Artificial Intelligence-Driven Physical Simulation and Animation Generation in Computer Graphics

Fei Wang
Research Department, Zhengzhou Professional Technical Institute of Electronics & Information, Zhengzhou, 451450, China

***Abstract*—This study explores an expression synthesis algorithm anchored in Generative Adversarial Networks (GAN) with attention mechanisms, achieving enhanced authenticity in facial expression generation. Evaluated on the MUG and Oulu-CASIA datasets, our method synthesizes six expressions with superior clarity (96.63±0.26 confidence for neutral expressions) and smoothness (SSIM >0.92 for video frames), outperforming StarGAN and ExprGAN in detail preservation and temporal stability. The proposed model demonstrates significant advantages in realism and identity retention, validated through quantitative metrics and comparative experiments.**

***Keywords*—GAN; computer graphics; expression synthesis; animation generation**

I. Introduction

With the advancement in computer graphics and artificial intelligence, facial expression synthesis technology has been applied in animation production, video games, human-computer interaction, etc. [1-2]. The effectiveness of expression synthesis, namely its realism and naturalness, directly influences the user's immersion and interactive experience. Hence, in-depth research into methods of facial expression synthesis becomes exceedingly crucial.

In the early stages, facial expression synthesis predominantly relied on parametric methods and muscle models. These methods were grounded in the biomechanical behavior of facial muscles, utilizing mathematical models to depict the impact of muscle movements on facial expressions [3]. While these methods were capable of encoding facial movement information to a certain extent, their complexity and inflexibility often resulted in the loss of detail, especially when handling rapid changes or a diversity of emotions. The Facial Action Coding System (FACS) is another classical approach for describing expressions [4]. It provides a standardized set of Action Units to precisely define facial expressions [5-6]. Although this method has its advantages in extracting static expressions, it still requires substantial manual intervention and complex annotation processes in dynamic sequence generation, making it difficult to meet the demands of rapid generation.

Recently, deep learning methods have garnered great attention in expression synthesis [7]. The new generation of deep generative models has increasingly adopted a strategy that combines expression feature extraction with driving images [8-9]. By utilizing high-dimensional facial feature information as conditional input or directly employing driving images as supervisory information, the generator can learn the intricate relationships and corresponding mappings of expressions. Furthermore, the introduction of attention mechanisms [10] enhances the precision and detail of the synthesis. In the evolution of Generative Adversarial Networks (GANs) [11], dynamic image generation models that incorporate temporal information, such as Temporal GANs [12], and Recurrent Neural Network architectures [13], are capable of effectively capturing temporal continuity, thereby enhancing the smoothness of the generated videos. This progress not only offers new approaches for expression animation generation but also provides additional possibilities for research in expression transfer and conversion.

The proposal of GANs has greatly propelled advancements in this field. Through adversarial training, GANs enable the generator to produce high-fidelity synthetic images with a quality approaching that of real images, and they exhibit outstanding performance in expression diversity. In this context, several GAN-based variants have emerged, such as conditional GAN [14], StarGAN [15], and ExprGAN [16]. Researchers have gradually realized the potential of combining expression features with deep learning methods. By extracting features from input facial images and integrating target expression information, the generator can produce corresponding expression changes while preserving individual characteristics. Existing methods (e.g., StarGAN and ExprGAN) perform poorly in expression detail and temporal continuity. By integrating channel and spatial attention mechanisms, our method significantly improves the naturalness and detail retention of generated expressions. This study proposes a GAN- and attention mechanism-based approach to address these issues and achieve high-quality expression animation generation.

In summary, this study explores an expression synthesis algorithm based on GAN and incorporates attention mechanisms. Section II details the proposed methodology, while Section III presents experiments and comparative results. Section IV concludes the study. Our method dynamically weights key features (e.g., mouth corners and nasolabial folds) through attention mechanisms, avoiding common issues such as blurring and artifacts in traditional methods. Through an in-depth study of the combination of expression feature extraction and dynamic generation, we can achieve higher breakthroughs in realism, detail, and generation effectiveness.

II. Method Introduction

A. Generative Adversarial Networks (GAN)

GANs generate realistic data and enhance discriminative capabilities through adversarial learning between a generator and a discriminator, ultimately making the generated data indistinguishable from real data.

For a synthetic model V, its training process can typically be distilled into an optimization problem.

$$\min_{G}\max_{D} V(G,D) = E_{x\sim P_{data}}(\log D(x)) + E_{z\sim P_z(z)}(\log(1-D(G(z)))) \quad (1)$$

where, G denotes the generator, which produces an output denoted as $D(x)$; D signifies the discriminator, yielding an output represented by $G(z)$; P_{data} refers to the distribution of the genuine data x; while z embodies noise data that conforms to the random distribution P_z. The $D(G(z))$ indicates the discriminator's predicted probability regarding the data generated by the generator. The first term $E_{x\sim P_{dalu}}$ represents the probability of correct classification for authentic samples by the discriminator. Conversely, the second term $E_{z\sim P_2(z)}$ denotes the probability of incorrect classification for generated samples by the discriminator.

In the n-th iteration of the network iteration, k pairs of training data $\{z^{(1)},...,z^{(m)}\}$ are randomly sampled from the noise distribution $p_{\text{data}}(x)$, and m samples $\{x^{(1)},...,x^{(m)}\}$ are randomly drawn from the data synthesis distribution.

Each iteration of the network randomly obtains training data from the prior noise distribution $p_g(z)$ and randomly draws a sample from the data synthesis distribution. Then it updates the weights of the generator and discriminator networks using the following formulas,

$$\nabla_{\theta_d}\frac{1}{m}\sum_{i=1}^{m}[\log D(x^{(i)}) + \log(1-D(G(z^{(i)})))] \quad (2)$$

$$\nabla_{\theta_g}\frac{1}{m}\sum_{i=1}^{m}\log(1-D(G(z^{(i)}))) \quad (3)$$

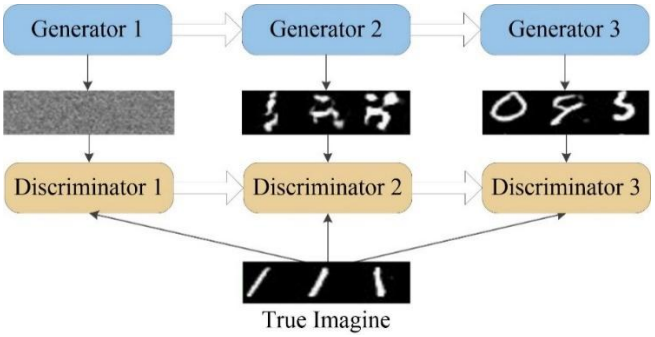

Fig. 1. Illustrates the training process of GAN networks.

In the training of GANs, shown in Fig. 1, the process commences with a randomly initialized first generation of the generator and discriminator. During the initial phase, the images produced by the generator exhibit low quality, while the discriminator begins to learn the distinction between authentic data and counterfeit data. Following the training of the generator, a second generation is obtained alongside the training of a new discriminator. This iterative process continues, ultimately leading the generator to produce images of near-perfection, rendering the discriminator unable to discern the genuine from the false, thereby achieving Nash equilibrium and enabling successful training of the generative network.

Conditional Generative Adversarial Networks (CGANs) represent an enhancement over traditional GANs, as shown in Fig. 2. By incorporating additional conditional information to direct the generation process, CGANs ensure that the generated data aligns more closely with specific demands or characteristics. The objective function is articulated as follows:

$$\min_{G}\max_{D} V(G,D) = E_{x\sim P_{data}}[\log D(x|y)] + E_{z\sim P_z(z)}[\log(1-D(G(z|y)))] \quad (4)$$

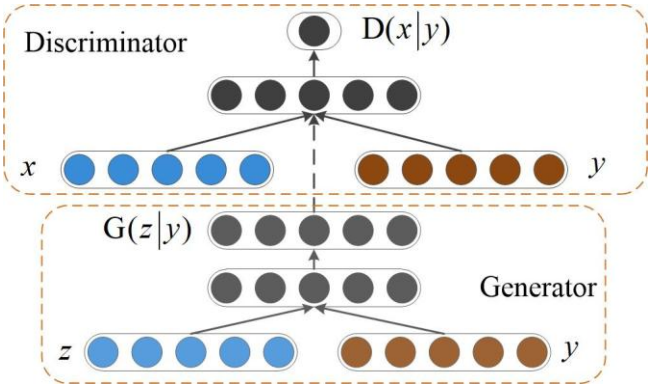

Fig. 2. The architecture of the Conditional GAN.

B. Integrating the Attention Mechanism

The Attention Mechanism is an approach that emulates human visual focus [10]. Its purpose is to enhance the efficiency and effectiveness of models when processing information, particularly in sequential data and image processing. Channel Attention and Spatial Attention are two crucial forms of the attention mechanism, primarily employed to boost the performance of Convolutional Neural Networks (CNNs) in image processing tasks. They augment the model's focus on significant features through different methodologies.

Channel Attention primarily concentrates on the channel dimension of the input feature maps. It endeavors to emphasize important features and suppress insignificant ones by assigning a weight to each channel, thereby strengthening the feature representation capacity.

The input feature map is denoted as $M=[m_1,m_2,...,m_C]$, with the i-th channel (out of a total of C channels) being represented as $m_i \in R^{h\times w}$. Here, h,w signify the height and width of the image, respectively. The channel statistics $P \in R^{1\times 1\times C}$ are achieved through global pooling operations, which compact the spatial dimensions, thereby embedding the global spatial information into vector p. The k-th element of this vector is shown as follows:

$$p_k = F_{GP}(m_k) = \frac{1}{h \times w} \sum_{i}^{h} \sum_{j}^{w} m_k(i, j) \tag{5}$$

where, F_{GP} denotes the global pooling function, while (i, j) signifies the spatial location. Subsequently, a fully connected layer is employed to encode the channel information.

$$\hat{p} = \sigma(W_{\text{up}} \delta(W_{\text{down}} p)) \tag{6}$$

where, σ and δ denote the sigmoid and rectified linear unit activation functions, respectively.

A feature map $\hat{M}_{ca}$ weighted by channel attention enhances significant features and suppresses less pertinent ones.

$$\hat{M}_{ca} = [\hat{p}_1 m_1, \hat{p}_2 m_2, ..., \hat{p}_C m_C] \tag{7}$$

Spatial attention focuses on the spatial dimensions of feature maps [17]. We denote the input image as $M = [m^{1,1}, m^{1,2}, ..., m^{i,j}, ..., m^{h,w}]$. A projection vector q is employed to capture spatial information, with W_{sq} representing the weights of the convolutional layer used for spatial compression operations. By applying the spatial attention weights to the input feature map through element-wise multiplication, we obtain a new feature map $\hat{M}_{sa}$, thereby amplifying the significance of specific spatial regions.

$$q = \sigma(W_{sq} * M) \tag{8}$$

$$\hat{M}_{sa} = [(q_{1,1}) m^{1,1}, (q_{1,2}) m^{1,2}, ..., (q_{i,j}) m^{i,j}, ..., (q_{h,w}) m^{h,w}] \tag{9}$$

Channel-wise attention and spatial attention can be integrated to create a more powerful attention mechanism, significantly enhancing the performance of Neural Networks in image processing tasks. This approach enables a more comprehensive capture of essential features within images, regardless of their orientation within the channel or spatial dimensions.

C. Facial Expression Synthesis Algorithm Model

This research employs CGANs for the synthesis of facial expression intensity. Based on the intensity of expressions ranging from weak to strong, data is manually categorized into four levels: neutral, weak, medium, and strong, and represented as $\mathbf{z} = [z_0, z_1, z_2, z_3]$. Initially, a generator G and a discriminator D are constructed. The generator G's task is to, given a source image x_s and its intensity label, produce a new image x_t featuring the targeted expression intensity. Inspired by cGANs, this algorithm uses the expression intensity label as a constraint to control the synthesis of expression intensity. Additionally, changes in expression intensity are usually very subtle; to enhance the network's learning capacity in handling intensity variations, a fusion attention module is incorporated into the generator G.

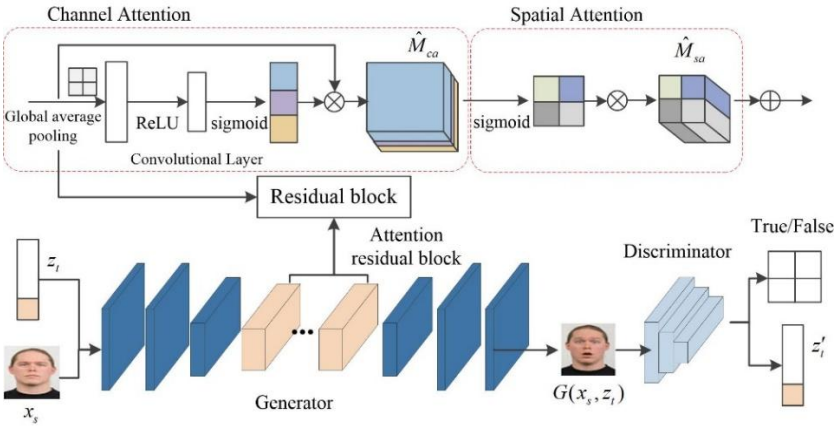

Fig. 3. Facial expression synthesis algorithm with attention mechanism.

The algorithm proposed integrates Conditional GAN and attention mechanisms, as shown in Fig. 3. By concatenating preprocessed original images of 128×128 pixels with target intensity labels, it undergoes processing through downsampling convolutional layers, residual modules infused with attention, and upsampling transpose convolutional layers, ultimately producing generative images that embody the desired expression intensity. Concurrently, the discriminator network enhances the PatchGAN structure by incorporating an auxiliary classifier, which not only differentiates between real and fake images but also assesses expression intensity, thereby augmenting both the accuracy and realism of the generated images.

This study devised various loss functions to constitute the final objective function. Through the optimization of this objective function, both the generator network and the discriminator network engage in adversarial learning, fine-tuning network parameters to achieve an optimal model, thereby synthesizing lifelike facial images that convey specified expression intensities. In the conclusion, neutral-expression facial images were employed as test samples, generating faces with varying expression intensities while preserving identity information. The principles of reconstruction loss, pixel loss, and identity retention loss functions are illustrated in Fig. 4.

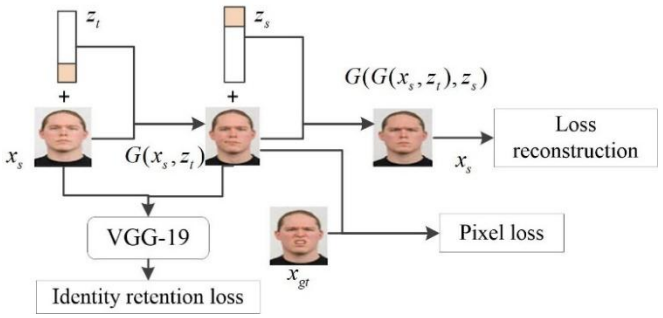

Fig. 4. Diagram of the loss function principle.

1) The adversarial loss is employed to assess the dissimilarity between generated images and real images. Here, an adversarial loss function analogous to the conditional GAN loss function is devised.

$$L_{adv}^{G} = E_{x_s, z_t}[\log(1 - D_{st}(G(x_s, z_t)))] \tag{10}$$

$$L_{adv}^{D} = -E_{x_s}[\log D_{st}(x_s)] - E_{x_s, z_t}[\log(1 - D_{st}(G(x_s, z_t)))] \tag{11}$$

where, L_{adv}^{G} and L_{adv}^{D} represent the adversarial losses of the generator G and the discriminator D, respectively. $G(x_s, z_t)$ denotes the image synthesized from the source image x_s and the target intensity label z_t, while $D_{st}(x)$ indicates the probability of the authenticity of the image x.

2) The formula for calculating the reconstruction loss L_{rec} is as follows:

$$L_{rec} = E_{x_s, z_s, z_t}\left\| x_s - G(G(x_s, z_t), z_s) \right\|_1 \tag{12}$$

where, z_s signifies the intensity label of the input image, while z_t represents the target label. $G(G(x_s, z_t), z_s)$ denotes the reconstructed image.

3) The formulation for the intensity classification loss function is delineated as follows:

$$L_{int}^{D} = E_{x_s, z_s}[-logD_{int}(z_s \mid x_s)] \tag{13}$$

$$L_{int}^{G} = E_{x_s, z_t}[-logD_{int}(z_t \mid G(x_s, z_t))] \tag{14}$$

Among them, $D_{int}(z_s \mid x_s)$ denotes the probability distribution of the source image z_s concerning its intensity label x_s; $D_{int}(z_t \mid G(x_s, z_t)$ reflects the probability distribution of the generated image concerning the target intensity z_t.

4) The computation formula for the expression intensity classification loss L_{pix} is as follows:

$$L_{pix} = E_{x_s, x_t, z_t}\left\| x_{gt} - G(x_s, z_t) \right\|_1 \tag{15}$$

where, x_{gt} denotes the real image with a facial expression intensity of z_t. The L1 norm is employed to calculate the difference between the generated image and its corresponding ground-truth.

5) The identity-preservation loss is utilized to ensure that the generated image retains the identity information of the original image, such as facial features and skin tone, while altering the intensity of the facial expression.

$$L_{id} = E_{x_s, z_t}\left\| \phi(x_s) - \phi(G(x_s, z_t)) \right\|_1 \tag{16}$$

where, $\phi(x_s)$ represents the input facial image x_s, from which identity features are extracted by the feature extractor ϕ. $\phi(G(x_s, z_t))$ denotes the identity features extracted from the generated facial image $G(x_s, z_t)$ by the feature extractor.

6) The comprehensive objective function is formulated by a weighted combination of the aforementioned five loss functions, enabling the model to maintain a balance among various types of losses throughout the training process.

$$L_G = L_{adv}^{G} + \lambda_{pix} L_{pix} + \lambda_{rec} L_{rec} + \lambda_{id} L_{id} + \lambda_{int} L_{int}^{G} \tag{17}$$

$$L_D = L_{adv}^{D} + \lambda_{int} L_{int}^{D} \tag{18}$$

where, $\lambda_{pix}, \lambda_{rec}, \lambda_{id}, \lambda_{int}$ signifies the weight coefficients. Through the iterative refinement of L_G and L_D, the ultimate result is the synthesis of photorealistic facial images with varying intensities of expression.

III. Experiment and Analysis

A. Data Set and Evaluation Index

This experiment utilizes the MUG dataset and the Oulu-CASIA dataset as sources of data. The MUG dataset includes sequences of six expressions from 86 subjects, while the Oulu-CASIA dataset comprises videos of 80 subjects under three lighting conditions, while the Oulu-CASIA database consists of expression videos recorded by 80 subjects under three distinct lighting conditions. The emotions involved include happiness, sadness, anger, disgust, surprise, and fear. All images were aligned using 68 key points (Fig. 5) and cropped to 128×128 resolution to eliminate lighting and pose interference. This study divided the training set according to 7:3, and ensured that each expression was representative in both the training set and the test set. The PyTorch deep learning framework was employed for the experiment.

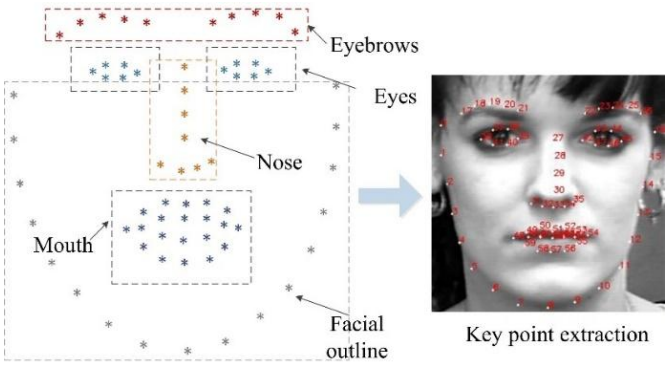

Fig. 5. Extraction of facial feature key points.

In this experiment, we first preprocess all facial images within the dataset. A keypoint detection algorithm identifies and extracts 68 key points from the facial images (see Fig. 5), followed by alignment. Subsequently, the images are cropped to a dimension of 128×128×3. Manual annotation is then conducted, classifying the expressions into four levels: neutral, weak, moderate, and strong. Taking "happiness" as an example, the classification results are shown in Table I.

TABLE I. CLASSIFICATION OF EXPRESSIONS IN THE TWO DATASETS

Database	MUG		Oulu-CASIA	
	Training set	Testing set	Training set	Testing set
Neutral	4191	420	498	56
Subtle	1514	-	469	-
Moderate	2410	-	482	-
Intense	2495	-	573	-

During the experiment, we utilized Face++'s online facial verification API to authenticate face images generated with varying intensities of expressions. By calculating the confidence level (ranging from 0 to 100) between the input image and the generated images, we evaluated the retention of identity information in the synthetic images. The misidentification rate was set at >78, and we compared the confidence levels of the input image with those of the generated neutral, mild, medium, and strong expression images, respectively.

B. Expression Synthesis Results and Evaluation

In this study, we generated different intensities of happy expressions based on neutral face images. Through experiments conducted on the MUG and Oulu-CASIA datasets, we evaluated the model's performance at various intensities. The results are depicted in Fig. 6. The experimental results indicate that at low intensity "happy" expressions, the generated face images only exhibit a slight smile, with a subtle upward trend of the corners of the mouth. As the intensity of the expression increases, the medium intensity "happy" expression shows marked changes, characterized by the mouth gradually opening, revealing half of the teeth, and the formation of nasolabial folds. When the expression intensity reaches its peak, the facial expression generated by the model exhibits the most intense emotional characteristics, with the corners of the mouth rising to their maximum extent, almost completely exposing all teeth, and the nasolabial folds becoming more pronounced. These results demonstrate that the proposed model can effectively and accurately simulate different intensities of "happy" expressions, showcasing excellent performance and expressive capability.

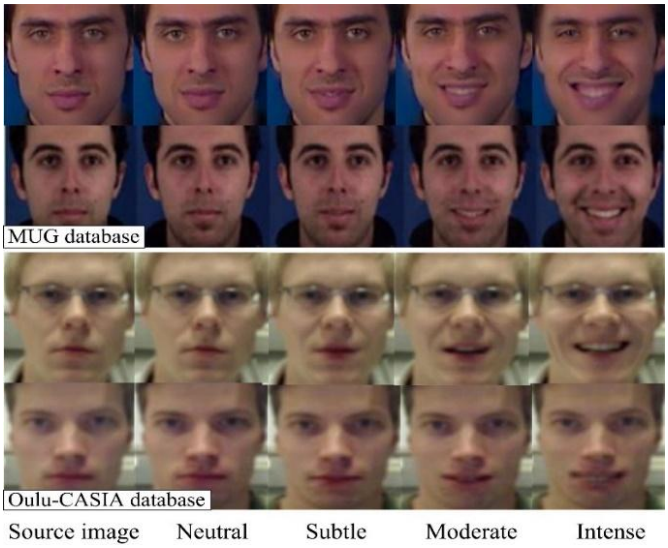

Fig. 6. Examples of expression synthesis results.

We conducted a quantitative evaluation of the "happy" expression synthesis results using confidence levels, shown in Table II.

TABLE II. FACIAL VERIFICATION CONFIDENCE LEVELS ON TWO DATASETS

Confidence	MUG	Oulu-CASIA
Neutral	96.63±0.26	97.0±0.52
Subtle	96.25±0.34	96.57±0.24
Moderate	95.62±0.25	95.67±0.42
Intense	94.25±0.76	94.47±0.24

Based on the results presented in Table II, we conducted an analysis of the "happy" expression synthesis performance across both the "MUG" and "Oulu-CASIA" datasets. Overall, both datasets exhibited a high degree of confidence in synthesizing happy expressions, albeit with certain distinctions. Under neutral expressions, Oulu-CASIA demonstrated slightly superior performance compared to MUG, indicating that both datasets accurately capture neutral expressions, with Oulu-CASIA excelling slightly more. As the intensity of the expression increased, the confidence level of MUG gradually decreased, particularly when rendering more intense happy expressions, whereas Oulu-CASIA maintained a commendable synthesis effect. In summary, Oulu-CASIA exhibited marginally superior stability and accuracy in "happy" expression synthesis.

C. Comparative Experiments

A comparative experiment was conducted to evaluate the image synthesis performance of the proposed method against StarGAN [15] and ExprGAN [16] methods on the MUG dataset. The synthesis effects of six expressions across different models are illustrated in Fig. 7.

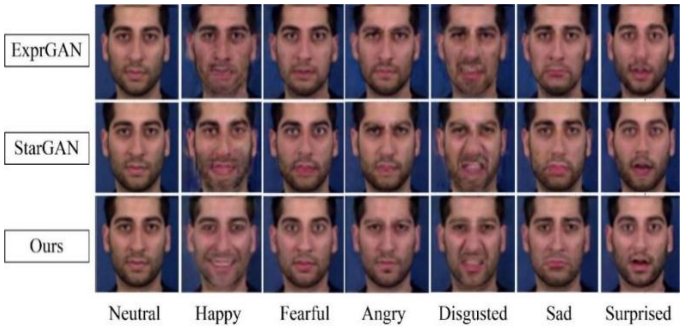

Fig. 7. Comparative synthesis effects of expressions across different models.

The proposed method significantly outperformed StarGAN and ExprGAN in expression synthesis. Specifically, when synthesizing sad expressions, our method not only accurately depicted the pouting mouth and sorrowful eyes but also rendered the overall expression more realistically and naturally. In contrast, StarGAN and ExprGAN exhibited more moderate performance in expression details. In the transformed happy expressions, the upturned corners of the mouth and the changes in nasolabial folds were less pronounced, lacking key muscular movement characteristics. StarGAN demonstrated certain limitations in expression synthesis, particularly in the synthesis of disgust expressions, which tended to be confusable with anger. This confusion was evidenced by the presence of most anger expression details in disgust expressions, leading to potential misclassification. Additionally, in the synthesized happy expressions, the mouth area exhibited noticeable blurriness, lacking critical information such as teeth exposure, which affected the overall authenticity of the expression. ExprGAN also encountered certain issues in expression synthesis, such as the occasional appearance of artifacts around the eyes and mouth during arbitrary expression synthesis, which impaired the overall image clarity and realism. The presence of these artifacts resulted in expressions appearing less natural and with less adequate detail.

The face verification confidence levels on the MUG dataset under different methods are shown in Table III. The proposed method, which integrates Conditional GANs and attention mechanisms, significantly enhanced the accuracy and naturalness of expression synthesis. Compared to traditional expression synthesis methods, the proposed method delivered more refined performance in key expression details (such as the upturning of the corners of the mouth and changes in nasolabial grooves), avoiding blurriness and artifacts. Experimental results indicate that this method achieved exceptionally high confidence levels in the synthesis of various expressions, particularly “happy” expressions, significantly outperforming existing methods like StarGAN and ExprGAN. As shown in Table III, our method achieves significantly higher identity preservation confidence (96.63±0.26) on the MUG dataset compared to ExprGAN (67.01±0.32), as the latter tends to lose identity features (e.g., skin tone and facial contours) during cross-intensity synthesis. This approach not only maintained stable performance across varying expression intensities but also more accurately captured muscle movement characteristics, resulting in more authentic and vivid generated expressions.

TABLE III. FACE VERIFICATION CONFIDENCE LEVELS ON THE MUG DATASET UNDER DIFFERENT METHODS

Confidence	Ours	StarGAN	ExprGAN
Neutral	96.63±0.26	96.34±6.25	67.01±0.32
Subtle	96.25±0.34	96.49±6.29	66.57±0.31
Moderate	95.62±0.25	94.07±6.15	64.46±0.42
Intense	94.25±0.76	92.73±6.22	64.03±0.38

D. Facial Expression Video Synthesis Effect

Based on the proposed method utilizing conditional Generative Adversarial Networks, we aim to synthesize facial expression videos. By testing images not included in the training dataset, we first extract facial features and construct a feature map of expressions, which are then employed to drive the synthesis of video frames. To evaluate the continuity and stability of the synthesized video frames, we measure the video's smoothness by calculating the Structural Similarity Index (SSIM) between consecutive frames. Experimental results, shown in Fig. 8, indicate that the video frames generated by our method display a natural appearance in terms of brightness and variations in expression, exhibiting commendable inter-frame continuity with no significant expression disjunctions or abrupt changes in brightness. We have conducted comparative experiments pitting our method against existing approaches (StarGAN and ExprGAN). The results demonstrate that our method surpasses the comparative techniques in the authenticity, smoothness, and continuity of the synthesized videos, with a more stable SSIM curve indicative of steadier frame transitions.

In summary, through systematic experimental validation, the proposed method based on CGANs has showcased exceptional performance in facial expression video synthesis, capable of producing high-quality, smooth, continuous facial expression videos.

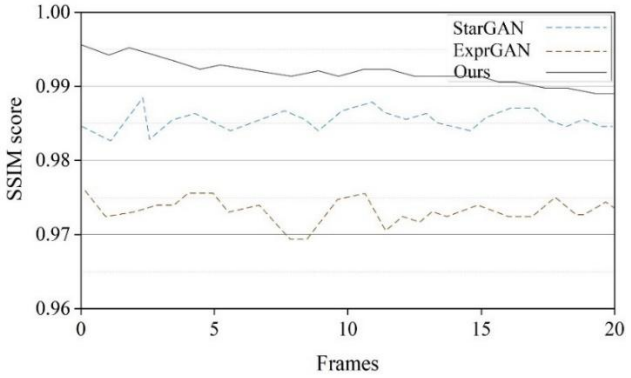

Fig. 8. Evaluation of continuity in synthesized expression videos.

IV. CONCLUSION

This study presents an algorithm for facial expression synthesis founded on Generative Adversarial Networks (GAN) combined with an attention mechanism. By effectively integrating feature fusion, we have significantly mitigated the loss of detail in the generated images, thereby enhancing the authenticity of the synthesized facial expressions. The experimental results reveal that the animated expressions

produced by our method not only exhibit a more natural and fluid visual appeal but also demonstrate marked improvements in clarity and detail retention compared to previous Generative Adversarial Networks. Particularly, in the synthesis experiments involving six distinct expressions, our algorithm has shown heightened accuracy and stability, rendering the generated expressions remarkably akin to genuine human expressions. Furthermore, the incorporation of the attention mechanism has endowed the model with greater flexibility in feature extraction, allowing for better preservation of key features and effectively elevating the synthesis quality. This study has two limitations: 1) The model requires aligned facial key points, limiting its applicability to unconstrained images; 2) Training on small datasets may affect generalization of rare expressions. In future work, we intend to explore more efficient model architectures, optimize training strategies, and further enhance model performance in more complex and dynamic facial expression synthesis. Additionally, the consideration of incorporating more context information and user interaction mechanisms to bolster the flexibility and adaptability of facial expression generation in practical applications will serve as a vital direction for our research.

ACKNOWLEDGEMENT

Henan Provincial Education Science Planning Projects: Research on the Practical Path of Applied Graphic Design Education in Higher Education Institutions from the Perspective of Artificial Intelligence. Project Approval Number: 2024YB0556.

REFERENCES

[1] Xie H X, Lo L, Shuai H H, et al. An overview of facial micro-expression analysis: Data, methodology and challenge. IEEE Transactions on Affective Computing, 2022, 14(3): 1857-1875.

[2] Fan D P, Huang Z, Zheng P, et al. Facial-sketch synthesis: A new challenge. Machine Intelligence Research, 2022, 19(4): 257-287.

[3] Ito J, Moriyama H, Shimada K. Morphological evaluation of the human facial muscles. Okajimas Folia Anatomica Japonica, 2006, 83(1): 7-14.

[4] Gupta T, Haase C M, Strauss G P, et al. Alterations in facial expressions of emotion: Determining the promise of ultrathin slicing approaches and comparing human and automated coding methods in psychosis risk. Emotion, 2022, 22(4): 714.

[5] Krumhuber E G, Skora L I, Hill H C H, et al. The role of facial movements in emotion recognition. Nature Reviews Psychology, 2023, 2(5): 283-296.

[6] Ye Y, Song Z, Zhao J. High-fidelity 3D real-time facial animation using infrared structured light sensing system. Computers & Graphics, 2022, 104: 46-58.

[7] Karnati M, Seal A, Bhattacharjee D, et al. Understanding deep learning techniques for recognition of human emotions using facial expressions: A comprehensive survey. IEEE Transactions on Instrumentation and Measurement, 2023, 72: 1-31.

[8] Yan Y, Huang Y, Chen S, et al. Joint deep learning of facial expression synthesis and recognition. IEEE Transactions on Multimedia, 2019, 22(11): 2792-2807.

[9] Xia Y, Zheng W, Wang Y, et al. Local and global perception generative adversarial network for facial expression synthesis. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 32(3): 1443-1452.

[10] Zhang F, Zhang T, Mao Q, et al. A unified deep model for joint facial expression recognition, face synthesis, and face alignment. IEEE Transactions on Image Processing, 2020, 29: 6574-6589.

[11] Creswell A, White T, Dumoulin V, et al. Generative adversarial networks: An overview. IEEE signal processing magazine, 2018, 35(1): 53-65.

[12] Wang J, Chen Y, Gu Y. A wearable-HAR oriented sensory data generation method based on spatio-temporal reinforced conditional GANs. Neurocomputing, 2022, 493: 548-567.

[13] Kanagachidambaresan G R, Ruwali A, Banerjee D, et al. Recurrent neural network. Programming with TensorFlow: Solution for Edge Computing Applications, 2021: 53-61.

[14] Cho J, Yoon K. Conditional activation GAN: improved auxiliary classifier GAN. IEEE Access, 2020, 8: 216729-216740.

[15] Liu Y, Wang X, Yuan C, et al. AttGAN: attention gated generative adversarial network for spatio-temporal super-resolution of ocean phenomena. International Journal of Digital Earth, 2024, 17(1): 2368705.

[16] Tov O, Alaluf Y, Nitzan Y, et al. Designing an encoder for stylegan image manipulation. ACM Transactions on Graphics (TOG), 2021, 40(4): 1-14.

Lu E, Hu X. Image super-resolution via channel attention and spatial attention. Applied Intelligence, 2022, 52(2): 2260-2268.