

Facial Expression Generation Based on Multi-scale Mixed Attention

Renjie Liu
 Southwest University
 College of Electronic and Information Engineering
 Chongqing, China
 493578750@qq.com

Guangyuan Liu*
 Southwest University
 College of Electronic and Information Engineering
 Chongqing, China
 *Corresponding author: liugy@swu.edu.cn

Binghui Hu
 Southwest University
 College of Electronic and Information Engineering
 Chongqing, China
 447378468@qq.com

Yinghao Qiao
 Southwest University
 College of Electronic and Information Engineering
 Chongqing, China
 669592192@qq.com

Abstract—Facial expression generation involves creating facial images with specific expressions using computational methods and finds extensive applications in face editing, film production, and data augmentation. The advent of Generative Adversarial Networks (GANs) has led to significant advancements in facial expression generation. However, images generated by these methods often suffer from issues such as overlap and blurriness, resulting in a lack of realism. To address these challenges, this paper introduces a Multi-scale Mixed Attention Generative Adversarial Network (MMA-GAN) aimed at producing high-quality facial expression images. The proposed MMA-GAN incorporates global residual connections at the beginning and end of the generator to preserve skin color and ignore irrelevant background content. Additionally, a multi-scale mixed attention module is integrated within the generator to adaptively learn features of key regions, thereby enhancing the learning of critical areas in the images. Experiments conducted on the publicly available AffectNet dataset validate the effectiveness of the MMA-GAN model. Results indicate that MMA-GAN outperforms related methods in both qualitative assessments and quantitative analysis metrics.

Keywords-generative adversarial network(GAN);expression generation; attention mechanism;

I. INTRODUCTION

With the rapid development of internet technology, enhancing human-computer interaction experiences has become a significant hallmark of technological progress. In this process, facial expression generation technology [1], as a cutting-edge technology, has become increasingly important. Facial expressions are not only an intuitive medium for human emotional expression but also an indispensable part of human-computer interaction. Through highly realistic simulations of facial expressions, this technology can greatly enhance the expressiveness of virtual characters or robots, enabling them to convey emotions and intentions more naturally and accurately, thereby improving the user's interaction experience. Whether in virtual reality, augmented reality, or in fields such as video games and social media, the application of facial expression

generation technology has opened up new possibilities, offering more rich and personalized ways of interaction.

Current facial synthesis technologies based on deep learning primarily focus on general facial generation tasks and often overlook the subtle variations in local features within facial expressions, which are particularly crucial in facial expression generation tasks. In fact, humans naturally focus their attention on areas of the face that change more significantly when recognizing and distinguishing different expressions. For example, changes in the mouth are most prominent when expressing happiness. However, traditional deep learning models have certain limitations in handling such details, often neglecting the specific changes in these key areas. This oversight can not only cause some areas of the image to become blurry but also reduce the overall quality of the generated image, making the generated facial expression images lack realism. Therefore, the task of facial expression generation requires a more refined and specialized approach to capture and reproduce these key changes in facial expressions.

II. DATA

A. Data Sources

The images in AffectNet [2] were obtained through search engines using relevant keywords that correspond to the eight basic emotions (anger, disgust, fear, happiness, sadness, surprise, neutral, and contempt) as well as other complex emotional expressions derived from psychological research. During the collection process, the research team implemented a series of steps to ensure the diversity and quality of the data, such as ensuring the diversity of image sources and the richness of emotional expressions.

B. Data preprocessing

The dataset employed was meticulously divided into two main parts: the training set and the test set. This division was completed during the data preprocessing stage, ensuring the independence of the training and evaluation processes.

Furthermore, to accurately capture facial expressions, we utilized advanced automatic face detection algorithms to identify and crop the facial regions in each image. To ensure the consistency of the input data, all facial expression images were uniformly resized to a resolution of 128×128 pixels and set to a 3-channel (RGB color mode). The primary purpose of the training set is for the training and validation processes of the network model, optimizing the model parameters and structure; meanwhile, the test set is used to evaluate the performance of the trained network model, which helps in verifying the model's generalization ability to unseen data.

III. METHOD

In this study, we propose a novel Generative Adversarial Network architecture named Multi-Scale Mixed Attention Generative Adversarial Network (MMA-GAN), which is capable of generating images with both continuous and discrete attribute features. Specifically, the input image x is combined with continuous attribute values ρ and discrete attribute labels c to generate images x' with subtle attribute variations. ρ is a real value in the interval $[0,1]$, representing the change in attribute strength, while c indicates the attribute labels that need to remain unchanged. The attribute vector is encoded in the form of $z=(c,\rho)$, where ρ serves as an adjustable continuous variable, which, when combined with c , forms the attribute encoding z . This combination is achieved through element-wise multiplication \odot . To achieve the above objectives, this paper utilizes the concept of CycleGAN [3]. The overall framework of MMA-GAN is shown in Fig.1.

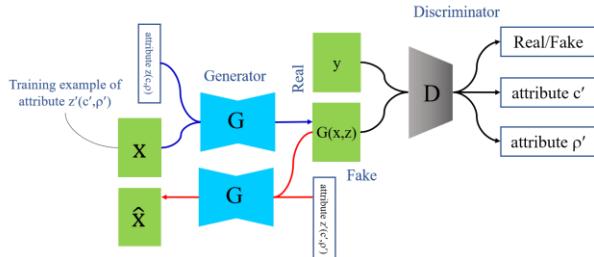


Figure 1 MMA-GAN Overall Framework

A. Multi-scale Mixed Attention Generator

To improve issues such as blurriness and lack of authenticity in generated facial expression images, this paper proposes a generator with multi-scale mixed attention residuals. This generative structure consists of a set of convolutions for down sampling, multi-scale mixed attention blocks, global residual connections, and deconvolution layers for up sampling, with the specific structure as shown in Fig.2. The input channel number of the generator is $N+3$, which is defined by the channel number of the input image and the dimensionality of labels such as the number of facial expression categories.

To make the model focus only on the parts that need to be changed according to the text, without generating parts unrelated to the expression, this paper adds a Global Residual Connection (GRC) Module at the beginning and end of the model. Since the input and output images share the same theme, the global residual connection retains the precise details of the

input image. This eliminates the need for the generator to generate the input image with the desired expression by forcing only the details related to the expression to be retained. Residual connections are typically applied at the feature level, but here, the paper uses global residual connections at the RGB image level.

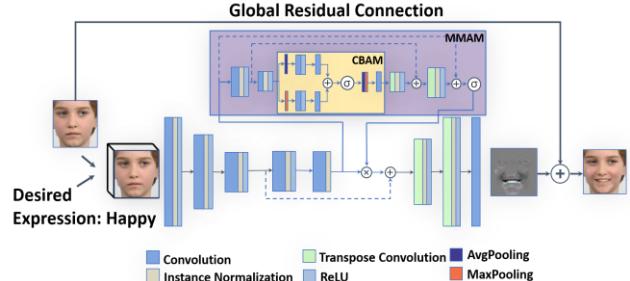


Figure 2 Generator Structure Diagram

To enhance the generator G's focus on key areas during iterations, this paper introduces a Multi-Scale Mixed Attention Module (MMARM). The core of MMARM includes an encoder-decoder network, mixed attention module, and residual connections. The encoder extracts features at various scales, and the decoder uses these features to recreate the feature map, aiming for an output matching the original's dimensions. However, down sampling in the encoding phase can lose details, making it challenging for the decoder to fully reconstruct the original feature map. Introducing residual connections between corresponding encoding and decoding layers helps recover significant facial feature details. Additionally, a mixed attention module, consisting of channel and spatial attention mechanisms, is added. This module enhances focus on important features and locations, further improving the model's ability to concentrate on critical areas.

B. Discriminator

In this section, we explore a design for a discriminator inspired by the concept of PatchGAN [4]. Traditional Generative Adversarial Network (GAN) [5] discriminators typically map an input image to a single real number, representing the probability of the image being a real sample. In contrast, the approach of PatchGAN maps the input to a matrix, where each element of the matrix represents the probability of the corresponding image patch being a real sample. By calculating the average of these probability values, we can obtain the discriminator's final assessment of the overall image authenticity.

C. Loss function

Adversarial loss [6]. Adversarial loss is primarily based on the discriminator and the generator. The method for calculating the loss function is as follows:

$$\mathcal{L}_{adv} = \mathbb{E}_x [\mathcal{D}_{src}(x)] - \mathbb{E}_{x,z} [\mathcal{D}_{src}(G(x, z))] \quad (1)$$

Reconstruction loss. The adversarial loss function can ensure the realism of the generated images but cannot guarantee that the generated images retain the content of the input images.

To address this issue, the reconstruction loss function is defined as follows:

$$\mathcal{L}_{rec} = \mathbb{E}_{x,z} \left[\| x - G(G(x,z), \mathcal{D}_{coor}(x)) \|_1 \right] \quad (2)$$

Category loss. The purpose of category loss is to generate facial expression images with target labels. Therefore, the goal is to ensure that the input image, when transformed into the output image $G(x, z)$, can be correctly classified into the target label c . To achieve this purpose, the definition of category loss in this paper is as follows:

$$\mathcal{L}_{cls}^D = \mathbb{E}_{x,c} \left[-\log \mathcal{D}_{cls}(c|x) \right] \quad (3)$$

$$\mathcal{L}_{cls}^G = \mathbb{E}_{x,c,\rho} \left[-\log \mathcal{D}_{cls}(c|G(x,z,\rho)) \right] \quad (4)$$

Conditional regression loss. The purpose of conditional regression loss is to enable D to correctly estimate the facial expression features z of both real and generated images. To achieve this purpose, the definition of conditional regression loss in this paper is as follows:

$$\mathcal{L}_{inf_o} = \mathbb{E}_{x,z} \left[\| \mathcal{D}_{coor}(G(x,z)) - z \|_2^2 \right] \quad (5)$$

Emotion intensity loss. The purpose of emotion intensity loss is to ensure that interpolation is semantically meaningful. The definition of emotion intensity loss in this paper is as follows:

$$\mathcal{L}_\rho = \mathbb{E}_{x,z} \left[\| \hat{\rho}(G(x,z)) - \rho \|_2^2 \right] \quad (6)$$

The complete loss function is represented by the following equation:

$$\mathcal{L}_G = \mathcal{L}_{adv} + \lambda_{cls} \mathcal{L}_{cls}^f + \lambda_{rec} \mathcal{L}_{rec} + \lambda_{inf_oG} \mathcal{L}_{inf_o} + \lambda_\rho \mathcal{L}_\rho \quad (7)$$

$$\mathcal{L}_D = -\mathcal{L}_{adv} + \lambda_{cls} \mathcal{L}_{cls}^{real} + \lambda_{inf_oD} \mathcal{L}_{inf_o} \quad (8)$$

IV. RESULTS AND VALIDATION

A. Qualitative Evaluation

Fig.3 illustrates six basic facial expression images generated by MMA-GAN. As shown in Fig.6, the proposed GRA-GAN can generate authentic facial expression images of different categories, effectively addressing issues of local region overlap and blurriness, distinctly reflecting various emotions.

Fig.3 results indicate that the MMA-GAN proposed in this chapter, through the multi-scale mixed attention network module, can better adapt to the task of facial expression generation. It can solve the phenomena of partial area overlap and blurriness, enhance the generation of facial details, thereby making the generated expressions more realistic.

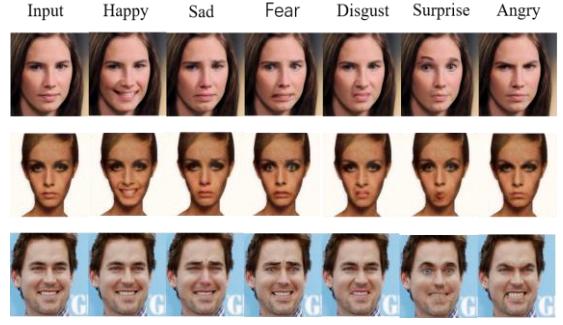


Figure 3 MMA-GAN Generated Facial Expression Images Example

B. Quantitative Evaluation

In this section, common image evaluation metrics are adopted as the criteria for assessing the quality of generated images, which include Fréchet Inception Distance (FID) [7], Peak Signal to Noise Ratio (PSNR) [8], Learned Perceptual Image Patch Similarity (LPIPS) [9], and Structural Similarity (SSIM) [10]. FID mainly compares the similarity between input and generated images from the feature level, with lower scores indicating higher similarity. LPIPS is a measure based on deep learning to assess the perceptual similarity between two images. Unlike traditional metrics that rely on pixel differences, LPIPS focuses on differences in image textures, patterns, or structures that the human visual system is likely to perceive. This metric evaluates their visual similarity by comparing differences in high-level feature representations of the images. SSIM is a metric for measuring the similarity between two images, often used in generative adversarial networks to assess the quality of generated images. Considering aspects of image brightness, contrast, and structure, SSIM provides a better reflection of human perception of image quality compared to traditional pixel-based error metrics, such as Mean Squared Error (MSE).

Table 1 MMA-GAN Compared to Other Models in Terms of FID, PSNR, SSIM, and LPIPS

Method	FID ↓	PSNR↑	SSIM↑	LPIPS↓
CycleGAN	14.34	20.34	0.82	0.0300
StarGAN	9.84	24.35	0.84	0.0180
MMA-GAN	6.82	29.89	0.87	0.0069

From Table 1, it is evident that MMA-GAN outperforms the other two models in terms of the four image quality assessments. The addition of global residual connections in MMA-GAN is intended to minimize unnecessary facial expression changes during the generation process, which results in a favorable display of global image quality. The introduction of the multi-scale mixed attention residual module enhances the model's ability to focus on parts of the face that require expression modifications. This module combines feature maps of different scales and adaptively weights these features through an attention mechanism, allowing the model to concentrate on both local details and the global structure simultaneously. This mechanism enables the network to modify specific facial regions, such as eyebrows and mouth, in response to complex

expression dynamics on a fine-grained level, while also maintaining the overall consistency of facial expressions on a more macro level. As for the standards of PSNR, SSIM, and LPIPS — which evaluate pixel-level detail, contrast in image structure, and perceptual similarity respectively — MMA-GAN employs a multi-scale mixed attention residual module that effectively enhances the quality of generated images in the spatial dimension, thereby ensuring a more refined consistency with the original image in terms of spatial layout.

C. Ablation Study

In this section, the paper conducted ablation experiments to assess the contributions of the MMAR module and the GRC module. We compared three versions of the model: the original model (Ours w/o MMAR & GRC), the model with only the MMAR module (Ours w/o GRC), the model with only the GRC module (Ours w/o MMAR), and the model with both GRC & MMAR modules (Ours). This research explores the impact of different modules on the performance of facial expression generation through ablation studies, with the results presented in Fig.4. Preliminary observations of the output from the original baseline model show the presence of overlapping and blurring, which limits the model's effectiveness in generating detailed facial expressions. To enhance the quality of the generated images, the MMAR module was introduced. The results after integrating the MMAR module (the second row of Fig.4) indicate a significant improvement in the presentation of facial details.

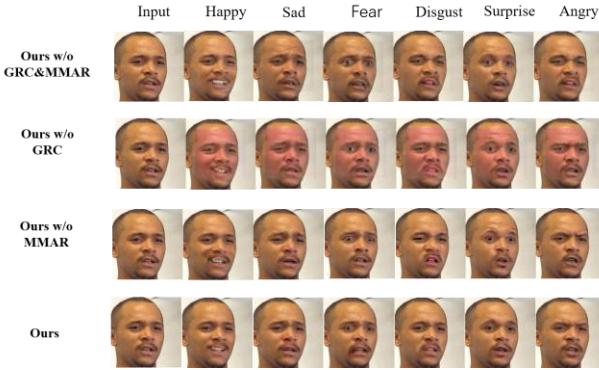


Figure 4 Example Results of MMAR-GAN Ablation Study

Nonetheless, there is still a lack of sufficient global constraints, leading to improved details but suboptimal overall image quality. Furthermore, the study explores the effect of adding the GRC module to the original baseline model. The results after the incorporation of the GRC module (the third row of Fig.4) show a marked improvement in the overall image quality, but still fall short in detail constraints. Particularly in

areas of the face with significant expression changes, such as the mouth, while the addition of the GRC module mitigated the overlapping and blurring and enhanced the realism and naturalness of the image, the precise reproduction of complex

expressions, especially the clarity of the mouth, requires further enhancement.

V. CONCLUSIONS

To address the issue of low-quality facial expression image generation, this paper proposes a Mixed Multi-Attention Generative Adversarial Network (MMA-GAN). The proposed GRA-GAN enhances the details of generated facial expression images by embedding a mixed attention mechanism within the generative adversarial network, which focuses on feature transmission from both the channel and spatial dimensions. This improves the clarity of facial regions and reduces overlap and blurring, thereby increasing the realism of the generated facial images. To further strengthen the model's learning capacity for key areas, a multi-scale mixed attention module is proposed to enhance the learning ability during feature transmission and strengthen the connections between residual blocks. This allows for better preservation of critical information during the transmission process in the residual blocks. The task of generating facial expressions in natural environments is challenging, as natural environment images often have many disturbances, such as lighting, poses, and occlusions. High-quality facial expression generation for datasets in natural environments is a major research direction for the future.

REFERENCES

- [1] Naga P, Marri S D, Borreo R. Facial emotion recognition methods, datasets and technologies: A literature survey[J]. Materials Today: Proceedings, 2023, 80: 2824-2828.
- [2] Mollahosseini A, Hasani B, Mahoor M H. Affectnet: A database for facial expression, valence, and arousal computing in the wild[J]. IEEE Transactions on Affective Computing, 2017, 10(1): 18-31.
- [3] Martin B, Edwards K, Jeffrey I, et al. Experimental microwave imaging system calibration via cycle-GAN[J]. IEEE Transactions on Antennas and Propagation, 2023.
- [4] Chen G, Zhang G, Yang Z, et al. Multi-scale patch-GAN with edge detection for image inpainting[J]. Applied Intelligence, 2023, 53(4): 3917-3932.
- [5] Brophy E, Wang Z, She Q, et al. Generative adversarial networks in time series: A systematic literature review[J]. ACM Computing Surveys, 2023, 55(10): 1-31.
- [6] Oikarinen T, Zhang W, Megretski A, et al. Robust deep reinforcement learning through adversarial loss[J]. Advances in Neural Information Processing Systems, 2021, 34: 26156-26167.
- [7] Soloveitchik M, Diskin T, Morin E, et al. Conditional frechet inception distance[J]. arXiv preprint arXiv:2103.11521, 2021.
- [8] Suriyan K, Ramaingam N, Rajagopal S, et al. Performance analysis of peak signal-to-noise ratio and multipath source routing using different denoising method [J]. Bulletin of Electrical Engineering and Informatics, 2022, 11(1): 286-292.
- [9] Ghildyal A, Liu F. Shift-tolerant perceptual similarity metric[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 91-107.
- [10] Bakurov I, Buzzelli M, Schettini R, et al. Structural similarity index (SSIM) revisited: A data-driven approach[J]. Expert Systems with Applications, 2022, 189: 116087.