ApprGAN: Appearance-Based Generative Adversarial Network for Facial Expression Synthesis

DOI:
10.1049/iet-ipr.2018.6576

Document Version
Accepted author manuscript

Link to publication record in Manchester Research Explorer

Citation for published version (APA):
Peng, Y., & Yin, H. (2019). ApprGAN: Appearance-Based Generative Adversarial Network for Facial Expression Synthesis. IET Image Processing. https://doi.org/10.1049/iet-ipr.2018.6576

Published in:
IET Image Processing

Citing this paper
Please note that where the full-text provided on Manchester Research Explorer is the Author Accepted Manuscript or Proof version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version.

General rights
Copyright and moral rights for the publications made accessible in the Research Explorer are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Takedown policy
If you believe that this document breaches copyright please refer to the University of Manchester’s Takedown Procedures [http://man.ac.uk/04Y6Bo] or contact uml.scholarlycommunications@manchester.ac.uk providing relevant details, so we can investigate your claim.
Abstract: Facial expression synthesis has drawn increasing attention in computer vision, graphics and animation. Recently generative adversarial nets (GANs) have become a new perspective for face synthesis and have had remarkable success in generating photorealistic images and image-to-image translation. In this paper, we present an appearance-based facial expression synthesis framework, ApprGAN, by combining shape and texture and introducing cycle-consistency and identity mapping into the adversarial learning. Specifically, given an input face image, a pair of shape and texture generators are trained for synthetic shape deformation and expression detail generation, respectively. Extensive experiments on expression synthesis and cross-database synthesis were conducted, together with comparisons with the existing methods. Results of expression synthesis and quantitative verification on various databases show the effectiveness of ApprGAN in synthesising photorealistic and identity-preserving expressions and its marked improvement over the existing methods.

1 Introduction

Facial expression synthesis or transfer aims at altering the expression of a specified person while still preserving personal identity and facial features of the subject. In spite of considerable effort from academia and industry, expression synthesis and analysis remain challenging due to the subtlety, complexity and variability of human expression [1, 2]. Synthesising photorealistic images with various expressions is of a great value in many application scenarios such as human-machine interaction, face editing, emotion analysis, animation and medical care [3].

The past few decades have witnessed numerous techniques being developed for expression synthesis. Existing methods can be divided into two categories: traditional handcrafted methods and recent deep learning based generative models. The majority of the early studies handle faces by image warping and mapping, facial component editing and face decomposition [4–7]. These methods sometimes can produce natural, high-resolution expressions, but usually require paired training samples and rely highly on accurate extraction of landmarks. With the recent advances of deep learning, generative adversarial nets (GANs) have achieved remarkable results in several applications [8–10]. The main idea is to simultaneously train paired discriminator and generator, to produce indistinguishable fake faces from real ones. The generator finds low-dimensional representations of faces in a latent space and manipulates them to synthesise new images. Despite their ability to generate impressive or promising results in most cases, synthesised images sometimes lack fine details and tend to be blurry or of low-resolution [10]. Besides, both shape and texture attributes are usually encoded in one latent feature space, making it hard to take control of subtle expression changes.

In this paper, we propose an appearance-based GAN, ApprGAN, to synthesise expressions by combining the conventional warping based methods with the generative models. We derive shape and texture attributes of faces and construct two GAN-based networks for geometry deformation and expression detail generation, respectively. We also incorporate reconstruction and identity mapping losses in the adversarial learning to stabilise and improve GAN training. Expressions are synthesised from neutral faces on both geometry attributes and expression details. To evaluate quality of these generated images, correlation coefficients and normalised landmark distances are calculated between synthesised and ground truth images. The effectiveness of the proposed method has been verified on three benchmark databases including the Bosphorus [11], extended AU-Coded Cohn-Kanade (CK+) [12, 13], and MUG [14] facial expression databases. The main contributions of the current work are summarised as follows:

1) An appearance-based generative adversarial network, ApprGAN, for facial expression synthesis with combination of shape and texture generative models;
2) Incorporation of cycle consistency and identity mapping losses into the generative adversarial learning to stabilise and improve image synthesis;
3) Synthesis of photorealistic and identity-preserving expressions with good generalisation across databases;
4) Improved quantitative measures of synthesised expressions compared to state-of-the-art methods.

The remaining of the paper is organised as follows. Related work is reviewed in Section 2. Section 3 describes the formulation of objective function, the proposed expression synthesis framework and expression verification method. Section 4 presents experimental results and analysis, followed by conclusions in Section 5.

2 Related Work

Facial expression describes facial motions and conveys important visual information in communication and understanding human behaviours [15]. Automatic analysis of facial expression has hence attracted a great deal of attention with a wide range of applications such as expression classification [16, 17] and face recognition [18, 19]. As a form of nonverbal communication, facial expressions convey emotion states through muscle movements beneath the skin. In 1872, Darwin first demonstrated the universal nature of expression and linked human emotions with mental states and body movements. Later psychologists proposed and developed the anatomy-based facial action coding system (FACS) [20, 21] that could comprehensively interpret all possible facial expressions. The FACS considers only emotion related facial motions and describes them by combinations of 44 individual and visually distinguishable action units (AUs). According to the intensity of AUs, expression intensities are annotated in five different levels, from trace to maximum. Seven prototypic facial expressions are typically defined as neutral, anger, disgust, fear, happiness, sadness and surprise [22].
In the last two decades, there have been various approaches to generating realistic facial expressions in computer vision [3][6][25], most of which are either morph-based or example-based. Liu et al. [4] proposed to calculate an expression ratio image (ERI) of a subject to capture the illumination change resulting from expression transition. With geometric warping, the ERI could be mapped to a different person’s face to obtain facial expression synthesis. Wang and Ahuja [17] used a principal component analysis based high-order singular decomposition approach to extract separate subspaces (identity, expression and feature subspaces) from a collection of images displaying different expressions. Expressions for a new subject could be synthesised using the learned subspace model. Ghent and McDonald [25] constructed facial expression shape and texture models to generate expression mapping functions. In conjunction with the radial basis function networks, neutral images could be mapped to alternative expressions of the same subjects. Zhang et al. [6] proposed an expression synthesis framework based on geometry to synthesise expression texture from facial landmark movements. Given a face image and its facial landmarks, expression were synthesised within 14 face subregions and blended along the subregion boundaries. Huang and De la Torre [26] developed a bilinear kernel reduced rank regression based method to learn a nonlinear mapping from neutral face to various expressions. The input neutral image was further combined with training data for preserving person-specific expression details against untrained illumination conditions. It is often difficult for these warping based methods to synthesise high-resolutional faces and unseen facial features (such as teeth and inner mouth areas, beard and person-specific expression details).

Recently deep generative models have achieved impressive success in a wide range of applications such as image generation [27], image-to-image style translation [28][30], and super-resolution [31][32]. The advances have also proved their usefulness in facial expression synthesis and transfer. Susskind et al. [33] trained a deep belief net capable of converting facial attribute labels into realistic face images as well as generating faces displaying specific identities and facial actions. Li et al. [34] built a convolutional neural network to generate facial images with input source images and reference attribute labels. Radford et al. [27] introduced deep convolutional GANs for unsupervised image generation. This approach could learn hierarchical representations of input and be able to generate expressive faces. Choi et al. [6] proposed a multi-domain image-to-image translation approach called StarGAN to synthesise images across domains using a single model, which could also be applied in facial expression transfer. Ding et al. [9] introduced the ExprGAN for facial expression editing with controllable expression intensity. In addition to the network, an expression controller was employed to learn expression code without the need for training data with intensity values. Song et al. [10] built a geometry-guided GAN architecture for synthesising images with continuously adjusting expressions. Geometry information was treated as an additional image channel to be concatenated with the input face to guide the transfer of facial expressions. Similarly, Qiao et al. [15] designed a geometry-contrastive GAN for transferring continuous emotions from one subject to another. Contrastive learning from geometry information was integrated in embedding network. Recently Pumarola et al. [35] presented AU-based GAN model for synthetic facial animation, which enables controlling the magnitude of activations of each AU. They also embedded an attention module to allow focusing only on image regions relating to specific expressions, making the network robust to changing backgrounds and illumination conditions.

3 Methods

This section describes the formulation of the proposed expression synthesis framework, its objective function, and expression verification method. We incorporate reconstruction and identity mapping losses in the adversarial learning to define the objective function and present the pipeline of the proposed ApprGAN. To evaluate the quality of synthesised expressions, we calculate the correlation coefficients and landmark distances.

3.1 Objective Function

Inspired by CycleGAN [29] and DTN [34], we define our loss functions containing three terms: adversarial loss for pushing the distribution of generated images to match those of the target domain; reconstruction loss for preventing unstable and unbalanced GAN training; and identity mapping loss for preserving texture composition and shape identity information.

Adversarial Loss: As a type of generative models, GANs learn a mapping from input to target domains based on the min-max game theory. A GAN consist of two adversarial networks: a discriminator that tries to classify between real and fake samples, and a generator that simultaneously trained to synthesise realistic fake samples to fool the discriminator. Given two image domains $X = \{x\}$ and $Y = \{y\}$, the generator $G_{X→Y}$ is optimised to map from input to target domains and meanwhile, the discriminator $D_Y$ aims to distinguish between samples $\{y\}$ and synthesised $\{G_{X→Y}(x)\}$. The adversarial loss is defined as

$$
\mathcal{L}_{GAN}(G_{X→Y}, D_Y, X, Y) = \mathbb{E}_{y \sim \text{data}(y)}[\log D_Y(y)] + \mathbb{E}_{x \sim \text{data}(x)}[\log (1 - D_Y(G_{X→Y}(x))]),
$$

in which $G_{X→Y}$ aims to minimise the objective function $\mathcal{L}_{GAN}(G_{X→Y}, D_Y, X, Y)$ against $D_Y$ attempting to maximise it. Similarly, the adversarial loss for the translation $Y → X$ can be expressed as $\mathcal{L}_{GAN}(G_{Y→X}, D_X, X, Y)$.

Reconstruction Loss: With the adversarial loss only, training of GANs is still remarkably difficult and suffers from many problems such as non-convergence, unbalance between generator and discriminator, diminished gradient, overfitting and so on. Furthermore, there is still little theoretical explanation of the unstable behaviour of GAN training [33]. To stabilise and improve the learning process, a reconstruction loss is introduced to enforce the learned mapping functions $X → Y$ and $Y → X$ to be forward-backward cycle consistent [29]. For each image $x$ in $X$, the forward cycle consistency, $G_{Y→X}(G_{X→Y}(x)) \approx x$, indicates that the reconstructed image $G_{Y→X}(G_{X→Y}(x))$ should be able to closely match the input. Also for each image $y$ in $Y$, the backward cycle consistency should also be satisfied, $G_{X→Y}(G_{Y→X}(y)) \approx y$. Thus, a $L_1$ norm based reconstruction loss can be written as

$$
\mathcal{L}_{rec}(G_{X→Y}, G_{Y→X}) = \mathbb{E}_{x \sim \text{data}(x)}[\|G_{Y→X}(G_{X→Y}(x)) - x\|_1] + \mathbb{E}_{y \sim \text{data}(y)}[\|G_{X→Y}(G_{Y→X}(y)) - y\|_1].
$$

Identity Mapping Loss: In facial expression synthesis, we need suitable control for expression changes without altering the facial texture and identity information. In order to preserve face identity and texture composition, we adopt an additional identity mapping loss [27] between input and output as a regulariser. The identity mapping loss encourages generators $G_{X→Y}$ and $G_{Y→X}$ to be near identity matrices when real samples from target domains $y$ in $Y$ and $x$ in $X$ are fed as input, respectively. The loss term can be introduced as

$$
\mathcal{L}_{id}(G_{X→Y}, G_{Y→X}) = \mathbb{E}_{x \sim \text{data}(x)}[\|G_{Y→X}(x) - x\|_1] + \mathbb{E}_{y \sim \text{data}(y)}[\|G_{X→Y}(y) - y\|_1].
$$

Full Objective Function: The full objective function for the expression synthesis framework is the sum of all three loss terms as

$$
\mathcal{L}_{full}(G_{X→Y}, G_{Y→X}, D_X, D_Y) = \mathcal{L}_{GAN}(G_{X→Y}, D_Y, X, Y) + \mathcal{L}_{GAN}(G_{Y→X}, D_X, X, Y) + \lambda_1 \mathcal{L}_{rec}(G_{X→Y}, G_{Y→X}) + \lambda_2 \mathcal{L}_{id}(G_{X→Y}, G_{Y→X}),
$$

where $\lambda_1$ and $\lambda_2$ are ratios controlling the balance between these three terms. Generators $G_{X→Y}$, $G_{Y→X}$ and discriminators $D_X$, $D_Y$.
are trained by solving
\[
\arg \min_{G_{x \rightarrow y}} \max_{D_y} \mathcal{L}_{\text{full}}(G_{X \rightarrow Y}, G_{Y \rightarrow X}, D_{X}, D_{Y}) .
\] (5)

In practice, to further stabilise the training procedure and generate higher quality images, we replace the negative log likelihood adversarial loss $\mathcal{L}_{\text{GAN}}$ by a least-square loss \[39\], $\mathcal{L}_{\text{LSGAN}}$. Instead of solving $\min_{G_{x \rightarrow y}} \max_{D_y} \mathcal{L}_{\text{GAN}}(G_{X \rightarrow Y}, D_{Y}, X, Y)$, we optimise the generator and discriminator by
\[
\min_{D_y} \mathcal{L}_{\text{LSGAN}}(D_y) = \min_{D_y} \left\{ \mathbb{E}_{x \sim p_{\text{data}}(x)} [D_y(G_{X \rightarrow Y}(x))^2] + \mathbb{E}_{y \sim p_{\text{data}}(y)} [D_y(G_{Y \rightarrow X}(y) - 1)^2] \right\} .
\] (6)
\[
\min_{G_{x \rightarrow y}} \mathcal{L}_{\text{LSGAN}}(G_{X \rightarrow Y}) = \min_{G_{x \rightarrow y}} \mathbb{E}_{x \sim p_{\text{data}}(x)} [(D_y(G_{X \rightarrow Y}(x) - 1)^2].
\] (7)

In the $\mathcal{L}_{\text{LSGAN}}$, $L_2$ norm is used instead for the log losses of the original GAN, preventing vanishing gradients when the samples lie in the correct side but far away from the real data.

Due to that the objective function contains mapping functions in both directions $X \rightarrow Y$ and $Y \rightarrow X$, networks trained with loss $\mathcal{L}_{\text{full}}(G_{X \rightarrow Y}, G_{Y \rightarrow X}, D_{X}, D_{Y})$ can be applied to generate images for both input and target domains. Furthermore, since the objective function is defined on two domains rather than image pairs, there is no requirement for the existence of paired training samples in the input and target domains.

In expression synthesis, we take neutral expressions as the input domain and train networks for modelling other six basic expressions. Each pair of the expressions is separately trained as neutral and anger, neutral and disgust, neutral and fear, neutral and happiness, neutral and sadness, and neutral and surprise. The formulation of the loss function for mapping from neutral to happiness images is shown in Fig.1 and the translation from happiness to neutral shares similar structure and formulation, so are the other expression pairs.

### 3.2 Expression Synthesis Framework

The active appearance model (AAM) \[40\] extracts both shape attributes and texture properties of a face. An unseen face is represented by both shape and texture variations learned from training images. In expression synthesis, the statistical shape model and shape-free texture are commonly used for modelling shape and texture changes caused by skin deformation, respectively \[41\] \[42\].

In this paper, we derive shape attributes and texture information from each training sample as a 1-D vector $S$ and a 2-D image $T$, respectively. Shape and texture for different expressions are then produced and combined to form generated expressions. The general framework of expression synthesis is shown in Fig 2.

In order to reduce landmark discrepancies across subjects caused by scale, pose and viewpoint, shape alignment is necessary. For each face, its shape is represented with a set of $r$ landmarks located around key areas
\[
S = [(l_1^{(1)}, \beta_1^{(1)}), (l_2^{(2)}, \beta_2^{(2)}), \cdots, (l_m^{(t)}, \beta_t^{(t)})],
\] (8)

where $(l_i^{(k)}, \beta_i^{(k)})$ represents coordinates of $\beta$-th landmark along $x$ and $y$ axes, respectively. Translation, rotation and scaling are used to align all the shapes so that eyes are located at the same positions. Given a set of aligned shapes, we split the landmarks of each shape into four groups based on facial features: eyebrows and eyes, nose, mouth and face contour. In addition, for texture attributes, we warp all faces in each expression category so that their landmarks match the mean shape of that expression. According to the distribution of landmarks, Delaunay triangulation divide faces into small triangles based on the landmarks. Piece-wise affine transformation is adopted to warp texture from the original shape to the mean shape in corresponding triangles.

Given an input neutral image and its shape and texture attributes $S^m$ and $T^m$, the trained shape and texture generators synthesise $S^t$ and $T^t$ for different expressions. We warp the generated texture $T^t$ to the synthesised shape $S^t$ to obtain the final synthesised expression.

In photorealistic expression synthesis, inner mouth region is person-specific and cannot be ignored in certain expressions especially happiness. Since there is no landmark labelled within the mouth, synthesised inner mouth region from the training set

---

IET Research Journals, pp. 1–10
© The Institution of Engineering and Technology 2015

3
Algorithm 1 The proposed ApprGAN algorithm.

**Input**: A training set containing neutral and target expressions, and a test set of neutral images.

**Output**: Subjects from the test set with the target expressions.

1. (Pre-processing) Extract shape vector $S$ and texture attribute $T$ from each image of training and test sets for shape generation and texture synthesis, respectively.
2. Build the ApprGAN network with objective function according to Eq. 4 from the training set.
3. Train shape and texture generators and discriminators by solving Eqs. 5, 6 and 7.
4. For each test subject, derive synthesised shape and texture attributes using trained shape and texture generators, and obtain its target expression by warping the generated texture to synthesised shape.
5. If target expression is happiness, replace the synthesised inner mouth shape with one of the training subjects having the most similar mouth shape.
6. Calculate normalised landmark distances and correlation coefficients between the synthesised and ground truth shape and texture attributes using Eqs. 9 and 10 respectively.

![Diagram](image)

**Fig. 2**: General process of expression synthesis.

Table 1 Network architecture used for shape and texture generators.

| Shape Generator | Texture Generator |
|-----------------|-------------------|
| Input 204 × 1   | Input 200 × 200   |
| 32 × 7 × 1 conv., stride 1, ReLU | 64 × 7 × 7 conv., stride 1, ReLU |
| 64 × 3 × 1 conv., stride 2, ReLU | 128 × 3 × 3 conv., stride 2, ReLU |
| 128 × 3 × 1 conv., stride 2, ReLU | 256 × 3 × 3 conv., stride 2, ReLU |
| Residual block, 128 filters | Residual block, 256 filters |
| Residual block, 128 filters | Residual block, 256 filters |
| Residual block, 128 filters | Residual block, 256 filters |
| Residual block, 128 filters | Residual block, 256 filters |
| Residual block, 128 filters | Residual block, 256 filters |
| 64 × 3 × 1 conv., stride 1/2, ReLU | 128 × 3 × 3 conv., stride 1/2, ReLU |
| 32 × 3 × 1 conv., stride 1/2, ReLU | 64 × 3 × 3 conv., stride 1/2, ReLU |
| Reflection padding 3 × 1 | Reflection padding 3 × 3 |
| 1 × 7 × 1 conv., stride 1, ReLU | 1 × 7 × 7 conv., stride 1, ReLU |

3.3 **Expression Verification on Synthesised Images**

To evaluate the quality of synthesised expressions, we compare synthesised images with the ground truth images on both shape and texture attributes [43]. For shape attributes, normalised distance is computed between the ground truth landmarks $\{(l_{m, i}^n, l_{m, j}^n)\}$ and the synthesised landmarks $\{(l_{m, i}^g, l_{m, j}^g)\}$.

$$
\text{dist} = \sum_{\beta=1}^{z} \left( \left( \frac{|(l_{m, i}^g)^{(\beta)} - (l_{m, i}^n)^{(\beta)}|}{\text{width}} \right)^2 + \left( \frac{|(l_{m, j}^g)^{(\beta)} - (l_{m, j}^n)^{(\beta)}|}{\text{height}} \right)^2 \right),$$

(9)

where $|·|$ is the absolute function, and $\text{width}$ and $\text{height}$ refer to the largest distances between each ground truth shapes in $x$ and $y$ axes, respectively.

To measure the texture similarity, we calculate the correlation coefficients between the synthesised texture $T^g$ and its corresponding ground truth texture $T^x$ as

$$
\rho = \frac{\sqrt{\sum_{i, j} (T^x(i, j) - \mu^x)(T^g(i, j) - \mu^g)}}{\sqrt{\sum_{i, j} (T^x(i, j) - \mu^x)^2 \sum_{i, j} (T^g(i, j) - \mu^g)^2}},

(10)

where $(i, j)$ specifies pixel locations in the texture images, and $\mu^g$ and $\mu^x$ are the mean values of $T^g$ and $T^x$, respectively.
Table 2: Synthesised expressions on Bosphorus database when trained using subjects from Bosphorus, CK+, MUG, combined CK+ and MUG, or combined Bosphorus, CK+ and MUG databases.

| Test Training | Synthesised Expressions |
|---------------|-------------------------|
| Input Neutral | ![Image](image_url) |
| Bosphorus     | ![Image](image_url) |
| CK+           | ![Image](image_url) |
| Bosphorus     | ![Image](image_url) |
| MUG           | ![Image](image_url) |
| CK+, MUG      | ![Image](image_url) |
| Bosphorus, CK+, MUG | ![Image](image_url) |

4 Experiments and Results

4.1 Databases

We conducted several experiments on three publicly available databases: Bosphorus, CK+, and MUG facial expression databases.

The Bosphorus database [11] consists of 105 subjects (60 men and 45 women, aged between 25 and 35) with a total number of 4666 images. Images are of a high resolution of 1600 × 1200 pixels with variations in expression, pose, and occlusion. Each subject has up to 35 expressions, including 7 prototypical expressions (neutral and 6 basic emotions), and each image has 24 manually labelled landmarks per image.

The CK+ database [12, 13] is a collection of videos of 123 subjects performing various expressions: anger, contempt, disgust, fear, happiness, sadness, and surprise. There are totally 593 video sequences starting with neutral images (the first frames) and ending with expressions at the maximum level (the last frames). Video frames are of sizes of 640 × 480 or 640 × 490 pixel arrays and annotated with 68 facial landmarks in sizes of s.

The MUG database [14] contains image sequences of 86 subjects (51 men and 35 women, aged between 20 and 35) performing various expressions (neutral and 6 basic emotions), among which only images of 52 participants are publicly available. Video sequences are of sizes of 640 × 480 or 640 × 490 pixel arrays and annotated with 68 facial landmarks in sizes of s.

4.2 Network Architecture and Training settings

We adopted 10-fold cross-validation for expression synthesis due to limited number of subjects. We randomly split each database into 10 groups and built 10 pairs of training and test sets for each database. For example, the happiness face of subject ξ in group 1 was synthesised using generators trained with images from all other groups 2-10. Therefore, there was no overlapping between the training and test identities and images. In texture synthesis, each input image was resized to 200×200. The input vectors for synthesising shapes were made of coordinates of 83 landmarks (8 points for inner mouth shape), and zeros inserted between four landmark groups, producing vectors with a length of 204.

Generator Architectures: The generator network has a similar architecture to the image transformation networks [31], which have shown impressive results for both style transfer and image super resolution. We adopt 6 residual blocks, fractionally strided convolutions with stride 1/2, reflection padding and instance normalisation. Network architectures used for shape and texture generators are displayed in Table 1. Each residual block contains two 3×1 (in shape generator) or 3×3 (in texture generator) convolutional layers with the same number of filters and an instance normalisation layer. To reduce artifacts, spatial reflection padding was used in convolutional blocks so that input and output of each convolutional layer have the same size.

Discriminator Architectures: In constructing discriminators, we adopted a Markovian discriminator - PatchGAN [28]. PatchGAN assumes only local dependency between pixels and models images as Markov random fields. Instead of targeting the whole image, this patch based discriminator aims to classify whether each overlapping patch in an image is real or fake. We used 34×1 and 70×70 PatchGANs as discriminators in synthesising shape and texture, respectively. Let 32C4×1 represent a 4×1 convolutional layer with 32 filters, stride 2 and leaky ReLUs having a slope of 0.2, the discriminator architectures are: 32C4×1—64C4×1—128C4×1 for shape discriminator and

http://www.faceplusplus.com/
Table 3 Synthesised expressions on CK+ database when trained using subjects from CK+, Bosphorus, MUG, combined Bosphorus and MUG, or combined Bosphorus, CK+ and MUG databases.

| Test | Training | Synthesised Expressions |
|------|----------|-------------------------|
|      |          | Anger | Disgust | Fear | Happiness | Sadness | Surprise |
| Input Neutral | CK+ | ![Anger](image) | ![Disgust](image) | ![Fear](image) | ![Happiness](image) | ![Sadness](image) | ![Surprise](image) |
| | Bosphorus | ![Anger](image) | ![Disgust](image) | ![Fear](image) | ![Happiness](image) | ![Sadness](image) | ![Surprise](image) |
| | CK+ | ![Anger](image) | ![Disgust](image) | ![Fear](image) | ![Happiness](image) | ![Sadness](image) | ![Surprise](image) |
| | MUG | ![Anger](image) | ![Disgust](image) | ![Fear](image) | ![Happiness](image) | ![Sadness](image) | ![Surprise](image) |
| | Bosphorus, MUG | ![Anger](image) | ![Disgust](image) | ![Fear](image) | ![Happiness](image) | ![Sadness](image) | ![Surprise](image) |
| | Bosphorus, CK+, MUG | ![Anger](image) | ![Disgust](image) | ![Fear](image) | ![Happiness](image) | ![Sadness](image) | ![Surprise](image) |
| Ground Truth Images | | ![Anger](image) | ![Disgust](image) | ![Fear](image) | ![Happiness](image) | ![Sadness](image) | ![Surprise](image) |

64C4×4−128C4×4−256C4×4−512C4×4 for texture discriminator, respectively. After these convolutional layers, a final layer was added to produce a 1D output. Instance normalisation was used in each convolutional layer except for the first one.

**Training Details:** The standard adversarial training of discriminators only focuses on the latest generated image, which may cause divergence and introduce some artifacts [44]. Based on these observations, we instead updated discriminators using a history of previously generated images. During the adversarial training, the size of image buffer was the same as the number of training images. For all experiments, we set $\lambda_2 = 0.5\lambda_1$, and $\lambda_1 = 10$ for shape generation and $\lambda_1 = 5$ for texture synthesis, as shape is more important than texture in regularising expression synthesis, supported by the experimental results. The networks were all trained from scratch using the Adam solver [45] with a mini-batch size of 1, and weights were initialised from a Gaussian distribution with 0 mean and 0.02 standard deviation. The learning rate started at 0.0002 was kept the same for the first 100 epochs and then linearly decreased to 0 over the next 100 epochs. Horizontal flipping was employed to augment data.

### 4.3 Expression Synthesis Results

For each database, we adopted 10-fold cross-validation strategy to split training and test pairs and train the networks.

For the Bosphorus database, 62 subjects were selected with each having a neutral image and six basic expressions. Table 2 displays the input neutral image in the 1st row and the ground truth images at the bottom row. Synthesised six basic expressions are shown in the 2nd row: anger, disgust, fear, happiness, sadness and surprise, from left to right.

For the CK+ database, we selected video sequences displaying six basic emotions: anger (33), disgust (46), fear (20), happiness (50), sadness (18) and surprise (45), in which videos in a certain expression were all from different subjects. The first and last frames in each video sequence were considered as the neutral image and the most expressive image respectively. With input neutral images in the 1st row, the 2nd and bottom rows of Table 3 illustrate the synthesised expressions and the corresponding ground truth images, respectively (from left to right: anger, disgust, fear, happiness, sadness and surprise).

On the MUG database, video sequences of 51 subjects were selected, from which the first frames were taken as the neutral images and the expression images at its peak intensity were used. Table 4 illustrates the input neutral images (in the 1st row), synthesised anger, disgust, fear, happiness, sadness, and surprise expressions (from left to right in the 2nd row), and the ground truth expressions at peak intensity (in the bottom row).

### 4.4 Expression Verification Results

To measure the quality of the synthesised expressions, we calculated the normalised distances, Eq. (9), for shape attributes and correlation coefficients, Eq. (10), for texture attributes between pairs of synthesised and ground truth images. Results for the Bosphorus, CK+ and MUG databases are presented in Tables 5, 6 and 7, respectively. The average correlation coefficients and normalised distances from these tables are $\rho_{\text{Bosphorus}} = 0.972 \pm 0.014$, $\text{dist}_{\text{Bosphorus}} = 2.061 \pm 0.681$; $\rho_{\text{CK+}} = 0.953 \pm 0.020$, $\text{dist}_{\text{CK+}} = 1.879 \pm 0.456$; and $\rho_{\text{MUG}} = 0.975 \pm 0.009$, $\text{dist}_{\text{MUG}} = 1.835 \pm 0.534$, showing the closeness of the synthesised and the ground truth expressions.

Results of expression verification experiments on various databases show that the proposed ApprGAN can synthesise good expressions, while preserving shape, texture details and identity information.
Table 4 Synthesised expressions on MUG database when trained using subjects from MUG, Bosphorus, CK+, combined Bosphorus and CK+, or combined Bosphorus, CK+, and MUG databases.

| Test | Training | Synthesised Expressions |
|------|----------|-------------------------|
| Anger | Disgust | Fear | Happiness | Sadness | Surprise |
| Input Neutral | ![Image](image1) | ![Image](image2) | ![Image](image3) | ![Image](image4) | ![Image](image5) | ![Image](image6) |
| MUG | ![Image](image7) | ![Image](image8) | ![Image](image9) | ![Image](image10) | ![Image](image11) | ![Image](image12) |
| Bosphorus | ![Image](image13) | ![Image](image14) | ![Image](image15) | ![Image](image16) | ![Image](image17) | ![Image](image18) |
| MUG | ![Image](image19) | ![Image](image20) | ![Image](image21) | ![Image](image22) | ![Image](image23) | ![Image](image24) |
| CK+ | ![Image](image25) | ![Image](image26) | ![Image](image27) | ![Image](image28) | ![Image](image29) | ![Image](image30) |
| Bosphorus, CK+ | ![Image](image31) | ![Image](image32) | ![Image](image33) | ![Image](image34) | ![Image](image35) | ![Image](image36) |
| Bosphorus, CK+, MUG | ![Image](image37) | ![Image](image38) | ![Image](image39) | ![Image](image40) | ![Image](image41) | ![Image](image42) |

Ground Truth Expressions

Table 5 Correlation coefficients and normalised distances between synthesised and ground truth texture, landmarks on Bosphorus database.

| Expression | Expression Verification (Mean ± Standard Deviation) | Correlation Coefficients | Normalised Distances |
|------------|----------------------------------------------------|--------------------------|---------------------|
| Anger      | 0.975 ± 0.012                                    | 2.054 ± 0.923            |
| Disgust    | 0.965 ± 0.015                                    | 2.004 ± 0.684            |
| Fear       | 0.972 ± 0.014                                    | 2.160 ± 0.639            |
| Happiness  | 0.970 ± 0.021                                    | 1.884 ± 0.525            |
| Sadness    | 0.977 ± 0.011                                    | 1.990 ± 0.695            |
| Surprise   | 0.974 ± 0.011                                    | 2.072 ± 0.622            |
| Average    | 0.972 ± 0.014                                    | 2.061 ± 0.681            |

Table 6 Correlation coefficients and normalised distances between synthesised and ground truth texture, landmarks on MUG database.

| Expression | Expression Verification (Mean ± Standard Deviation) | Correlation Coefficients | Normalised Distances |
|------------|----------------------------------------------------|--------------------------|---------------------|
| Anger      | 0.981 ± 0.006                                    | 2.019 ± 0.899            |
| Disgust    | 0.968 ± 0.014                                    | 1.884 ± 0.572            |
| Fear       | 0.978 ± 0.007                                    | 1.766 ± 0.430            |
| Happiness  | 0.976 ± 0.010                                    | 1.571 ± 0.319            |
| Sadness    | 0.976 ± 0.009                                    | 1.727 ± 0.443            |
| Surprise   | 0.970 ± 0.010                                    | 2.046 ± 0.540            |
| Average    | 0.975 ± 0.009                                    | 1.835 ± 0.534            |

4.5 Cross-database Generalisation Results

Expression synthesis across databases was also implemented to test the generalisation ability of the proposed ApprGAN. Taken one database as training set, we tested the performance of trained model on other databases. Experiments were conducted across the Bosphorus, CK+, MUG databases. For example, we synthesised expressions for subjects in Bosphorus database using the models trained on CK+, MUG, combined CK+ and MUG, and combined Bosphorus, CK+ and MUG respectively. Results are shown in the 3rd to 6th rows of Table 4. There was no overlapping images between the training and test sets. Correlation coefficients and normalised distances were also calculated between synthesised and ground truth expressions to verify the robustness of the proposed method. Similar experimental protocols were adopted for the Bosphorus and CK+ databases.

Similarly, cross-synthesised expressions are presented in Tables 5 and 6 for the CK+ and MUG databases respectively. The 3rd to the 6th rows of these Tables display synthesised expressions when trained on various databases, and the corresponding ground truth expression images are shown in the bottom rows (from left to right: anger, disgust, fear, happiness, sadness and surprise). Table 7 lists the expression verification results in terms of correlation and
The correlation coefficients and normalised distances are 0.966 ± 0.011 for subjects in Bosphorus, 0.962 ± 0.015, 2.472 ± 0.811 for subjects in Bosphorus, 0.962 ± 0.015, 2.267 ± 0.617 for subjects in CK+, and 0.973 ± 0.010, 1.980 ± 0.501 for subjects in MUG databases, respectively. Compared to results in Tables 5, 6 and 7, it is noted that despite of training images from different databases, cross-synthesised expressions closely assembles the ground truth images.

Table 8 Results of expression verification experiments for cross-database synthesis.

| Test | Training | Correlation Coefficients | Normalised Distances |
|------|----------|--------------------------|----------------------|
| Bosphorus | CK+ | 0.965 ± 0.013 | 2.428 ± 0.895 |
| | MUG | 0.973 ± 0.013 | 2.105 ± 0.815 |
| | CK+, MUG | 0.970 ± 0.013 | 2.228 ± 0.891 |
| CK+ | MUG | 0.965 ± 0.015 | 1.923 ± 0.530 |
| | Bosphorus, MUG | 0.966 ± 0.012 | 1.965 ± 0.554 |
| | Bosphorus, CK+, MUG | 0.970 ± 0.011 | 1.983 ± 0.573 |
| MUG | Bosphorus | 0.980 ± 0.006 | 1.864 ± 0.429 |
| | CK+ | 0.975 ± 0.005 | 2.106 ± 0.522 |
| | Bosphorus, CK+ | 0.977 ± 0.006 | 1.817 ± 0.448 |
| | Bosphorus, CK+, MUG | 0.981 ± 0.006 | 1.833 ± 0.446 |

Table 9 Expression synthesis results on MUG database by various methods.

| Methods | Anger | Disgust | Fear | Happiness | Sadness | Surprise |
|---------|-------|---------|------|-----------|---------|----------|
| Ground Truth Expressions | ![Image](image1.png) | ![Image](image2.png) | ![Image](image3.png) | ![Image](image4.png) | ![Image](image5.png) | ![Image](image6.png) |
| Proposed ApprGAN | ![Image](image7.png) | ![Image](image8.png) | ![Image](image9.png) | ![Image](image10.png) | ![Image](image11.png) | ![Image](image12.png) |
| StarGAN [8] | ![Image](image13.png) | ![Image](image14.png) | ![Image](image15.png) | ![Image](image16.png) | ![Image](image17.png) | ![Image](image18.png) |
| CycleGAN [29] | ![Image](image19.png) | ![Image](image20.png) | ![Image](image21.png) | ![Image](image22.png) | ![Image](image23.png) | ![Image](image24.png) |
| Eigentransformation Based Approach [43] | ![Image](image25.png) | ![Image](image26.png) | ![Image](image27.png) | ![Image](image28.png) | ![Image](image29.png) | ![Image](image30.png) |
| Input Neutral | ![Image](image31.png) | ![Image](image32.png) | ![Image](image33.png) | ![Image](image34.png) | ![Image](image35.png) | ![Image](image36.png) |

Results of the experiments across databases show the effectiveness of the proposed ApprGAN for generalising to other databases. Training on the combined Bosphorus, CK+ and MUG databases further improve generalisation ability and produce more realistic expressions. Although that different illumination conditions exist and subjects in these databases appear to perform some expressions in slightly different ways, the proposed ApprGAN still generate natural expressions.
Table 10: Expression verification result comparison of the proposed ApprGAN with the state-of-the-art methods.

| Databases | Expression Verification (Mean±Standard Deviation) | Eigentransformation Based Expression Synthesis [43] | CycleGAN [29] | StarGAN [8] | Proposed ApprGAN |
|-----------|-----------------------------------------------|-----------------------------------------------|--------------|--------------|-----------------|
| Bosphorus | Correlation Coefficients: 0.971 ± 0.014 | Landmark Distances: 1.765 ± 0.572 | 0.954 ± 0.021 | 0.995 ± 0.017 | 0.972 ± 0.014 |
| CK+       | Correlation Coefficients: 0.969 ± 0.013 | Landmark Distances: 1.931 ± 0.572 | 0.943 ± 0.022 | 0.948 ± 0.020 | 0.953 ± 0.020 |
| MUG       | Correlation Coefficients: 0.978 ± 0.009 | Landmark Distances: 1.542 ± 0.480 | 0.951 ± 0.021 | 0.957 ± 0.018 | 0.975 ± 0.009 |

Table 11: Cross-database expression verification result comparison of the proposed ApprGAN and the eigentransformation based approach [43].

| Test | Training | Eigentransformation Based Expression Synthesis [43] | Proposed ApprGAN |
|------|----------|-----------------------------------------------|-----------------|
|      |          | Correlation Coefficients (Mean±Standard Deviation) | Landmark Distances (Mean±Standard Deviation) | Correlation Coefficients (Mean±Standard Deviation) | Landmark Distances (Mean±Standard Deviation) |
| CK+  |          | 0.915 ± 0.036 | 2.510 ± 0.725 | 0.958 ± 0.016 | 2.810 ± 0.774 |
| Bosphorus |          | 0.950 ± 0.017 | 2.401 ± 0.716 | 0.963 ± 0.017 | 2.222 ± 0.672 |
| CK+, MUG |          | 0.960 ± 0.016 | 2.426 ± 0.723 | 0.963 ± 0.016 | 2.606 ± 0.803 |
| Bosphorus |          | 0.950 ± 0.018 | 2.446 ± 0.694 | 0.957 ± 0.017 | 2.471 ± 0.621 |
| CK+  |          | 0.950 ± 0.019 | 2.467 ± 0.698 | 0.950 ± 0.023 | 2.184 ± 0.581 |
| MUG   |          | 0.957 ± 0.018 | 2.396 ± 0.682 | 0.957 ± 0.018 | 2.262 ± 0.605 |
| Bosphorus, CK+, MUG |          | 0.962 ± 0.015 | 2.462 ± 0.726 | 0.962 ± 0.015 | 2.472 ± 0.811 |
| CK+  |          | 0.960 ± 0.013 | 2.046 ± 0.516 | 0.965 ± 0.012 | 2.030 ± 0.494 |
| Bosphorus, CK+ |          | 0.950 ± 0.023 | 2.057 ± 0.517 | 0.959 ± 0.019 | 2.306 ± 0.552 |
| MUG   |          | 0.964 ± 0.013 | 2.063 ± 0.504 | 0.964 ± 0.013 | 1.972 ± 0.504 |
| Bosphorus, CK+, MUG |          | 0.973 ± 0.010 | 1.980 ± 0.501 |              |              |

4.6 Comparisons and Discussions

We first compared the proposed method ApprGAN with the eigentransformation based expression synthesis approach [43]. Experiments were conducted under the same protocols (image acquisition, partition and pre-processing) with only neutral images and expressions at maximum intensity considered. Both methods can generate natural expressions and generalise well across databases. Unlike that the eigentransformation based method requires paired training samples, the proposed method can be applied as an unsupervised way in a wide range of scenarios. The proposed ApprGAN was also compared with the original CycleGAN [29] and the StarGAN [8]. Experimental protocols are the same for all the methods, and all training parameters were the same to their original papers except that we trained the StarGAN for longer, i.e. 1600 epochs to compensate for limited number of training samples. Images were fed directly to the CycleGAN and StarGAN, without separated as shape and texture attributes; this causes unnatural expressions and artifacts in some cases. Table[9] shows the expression synthesis results of these four methods together with ground truth expressions. It can be observed that the two image-based approaches CycleGAN and StarGAN could not capture some subtle expression changes (fear and sadness) and caused some artifacts with intensive expressions (disgust, happiness and surprise). Nevertheless, the eigentransformation based method and the proposed ApprGAN produced photorealistic facial expressions with intensity variations.

To be able to conduct quantitative analysis, we extracted 83 landmarks [10] and derived texture attributes by affine warping from the synthesised images using CycleGAN and StarGAN. Quantitative expression verification results of these four methods are listed in Table[10] with the proposed ApprGAN significantly outperforming CycleGAN and StarGAN, further validating the effectiveness of combining shape and texture in the generative adversarial model. Further experiments on cross-database synthesis and comparisons of the proposed ApprGAN with the eigentransformation based method were also conducted. As shown in Table[11] the proposed ApprGAN outperformed the eigentransformation based method markedly in both texture and shape synthesis in most cases.

The proposed method yields compelling results in many cases and generalises well across databases, though there are still several limitations. 1) The proposed method relies on good acquisition of landmarks to derive shape and texture attributes for pre-processing. 2) The lack or limited number of samples in the training set can cause poor performance. 3) It is hard to generate expressions for subjects with specific features (such as glasses, beard and unusual teeth). Models trained on a small database can produce unstable results with some artifacts. In addition, with much larger number of parameters, the training process is much more time-consuming than the eigentransformation based method.

5 Conclusions

In this paper, we have proposed an expression synthesis framework based on the generative adversarial learning. With the combination of geometry and appearance information, we separately train shape and texture models for generating facial expressions from a single image. Associated with the basic adversarial loss in GANs, we further incorporate reconstruction and identity mapping losses into the objective function. The combination of these three losses stabilise the network training and meanwhile maintaining personal identity and facial features of the input, which is important to precisely synthesising and interpreting facial expressions. The proposed ApprGAN yields realistic synthesised expressions. It has also been shown to generalise well across databases with differences in expressing emotions and illumination conditions. Comparisons with the eigentransformation based method, CycleGAN and StarGAN have demonstrated the advantages of the proposed method both visually and quantitatively. Future work will focus on developing robust expression synthesis approach against pose variations and illumination changes.

6 References

1. B. Fusel, J. Luettin. “Automatic facial expression analy- sis: a survey”, Pattern Recognit., 36(1), pp. 259–275, (2003).
2. C. Shan, S. Gong, P. W. McOwan. “Facial expression recognition based on local binary patterns: A comprehensive study”, Image Vis. Comput., 27(6), pp. 803–816, (2009).
3. Y. Rahulamathavan, R. C.-W. Phan, J. A. Chambers, D. J. Parish. “Facial expression recognition in the encrypted domain based on local fisher discriminant analysis”, IEEE Trans. Affect. Comput., 4(1), pp. 83–92, (2013).
4. Z. Liu, Y. Shan, Z. Zhang. “Expression expression mapping with ratio images”, Proc. Conf. Computer Graphics and Interactive Techniques, pp. 271–276, (2001).
