E}_l\left[\log D_l(l)\right]\\ &{+}\mathbb{E}_{x,z_1}\left[\log(1{-}D_l(G_l(x,z_1,l_e)))\right],\end{aligned} \quad (3)$$

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.

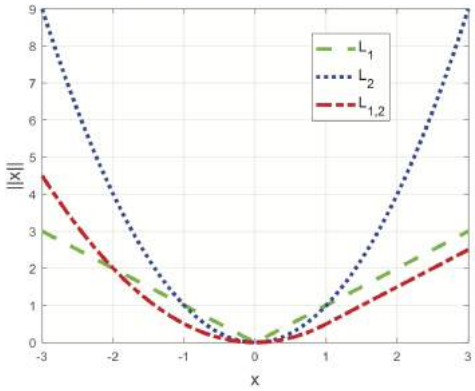

Fig. 2: Comparison of L_1, L_2 and $L_{1,2}$ losses on the image translation task.

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,

$$\mathcal{L}_{Landmark}(G_l)=\mathbb{E}_{l,\hat{l}}\left[\frac{1}{68}\sum_{i=1}^{68}\sqrt{(p_i-\hat{p}_i)^2+(q_i-\hat{q}_i)^2}\right], \tag{4}$$

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

Thus, the final loss of the first stage is,

$$\mathcal{L}_1=\mathcal{L}_{cGAN}(G_l,D_l)+\lambda_1\mathcal{L}_{Landmark}(G_l), \tag{5}$$

where the parameter λ_1 controls the relative importance of both $\mathcal{L}_{cGAN}(G_l,D_l)$ and $\mathcal{L}_{Landmark}(G_l)$.

B. 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 $[x,G_l(x,z_1,l_e),l_e]=\widehat{x}$, thus the adversarial objective of the second stage can be expressed as,

$$\begin{aligned}\mathcal{L}_{LGAN}(G_e,D_e)&=\mathbb{E}_{x,y}\left[\log D_e(x,y)\right]\\&+\mathbb{E}_{\widehat{x},x,z_2}[\log(1-D_e(x,G_e(\widehat{x},z_2)))],\end{aligned} \tag{6}$$

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 $L1$ and $L2$ distances can be defined as $\mathcal{L}_{L1}(G_e)=\mathbb{E}_{\widehat{x},y,z_2}\left[||y-G_e(\widehat{x},z_2)||_1\right]$ and $\mathcal{L}_{L2}(G_e)=\mathbb{E}_{\widehat{x},y,z_2}\left[||y-G_e(\widehat{x},z_2)||_2\right]$, respectively.

In this paper, we also explore an alternative option and propose a robust version loss in Figure 2, i.e., $\mathcal{L}_{L1,2}$, which takes both $L1$ and $L2$ losses into consideration simultaneously. The proposed $\mathcal{L}_{L1,2}$ loss is less sensitive to outliers than the $L2$ loss, and in some cases it could prevent exploding gradients [57]. Moreover, we observe that the $\mathcal{L}_{L1,2}$ loss can provide more image details than the $L1$ loss in the preliminary experiments. Thus, the definition of the proposed $\mathcal{L}_{L1,2}$ is:

$$\mathcal{L}_{L1,2}(G_e)=\begin{cases}0.5*\mathcal{L}_{L_2}(G_e) & if\ |y-G_e(\widehat{x},z_2)|<1,\\ \mathcal{L}_{L_1}(G_e)-0.5 & if\ |y-G_e(\widehat{x},z_2)|\geq 1.\end{cases} \tag{7}$$

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,

$$\mathcal{L}_{identity}(G_e)=\mathbb{E}_{\widehat{x},y,z_2}\left[||F(y)-F(G_e(\widehat{x},z_2))||_1\right], \tag{8}$$

where F is a face feature extractor. We employ [58] to compare the identity distance between the target face y and the generated face $G_e(\widehat{x},z_2)$, and try to minimize the distance $F(y)-F(G_e(\widehat{x},z_2))$ 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:

$$\mathcal{L}_2=\mathcal{L}_{LGAN}(G_e,D_e)+\lambda_2\mathcal{L}_{L_{1,2}}(G_e)+\lambda_3\mathcal{L}_{identity}(G_e), \tag{9}$$

where the parameter λ_2 and λ_3 control the relative importance of $\mathcal{L}_{LGAN}(G_e,D_e)$, $\mathcal{L}_{L_{1,2}}(G_e)$ and $\mathcal{L}_{identity}(G_e)$. 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 $(G_e(\widehat{x},z_2),x)$ and real (y,x) pairs, as shown in Figure 1. Thus, the loss functions for the discriminator D_l and D_e are: $\mathcal{L}(D_l)=\frac{1}{2}\left[\mathcal{L}_{bce}(D_l(l),1)+\mathcal{L}_{bce}(D_l(G_l(x_1,z_1,l_e),0)\right]$, and $\mathcal{L}(D_e)=\frac{1}{2}[\mathcal{L}_{bce}(D_e(y,x),1)+\mathcal{L}_{bce}(D_e(G_e(\widehat{x},z_2),x),0)]$, respectively, where $\mathcal{L}_{bce}$ 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.

C. 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 Figure 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_e, we adopt PatchGAN proposed in [8], PatchGAN tries to classify if each 70×70 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_e and D_e fixed, and then train G_e and D_e 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_l 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 256×256. Each model is trained 200 epochs with batch size 1. We employ the Adam [60] with momentum terms β_1=0.5 and β_2=0.999 as our solver. The initial learning rate for Adam is 0.0002.

We set S_l=4, λ_1=2, λ_2=100 and λ_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].

IV. Experiments

A. 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 217 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 resting 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 $\times$ 7 expressions $\times$ 4 images = 2,800 images in total. For each expression, we randomly choose 10 subjects as testing data and the resting 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] is a dataset of stylized characters with 7 basic facial expressions, which contains about 50,000 images modeled using the MAYA software. In our experiments, we randomly select 1,200 faces from each expression, and then randomly choose 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 over 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 are captured from five cameras with different angles and each subject is asked to show three different gaze directions. For each emotion, we select 67% images as training data and the rest 33% images as testing data. Similar to C2GAN [12], all the images are rescaled to $256\times256\times3$ 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.

B. 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 I shows the results using different components of LandmarkGAN on the

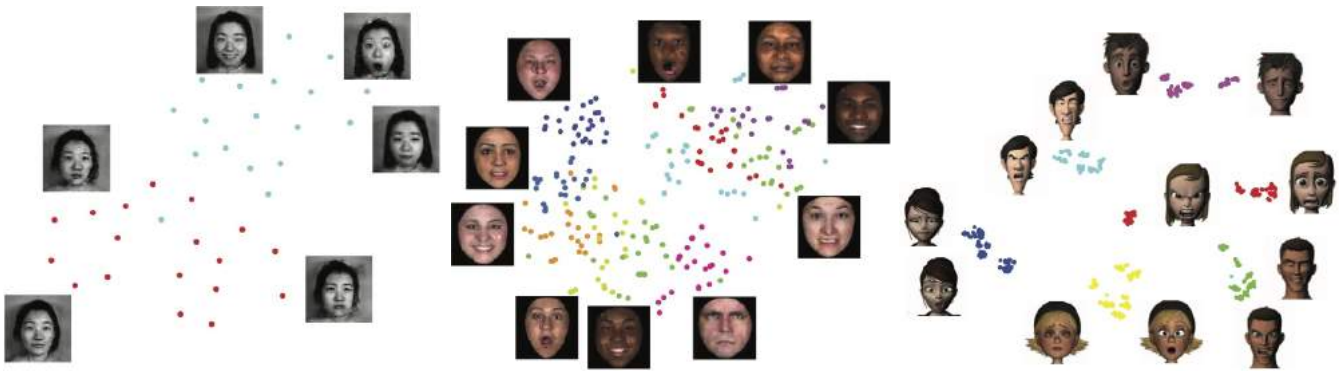

Fig. 3: Identity feature space of JAFFE (*left*), 3DFE (*middle*), and FERG (*right*). Each color represents a subject.

TABLE I: Quantitative results of different components on JAFFE, 3DFE, and FERG.

Setting	PSNR ↑			SSIM ↑			IS ↑		
	JAFFE	3DEF	FERG	JAFFE	3DFE	FERG	JAFFE	3DFE	FERG
LandmarkGAN - D - I - LR	16.8467	22.3147	32.9560	0.4968	0.8737	0.9813	1.4508	1.7392	1.5576
LandmarkGAN - D - I	17.2185	22.8473	36.4789	0.5197	0.8903	0.9884	1.4619	1.7565	1.5756
LandmarkGAN - LR	17.1502	22.8186	34.1283	0.5186	0.8881	0.9863	1.4522	1.7489	1.5846
LandmarkGAN - I	17.2613	23.1719	37.7746	0.5244	0.9021	0.9890	1.4858	1.7610	1.5937
LandmarkGAN - D	17.2852	23.8509	38.8294	0.5361	0.9133	0.9930	1.5508	1.7625	1.5960
LandmarkGAN ($L_{1,2}$, $l_e \rightarrow G_l$ & $l_e \rightarrow G_e$, w/ soft-crops)	**17.2911**	**23.8754**	**38.9132**	**0.5392**	**0.9332**	**0.9932**	**1.5708**	**1.7984**	**1.6103**
LandmarkGAN (L_1, $l_e \rightarrow G_l$ & $l_e \rightarrow G_e$, w/ soft-crops)	17.1892	23.4216	38.3760	0.5189	0.9174	0.9789	1.5687	1.7509	1.5762
LandmarkGAN (L_2, $l_e \rightarrow G_l$ & $l_e \rightarrow G_e$, w/ soft-crops)	17.2343	23.6743	38.7598	0.5276	0.9265	0.9862	1.5268	1.7653	1.5876
LandmarkGAN ($L_{1,2}$, $l_e \rightarrow G_l$, w/ soft-crops)	17.2807	23.8532	38.8750	0.5375	0.9324	0.9919	1.5698	1.7894	1.6062
LandmarkGAN ($L_{1,2}$, $l_e \rightarrow G_e$, w/ soft-crops)	17.0156	23.3642	37.5791	0.5108	0.9185	0.9868	1.5202	1.7530	1.5693
LandmarkGAN ($L_{1,2}$, $l_e \rightarrow G_l$ & $l_e \rightarrow G_e$, w/o soft-crops)	17.0726	23.6752	38.6709	0.5167	0.9298	0.9912	1.5649	1.7785	1.6096

JAFFE, 3DFE, and FERG 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 I. 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 with three different variants in Table I, 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 I, 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 I 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 II: Results of expression recognition accuracy (↑) with different training and testing settings.

Train/Test	Real/Real	Real/Syn.	Real+Nor./Real	Real+Syn./Real
JAFFE	34.44%	25.83%	35.84%	**38.89%**
3DFE	46.75%	35.33%	47.64%	**49.46%**
FERG	99.93%	95.63%	99.95%	**99.97%**

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 Figure 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

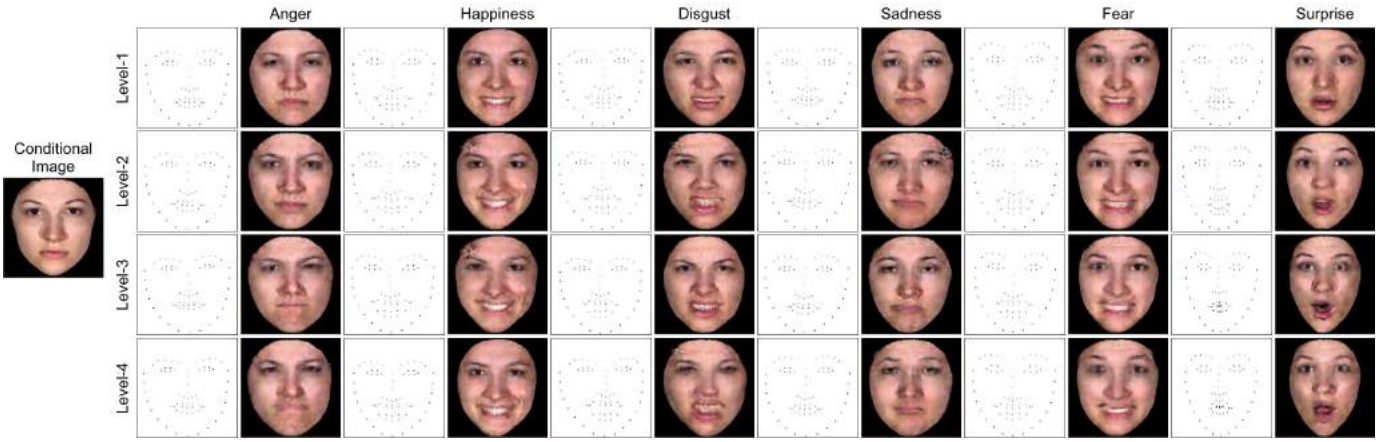

Fig. 4: The generated landmark and expression with different intensity levels on 3DFE.

Fig. 5: Different methods for facial expression-to-expression translation task on JAFFE, 3DFE, and FERG. From top to bottom: input, C2GAN [12], LandmarkGAN (Ours), and Ground Truth (GT). Note that the input of LandmarkGAN is only a single image. While C2GAN needs to input the target landmarks. The samples in this figure were randomly selected for visualization purposes.

TABLE III: Quantitative results of different methods on JAFFE, 3DFE and FERG.

Method	PSNR ↑			SSIM ↑			IS ↑		
	JAFFE	3DEF	FERG	JAFFE	3DFE	FERG	JAFFE	3DFE	FERG
GANimation [67]	15.3878	20.1987	37.9322	0.5176	0.8956	0.9875	1.3287	1.6720	1.4389
GANmut [24]	15.6543	20.3445	38.1489	0.5221	0.8998	0.9887	1.3351	1.6806	1.4670
AttentionGAN [42]	16.1260	21.0475	38.3368	0.5267	0.9006	0.9896	1.3514	1.7169	1.5268
C2GAN [12]	17.2442	22.4289	38.4789	0.5317	0.9147	0.9911	1.5695	1.7531	1.5882
LandmarkGAN (Ours)	**17.2911**	**23.8754**	**38.9132**	**0.5392**	**0.9332**	**0.9932**	**1.5708**	**1.7984**	**1.6103**

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.) and (ii) the classifiers trained on real images will also be able to classify the synthesized image correctly. The results are listed in Table II. 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. Figure 4 shows the results with different intensity levels of expressions on the 3DFE dataset. For example, in the last column of Figure 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,

$$\mathcal{L}_{cGAN}(G_l, D_l) = \mathbb{E}_l\left[\log D_l(l)\right] + \mathbb{E}_{x,z_1}\left[\log(1 - D_l(G_l(x, z_1, l_e, l_i)))\right], \tag{10}$$

where l_i is the one-hot label of expression intensity. Moreover, $\widehat{x}$ in Eq. (6) needs to be updated to $\widehat{x}=[x, G_l(x, z_1, l_e, l_i), l_e, l_i]$. Results in Figure 4 validate that the proposed LandmarkGAN discovers the expression intensity by manipulating the facial landmark. Note that Figure 4

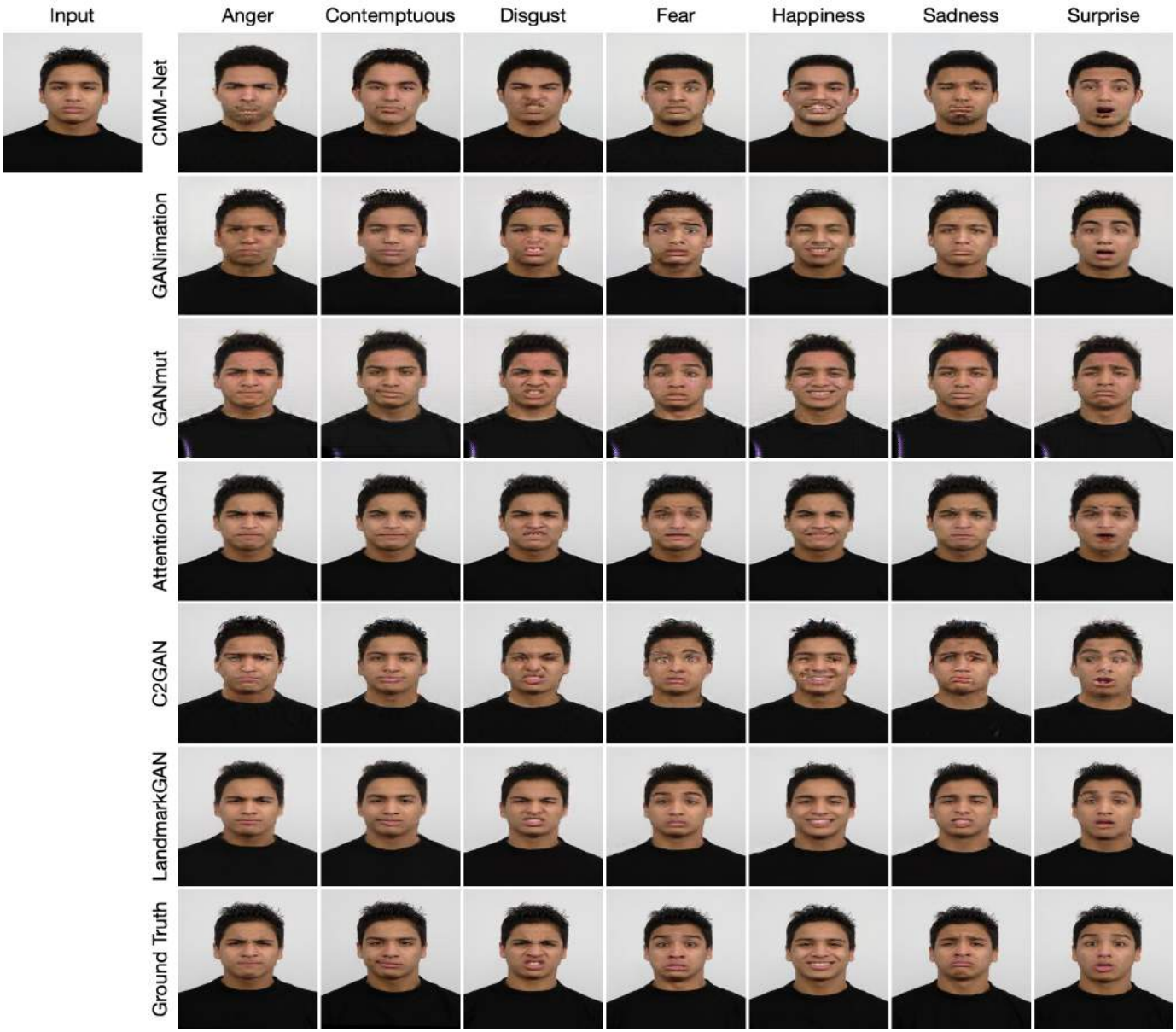

Fig. 6: Different methods for facial expression-to-expression translation task on RaFD. From top to bottom: CMM-Net [17], GANimation [67], GANmut [24], AttentionGAN [42], C2GAN [12], LandmarkGAN (Ours), and Ground Truth (GT). The samples in this figure were randomly selected for visualization purposes.

shows a very challenging task. We only need to input a neutral face image to generate face images with different intensity levels and expressions at the same time. To the best of our knowledge, our method is currently the only one that can achieve this goal. Finally, we note that the stronger the expression, the better the facial expression classification result we will achieve.

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 III that the proposed LandmarkGAN consistently achieves better performance than C2GAN, GANimation, GANmut, and AttentionGAN on all the JAFFE, 3DEF, and FERG datasets. We also show the visualization

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.

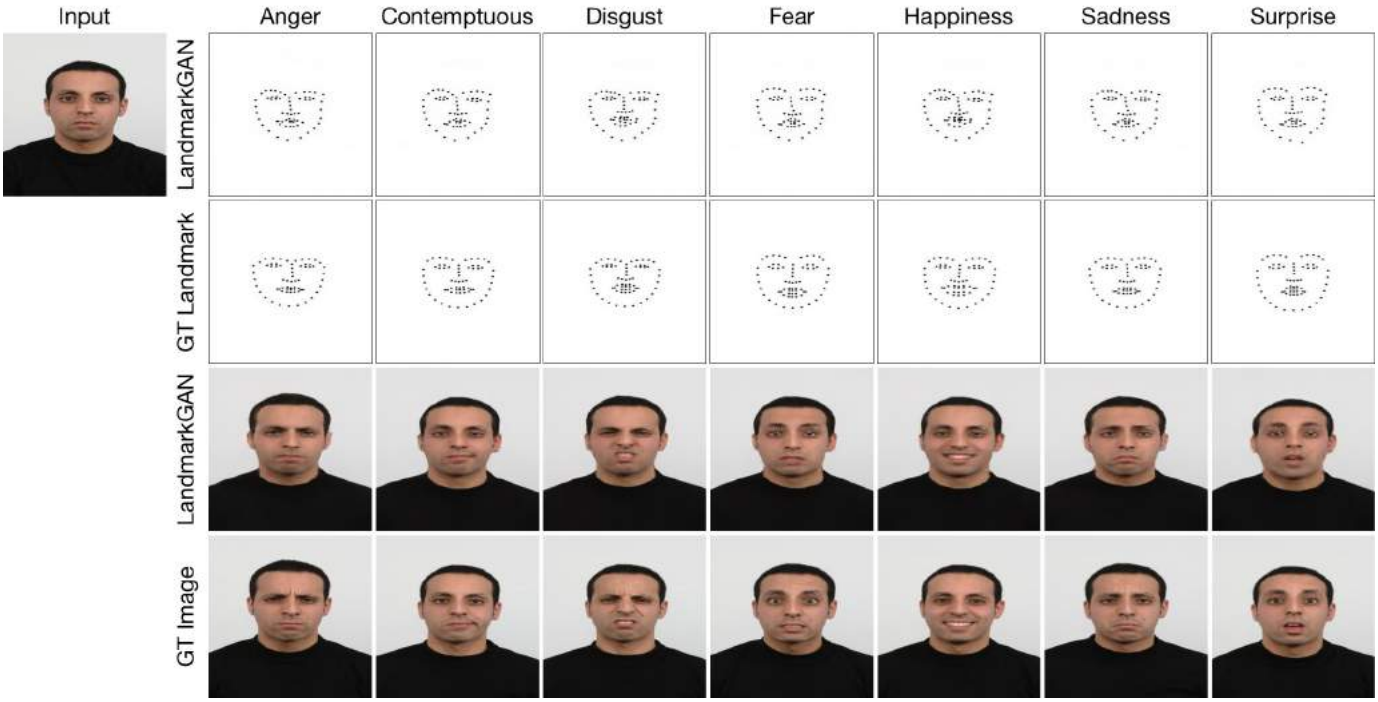

Fig. 8: The generated landmarks and images of LandmarkGAN on RaFD. We only need to input a single image to obtain the diversity landmarks and expressions, while C2GAN needs to input a image and the GT landmarks to generate diversity expressions.

results compared to the state-of-the-art method, i.e., C2GAN, in Figure 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 Figure 6 and Table IV. We can see that the proposed LandmarkGAN

TABLE IV: Quantitative results of different methods on RaFD.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
CMM-Net [17]	18.6521	0.8367	0.1282
GANimation [67]	20.8765	0.8432	0.1042
GANmut [24]	21.1348	0.8468	0.0982
AttentionGAN [42]	21.4586	0.8565	0.0967
C2GAN [12]	21.9192	0.8618	0.0934
LandmarkGAN (Ours)	**22.2132**	**0.8761**	**0.0928**

TABLE V: AMT score (↑) of different methods.

Method	JAFFE	3DFE	FERG	RaFD
C2GAN [12]	1.2	13.3	70.9	34.2
LandmarkGAN (Ours)	**2.5**	**19.7**	**82.9**	**39.1**

TABLE VI: L_2 distance (↓) for the landmark generation task.

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**

achieves better results compared with the leading baselines, which validates the effectiveness of the proposed method. Finally, in Figure 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 [39], [8] and conduct an perceptual studies to evaluate the quality of the generated images. Results are shown in Table V. 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 V 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 $L2$ distance between detected and generated landmarks as the evaluation metric. Quantitative results are provided in Table VI. Note that the generated landmarks have a lower $L2$ distance than [53] as our multi-task learning strategy. In Figure 8, we show some samples of the generated landmarks.

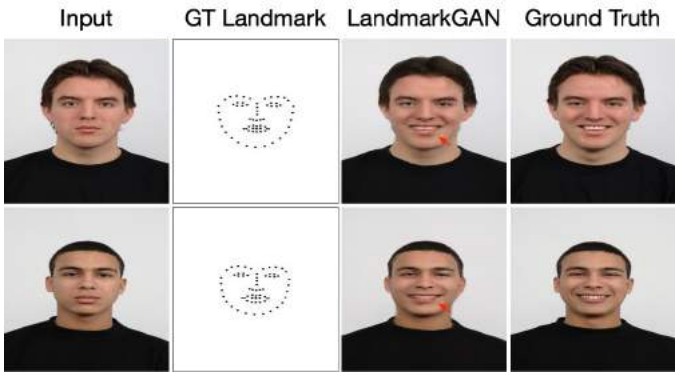

Fig. 9: Failure cases on RaFD.

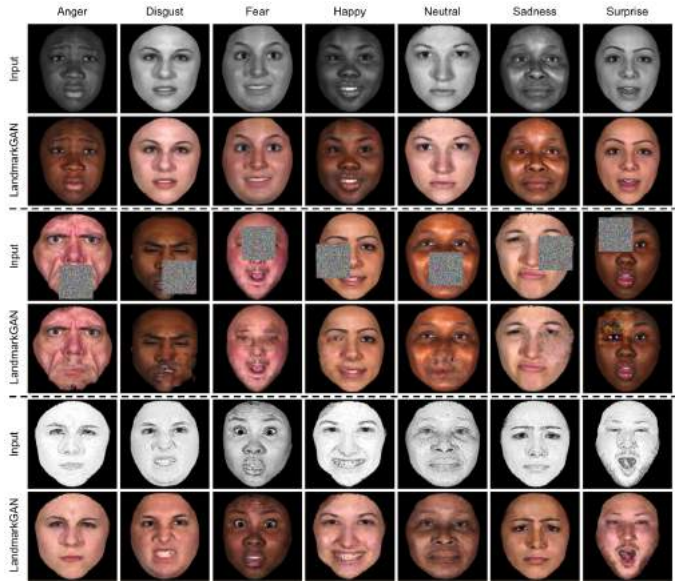

Fig. 10: Results of LandmarkGAN on three generative tasks, i.e., face colorization (*top*), face inpainting (*middle*), and sketch-to-face translation (*bottom*) on 3DFE.

We see that the generated 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 Figure 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 $256\times256\times3$ first, then we set the mask size as 96×96 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 Figure 10. We see that the proposed LandmarkGAN generates reasonable results on the three different generative tasks, validating the generalization ability of the proposed LandmarkGAN.

V. Conclusions

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 paper, 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, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NeurIPS*, 2014. 1, 3, 5

[2] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," in *ICLR*, 2018. 1, 3

[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 *CVPR*, 2018. 1, 3

[4] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *CVPR*, 2019. 1, 3

[5] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *ICML*, 2019. 1, 3

[6] A. Brock, J. Donahue, and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis," in *ICLR*, 2019. 1, 3

[7] T. R. Shaham, T. Dekel, and T. Michaeli, "Singan: Learning a generative model from a single natural image," in *ICCV*, 2019. 1, 3

[8] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *CVPR*, 2017. 1, 2, 3, 4, 5, 10

[9] A. Anoosheh, E. Agustsson, R. Timofte, and L. Van Gool, "Combogan: Unrestrained scalability for image domain translation," in *ICLR*, 2018. 1, 3

[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 *CVPR*, 2018. 1, 3

[11] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool, "Pose guided person image generation," in *NeurIPS*, 2017. 1, 3, 5

[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 *ACM MM*, 2019. 1, 2, 3, 5, 6, 7, 8, 10

[13] H. Tang, W. Wang, D. Xu, Y. Yan, and N. Sebe, "Gesturegan for hand gesture-to-gesture translation in the wild," in *ACM MM*, 2018. 1, 3, 5

[14] S. E. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee, "Learning what and where to draw," in *NeurIPS*, 2016. 1, 3

[15] S. Reed, A. van den Oord, N. Kalchbrenner, V. Bapst, M. Botvinick, and N. de Freitas, "Generating interpretable images with controllable structure," *Technical Report*, 2016. 1

[16] I. Korshunova, W. Shi, J. Dambre, and L. Theis, "Fast face-swap using convolutional neural networks," in *ICCV*, 2017. 1, 3

[17] W. Wei, A.-P. Xavier, X. Dan, R. Elisa, F. Pascal, and S. Nicu, "Every smile is unique: Landmark-guided diverse smile generation," in *CVPR*, 2018. 1, 3, 6, 8, 10

[18] X. Di, V. A. Sindagi, and V. M. Patel, "Gp-gan: Gender preserving gan for synthesizing faces from landmarks," in *ICPR*, 2018. 1, 3

[19] F. Qiao, N. Yao, Z. Jiao, Z. Li, H. Chen, and H. Wang, "Geometry-contrastive generative adversarial network for facial expression synthesis," *arXiv preprint arXiv:1802.01822*, 2018. 1, 3

[20] M. Zhang, J. Li, N. Wang, and X. Gao, "Compositional model-based sketch generator in facial entertainment," *IEEE TCYB*, vol. 48, no. 3, pp. 904–915, 2017. 3

[21] D. S. Tan, J. H. Soeseno, and K.-L. Hua, "Controllable and identity-aware facial attribute transformation," *IEEE TCYB*, 2021. 3

[22] J. Yu, X. Xu, F. Gao, S. Shi, M. Wang, D. Tao, and Q. Huang, "Toward realistic face photo-sketch synthesis via composition-aided gans," *IEEE TCYB*, 2020. 3

[23] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014. 3

[24] S. d'Apolito, D. P. Paudel, Z. Huang, A. Romero, and L. Van Gool, "GANmut: Learning interpretable conditional space for gamut of emotions," in *CVPR*, 2021. 3, 6, 7, 8, 10

[25] H. Tang, W. Wang, S. Wu, X. Chen, D. Xu, N. Sebe, and Y. Yan, "Expression conditional gan for facial expression-to-expression translation," in *ICIP*, 2019. 3

[26] H. Tang, X. Chen, W. Wang, D. Xu, J. J. Corso, N. Sebe, and Y. Yan, "Attribute-guided sketch generation," in *FG 2019*, 2019. 3

[27] Z. Xu, T. Lin, H. Tang, F. Li, D. He, N. Sebe, R. Timofte, L. Van Gool, and E. Ding, "Predict, prevent, and evaluate: Disentangled text-driven image manipulation empowered by pre-trained vision-language model," in *CVPR*, 2022. 3

[28] M. Tao, H. Tang, F. Wu, X.-Y. Jing, B.-K. Bao, and C. Xu, "Df-gan: A simple and effective baseline for text-to-image synthesis," in *CVPR*, 2022. 3

[29] S. Wu, H. Tang, X.-Y. Jing, H. Zhao, J. Qian, N. Sebe, and Y. Yan, "Cross-view panorama image synthesis," *IEEE TMM*, 2022. 3

[30] S. Wu, H. Tang, X.-Y. Jing, J. Qian, N. Sebe, Y. Yan, and Q. Zhang, "Cross-view panorama image synthesis with progressive attention gans," *Elsevier PR*, 2022. 3

[31] H. Tang, L. Shao, P. H. Torr, and N. Sebe, "Local and global gans with semantic-aware upsampling for image generation," *IEEE TPAMI*, 2022. 3

[32] B. Ren, H. Tang, and N. Sebe, "Cascaded cross mlp-mixer gans for cross-view image translation," in *BMVC*, 2021. 3

[33] H. Tang and N. Sebe, "Layout-to-image translation with double pooling generative adversarial networks," *IEEE TIP*, 2021. 3

[34] G. Liu, H. Tang, H. M. Latapie, J. J. Corso, and Y. Yan, "Cross-view exocentric to egocentric video synthesis," in *ACM MM*, 2021. 3

[35] H. Tang, S. Bai, and N. Sebe, "Dual attention gans for semantic image synthesis," in *ACM MM*, 2020. 3

[36] H. Tang, D. Xu, Y. Yan, P. H. Torr, and N. Sebe, "Local class-specific and global image-level generative adversarial networks for semantic-guided scene generation," in *CVPR*, 2020. 3

[37] G. Liu, H. Tang, H. Latapie, and Y. Yan, "Exocentric to egocentric image generation via parallel generative adversarial network," in *ICASSP*, 2020. 3

[38] H. Tang, D. Xu, N. Sebe, Y. Wang, J. J. Corso, and Y. Yan, "Multi-channel attention selection gan with cascaded semantic guidance for cross-view image translation," in *CVPR*, 2019. 3

[39] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *ICCV*, 2017. 3, 10

[40] T. Kim, M. Cha, H. Kim, J. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *ICML*, 2017. 3

[41] Z. Yi, H. Zhang, P. T. Gong *et al.*, "Dualgan: Unsupervised dual learning for image-to-image translation," in *ICCV*, 2017. 3

[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 TNNLS*, 2021. 3, 6, 7, 8, 10

[43] Y. Ge, Z. Li, H. Zhao, G. Yin, S. Yi, X. Wang *et al.*, "Fd-gan: Pose-guided feature distilling gan for robust person re-identification," in *NeurIPS*, 2018. 3

[44] H. Dong, X. Liang, K. Gong, H. Lai, J. Zhu, and J. Yin, "Soft-gated warping-gan for pose-guided person image synthesis," in *NeurIPS*, 2018. 3

[45] Y. Zhou, X. Han, E. Shechtman, J. Echevarria, E. Kalogerakis, and D. Li, "Makelttalk: speaker-aware talking-head animation," *ACM TOG*, vol. 39, no. 6, pp. 1–15, 2020. 3

[46] L. Chen, R. K. Maddox, Z. Duan, and C. Xu, "Hierarchical cross-modal talking face generation with dynamic pixel-wise loss," in *CVPR*, 2019. 3

[47] E. Zakharov, A. Shysheya, E. Burkov, and V. Lempitsky, "Few-shot adversarial learning of realistic neural talking head models," in *ICCV*, 2019. 3

[48] N. Dall'Asen, Y. Wang, H. Tang, L. Zanella, and E. Ricci, "Graph-based generative face anonymisation with pose preservation," in *ICIAP*, 2022. 3

[49] H. Tang and N. Sebe, "Total generate: Cycle in cycle generative adversarial networks for generating human faces, hands, bodies, and natural scenes," *IEEE TMM*, 2021. 3

[50] H. Tang, H. Liu, and N. Sebe, "Unified generative adversarial networks for controllable image-to-image translation," *IEEE TIP*, 2020. 3

[51] H. Tang, S. Bai, P. H. Torr, and N. Sebe, "Bipartite graph reasoning gans for person image generation," in *BMVC*, 2020. 3

[52] H. Tang, S. Bai, L. Zhang, P. H. Torr, and N. Sebe, "Xinggan for person image generation," in *ECCV*, 2020. 3

[53] Q. Sun, L. Ma, S. J. Oh, L. Van Gool, B. Schiele, and M. Fritz, "Natural and effective obfuscation by head inpainting," in *CVPR*, 2018. 3, 5, 10

[54] L. Song, Z. Lu, R. He, Z. Sun, and T. Tan, "Geometry guided adversarial facial expression synthesis," in *ACM MM*, 2018. 3

[55] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: an open source facial behavior analysis toolkit," in *WACV*, 2016. 3, 5

[56] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *CVPR*, 2016. 4

[57] R. Girshick, "Fast r-cnn," in *ICCV*, 2015. 4

[58] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, "Openface: A general-purpose face recognition library with mobile applications," *CMU School of Computer Science*, 2016. 4

[59] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *ICML*, 2015. 4

[60] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015. 5

[61] T. Nguyen, T. Le, H. Vu, and D. Phung, "Dual discriminator generative adversarial nets," in *NeurIPS*, 2017. 5

[62] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *NeurIPS*, 2019. 5

[63] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets," in *FG*, 1998. 5

[64] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3d facial expression database for facial behavior research," in *FG*, 2006. 5

[65] D. Aneja, A. Colburn, G. Faigin, L. G. Shapiro, and B. Mones, "Modeling stylized character expressions via deep learning." in *ACCV*, 2016. 5

[66] O. Langner, R. Dotsch, G. Bijlstra, D. H. Wigboldus, S. T. Hawk, and A. Van Knippenberg, "Presentation and validation of the radboud faces database," *Taylor & Francis Cognition and emotion*, 2010. 5

[67] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer, "Ganimation: Anatomically-aware facial animation from a single image," in *ECCV*, 2018. 5, 6, 7, 8, 10

[68] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE TIP*, vol. 13, no. 4, pp. 600–612, 2004. 5

[69] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *NeurIPS*, 2016. 5

[70] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018. 5

[71] I. Durugkar, I. Gemp, and S. Mahadevan, "Generative multi-adversarial networks," in *ICLR*, 2017. 5

[72] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *JMLR*, vol. 9, no. Nov, pp. 2579–2605, 2008. 6

[73] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015. 6

[74] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "First order motion model for image animation," *NeurIPS*, 2019. 10

[75] K. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar, "A lip sync expert is all you need for speech to lip generation in the wild," in *ACM MM*, 2020. 10

Hao Tang is currently a Postdoctoral with Computer Vision Lab, ETH Zurich, Switzerland. He received the master's degree from the School of Electronics and Computer Engineering, Peking University, China and the Ph.D. degree from Multimedia and Human Understanding Group, University of Trento, Italy. He was a visiting scholar in the Department of Engineering Science at the University of Oxford. His research interests are deep learning, machine learning, and their applications to computer vision.

Nicu Sebe is Professor in 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 Co-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 ICPR 2020. He is a general chair of ACM Multimedia 2022 and a program chair of ECCV 2024. He is a fellow of IAPR.