Complex & Intelligent Systems (2024) 10:4609–4624
https://doi.org/10.1007/s40747-024-01401-7

ORIGINAL ARTICLE

Semantic prior guided fine-grained facial expression manipulation

Tao Xue[1] · Jin Yan[1] · Deshuai Zheng[1] · Yong Liu[1]

Received: 9 August 2023 / Accepted: 22 February 2024 / Published online: 27 March 2024

Abstract
Facial expression manipulation has gained wide attention and has been applied in various fields, such as film production, electronic games, and short videos. However, existing facial expression manipulation methods often overlook the details of local regions in images, resulting in the failure to preserve local structures and textures of images. To solve this problem, this paper proposes a local semantic segmentation mask-based GAN (LSGAN) to generate fine-grained facial expression images. LSGAN is composed of a semantic mask generator, an adversarial autoencoder, a transformative generator, and an AU-intensity discriminator. Our semantic mask generator generates eye, mouth, and cheek masks of face images. Then, our transformative generator integrates target expression labels and corresponding facial region features to generate a vivid target facial expression image. In this fashion, we can capture expressions from target face images explicitly. Furthermore, an AU-intensity discriminator is designed to capture facial expression variations and evaluate quality of generated images. Extensive experiments demonstrate that our method achieves authentic face images with accurate facial expressions and outperforms state-of-the-art methods qualitatively and quantitatively.

Keywords Semantic segmentation · Fine-grained expression manipulation · Generative adversarial network

Introduction

Facial expression manipulation aims to convert facial expression of target images into source images while maintaining the image's original identity information. It has gained significant potential in various applications, such as film production, electronic games, and short videos. Researches in the field of artificial intelligence has made significant progress [1–4]. However, since face expression is complexed and would bring ambiguous for face structure, it is challenging to produce accurate facial expression images with photo-realistic textures and faithful structures.

State-of-the-art facial expression manipulation methods can be divided into two categories: message judgment methods and sign judgment methods.

(i) Message judgment methods focus on directly encoding expression features and thus learning expression patterns of face images. For example, StarGAN [5] is proposed to transfer expressions between different image domains under the guidance of discrete expression labels, e.g., happy, angry, or sad. However, message judgment methods fail to perform continuous expression editing and cannot guarantee the quality of generated images.
(ii) Sign judgment methods estimate facial action unit (AU) signals to synthesize expressions. For example, Pumarola et al. [6] utilized AU intensities as guidance to synthesize facial expression images. They incorporated attention mechanisms into the generator's last layer, enabling the manipulation of images with complex backgrounds. However, labeling AU for face datasets consumes enormous labor. Meanwhile, solely learning global AU features would limit local expression edition performance.

✉ Yong Liu
liuy1602@njust.edu.cn

Tao Xue
xt1111@njust.edu.cn

Jin Yan
yanjin@njust.edu.cn

Deshuai Zheng
zds@njust.edu.cn

[1] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China

Research [7] has shown that human attention naturally focuses on special facial regions when recognizing and distinguishing facial expressions. For example, the eyes play a vital role in fear analysis, while the mouth is crucial for identifying happiness. Driven by this analysis, we propose to extract local facial features of key facial regions, i.e., eye, mouth, and cheek regions, and then inject target AU signals into local facial features for facial expression manipulation. In this fashion, expression features could be integrated into corresponding facial regions purposefully and thus facilitate fine-grained facial expression manipulation.

Previous methods mainly focus on the whole face, neglecting local facial parts, resulting in overlapping and blurring of local facial regions in the generated results. Toward this problem, we propose a local semantic segmentation mask-based GAN (LSGAN). Our LSGAN captures texture details of key facial regions by designing several networks based on local semantic regions, and uses reconstruction networks to further preserve the structural information of the image. Based on the above, our LSGAN comprises a semantic mask generator (SMG), an adversarial autoencoder (AAE), a transformative generator (TG), and an AU-intensity discriminator (AUD). First, we design the SMG to generate masks of key facial regions of source facial images, i.e., eye, mouth, and cheek regions. Then, we propose the AAE to map facial masks into structured latent codes. Specifically, we introduce the ME-graphAU [8] to predict the AU intensity of target images. Afterward, our TG integrates target AU-intensity labels and corresponding source facial region codes to generate desired facial expression images. Furthermore, we design AUD to capture facial expression variations and evaluate the quality of generated images. During training, we introduce reconstruction losses to preserve generated faces' identity and structure information. The researches [9, 10] have made significant contributions to function optimization, and based on these studies, we adopt adaptive moment estimation (Adam) solver to optimize the loss function of each module.

The main contributions of our work are threefold:

- We propose a facial expression manipulation method, dubbed LSGAN, to generate target facial expression images with photo-realistic textures and faithful structures. LSGAN combines key facial region masks with target AU-intensity labels to achieve facial expression manipulation.
- We design a AAE to generate latent codes of facial semantic masks. In particular, our TG integrates latent codes with target expression labels to generate desired facial expression images, thus alleviating the correspondence ambiguity between source and target expression faces. We introduce self-reconstruction and cyclic reconstruction with same local network structure as generator to ensure the stability of our generated network and maintain the feature and structural invariance of unrelated regions.
- Our experiments demonstrate that LSGAN can achieve better facial expression manipulation performance. The average MSE of the 16 AU intensities in our method is 0.018, which is lower than the state-of-the-art methods. We also demonstrate the importance of the special facial region partition in facial expression synthesis.

Related work

Given two unpaired images (I_i, I_t), our main goal is to generate new image $\widetilde{I_i}$ with facial expression of I_t while preserving the identity information in the original image I_i. Due to the fact that differences in facial expressions often occur in key areas such as the eyes and mouth, rather than uniformly across the entire face, it has prompted us to perform attribute classification on faces to capture the features of these key regions. Subsequently, these key region features are considered and utilized during the process of facial expression manipulation. Therefore, the related work mainly includes two aspects: facial attribute classification and facial expression manipulation.

Facial attribute classification

Yang et al. [11] focused on facial attribute recognition through the utilization of deep convolution neural networks. The primary objective of this study is to attain a heightened response within facial regions, consequently generating candidate windows of faces. However, the complexity of the CNN structure results in significant time costs when implemented practically. To address this limitation, Zhang et al. [12] proposed a novel framework that integrated face detection and alignment tasks by employing unified cascaded CNNs and multi-task learning. This approach aims to streamline the process and improve efficiency. While the aforementioned method effectively acquires feature points for facial attributes, the resultant attribute regions often encompass redundant components. Aggarwal [13, 14] applied deep learning method to crop segmentation and achieved relatively complete region segmentation results.

To enhance accuracy and alleviate the impact of overlapping regions, Zhao et al. [15] utilized a semantic segmentation method for facial attribute classification. This article introduces a global pyramid pooling, which offers additional contextual information. Moreover, it also proposes a deep supervised optimization strategy designed for ResNet-based fully convolutional networks (FCNs). To tackle the drawbacks associated with employing hole convolution in semantic segmentation, Lin et al. [16] proposed a multi-path reinforcement network called RefineNet, which explicitly

incorporates all the information derived from the downsampling process and employs remote residual connections to achieve accurate high-resolution predictions. RefineNet only utilizes the residual layer of a conventional ResNet, thereby avoiding the computational costs associated with hole convolution.

To further reduce the computational time of semantic segmentation networks, Yu et al. [17] introduced a bilateral segmentation network, known as Bilateral Segmentation Network (BiSeNet), with the objective of striking a balance between accuracy and speed. As an enhancement to BiSeNet, Yu et al. [18] proposed BiSeNetV2, which features a more concise bilateral structure. This article introduces multiple auxiliary training branches to enhance the feature extraction capabilities of various shallow networks. Moreover, the authors designed an efficient feature fusion module to effectively integrate spatial detail information with high-level semantic information. However, the utilization of a simplistic feature extraction framework in their methods lead to a decline in accuracy. Additionally, there is potential for further improvement in enhancing the attention framework within the network.

Facial expression manipulation

Generative adversarial networks have emerged as the prevailing approach in facial expression manipulation. Expanding upon classic GAN architectures, numerous variants have been devised to further improve performance. One such variant is conditional GAN (cGAN), which introduces additional condition information to control the distribution of generated data. In recent years, there has been a proliferation of studies utilizing cGAN for facial expression synthesis.

Prominent approaches, such as StarGAN [5] and AttGAN [19], employ generator networks that utilize input images and target domain information to generate images in diverse domains. For instance, StarGAN [5] utilizes facial images and target facial attributes to enable attribute editing through a single generator and discriminator. On the other hand, AttGAN [19] adopts an encoder-decoder architecture similar to StarGAN but represents facial attributes using latent representations. Ding et al. [20] modeled the intensity of facial expressions to generate a wider range of expressions, but the global expression in the method only describes the overall facial emotion, resulting in limited ability to capture fine details. Geng et al. [21] used 3D Morphable Model (3DMM) to fit an image and then re-render the image with desired expression. Another model, namely LGP-GAN [22], employs a two-stage cascaded structure and integrates both local and global perception to generate facial expressions. However, the complexity of facial expressions, especially microexpression with complex local details, presents significant challenges for these methods.

The facial action coding system (FACS) [23] provides commonly used descriptors such as raised cheeks or depressed lips in expression manipulation approaches. Tools like OpenFace [24] have been developed to achieve the recognition of Action Units (AUs). Additionally, a recent work called ME-graphAU [8] introduces a deep learning-based approach for modeling AU relationships explicitly. This method aims to describe the intricate connections between different AUs and provide a more comprehensive understanding of facial expressions.

With the help of these tools, Pumarola et al. [6] introduced a technique that utilizes AUs to guide the synthesis of facial expressions. This approach allows for precise control over the strength of individual AUs and their combination to form a cohesive expression. However, it only learns global AU features, which limits its performance in editing local expression details. Wang et al. [25] added a path for predicting an appearance flow to align the input image to the target expression. Wu et al. [26] proposed a cascaded expression focal GAN that progressively modifies facial expressions by emphasizing local expression features. On the other hand, considering the distinct structured appearances of facial expressions, Song et al. [27] and Qiao et al. [28] proposed geometrically guided Gans that leverage facial markers to define the facial geometry and generate facial expressions. Nevertheless, aligning landmarks from source images with target images that possess distinct facial shapes is a significant challenge and frequently results in the presence of artifacts within the generated images. These approaches typically rely on global expressions, AU predictions, or landmark predictions and lack the ability to estimate them automatically. Ling et al. [29] improved generator architecture in GANimation and used relative AUs as input. With relative action units, the generator learns to only transform regions of interest which are specified by non-zero-valued relative AUs. To further preserve identity information and edit relevant areas, Wang et al. [30] added an attention module to the generator for facial expression manipulation, and obtained long-range dependencies in the image by using self attention blocks instead of direct skip connections. In order to improve the performance of expression transfer, Shao et al. [31] disentangled the input image into two fine-grained representations (AU-related and AU-free features), and proposed an EET framework to explicitly transfer fine-grained expressions by straightforwardly mapping the unpaired input to two synthesized images with swapped AU-related features. Tang et al. [32] introduced an end-to-end expression-guided GAN in their work, enabling the manipulation of fine-grained expressions and synthesis of continuous intermediate expressions between source and target expressions. However, these methods still overlooks the role of details in local regions in preserving the local structure and texture of the image.

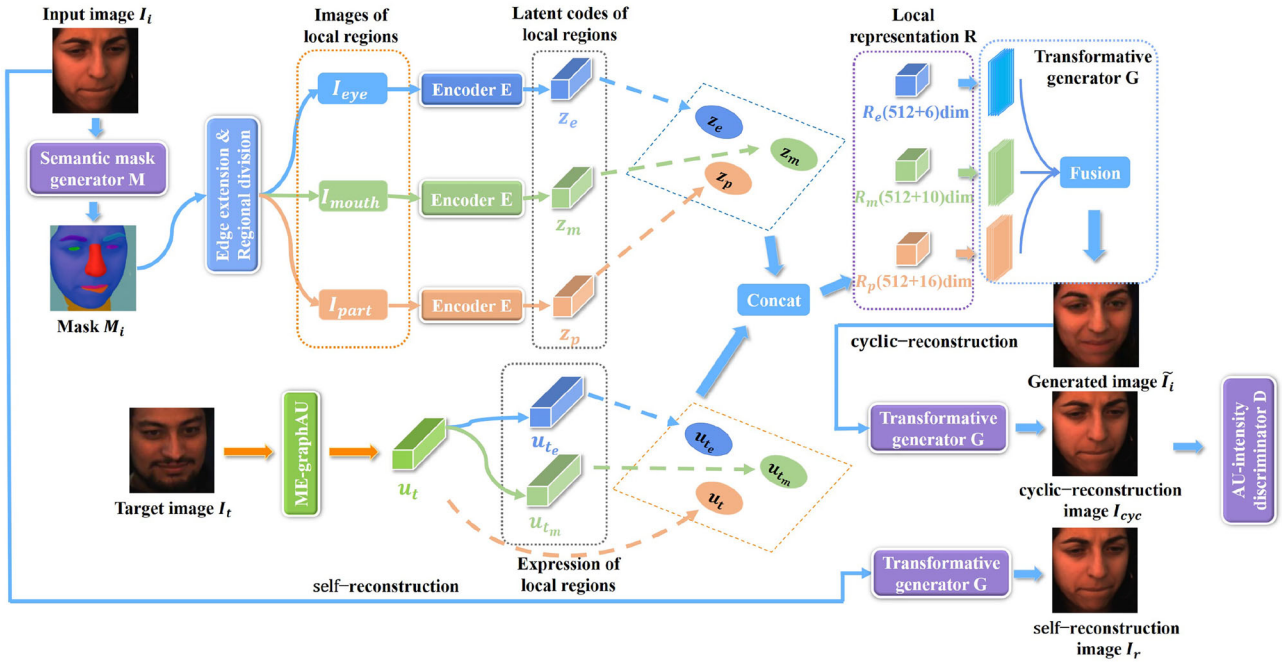

Fig. 1 The architecture of our framework, which consists of a mask generator, a AAE, a transformative generator and an AU-intensity discriminator. Given a source face image I_i and a target AU intensity vector u_t: $\{u_{t_e}, u_{t_m}\}$, we concatenate each latent codes from local region with corresponding AU vector to generate a new image $\tilde{I}_i$

Method

Our LSGAN consists of a semantic mask generator (SMG), an adversarial autoencoder (AAE), a transformative generator (TG), and an AU-intensity discriminator (AUD), as shown in Fig. 1. Our SMG receives input image I_i and produces the key facial region masks, i.e., I_{eye}, I_{mouth}, and I_{part}. Then, our AAE formulates local latent codes for the facial region masks. Afterward, our TG generates a new facial images with target expression. During this procedure, our AUD forces the generated facial expression images to lie on the same manifold as real frontal faces.

Semantic mask generator (SMG)

Through observation, it is found that facial expression manipulation often occurs in key facial regions, such as mouth and eyes. Therefore, we design SMG to locate eye and mouth regions and generate corresponding local facial part masks.

Our SMG consists of three modules, a spatial path module, a context path module, and a feature fusion module. The spatial path module encodes affluent spatial information, and the context path module provides sufficient receptive field. Our spatial path module consists of a convolutional layer and two basic blocks. Here, we introduce the Resnet [33] architecture, which includes a residual branch and a short-cut branch, to design our basic blocks. Our context path module includes two basic blocks and two attention refinement modules (ARM). Inspired by [34], we design the attention refinement module, which leverages pooling layers along the horizontal and vertical coordinate directions to capture contextual information. The components of ARM is shown in Fig. 2b. By performing pooling on the input feature map of size C×H×W in horizontal and vertical directions, we obtain the following feature maps:

$$z_c^h(h) = \frac{1}{W} \sum_{0 \le i < W} x_c(h, i),$$
$$z_c^w(w) = \frac{1}{H} \sum_{0 \le j < H} x_c(j, w). \tag{1}$$

Then, with the following formula $f = \delta(F_1(|z^h, z^w|))$, we concatenate z^h with z^w, and perform the F_1 operation on the concatenated result, which involves dimension reduction and activation using a 1×1 convolutional kernel. Along the spatial dimension, we split f into $f^h \in \mathbb{R}^{C/r \times H \times 1}$ and $f^w \in \mathbb{R}^{C/r \times 1 \times W}$. We perform dimension expansion using a 1×1 convolutional kernel, and finally apply the sigmoid activation function to obtain the final attention weights $g^h \in \mathbb{R}^{C \times H \times 1}$ and $g^w \in \mathbb{R}^{C \times 1 \times W}$ in both directions:

$$g^h = \sigma(F_h(f^h)),$$
$$g^w = \sigma(F_w(f^w)). \tag{2}$$

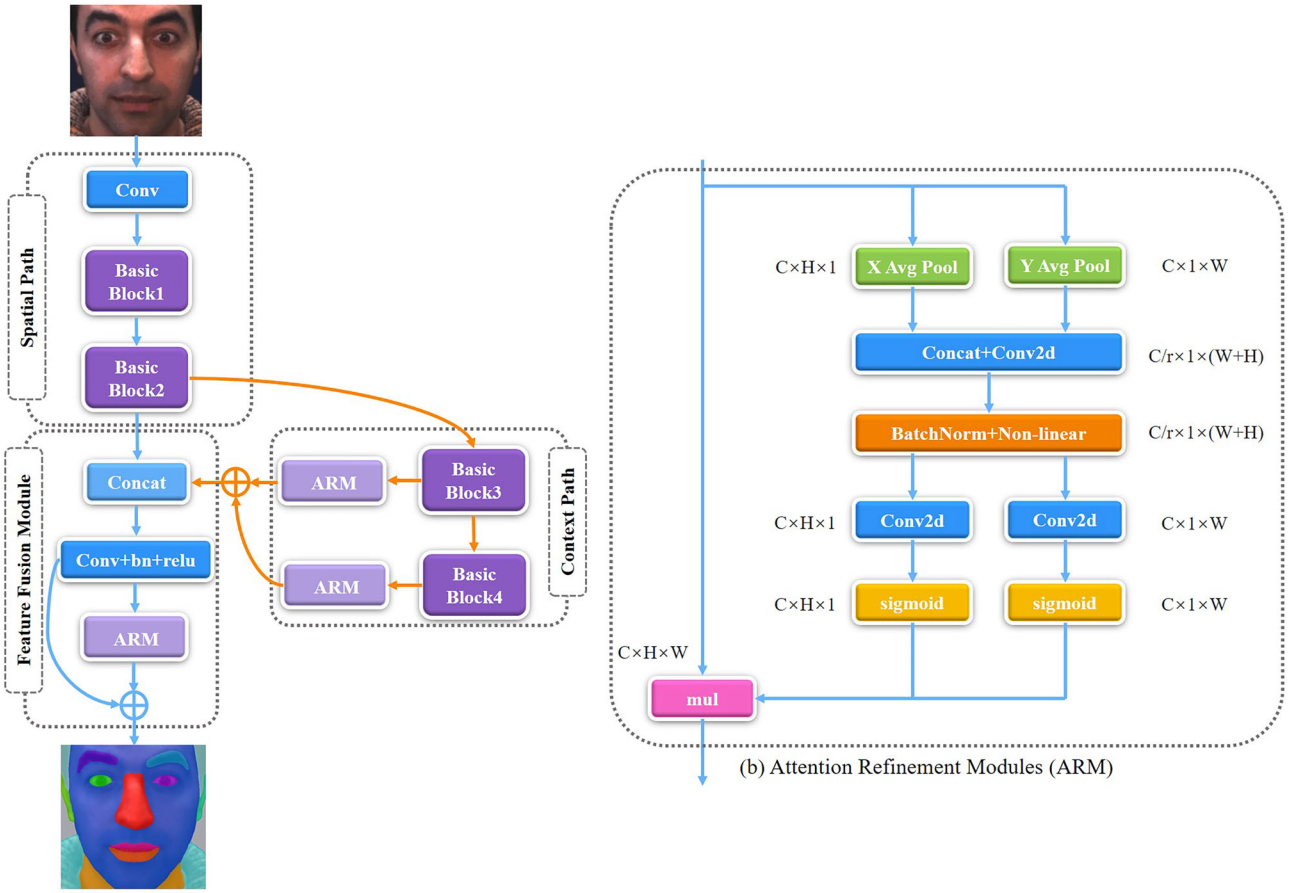

Fig. 2 An overview of SMG. **a** Network architecture. **b** Components of attention refinement modules

Finally, the output formula of ARM can be expressed as follows:

$$y_c(i,j) = x_c(i,j) \times g_c^h(i) \times g_c^w(j), \tag{3}$$

whereas, $x_c(i,j)$ and $y_c(i,j)$ correspond to the input and output features, respectively. Furthermore, our feature fusion module is composed of a convolutional layer and an attention refinement module. It fuses the output features of the former two modules and generates face segmentation results (see Fig. 2).

Since facial expression manipulation would cause large shape changes of facial component, we enlarge the size of eyes, eyebrows, and mouth areas and divide the source image I_i into eye region masks I_{eye}, mouth region masks I_{mouth} and cheek region masks I_{part}. In this way, we can provide facial semantic priors for following procedures.

Adversarial autoencoder (AAE)

After obtaining the local facial part masks, encoding their latent codes becomes crucial for our task. Therefore, we introduce an AAE [35] to encode the latent code for local facial part masks.

Our AAE is composed of an encoder E, a decoder, and a discriminator D_z, as shown in Fig. 3. The encoder generates latent codes $z = E(x) \sim q_z$ from source images x, and x can be the original image I_i or local image I_{eye}, I_{mouth}, etc. The decoder reconstructs input source images with 5 convolution layers and a fully-connected layer. The discriminator judges whether the latent code arises from the predicted code of the autoencoder or from a sampled distribution specified by the user. We employ a latent adversarial loss L_{adv}^z to learn the structured latent mapping between the latent space and Gaussian distribution:

$$\begin{aligned} L_{\text{adv}}^z = {} & \mathbb{E}_{z' \sim p_z}[\log D_z(z')] \\ & + \mathbb{E}_{z \sim p_{\text{data}}}[\log(1 - D_z(E(x)))], \end{aligned} \tag{4}$$

where z is sampled from the image domain p_{data} and z' follows the Gaussian distribution p_z. We pretrain the AAE and introduce the encoder into our LSGAN. In this manner, we can estimate latent codes for different regions, namely z_e for

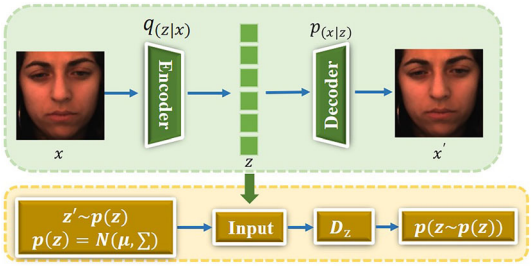

Fig. 3 An overview of AAE

the eye region, z_m for the mouth region, and z_p for sub critical regions.

After obtaining the latent code of each region, we proceed to divide the target AU-intensity vector u_t into two subsets: u_{t_e} for the eye region and u_{t_m} for the mouth region, while z_p is still cascaded with u_t. This paper uses the target AU-intensity vector as a conditional variable to construct $\widetilde{I}_i = G(z|u_t)$ to generate an image with the expected expression. Here z is the set of latent code for each region. We cascade the latent code with the corresponding AU vector to obtain representations of different regions: R_e for the eye region, R_m for the mouth region and R_p for sub critical regions.

Transformative generator (TG)

Our TG is proposed to construct facial images with target expression by explicitly exploiting facial semantic priors of source images and AU intensity of target images. Our TG comprises three distinct generation structures, as shown in the Fig. 1. The input to the generator is the local representation of different regions. Specifically, the upper layer is primarily responsible for facial expression changes in the eye region and surrounding areas, while the middle layer focuses on expression changes in the mouth area. The lower layer is utilized for expression changes in sub critical regions, such as the nose, cheeks, and chin. Each layer contains 6 upsampling residual blocks, resulting in respective outputs $\widetilde{I}_e$, $\widetilde{I}_m$, and $\widetilde{I}_p$. Subsequently, these outputs are fused based on their corresponding local masks to generate the final image $\widetilde{I}_i$.

To improve the structural stability of our TG, we reconstruct the original image I_i and generated image $\widetilde{I}_i$ based on the origin image's AU-intensity vector u_i. In this manner, we can obtain cyclic-reconstruction image I_{cyc} and self-reconstruction image I_r.

AU-intensity discriminator (AUD)

Our AUD has two tasks: (1) discriminating generated images by TG from real ones; and (2) assessing the expression intensity of the generated image relative to its target AU-intensity vector. We utilize D_{adv} to complete the first task. For the second task, we introduce D_{cls} to AUD to ensure the accurate transmission of AU changes throughout the generation process. The structure of our AUD is composed of 6 convolution layers with a stride of 2. Through the above design of our AUD, it can more effectively evaluate the authenticity and quality of generated images, while simultaneously controlling the changes in facial expressions.

Loss function

Our approach focuses on generating facial images that accurately reflect the expected facial expression while preserving the underlying identity structure of the original image. To this end, our generator loss function encompasses not only the expression vector loss, but also the loss of identity information. We pretrained two networks (G_{exp}, C_{id}) to obtain the identity and facial expression information of the current image. And the architecture of these two networks is inspired by the traditional Visual Geometry Group 19-layer (VGG19) network [36]. Furthermore, to enhance the stability of our generator network, we incorporate self-reconstruction loss and cyclic-reconstruction loss into our overall loss function. Self-reconstruction uses the origin image as input and output image. While the input of cyclic reconstruction is the generated image, and its output is the origin image. The network for image reconstruction is consistent with the original generated network which consists of three hierarchical networks. Reconstruction loss acts as a regularization term that helps prevent overfitting. By compelling the network to capture key features during the reconstruction process, the reconstruction loss helps enhance the network's ability to represent input data, thereby improving its stability. Furthermore, integrating the reconstruction loss with other losses allows for a balance between different objectives, enhancing the overall model's stability and generalization capabilities.

Adversarial loss

L_{adv} models the discriminator's ability to correctly distinguish real or false facial expression images. The adversarial loss is formulated as:

$$L_{\text{adv}} = \mathbb{E}_{I_i \sim p_{\text{data}}}[\log D(I_i)] + \mathbb{E}_{z \sim p_l}[\log(1 - D(\widetilde{I}_i))], \tag{5}$$

where I_i is the input image and corresponding latent code set is z. And our AU-intensity discriminator D is composed of D_{adv} and D_{cls}. D_{adv} discriminates generated images by TG from real ones while D_{cls} ensures the accurate transmission of AU changes.

Expression loss

Since the expression is decomposed into a set of intensity values of AUs, we need to use expression loss to align the distribution of expressions in the generated image with the target image's AUs.

$$L_{\text{au}} = -\frac{1}{d}\sum_{j=1}^{d}\sum_{q=0}^{m} ||G_{\text{exp}}(\widetilde{I}_i) - u_t||_2, \tag{6}$$

where G_{exp} was pretrained to generate each AU intensity from multiple levels. u_t is the facial expression of target image.

Identity loss

We use identity loss to preserve the identity information in the input image I_i:

$$L_{\text{id}} = -\sum_{k=1}^{n} \mathbb{1}_{k=l_i} \log\left(C_{\text{id}}^{(k)}(\widetilde{I}_i)\right), \tag{7}$$

where C_{id} is pretrained to construct a mapping between input images I_i and their identity labels l_i.

Self-reconstruction loss

To ensure the stability of image generation, generator G should be able to self-reconstruct I_i. $L1$ norm often results in structural distortion and image blurring, so we we apply an $L1$ loss and a MS-SSIM loss [37] to constrain self-reconstruction:

$$L_{\text{rec}} = ||I_r - I_i||_1 + (1 - \text{SSIM}(I_r, I_i)), \tag{8}$$

where I_r represents the image generated by self-reconstruction, which should be as similar to the input image I_i as possible.

Cyclic-reconstruction loss

To further ensure the integrity of identity information, cyclic-reconstruction losses have also been introduced into this work:

$$L_{\text{cyc}} = ||I_{\text{cyc}} - I_i||_1, \tag{9}$$

where I_{cyc} represents the image generated by cyclic reconstruction, its input is the generated image $\widetilde{I}_i$ by G.

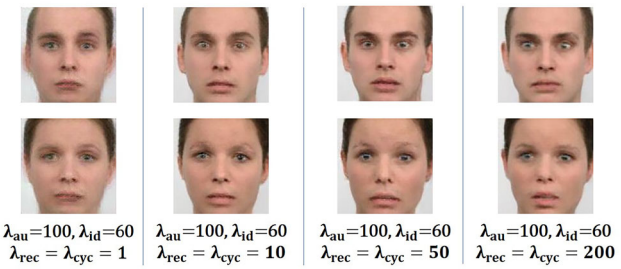

Fig. 4 Generating images under different hyper-parameters

Overall objective function

Combining the losses introduced above, the full objective function is formulated as:

$$L = L_{\text{adv}} + \lambda_{\text{au}} L_{\text{au}} + \lambda_{\text{id}} L_{\text{id}} + \lambda_{\text{rec}} L_{\text{rec}} + \lambda_{\text{cyc}} L_{\text{cyc}}, \tag{10}$$

where λ_{au}, λ_{id}, λ_{rec} and λ_{cyc} are the hyper-parameters that represent the weight of each loss function.

Experiment

Datasets and settings

Datasets

To evaluate the effectiveness and generalization of our approach, we conduct experiments on two widely used datasets, namely RaFD [38] and DISFA [39]. The RaFD dataset is composed of high-quality facial images of 67 models displaying eight distinct emotional expressions, namely anger, disgust, fear, happiness, sadness, surprise, contempt, and neutrality. Each expression is depicted in three different gaze directions across five camera angles.

The DISFA dataset, established in 2013, contains AU video samples obtained from 27 participants (15 males and 12 females) watching a 242-second video comprising nine segments meant to elicit various emotions. During the video recording, subjects' facial expressions were captured from the front with consistent environmental conditions, including lighting and background. The video resolution was set at 1024×768, with a frame rate of 20 fps. Each participant's data consisted of 4845 frames, each annotated by two FACS experts with start and end times for 12 types of AUs and corresponding intensity levels ranging from 0 to 5.

Implementation details

We implemented our network in PyTorch and the computer configuration used in the experiment is intel core I7-8700

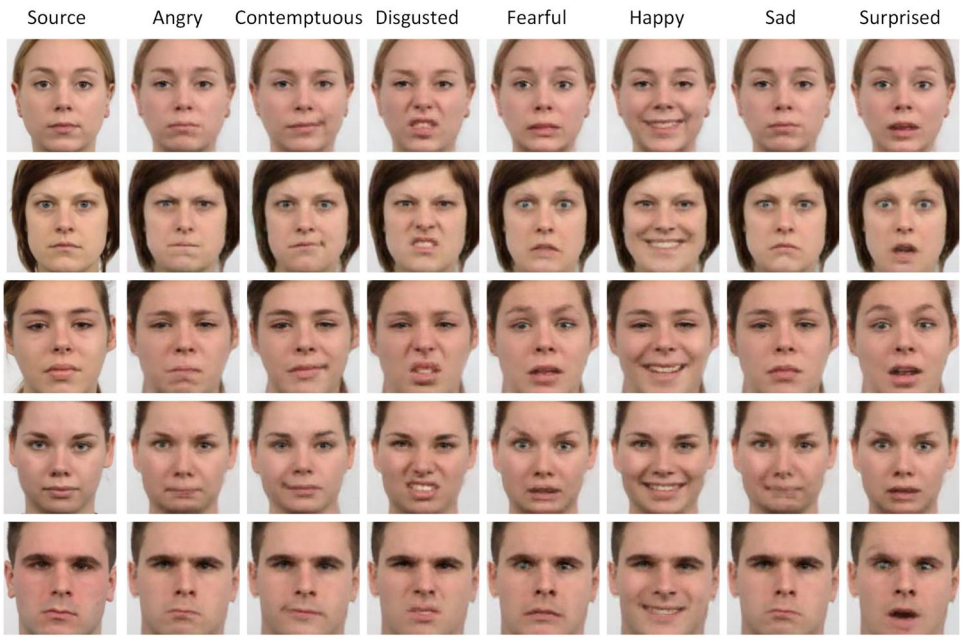

Fig. 5 Some generated images based on our LSGAN. The input source image has a neutral expression

CPU and NVIDIA 3090 GPU. In our experiments. Each image is cropped to the size of 128 × 128. Due to the lack of corresponding AU vectors in the RAFD dataset, we use ME-graphAU [8] to annotate intensities of 16 AUs (1, 2, 4, 5, 6, 7, 9, 10, 12, 14, 15, 17, 20, 23, 25 and 26) as continuous expression labels.

We set the hyper-parameters as: $\lambda_{au} = 100$, $\lambda_{id} = 60$, $\lambda_{rec} = 50$ and $\lambda_{cyc} = 50$. These parameters are subject to a simple constraint condition. Generally speaking, λ_{au} is approximately equal to the sum of λ_{rec} and λ_{cyc}. The value of λ_{id} falls within the range of [50, 100]. The generated results based on different parameters are shown in the Fig. 4, it can be seen that when λ_{rec} is set to 10, the details of the eyebrows and eyes appear incomplete. When λ_{rec} is 200, the changes in facial expressions are not sufficiently prominent.

The adversarial learning in E, G, D_z and D_{img} employs the Adam solver and a learning rate of 10^{-4}, while G_{exp} uses a learning rate of 2×10^{-4}. We train our framework for 600 and 16 epochs on RaFD and DISFA datasets, respectively, with a batch size of 8.

Evaluation metrics

To quantitatively evaluate our method for expression transfer, we introduce mean square error (MSE) and intraclass correlation coefficient (ICC) to measure the difference and correlation between the AU intensities of generated images with ground truth. To further compare the quality and structural similarity of images generated by different methods, we introduce evaluation indicators such as peak signal to noise ratio (PSNR) [40], structural similarity (SSIM) [40], Frechet inception distance (FID) [41] and LPIPS distance [42].

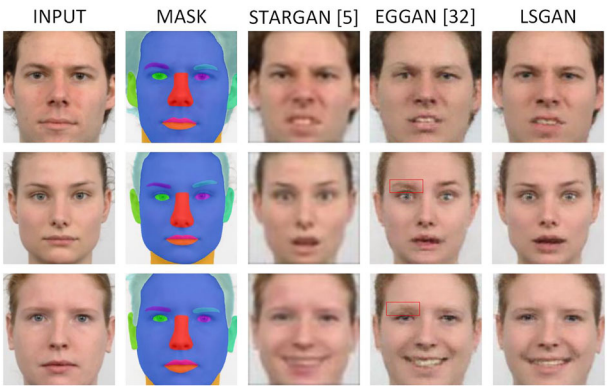

Fig. 6 Qualitative comparison of facial expression synthesis on RaFD database (target facial expression from top to bottom: disgusted, surprised and happy). The results of MASK are used in LSGAN

PSNR is a widely used metric for assessing the quality of an image. It quantifies the degree of distortion in an image by comparing the differences between the original and processed/compressed versions. The higher the PSNR value, the lower the level of distortion. The formula for calculating PSNR is expressed as follows:

$$\mathrm{PSNR} = 10 \cdot \log_{10}\left(\frac{\mathrm{MAX}^2}{\mathrm{MSE}}\right) \tag{11}$$

Here, MAX represents the maximum possible pixel value of the image (e.g., for an 8-bit image, MAX = 255), and MSE denotes the mean squared error, computed as the average of the squared differences between corresponding pixels of two images.

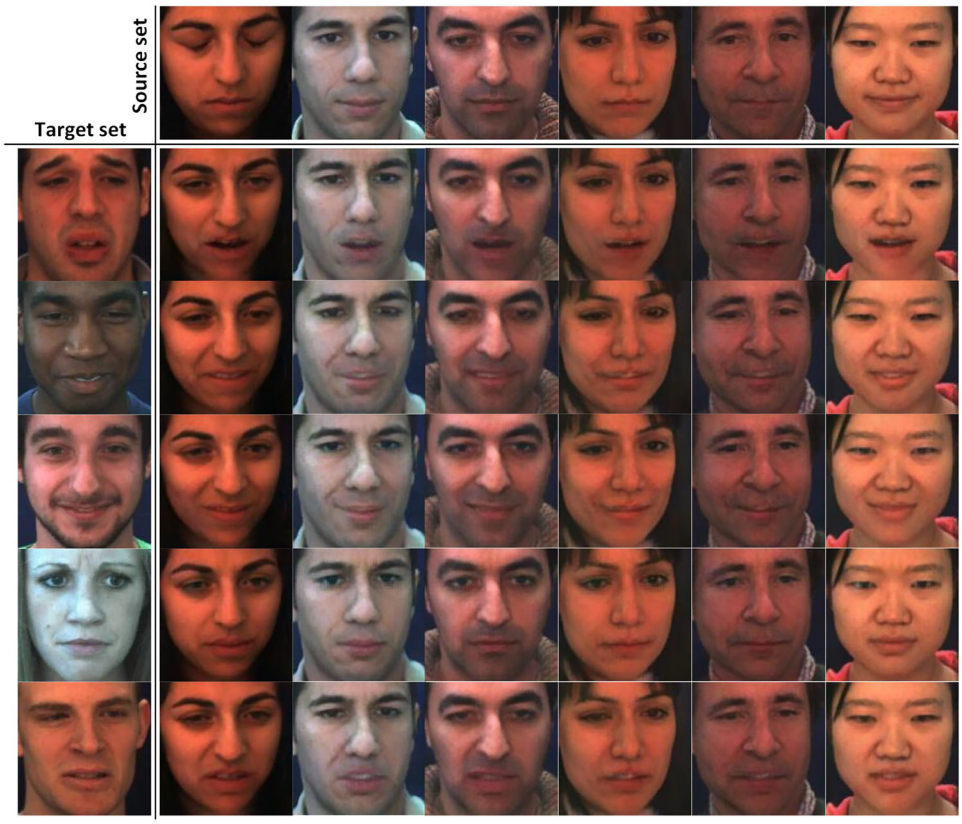

Fig. 7 Facial expression manipulation based on different target facial expression images of DISFA datasets

SSIM is another important metric for evaluating image quality. Unlike PSNR, which focuses solely on pixel-wise differences, SSIM also takes into account structural information and texture similarity. It provides a more comprehensive assessment of perceived image quality by considering both local and global image characteristics. SSIM is a metric based on three comparisons between samples x and y: luminance, contrast, and structure, expressed by the following equation:

$$\mathrm{SSIM}(x, y) = [l(x, y)^{\alpha} \cdot c(x, y)^{\beta} \cdot s(x, y)^{\gamma}], \tag{12}$$

where $l(x, y)$ represents luminance comparison, $c(x, y)$ represents contrast comparison (reflecting the magnitude of brightness changes in the image, i.e., the standard deviation of pixels), and $s(x, y)$ indicates structure comparison. The parameters α, β, and γ are constants.

FID is a measure commonly employed to assess the dissimilarity between two multivariate normal distributions. It is often used in evaluating the performance of generative models, such as GANs. The feature means μ_g and variances C_g of generated images, along with the means μ_r and variances C_r of real images, are used to compute the distance between feature vectors based on their means and variances. This distance is termed as FID, defined as:

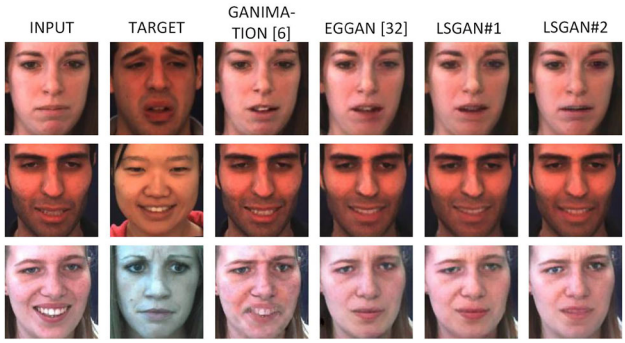

Fig. 8 Visual comparison of different expressions for EGGAN, GANIMATION, LSGAN#1 with Cross entropy loss in L_{au} and LSGAN#2 with MSE loss in L_{au} in the DISFA dataset

$$\mathrm{FID}\left(P_r, P_g\right) = ||\mu_r - \mu_g|| + T_r\left(C_r + C_g - 2\left(C_r C_g\right)^{1/2}\right). \tag{13}$$

Here, T_r is the trace operation (the sum of the elements on the main diagonal of a square matrix).

In contrast, LPIPS distance is a perceptual similarity metric and has been demonstrated to correlate well with human perceptual similarity. LPIPS distance has been widely utilized in various computer vision tasks, including image synthesis and style transfer. Given a reference block x from

Table 1 Quantitative evaluation of expression manipulation for STARGAN [5], GANIMATION [6], EGGAN [32], FADM [43] and our LSGAN

AU	MSE (lower is better)					ICC (higher is better)				
	STAR GAN [5]	GAN IMA TION [6]	EG GAN [32]	FADM [43]	LS GAN	STAR GAN [5]	GAN IMA TION [6]	EG GAN [32]	FADM [43]	LS GAN
1	0.037	0.041	0.050	0.033	**0.032**	0.790	0.725	0.658	**0.817**	0.799
2	0.026	0.043	0.056	**0.013**	0.022	0.771	0.563	0.427	**0.895**	0.793
4	**0.073**	0.090	0.100	0.181	0.086	**0.514**	0.492	0.283	0.014	0.430
5	0.109	0.033	0.034	**0.020**	0.023	0.538	0.825	0.785	**0.903**	0.885
6	0.007	0.004	0.025	0.031	**0.004**	0.942	0.976	0.772	0.701	**0.971**
7	**0.012**	0.019	0.018	0.114	0.024	**0.948**	0.932	0.925	0.485	0.898
9	0.026	**0.008**	0.021	0.070	0.010	0.663	**0.922**	0.740	0.036	0.893
10	**0.013**	0.024	0.029	0.067	0.021	**0.936**	0.894	0.857	0.647	0.904
12	0.008	0.001	0.007	0.007	**0.001**	0.956	**0.994**	0.961	0.962	0.994
14	0.004	0.007	0.007	0.006	**0.003**	0.435	0.053	0.100	0.001	**0.591**
15	0.004	0.001	0.001	0.002	**0.001**	0.381	0.522	0.667	0.249	**0.735**
17	0.032	0.045	0.034	0.038	**0.013**	0.360	0.384	0.462	0.380	**0.825**
20	0.001	0.002	0.002	**0.001**	0.003	0.501	0.337	0.231	**0.589**	0.086
23	0.003	0.004	**0.003**	0.005	0.004	**0.564**	0.320	0.511	0.357	0.414
25	0.014	0.049	0.026	0.040	**0.011**	0.954	0.854	0.922	0.871	**0.968**
26	0.047	0.040	0.054	0.044	**0.034**	0.630	0.667	0.556	0.667	**0.720**
Avg	0.026	0.026	0.029	0.042	**0.018**	0.680	0.654	0.616	0.536	**0.744**

Bold indicates the best results among various methods under the same evaluation metric in each AU
We compare MSE (lower is better) and ICC (higher is better) between 16 AU intensities of target images and generated images

Table 2 Quantitative comparison with FID (lower is better), PSNR (higher is better) and SSIM (higher is better) on the generated images of different methods

Method	STARGAN [5]	GANIMATION [6]	EGGAN [32]	LSGAN
FID	130.557	37.702	48.598	31.121
PSNR	20.587	21.352	20.086	23.417
SSIM	0.832	0.820	0.834	0.858

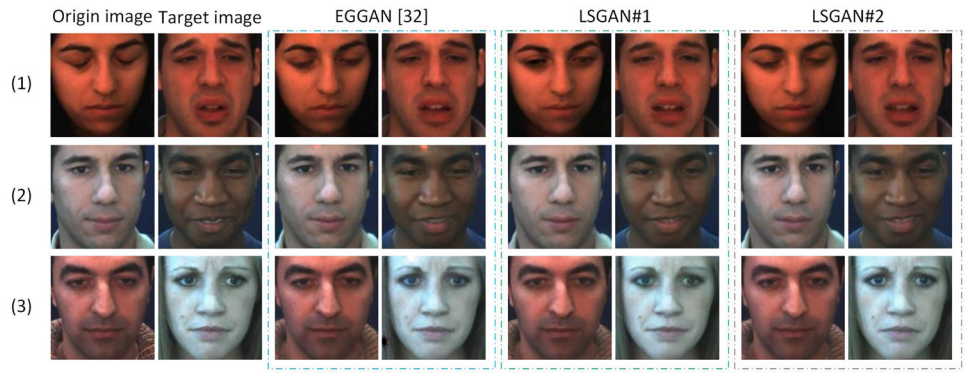

Fig. 9 Origin images and target expressions for quantitative evaluation. The original and target images are used to generate their corresponding reconstructed images

the ground truth image and a distorted block x_0 from a noisy image, the formula for calculating LPIPS is as follows:

$$d(x, x_0) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \left|\left| w_l \odot (\hat{y}^l_{hw} - \hat{y_0}^l_{hw}) \right|\right|_2^2. \quad (14)$$

Here, y^l and ${y_0}^l$ represent the feature maps of the l-th layer of the images.

Qualitative evaluations

We divide the images in the RAFD dataset into 8 discrete emotional expression categories: neutral, angry, contemptuous, disgusted, fearful, happy, sad, and surprised. Figure 5 shows the generation results of our LSGAN.

The results indicate that our method can generate various types of images with target expressions. We do further evaluation to compare the performance of our LSGAN with the current state-of-the-art approaches, as show in Fig. 6.

Among them, STARGAN [5] handles well for image translation in different domains, but its results are a little blurry with some artifacts. EGGAN [32] can achieve higher quality results but still ignores the structural integrity of local regions, such as the right eyebrow as shown in Fig. 6. Using the generator targeting local semantic regions and reconstruction networks that maintain facial structure, it can be seen that both global expressions and local muscle actions look natural in the generated images of our LSGAN. This proves that our method has more complete details of local regions while achieving expression transformation.

To conduct a more detailed analysis of AU-intensity vector, we do further experiments on DISFA dataset. For different target images and their expressions, the original image achieved expression manipulation while retaining its own identity features, as shown in the Fig. 7.

The comparison with other methods are shown in Fig. 8. Based on a large amount of training data, the images generated by various methods have good results. However, we find that GANIMATION [6] generates blurring and overlap around local areas when there is significant facial deformation, such as from opening the mouth to closing the mouth. By introducing reconstruction networks, our method has slight advantages in terms of structural integrity and local muscle actions, such as texture details in the eye region. This demonstrates the advantage of our method in controlling key facial region details.

Quantitative evaluations

To quantitatively evaluate the expression manipulation, we compute MSE and ICC between the AU intensities of generated images with ground truth. We use ME-graphAU to estimate the AU-intensity vector for each image. Since the comparison method only includes partial AU-intensity vector results, we choose 16 AU intensities near the eyes and mouth. As shown in Table 1, although the latest diffusion method [43] has shown good results in generating high-quality images, it is difficult to capture the detailed changes in facial expressions. It can be seen that our method is the most accurate in predicting the AU-intensity vector near the mouth area. Overall, our method achieves higher average ICC and lower average MSE of 16 AU intensities. This proves the effectiveness of our network based on different regions.

While considering facial expression transfer, we further conduct a comparative analysis of image quality among various methods. We use FID, PSNR and SSIM to analyze the

Table 3 Quantitative comparison with FID (lower is better) and LPIPS (lower is better) on reconstructed images

Method	FID (lower is better)			LPIPS (lower is better)		
	EGGAN [32]	LSGAN#1	LSGAN#2	EGGAN [32]	LSGAN#1	LSGAN#2
FIG (1)	33.5	34.9	29.2	0.0272	0.0341	0.0238
FIG (2)	103.2	55.2	52.9	0.0626	0.0530	0.0437
FIG (3)	50.7	55.6	39.3	0.0504	0.0503	0.0388
FIG (Avg)	35.7	27.9	24.3	0.0347	0.0330	0.0268

FIG (1), FIG (2), FIG (3) corresponds to the three pairs of images in Fig. 9

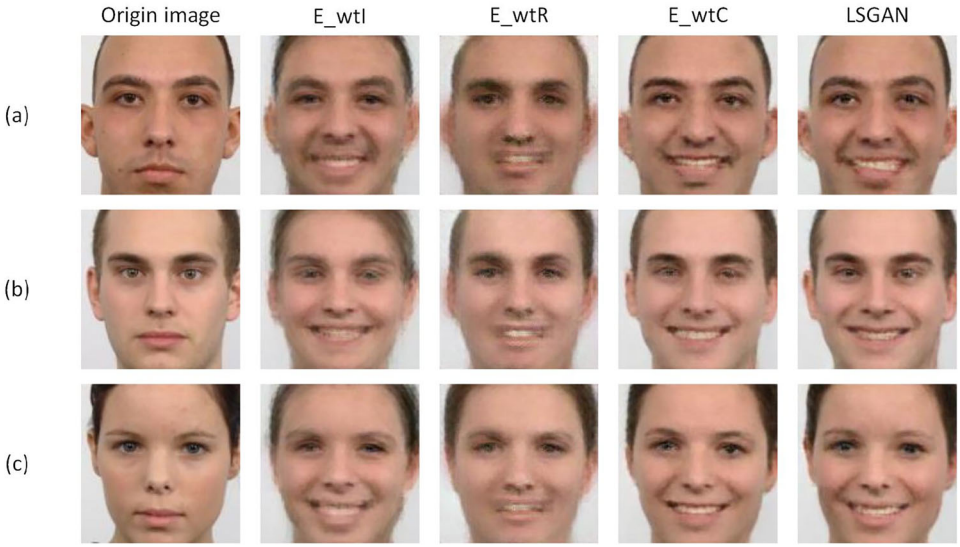

Fig. 10 Illustration of the effectiveness of different loss terms. LSGAN is trained without L_{id}, L_{rec}, L_{cyc}, respectively

quality of generated images, as shown in Table 2. It can be seen that compared to other methods, the images generated by LSGAN are closer to the image quality and structural similarity of the ground truth.

To further compare the stability and image quality of different networks, we use FID metrics and LPIPS distance to evaluate the reconstructed image on the DISFA datasets. We select 3 pairs of images for evaluation, as shown in Fig. 9.

The quantitative evaluation results of the reconstructed images in Fig. 9 are shown in Table 3 below.

Among them, FIG(1), FIG (2) and FIG (3) correspond to the three pairs of images in Fig. 9, respectively. And FIG (Avg) is the average results of 50 pairs of test images. It indicate that our method has better reconstruction performance, that is, our LSGAN can achieve more stable image generation while preserving the original image identity and structure. In addition, the results of LPIPS reflect that our network is able to maintain the features of invariant local regions well.

The ablation study

In this section, we evaluate the main components in our LSGAN. Specifically, we investigate the effects of different loss terms in our framework by examining their impact on image generation. To accomplish this, we train the network by removing one of three key loss terms: identity loss, self-reconstruction loss, and cyclic reconstruction loss, which are denoted as E_wtI, E_wtR, and E_wtC, respectively. Figure 10 shows some of the results with target facial expression of happiness.

Table 4 Quantitative comparison with FID (lower is better)

Source	E_wtI	E_wtR	E_wtC	LSGAN
FID (a)	59.9	179.9	44.5	41.4
FID (b)	77.1	89.3	40.8	23.4
FID (c)	65.0	69.9	44.5	31.8

$E\_wtI\ w/oL_{\text{id}}$ $E\_wtR\ w/oL_{\text{rec}}$ $E\_wtC\ w/oL_{\text{cyc}}$
FID (a), FID (b), FID (c) correspond to three sets of images in Fig. 10

We further evaluate the generated image and ground truth with FID metrics, and the results are shown in Table 4. In summary, the various losses used in our network are necessary.

To further analyze the effects of different network architectural modules on our LSGAN, we conduct the ablation study on LSGAN and its variants. The generated images with disgust expressions of our LSGAN and its variants are shown in Fig. 11. E_wtARM denotes the adoption of the conventional channel attention network SENET instead of ARM, whereas E_wtSMG signifies the utilization of global

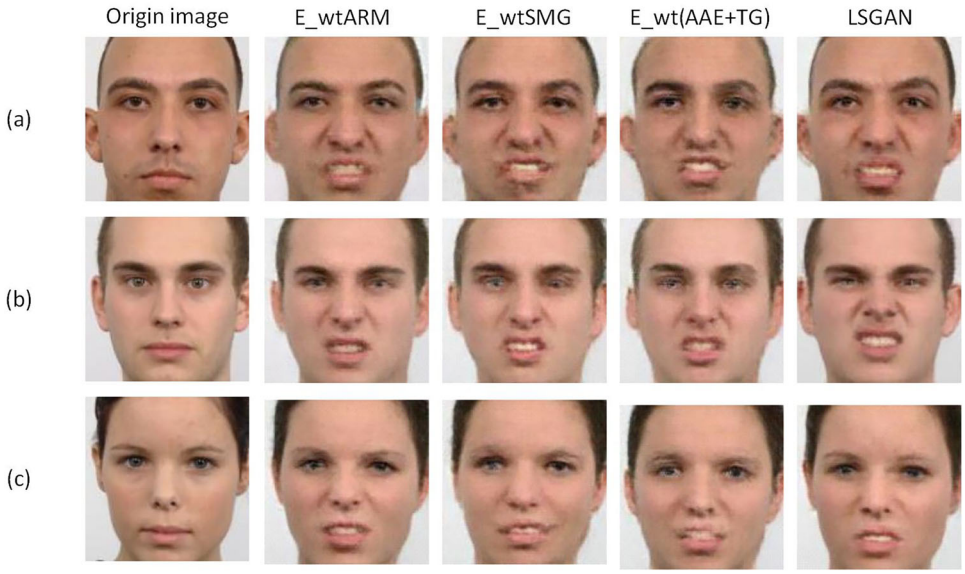

Fig. 11 Illustration of the effectiveness of different network structure module. We compare LSGAN with its three different network structure variants (E_wtARM, E_wtSMG, E_wt(AAE + TG)

Table 5 Quantitative comparison with FID (lower is better)

Source	E_wtARM	E_wtSMG	E_wt(AAE + TG)	LSGAN
FID (a)	47.8	54.8	68.2	41.1
FID (b)	41.7	71.9	62.6	27.5
FID (c)	63.2	71.3	68.8	52.4

FID (a), FID (b), FID (c) correspond to three sets of images in Fig. 11

Table 6 Quantitative evaluation of expression manipulation for our LSGAN and its variants on RAFD dataset

AU	MSE (lower is better)						
	E_wtI	E_wtR	E_wtC	LSGAN	E_wtARM	E_wtSMG	E_wt(AAE + TG)
1	0.062	0.178	0.032	**0.032**	0.046	0.048	0.122
2	0.058	0.106	**0.020**	0.022	0.046	0.056	0.085
4	0.104	0.141	0.100	**0.086**	0.141	0.100	0.192
5	0.023	0.316	0.025	**0.023**	0.051	0.034	0.178
6	0.006	0.012	0.005	**0.004**	0.006	0.025	0.021
7	0.036	0.038	**0.018**	0.024	0.028	0.018	0.062
9	0.024	0.058	0.020	**0.010**	0.015	0.021	0.044
10	0.023	0.035	0.018	0.021	**0.014**	0.029	0.054
12	0.002	0.017	0.002	**0.001**	0.001	0.007	0.008
14	0.005	0.006	0.004	**0.003**	0.006	0.007	0.009
15	**0.001**	0.008	0.001	0.001	0.001	0.001	0.009
17	0.015	0.083	0.033	**0.013**	0.017	0.033	0.079
20	**0.002**	0.004	0.003	0.003	0.003	0.002	0.004
23	0.004	0.011	**0.003**	0.004	0.003	0.003	0.006
25	0.047	0.014	0.016	**0.011**	0.020	0.026	0.023
26	0.058	0.099	**0.027**	0.034	0.038	0.051	0.095
Avg	0.029	0.070	0.020	**0.018**	0.027	0.029	0.062

Bold indicates the best results among various methods under the same evaluation metric in each AU
We compare MSE (lower is better) between 16 AU intensities of target images and generated images

Table 7 Quantitative evaluation of expression manipulation for our LSGAN and its variants on RAFD dataset

AU	ICC (higher is better)						
	E_wtI	E_wtR	E_wtC	LSGAN	E_wtARM	E_wtSL	E_wt(AAE + TG)
1	0.570	0.052	**0.802**	0.799	0.702	0.667	0.292
2	0.364	0.025	**0.821**	0.793	0.515	0.427	0.154
4	0.367	0.058	0.334	**0.430**	0.162	0.284	0.045
5	0.872	0.013	0.867	**0.885**	0.742	0.793	0.281
6	0.959	0.899	0.959	**0.971**	0.957	0.772	0.815
7	0.856	0.829	**0.929**	0.898	0.879	0.925	0.738
9	0.706	0.172	0.749	**0.893**	0.822	0.742	0.375
10	0.893	0.813	0.921	0.904	**0.935**	0.857	0.711
12	0.990	0.900	0.991	0.994	**0.995**	0.961	0.951
14	0.364	0.106	0.417	**0.591**	0.105	0.100	0.201
15	0.544	0.482	0.613	**0.735**	0.715	0.687	0.444
17	0.806	0.180	0.522	**0.825**	0.734	0.470	0.072
20	**0.361**	0.148	0.017	0.086	0.256	0.232	0.102
23	0.293	0.136	**0.627**	0.414	0.578	0.540	0.048
25	0.861	0.958	0.954	**0.968**	0.941	0.922	0.923
26	0.495	0.208	**0.784**	0.720	0.693	0.567	0.254
Avg	0.644	0.374	0.707	**0.744**	0.671	0.622	0.400

Bold indicates the best results among various methods under the same evaluation metric in each AU
We compare ICC (higher is better) between 16 AU intensities of target images and generated images

images alone for sentiment operations without incorporating local feature alignment. E_wt(AAE + TG) uses traditional encoding and decoding structures in LSGAN instead of a combination of AAE and TG. It can be seen that the variants of LSGAN have a decrease in the quality of generated images in key facial regions.

Similarly, we evaluate the generated image and ground truth with FID metrics, as shown in Table 5. It is evident that compared to its variants, LSGAN can alleviate issues such as low resolution and image artifacts.

We also quantitatively evaluate different variants of LSGAN in terms of transferring fine-grained expressions. Tables 6 and 7 present the MSE and ICC between the AU intensity of the ground truth and the generated image of LSGAN and its variants, respectively. The absence of reconstruction loss significantly diminishes the network's capacity to convey expression intensity labels. Moreover, when compared to conventional encoding and decoding architectures, the utilization of SMG, AAE and TG enables LSGAN to effectively align the features extracted from the input image with those of the target image. This alignment facilitates the precise transfer of nuanced facial expressions, enhancing our network's ability to faithfully capture fine-grained expression details. In summary, LSGAN showcases superior performance in fine-grained expression manipulation, as evidenced by its highest average ICC and lowest average MSE among its variants. This demonstrates the effectiveness of LSGAN in accurately manipulating facial expressions with fine details.

Conclusion

This paper introduces a novel approach for fine-grained facial expression manipulation, termed LSGAN. By integrating facial expression images generated from various facial regions, our approach is able to fully capture and leverage region-specific information while preserving the overall structural integrity of the image. At the same time, our method ensures that the AU intensity of the generated image is basically consistent with the target image. Our proposed method has been rigorously evaluated through both qualitative and quantitative analyses using publicly available databases, demonstrating its high performance in generating expression-specific facial images.

Acknowledgements This work was supported in part by National Natural Science Fund of China [Grant No. 61473155].

Data availability The relevant data and material in this article are available from the corresponding author.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Song X, Wu N, Song S, Stojanovic V (2023) Switching-like event-triggered state estimation for reaction-diffusion neural networks against dos attacks. Neural Process Lett 55:1–22
2. Song X, Wu N, Song S, Zhang Y, Stojanovic V (2023) Bipartite synchronization for cooperative-competitive neural networks with reaction-diffusion terms via dual event-triggered mechanism. Neurocomputing 550:126498
3. Peng Z, Song X, Song S, Stojanovic V (2023) Hysteresis quantified control for switched reaction-diffusion systems and its application. Complex Intell Syst 9(6):7451–7460
4. Zhang Z, Song X, Sun X, Stojanovic V (2023) Hybrid-driven-based fuzzy secure filtering for nonlinear parabolic partial differential equation systems with cyber attacks. Int J Adapt Control Signal Process 37(2):380–398
5. Choi Y, Choi M, Kim M, Ha J-W, Kim S, Choo J (2018) Stargan: unified generative adversarial networks for multi-domain image-to-image translation. In Proceedings of the IEEE conference on computer vision and pattern recognition. pp 8789–8797
6. Pumarola A, Agudo A, Martinez AM, Sanfeliu A, Moreno-Noguer F (2018) Ganimation: anatomically-aware facial animation from a single image. In Proceedings of the European conference on computer vision (ECCV). pp 818–833
7. Wegrzyn M, Vogt M, Kireclioglu B, Schneider J, Kissler J (2017) Mapping the emotional face. how individual face parts contribute to successful emotion recognition. PLoS ONE, 12(5):e0177239
8. Luo C, Song S, Xie W, Shen L, Gunes H (2022) Learning multi-dimensional edge feature-based AU relation graph for facial action unit recognition. In: proceedings of the thirty-first international joint conference on artificial intelligence. International joint conferences on artificial intelligence organization
9. Chauhan S, Singh M, Aggarwal AK (2021) Experimental analysis of effect of tuning parameters on the performance of diversity-driven multi-parent evolutionary algorithm. In: 2021 IEEE 2Nd international conference on electrical power and energy systems (ICEPES), pp 1–6
10. Chauhan S, Singh M, Aggarwal AK (2023) Designing of optimal digital IIR filter in the multi-objective framework using an evolutionary algorithm. Eng Appl Artif Intell 119:105803
11. Yang S, Luo P, Loy CC, Tang X (2016) From facial parts responses to face detection: A deep learning approach. In: IEEE International Conference on Computer Vision
12. Zhang K, Zhang Z, Li Z, Qiao Y (2016) Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process Lett 23(10):1499–1503
13. Aggarwal AK (2022) Biological tomato leaf disease classification using deep learning framework. Int J Biol Biomed Eng 16(1):241–244
14. Aggarwal AK, Jaidka P (2022) Segmentation of crop images for crop yield prediction. Int J Biol Biomed 7
15. Zhao H, Shi J, Qi X, Wang X, Jia J (2016) Pyramid scene parsing network. In: IEEE Computer Society
16. Lin G, Milan A, Shen C, Reid I (2017) Refinenet: multi-path refinement networks for high-resolution semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1925–1934
17. Yu C, Wang J, Peng C, Gao C, Yu G, Sang N (2018) Bisenet: bilateral segmentation network for real-time semantic segmentation. Springer, Cham
18. Yu C, Gao C, Wang J, Yu G, Shen C, Sang N (2020) Bisenet v2: bilateral network with guided aggregation for real-time semantic segmentation. Int J Comput Vis 129(2021):3051–3068
19. He Z, Zuo W, Kan M, Shan S, Chen X (2019) Attgan: facial attribute editing by only changing what you want. IEEE Trans Image Process 28(11):5464–5478
20. Ding H, Sricharan K, Chellappa R (2018) Exprgan: facial expression editing with controllable expression intensity. In: Proceedings of the AAAI conference on artificial intelligence, vol 32
21. Geng Z, Cao C, Tulyakov S (2020) Towards photo-realistic facial expression manipulation. Int J Comput Vis 128:2744–2761
22. Xia Y, Zheng W, Wang Y, Hui Y, Dong J, Wang F-Y (2021) Local and global perception generative adversarial network for facial expression synthesis. IEEE Trans Circ Syst Video Technol 32(3):1443–1452
23. Cohn JF, Ekman P (2005) Measuring facial action. New Handbook Methods Nonverbal Behav Res 525:1
24. Baltrusaitis T, Zadeh A, Lim YC, Morency L-P (2018) Openface 2.0: Facial behavior analysis toolkit. In: 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018), pp 59–66,
25. Wang J, Zhang J, Lu Z, Shan S (2019) Dft-net: disentanglement of face deformation and texture synthesis for expression editing. In: 2019 IEEE International Conference on Image Processing (ICIP), pp 3881–3885
26. Wu R, Zhang G, Lu S, Chen T (2020) Cascade ef-gan: Progressive facial expression editing with local focuses. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 5021–5030
27. Song L, Lu Z, He R, Sun Z, Tan T (2018) Geometry guided adversarial facial expression synthesis. In: Proceedings of the 26th ACM international conference on Multimedia. pp 627–63,
28. Qiao F, Yao N, Jiao Z, Li Z, Chen H, Wang H (2018) Emotional facial expression transfer from a single image via generative adversarial nets. Comput Anim Vir Worlds 29(3–4):e1819
29. Ling J, Xue H, Song L, Yang S, Xie R, Gu X (2020). Toward fine-grained facial expression manipulation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16, Springer, pp 37–53
30. Wang F, Xiang S, Liu T, Fu Y (2021) Attention based facial expression manipulation. In: 2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), pp 1–6
31. Shao Z, Zhu H, Tang J, Xuequan L, Ma L (2021) Explicit facial expression transfer via fine-grained representations. IEEE Trans Image Process 30:4610–4621
32. Tang J, Shao Z, Ma L (2021) Eggan: Learning latent space for fine-grained expression manipulation. IEEE Multimed 28(3):42–51
33. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 770–778
34. Hou Q, Zhou D, Feng J (2021) Coordinate attention for efficient mobile network design. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp 13713–1372
35. Makhzani A, Shlens J, Jaitly N, Goodfellow I, Frey B (2015) Adversarial autoencoders. arXiv preprint arXiv:1511.05644

36. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556
37. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 13(4):600–612
38. Langner O, Dotsch R, Bijlstra G, Wigboldus Daniel HJ, Hawk Skyler T, Knippenberg AD Van (2010) Presentation and validation of the radboud faces database. Cogn Emotion 24(8):1377–1388
39. Mavadati SM, Mahoor MH, Bartlett K, Trinh P, Cohn JF (2013) Disfa: a spontaneous facial action intensity database. IEEE Trans Affect Comput 4(2):151–160
40. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 13(4):600–612
41. Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S (2017) Gans trained by a two time-scale update rule converge to a local nash equilibrium. Adv Neural Inf Process Syst 30
42. Zhang R, Isola P, Efros AA, Shechtman E, Wang O (2018) The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 586–595
43. Zeng B, Liu X, Gao S, Liu B, Li H, Liu J, Zhang B (2023) Face animation with an attribute-guided diffusion model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 628–637