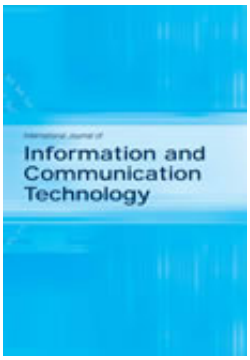
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Three-dimensional facial image modelling and animation generation method integrated with emotion recognition

Xiaowen Guo

DOI: 10.1504/IJICT.2025.10075321

Article History:

Received:	29 September 2025
Last revised:	28 October 2025
Accepted:	31 October 2025
Published online:	15 January 2026

Three-dimensional facial image modelling and animation generation method integrated with emotion recognition

Xiaowen Guo

College of Art and Creativity,
Anhui University of Applied Technology,
Hefei, 230000, China
Email: guoxiaowen2000@163.com

Abstract: This study proposes a method for 3D facial modelling and animation generation integrated with emotion recognition. This method deeply combines high-precision emotion recognition with animation driving to improve the naturalness of facial expressions, emotional accuracy, and real-time interaction capabilities. Experimental results show that the average accuracy rate of 3D facial emotion recognition increases from 89.1% to 92.1%, with happiness (HA) reaching 98.9%, verifying the model's high reliability and stability. In animation generation experiments, the emotion recognition accuracy reaches 90%, emotional consistency is 88%, and the average frame generation time is 48 ms/frame, all outperforming the control model. The research innovation and contribution lie in proposing a systematic integration strategy of emotion recognition with 3D modelling and animation generation. This enriches the theoretical framework of facial animation generation, achieving a balance among accuracy, naturalness, and efficiency, and providing an efficient and feasible technical solution.

Keywords: emotion recognition; 3D facial image; facial modelling; animation generation.

Reference to this paper should be made as follows: Guo, X. (2025) 'Three-dimensional facial image modelling and animation generation method integrated with emotion recognition', *Int. J. Information and Communication Technology*, Vol. 26, No. 52, pp.117–134.

Biographical notes: Xiaowen Guo earned her Bachelor's in Art and Design from Soochow University, Suzhou, Jiangsu, China in 2006, and Master's in Fine Arts from Anhui University of Finance and Economics, Bengbu, Anhui, China in 2011. Since 2012, she has been serving as a faculty member at the College of Art and Creativity, Anhui University of Applied Technology. In 2020, she was appointed as an Associate Professor. Her research interests encompass 3D modelling, 3D game scene construction, engineering animation, and unreal engine.

1 Introduction

In fields such as computer graphics, virtual reality, and human-computer interaction, 3D facial modelling and animation generation techniques have long been research hotspots. With the rapid development of virtual reality, augmented reality, and the digital human

industry, users have raised higher demands for the realism and naturalness of virtual avatars (Wang and Shi, 2023). The face is not only a crucial feature for identity recognition but also a core medium for emotional expression, where subtle muscle movements often convey complex emotional information (Di et al., 2021). However, traditional 3D facial modelling and animation generation methods primarily focus on geometric structures and motion drives, lacking modelling and expression at the emotional level. For instance, expression-driven approaches based on skeletal rigging or blendshapes typically achieve facial transformations through linear interpolation, which can reproduce basic emotions such as smiling or frowning. Yet, these methods often appear rigid and lifeless when expressing nuanced emotions or dynamic transitions. In emotional interaction scenarios involving virtual characters – such as comforting, surprising, or embarrassing moments – traditional animation techniques frequently limit themselves to large-scale movements of the mouth corners or eyebrows. They fail to capture subtle expressions like eye changes or muscle tremors. This results in virtual characters appearing stiff and emotionally detached during interactions. Such limitations are particularly pronounced in scenarios requiring natural emotional feedback, like virtual streaming, immersive interactions, remote education, and medical assistance. These become a critical bottleneck hindering the improvement of virtual humans' realism and affinity (Tu et al., 2021; Song and Kwon, 2024).

With advances in deep learning and computer vision, data-driven methods for emotion recognition and 3D face reconstruction have gradually emerged. Researchers have explored applying models such as convolutional neural networks (CNNs) and graph neural networks to facial feature extraction and emotion classification tasks, thereby improving the realism of face modelling and animation generation (Chen and Chen, 2023). Meanwhile, the progress of 3D scanning devices and high-resolution acquisition technologies has enabled multimodal data integration, laying a stronger foundation for dynamic facial modelling (Jiang et al., 2024). Nevertheless, most current studies still face two challenges: first, emotion recognition and 3D modelling are often performed separately, making it difficult to achieve seamless integration of model-driven techniques and emotional expressiveness (Li et al., 2022); second, the lack of fine-grained emotional control in animation generation hinders the production of natural and fluid facial dynamics (Wu et al., 2022). Thus, embedding emotion recognition mechanisms into 3D facial modelling and animation generation to achieve deep integration of geometric modelling and emotion-driven control has become an urgent problem to solve.

Despite recent advancements in 3D facial animation technology, there remain notable shortcomings in emotional expression, primarily manifested in limited emotional recognition accuracy, unnatural facial animations, and constrained real-time interaction performance. These limitations hinder virtual humans from achieving ideal emotional transmission and interactive experiences. Existing approaches predominantly focus on geometric modelling or expression-driven methods, lacking systematic integration of emotional information. This makes it difficult to fully resolve the conflict between emotional expression and animation naturalness. To address this issue, this study proposes a 3D facial modelling and animation generation method incorporating emotional recognition. By driving animation generation with high-precision emotional recognition, the method achieves natural facial expressions and emotional accuracy while optimising real-time performance. The innovation lies in closely integrating emotional recognition with 3D modelling and animation generation to form a systematic framework. This enhances the emotional expressiveness and interactive experience of

animations technically and enriches the research paradigm of facial animation generation theoretically. Compared to existing methods, this study demonstrates significant advantages in emotional expression accuracy, animation naturalness, and real-time interaction performance. This provides a feasible technical support for applications such as virtual reality, online education, digital entertainment, and intelligent interaction.

2 Related work

In recent years, 3D facial modelling and animation generation have emerged as critical technologies in computer graphics and intelligent interaction, gradually becoming the core enabling methods for virtual reality, digital entertainment, and intelligent human-computer interaction applications. Early studies primarily focused on geometric modelling and motion capture. For example, Pourebadi et al. (2022) proposed the 3D morphable model (3DMM), which enabled a parametric description of facial structures and allowed flexible reconstruction with only a small number of parameters. Kopalidis et al. (2024) improved the efficiency and feasibility of animation driving through real-time facial capture methods. With the rise of deep learning, international scholars further explored single-image 3D reconstruction and animation generation methods based on CNNs and generative adversarial networks (GANs). Javanmardi et al. (2024) introduced an end-to-end reconstruction framework that significantly improved modelling accuracy. Building on psychological foundations, Mejia-Escobar et al. (2023) applied the basic emotion theory to provide a basis for facial expression classification. Subsequently, Esmaeili and Kiani (2024) constructed large-scale emotion recognition models using deep networks, advancing the automation of affective computing. Nevertheless, although breakthroughs have been made in both modelling and recognition, the two dimensions have often remained independent. This lack of organic integration in 3D animation generation results in facial animations that appear realistic in appearance but fail to effectively convey complex emotional states.

In China, researchers have also actively promoted the integration of 3D facial modelling and emotion recognition. Zhang and Qian (2025) proposed an expression modelling method based on facial action units, which captured emotional variations in detail but showed limitations in achieving smooth dynamic transitions. Wang et al. (2021) combined deep neural networks with geometric priors to reconstruct 3D facial expressions in a low-dimensional parameter space, thereby improving model stability and generalisability. Another study by Zhang and Pu (2024) attempted to directly couple emotion recognition results with animation-driving parameters to enhance the naturalness and expressiveness of virtual human animations. In addition, some scholars have adopted multimodal fusion approaches. For example, Diao et al. (2024) integrated speech and facial expression features to improve the robustness of emotion recognition, providing richer input for animation generation. Despite these efforts, two major challenges remain in domestic studies. First, most work continues to prioritise geometric modelling and motion-driven techniques, while treating emotion as an auxiliary factor rather than embedding it deeply into the modelling and animation process (Xu et al., 2022). Second,

existing fusion methods are often restricted to single modalities or limited parameter mappings, lacking the ability to dynamically model complex expressions and affective states. As a result, while realism in facial animation has improved, gaps remain in naturalness, subtlety, and interactivity (Zeng et al., 2022). Table 1 is the comparison of research literature on 3D facial modelling, animation, and emotion recognition.

Globally, extensive progress has been achieved in 3D facial modelling, animation generation, and emotion recognition. On the modelling side, both geometric methods and deep learning approaches have advanced accuracy and efficiency. On the recognition side, multimodal fusion and deep neural networks have greatly enhanced classification and prediction performance. However, a significant disconnect still persists in the full pipeline linking 'geometric modelling – emotion recognition – animation generation'. Most studies achieve facial shape reconstruction while overlooking the naturalness and dynamics of emotional expression, whereas stand-alone emotion recognition methods struggle to form effective mappings with 3D animation generation. Therefore, future research trends should focus on developing unified frameworks that deeply integrate emotion recognition with 3D modelling, achieving organic coupling from expression recognition to dynamic animation generation. Against this background, this study proposed a 3D facial modelling and animation generation method incorporating emotion recognition. The approach aimed to bridge this gap by maintaining modelling accuracy while substantially enhancing the emotional expressiveness and interactive naturalness of virtual facial animation, thereby providing more reliable and efficient technological support for applications in virtual reality, digital humans, and intelligent interaction.

3 Model for 3D facial image modelling and animation generation with emotion recognition

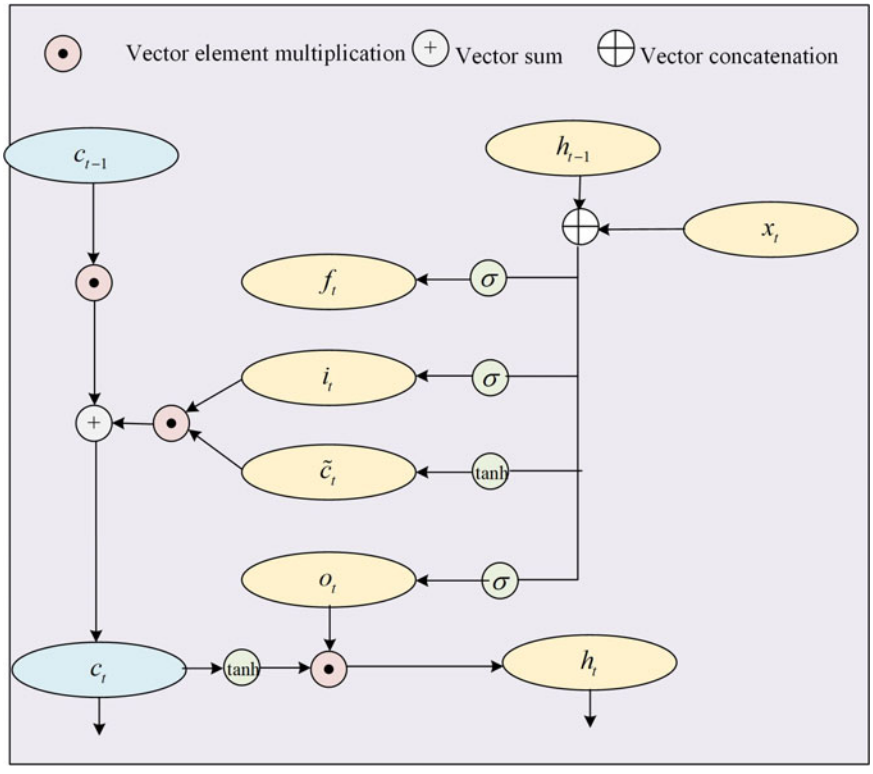
3.1 Emotion recognition and the LSTM algorithm

Emotion recognition is an important research direction in the field of artificial intelligence and human-computer interaction. Its core goal is to automatically determine an individual's emotional state by analysing multimodal information such as facial expressions, speech, body movements, and physiological signals (Khare et al., 2024; Lei and Kim, 2021). In recent years, the introduction of deep learning has significantly promoted the development of this field. Among them, long short-term memory (LSTM) is widely used in speech and video expression emotion recognition due to its gate mechanism (Guo et al., 2023). It can effectively process time series data, overcome the gradient vanishing and explosion problems of RNN, and capture dynamic information such as pitch, energy, and expression changes by modelling temporal features, improving the accuracy and robustness of recognition. This type of method has demonstrated significant value in applications like virtual reality, intelligent customer service, mental health assessment, educational counselling, and intelligent driving, providing strong support for achieving more natural human-computer interaction (Sun et al., 2024). The structure of LSTM is illustrated in Figure 1.

Table 1 Comparison of research literature on 3D facial modelling, animation, and emotion recognition

Source of literature	*Research method*	*Main contributions*	*Advantages*	*Disadvantages*
Pourebadi et al. (2022)	3DMM	Provide parameterised facial descriptions and support flexible reconstruction	Parameterisation and flexible reconstruction	Not combined with emotion recognition
Kopalidis et al. (2024)	Real-time facial capture	Verify the efficiency and feasibility of animation drivers, and promote the implementation of technology	High driving efficiency and applicability	No emotional information transmission
Javanmardi et al. (2024)	Convolution + GANs end to end framework	Improve modelling accuracy and promote deep learning applications	High precision, end-to-end process	No emotion recognition and lack of emotional expression
Mejia-Escobar et al. (2023)	Based on basic emotion theory	Provide a psychological foundation for facial expression classification to support research on emotion recognition	Establish theoretical foundation and enhance scientificity	Only theoretical, without actual modelling/animation generation
Esmaeili and Kiani (2024)	Deep network large-scale facial expression recognition model	Promote recognition automation and enhance scale and efficiency	Automated identification and large-scale processing	Not integrated with modelling/ animation and unable to serve emotional expression
Zhang and Qian (2025)	Facial expression modelling based on facial action units	Elaborate on emotional changes and strengthen the connection with emotions	Capture emotions in detail and depict accurately	Smooth transition of dynamic expressions
Wang et al. (2021)	Deep neural networks + geometric prior for facial expression reconstruction	Improve model stability/generalisation and optimise reconstruction performance	Strong stability and generalisation, and good reconstruction	Lack of deep emotional integration and limited ability to express oneself
Zhang and Pu (2024)	Coupling of emotion recognition results with animation parameters	Enhance the naturalness and emotional expression of virtual human animation	Establish emotional animation connection and strong expression	Direct coupling and no system framework
Diao et al. (2024)	Multi-modal fusion of voice and facial expressions	Improve the robustness of emotion recognition and provide rich driving information	Robust identification and abundant driving information	Not deeply embedded in the process and insufficient fusion depth

Figure 1 LSTM structure (see online version for colours)



In Figure 1, at time step t, the inputs to the neuron consist of three components: c_{t-1}, h_{t-1}, and x_t represent the outputs from the previous time step, x_t also denotes the external sequence input at time t. C refers to the control unit, and h denotes the hidden layer. After undergoing a series of computational operations, these three inputs produce two outputs, c_t and h_t. The computation process is as equations (1)–(4):

$$i = \sigma\left(W_i x_t + U_i h_{t-1} + b_i\right) \tag{1}$$

$$f_t = \sigma\left(W_f x_t + U_f h_{t-1} + b_f\right) \tag{2}$$

$$o_t = \sigma\left(W_o x_t + U_o h_{t-1} + b_o\right) \tag{3}$$

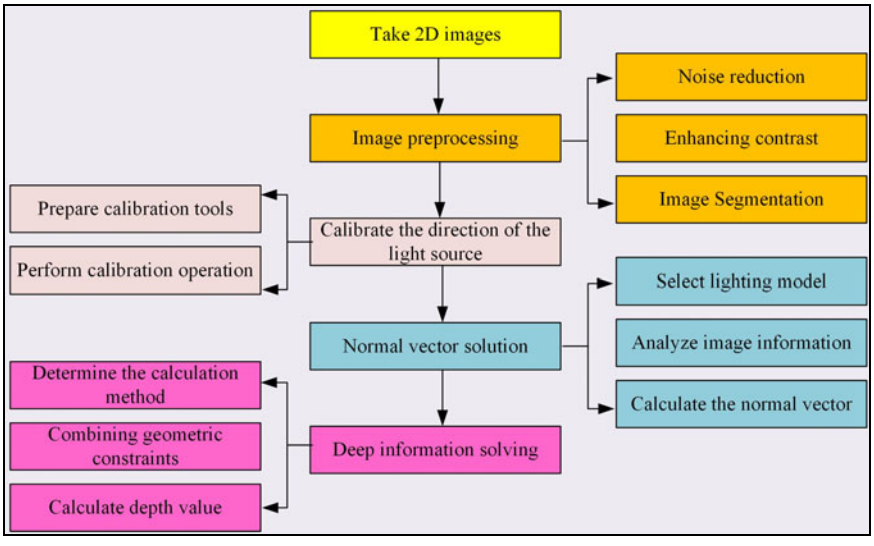
$$\tilde{c}_t = \tanh\left(W_c x_t + U_c h_{t-1} + b_c\right) \tag{4}$$

3.2 3D facial image modelling with integrated emotion recognition

In the study of 3D facial modelling and animation generation, incorporating emotion recognition enhances the realism of the model and improves the naturalness of human-computer interaction. The workflow of the 3D facial image algorithm begins with capturing 2D images. These images are pre-processed through denoising, contrast enhancement, and segmentation. Calibration tools are then prepared to determine the light source direction. Subsequently, an illumination model is selected, image

information is analysed, and normal vectors are computed. Finally, depth information is obtained by applying geometric constraints and calculating depth values (Zhang et al., 2023). The workflow of the 3D facial image algorithm is illustrated in Figure 2.

Figure 2 3D facial image algorithm flowchart (see online version for colours)



3.2.1 Light source direction determination

In the process of 3D facial modelling, determining the direction of the light source is the basis for subsequent normal vector and depth reconstruction. This study uses a forward projection camera model to determine the direction of the light source through a combination of experimental calibration and geometric calculation. At the hardware level, a fixed position point light source or LED ring array can be used, and the relative position between the camera and the light source can be measured on the acquisition device through a calibration board or a small black ball reflection point. The system uses the reflection law to derive the direction vector of the light source based on the relationship between the camera angle and the position of the reflection point, achieving accurate calibration.

In the acquisition system designed for this study, the camera model follows an orthographic projection. Let p represent the highlight point on a small black sphere, L the unit vector of the light source direction, R the unit vector of the reflection direction, V the unit vector of the camera view direction, and N_p the unit normal vector at the highlight point. When the reflection direction coincides with the camera view direction ($R = V$), the highlight point becomes visible in the field of view. At this moment, the reflection direction vector is a unit vector, $R = [0, 0, 1]$. Since the light source direction L and the reflection direction R are symmetric with respect to the normal vector N_P, the light source direction can be determined using equation (5):

$$R \cdot N_p = |R| \cdot |N_p| \cdot \cos\alpha \tag{5}$$

Given that L, R, and N_p are all unit vectors, as shown in equation (6):

$$L = 2 \cdot N_p \cdot R \cdot N_p - R \tag{6}$$

$R = V$ with the camera view direction V known, it is sufficient to compute the unit normal vector N_p at the highlight point to determine the light source direction L.

3.2.2 Normal vector calculation

After determining the direction of the light source, the surface normal vector needs to be calculated based on the pixel brightness information. This study uses the illumination reflection model to establish the relationship between brightness and normal vector, and converts RGB images to HSI colour space to extract brightness component I. Based on the Lambert reflection model, the study converts the dot product relationship between brightness and light source direction into a linear equation system. For multi-light source conditions, the least squares method can be used to solve the unit normal vector and reflectance of each pixel. Under single light source conditions, the shape-from-shading algorithm is used to obtain the surface normal direction through nonlinear optimisation iteration. To eliminate noise and mirror interference, the system performs brightness normalisation, shadow area removal, and mirror highlight separation before calculation, and improves the stability and continuity of the normal solution through bilateral filtering or robust estimation.

To calculate the unit normal vector of the image, in addition to the light source direction, pixel intensity values are required. In the hue, saturation, intensity (HSI) colour space, I represents the intensity of a pixel. Therefore, in this study, the red, green, blue (RGB) colour space was converted into the HSI colour space. The process of conversion from RGB to HSI is shown in equation (7).

$$\begin{cases} H = \arccos\left\{ \dfrac{[(R-G)+(R-B)/2]}{\sqrt{(R-G)^2+(R-B)(G-B)}} \right. \\ S = 1 - \dfrac{3}{(R+G+B)}[\min(R, G, B)] \\ I = \dfrac{(R+G+B)}{3} \end{cases} \tag{7}$$

3.2.3 Depth value calculation of the facial surface

After obtaining the pixel normal vector field, the depth values of each pixel can be calculated based on the geometric relationship between the normal direction and the depth gradient. This study converts the normal component into an image gradient and uses an integration method based on Poisson reconstruction for global depth restoration. By constructing a sparse linear equation system and solving it with boundary conditions, a continuous and smooth 3D facial surface model can be obtained. To enhance the accuracy and stability of the model, the system performs edge preserving smoothing on the normal vector field before integration, and adds regularisation constraints during solution to suppress error accumulation. Finally, by matching the reconstructed depth data with the 2D image coordinates, a 3D face model with high geometric accuracy and natural details can be generated. This provides a reliable foundation for subsequent expression modelling and animation generation.

Assume that a point $(x, y, z_{(x,y)})$ on the facial expression image is given, along with its neighbouring points $(x + 1, y, z_{(x+1,y)})$ and $(x, y + 1, z_{(x,y+1)})$. Let N denote the unit normal vector at this surface point. For an image of size $m \times n$, a sparse matrix M of size $(2 \times m \times n) \times (m \times n)$ can be obtained, composed of N_z. At the same time, a vector V of size $(2 \times m \times n) \times 1$ can be derived, composed of N_x and N_y. Based on the constraint relationship, the equations (8) and (9) assumption is made:

$$MZ = V \tag{8}$$

$$M^T MZ = M^T V \tag{9}$$

Assume $M^T M = A$ and $M^T V = B$, therefore, as shown in equation (10):

$$AZ = b \tag{10}$$

By solving for Z, the depth value of each point on the facial surface can be obtained.

In the process of 3D facial expression recognition, the accuracy of feature extraction directly affects the reliability of emotion classification. To fully characterise the geometric and texture features of the face during expression changes, this study proposes a hybrid feature extraction method based on keypoint localisation. Firstly, a 3D facial keypoint detection algorithm is used to automatically extract 20 main feature points of the face. They include the tip of the nose, the two ends of the nose, the upper and lower ends of the lips, the corners of the mouth, eight feature points of the eyes (two pairs of upper and lower, left and right corner points each), the midpoint and two end points of the eyebrows, and the chin point. These key points cover the most significant areas of facial expression changes, such as the muscles around the mouth, eyes, and eyebrows, which are the main driving areas for emotional changes.

Based on the above 20 feature points, this study extracts 23 geometric distance features to quantify the degree of deformation of key facial muscle groups. Among them, the distance between eyebrows and the height of the eyebrow arch reflect the contraction of the frontal and frowning muscles, which can be used to identify expressions such as anger and surprise. The distance between the inner and outer corners of the eyes, as well as the distance between the upper and lower eyelids, reflects the movement of the orbicularis oculi muscle and is related to emotions such as squinting and opening the eyes. The distance changes from the tip of the nose to the corners of the mouth and from the nose to the corners of the mouth reflect the activity of the zygomatic muscle and the levator laboris muscle, which are related to expressions such as smiling and disgust. The distance between the corners of the mouth, the distance from the upper lip to the lower lip, and the distance from the corners of the mouth to the chin reflect the movement status of the orbicularis oris muscle and lower jaw, and can effectively distinguish different emotions such as happiness, sadness, and surprise. In addition, cross regional distance features such as nose tip to chin, eye corner to mouth corner, and eyebrow to mouth corner are calculated to reflect changes in overall facial expression intensity. All distances are calculated using Euclidean distance in three-dimensional space to ensure minimal impact of pose changes on feature values.

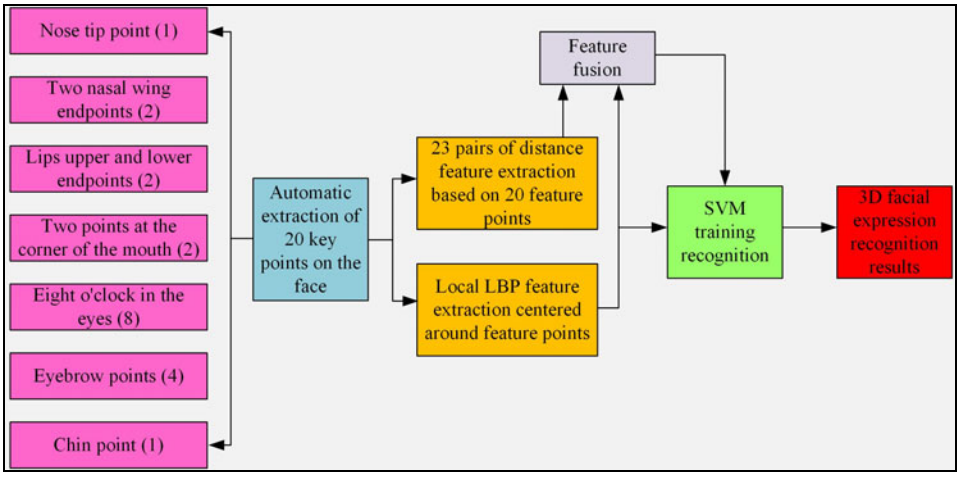
Based on these 20 landmarks, two types of features were extracted:

1 twenty-three distance features between landmark pairs

2 local binary pattern (LBP) features centred on the landmarks.

These features were then fused and classified using a support vector machine (SVM). The final output produced the 3D facial expression recognition results. The workflow of the emotion recognition-based 3D facial image recognition process is shown in Figure 3.

Figure 3 Workflow of 3D facial image recognition based on emotion recognition (see online version for colours)



3.3 Emotion recognition-integrated animation generation method

In research on 3D facial animation generation, effectively integrating emotion recognition results into the modelling and rendering pipeline is critical to enhancing the naturalness and interactivity of animations (Debnath et al., 2022). The proposed algorithm for 3D facial animation generation based on emotion recognition first takes a facial image as input. The facial expression is classified (e.g., smile, sadness, surprise, blinking), and the corresponding emotion category is obtained. Next, the image undergoes deformation pre-processing and triangulation to establish the geometric structure of the face. After triangulation, a coordinate transformation is applied to the pixels within each triangle. First, an affine transformation mapping is constructed for each triangle using the coordinates of its three vertices. Then, all internal pixels are transformed according to the mapping of the vertices. The initial step in coordinate transformation is to establish a mapping relationship, namely, the feature mapping function between the original image and the target deformed image, as shown in equation (11):

$$\begin{pmatrix} u \\ v \\ l \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} x \\ y \\ l \end{pmatrix} \tag{11}$$

x, y denote the coordinates of any pixel in the source triangle, u, v denote the coordinates of the corresponding pixel in the target triangle, and a_{ij}, $i = 1, 2$ and $j = 1, 2, 3$ are the undetermined coefficients in the coefficient matrix M. Thus, M is defined as equation (12):

$$M = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix} \tag{12}$$

Given the coordinates of the source triangle vertices (x_1, y_1), (x_2, y_2), and (x_3, y_3) and the target triangle vertices (u_1, v_1), (u_2, v_2), and (u_3, v_3), matrix M can be determined, as shown in equation (13).

$$M = \begin{pmatrix} u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{pmatrix} \bullet \begin{pmatrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \\ 1 & 1 & 1 \end{pmatrix} \tag{13}$$

The grey value of a mapped point is calculated using the neighbourhood mean interpolation algorithm expressed in equation (14). Specifically, the grey value of pixel x, y is the average of its N nearest integer-coordinate neighbouring pixels (x_i, y_i), $i = 1, \ldots, N$, as shown in equation (14).

$$gray(x, y) = \frac{1}{N} \sum_{i=1}^{N} (x_i, y_i) \tag{14}$$

This algorithm not only preserves the accuracy of geometric modelling but also authentically drives the recognised emotional features into the animation, thereby enhancing the realism of virtual facial expressions and improving the interactive experience. The framework of the facial animation generation algorithm based on emotion recognition is shown in Figure 4.

4 Experimental design and result analysis of 3D facial image modelling and animation generation method integrating emotion recognition

To validate the effectiveness of the proposed algorithm, the publicly available BU-3DFE 3D facial database was selected as the experimental dataset. This database contains 3D expression data from 100 participants, including 44 males and 56 females. During data collection, participants were instructed to display seven prototypical expressions: neutral (NE), anger (AN), disgust (DI), fear (FE), happiness (HA), sadness (SA), and surprise (SU). Except for the neutral expression, which had only one intensity level, the other six emotions were divided into four levels of intensity: low, moderate, high, and peak. Consequently, each subject contributed 25 3D facial models, providing rich and diverse data for experimentation. In the experimental design, only the 3D facial models with the highest intensity levels were used. To ensure representativeness and reduce potential bias from gender or individual differences, 60 models were randomly selected from the 100 available samples, with neutral expressions excluded.

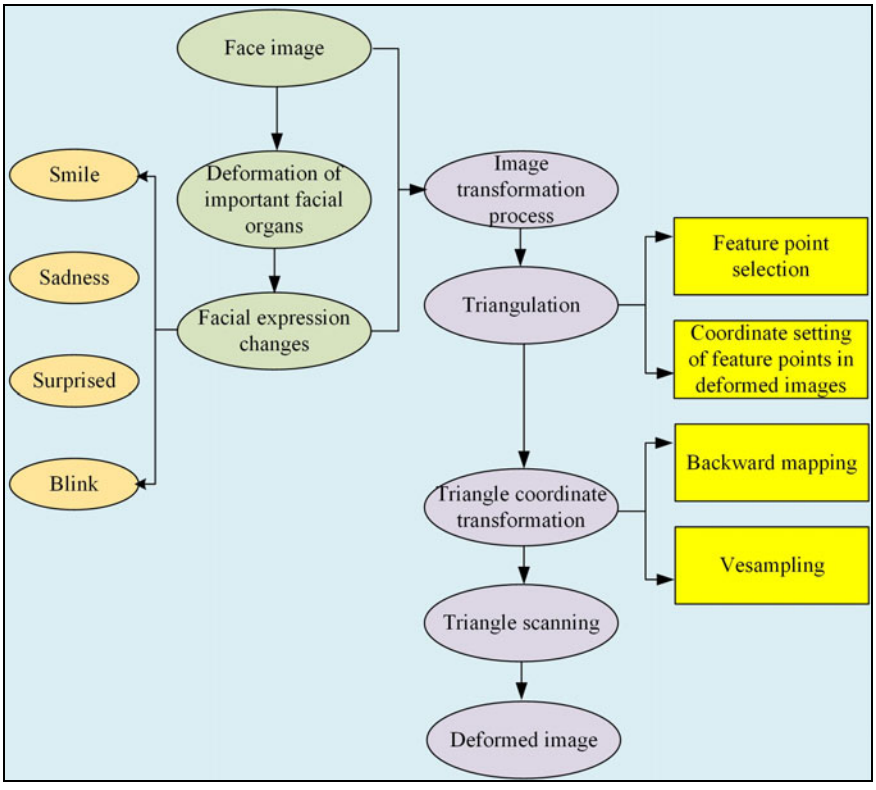
The application experiment of 3D facial image modelling with emotion recognition aimed to explore its effectiveness in emotion classification. The SVM algorithm was used to compare different types of features. Two feature sets were selected:

1 distance features between facial landmarks, since expression changes cause significant geometric variation in key points, directly reflecting facial movement patterns

2 combined features that integrate distance features with local LBP features extracted around the landmarks.

The fused features capture local texture information surrounding the landmarks, reflecting subtle muscle movements. By analysing and integrating these two types of features, the experiment performed emotion recognition on 3D facial images and evaluated recognition accuracy to validate the effectiveness of the proposed modelling method. The experimental results are presented in Figures 5 and 6.

Figure 4 Framework of facial animation generation algorithm based on emotion recognition (see online version for colours)



As shown in Figure 5, the overall accuracy of 3D facial emotion recognition based on distance features was relatively high, effectively reflecting geometric variations of the face under different emotional states. Among them, happiness (HA) and surprise (SU) achieved the highest recognition accuracies, reaching 91.7% and 96.5%, respectively, with almost no confusion with other emotions. This indicates that the geometric features of these two emotions are particularly distinct. The recognition accuracies of anger (AN), disgust (DI), fear (FE), and sadness (SA) also exceeded 80%, with an average accuracy of 89.1%. Overall, this method performed well for emotions with high discriminability, although certain misclassifications occurred among emotions with relatively similar facial expressions.

Figure 5 Accuracy of 3D facial emotion recognition based on distance features (unit: %) (see online version for colours)

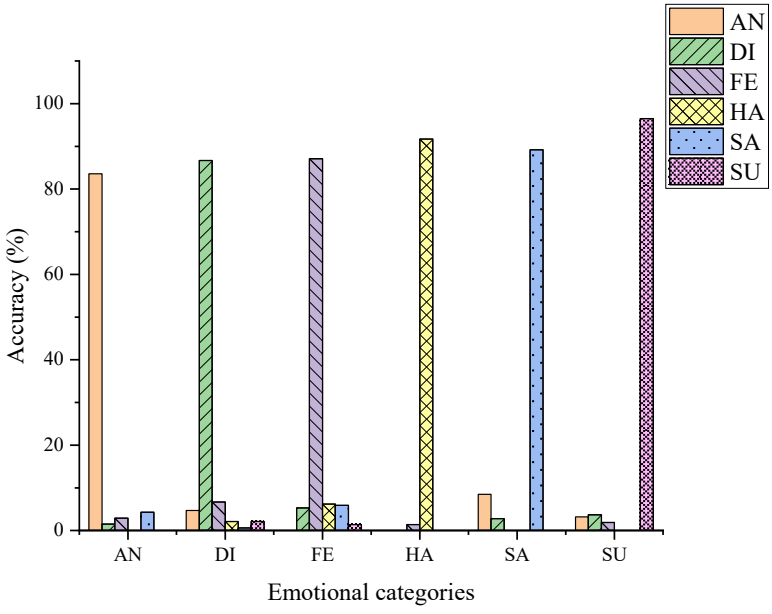
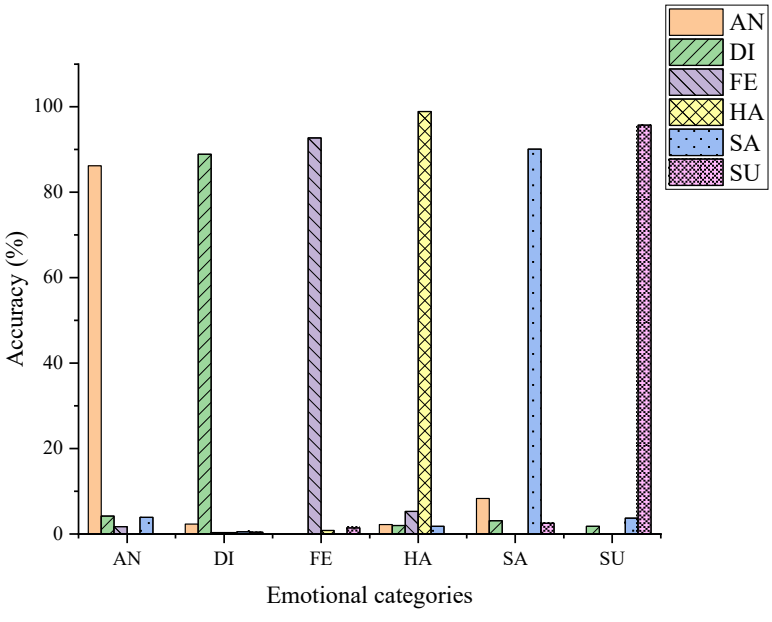


Figure 6 Accuracy of 3D facial emotion recognition based on fused features (unit: %) (see online version for colours)

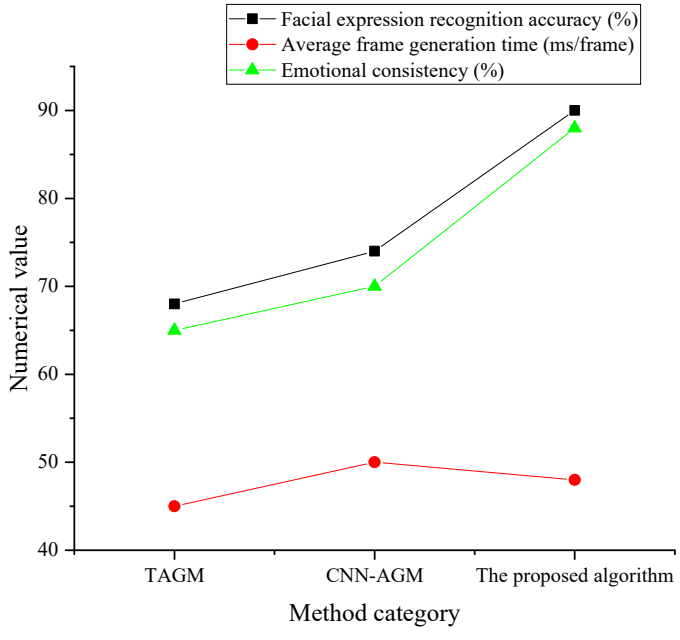


As illustrated in Figure 6, recognition accuracies based on fused features showed some variation across emotions but maintained a high overall level. Happiness achieved the

highest accuracy at 98.9%, followed by surprise (95.7%), fear (92.7%), sadness (90.1%), and disgust (88.9%). Anger had the lowest accuracy at 86.2%, though still within an acceptable range. Except for a few cases, misclassification rates among different emotions remained generally low, and the overall average accuracy reached 92.1%. These results demonstrate that the proposed model effectively captured the feature differences of various 3D facial emotions, ensuring high recognition reliability and stability. Moreover, the results confirmed that combining geometric distance features with local texture patterns substantially enhanced robustness, allowing the model to handle subtle and fine-grained emotional variations that often challenge single-feature approaches. This improvement is particularly significant for complex or ambiguous expressions, where the interaction of multiple cues provides more discriminative power.

In order to further verify the effectiveness of the animation generation method driven by emotion recognition, this experiment compares it with two baseline models: the traditional animation generation method without emotion recognition (TAGM) and the animation generation method based on CNN-AGM. The evaluation takes into account key indicators such as expression recognition accuracy, frame generation efficiency and emotional consistency, and comprehensively evaluates the performance of different models in animation quality and emotional expression ability. By directly integrating emotion recognition into the generation process, the framework not only improves the fidelity of animated facial expressions, but also achieves stronger consistency between emotion recognition and rendering output, highlighting its advantages in practical applications.

Figure 7 Performance comparison between the proposed emotion recognition–based animation generation model and traditional methods (see online version for colours)



As shown in Figure 7, the proposed animation generation model integrating emotion recognition outperformed the traditional methods across all metrics. Specifically, in expression recognition accuracy, the proposed method achieved 90%, significantly higher than TAGM (68%) and CNN-AGM (74%). This demonstrates that the emotion recognition module substantially enhanced the accuracy of facial expression representation in animations, particularly for subtle or complex emotional cues that are typically challenging for conventional models. In terms of average frame generation time, the proposed method required 48 ms/frame, slightly higher than TAGM (45 ms/frame) but lower than CNN-AGM (50 ms/frame), indicating that efficiency was maintained while improving performance. Moreover, the model exhibited stable computational behaviour across diverse animation sequences, ensuring consistent real-time performance. For emotional consistency, the proposed method reached 88%, a considerable improvement compared to TAGM (65%) and CNN-AGM (70%). This further confirmed the significant role of emotion recognition in enhancing the consistency of emotional expression in animations and in preserving the coherence between predicted emotions and rendered outputs. Overall, the proposed method achieved a favourable balance among accuracy, emotional expressiveness, and generation efficiency, demonstrating its practical applicability for high-fidelity 3D facial animation in virtual reality, digital entertainment, and interactive human–computer applications.

5 Discussion

The experimental results showed that the recognition accuracy of the model on highly distinguishable emotions such as ‘happy’ and ‘surprised’ was significantly higher than other emotions. This was closely related to the obvious and consistent facial geometric changes corresponding to these expressions, such as the upward movement of the mouth corners and the widening of the eyes, which were easy to capture through keypoint distance. On the contrary, emotions like anger, disgust, fear, and terror have subtle changes in the eyebrows, corners of the mouth, and around the eyes, and there are some similar features between different emotions. Therefore, even if the overall recognition rate exceeds 80%, there is still a certain degree of misclassification phenomenon. After integrating texture features, the overall accuracy of various emotion recognition increased to 92.1%. This indicates that local muscle movement and facial texture information can effectively supplement the shortcomings of geometric features and enhance the model’s perception ability of subtle expressions. In animation generation, the introduction of emotion recognition module significantly improves the accuracy and consistency of facial expressions, and ensures that the average frame generation time is within an acceptable range, achieving a balance between naturalness and efficiency. However, the model still has limitations in capturing micro expressions and adapting to individual differences, and its robustness and generative performance may be affected under complex lighting, non-standard poses, or extreme emotional states. Therefore, although this method performs outstandingly in high discriminative emotion and animation consistency, there is still room for improvement in micro expression recognition, complex scene adaptation, and computational efficiency optimisation.

6 Conclusions

The proposed 3D facial modelling and animation generation method, enhanced with emotion recognition, effectively captured facial features under different emotional states and applied them to animation generation, thereby achieving high-precision expression rendering and emotional consistency. Experimental results demonstrated that distance feature-based 3D facial emotion recognition achieved high accuracy for highly distinguishable emotions such as happiness and surprise. Furthermore, by fusing multiple features, the overall average accuracy increased to 92.1%, indicating that the model reliably reflected both geometric and textural variations of the face under diverse emotional states. In terms of animation generation, the emotion recognition–integrated model outperformed both traditional methods and CNN-based approaches across key evaluation metrics, including expression recognition accuracy, emotional consistency, and generation efficiency. Specifically, the proposed method achieved 90% accuracy in expression recognition, 88% in emotional consistency, and an average frame generation time of 48 ms/frame, striking a favourable balance among accuracy, naturalness, and efficiency. These findings confirm that embedding an emotion recognition module into 3D facial modelling and animation generation significantly enhances the emotional expressiveness of animations, producing facial expressions that are both realistic and natural while maintaining strong computational performance. In practical applications, this method has broad potential value: in the fields of virtual anchors and digital entertainment, it can generate emotional character expressions, enhance audience immersion and interactive experience. In psychological health assessment and emotion monitoring, accurately capturing micro expression changes can assist doctors or psychological counsellors in judging individual emotional states, supporting emotional intervention and psychological therapy. In educational training and remote teaching, virtual teachers or peer roles with emotional expression abilities can enhance learners' sense of participation and learning motivation. In medical rehabilitation scenarios, emotionally consistent virtual humans can be used for facial motor function training, language therapy, and patient psychological counselling. In the interaction between games and virtual reality, vivid and natural virtual characters can enhance players' immersion and interactive experience. In social robots and intelligent assistants, facial expressions driven by emotion recognition can improve the quality of communication between humans and robots, enabling virtual assistants to have higher affinity and emotional feedback capabilities. In addition, this method also has practical value in remote collaboration, meetings, and virtual meeting systems, as it can accurately convey participants' emotional states through virtual avatars, thereby improving communication efficiency and team collaboration effectiveness. The multi feature fusion 3D facial emotion recognition and animation generation method proposed here significantly improves the accuracy and stability of emotion recognition, and effectively ensures the naturalness and consistency of expression animations, while performing well in computational efficiency. This method provides a theoretical and technical foundation for future applications like multimodal virtual human interaction, high fidelity animation generation, mental health analysis, education and training, medical rehabilitation, and intelligent assistants. It also provides a systematic practical paradigm for the combination of 3D facial emotion recognition and animation generation.

Despite these advancements, challenges remain due to the inherent complexity and diversity of emotional expression. On the one hand, the model still faced limitations in distinguishing emotions with high similarity, such as anger, disgust, and fear. On the other hand, its generalisation and adaptability need further validation when applied to complex expressions, extreme head poses, or large-scale, heterogeneous datasets. Additionally, the current method primarily relied on 3D geometric features and limited texture information, without fully leveraging multimodal cues such as speech, semantics, and gestures, which constrained performance in multimodal virtual human interaction and high-complexity scenarios. Future research could therefore proceed along two directions: enhancing model generalisation and expanding multimodal information integration. First, incorporating richer geometric and textural features alongside speech and semantic cues could improve accuracy and robustness, particularly for highly similar emotions. Second, optimising the animation generation algorithm could further improve efficiency and stability when handling complex expressions, extreme poses, and real-time interactive scenarios. By addressing these challenges, the proposed framework may be extended to more demanding applications in immersive digital entertainment, high-complexity virtual interactions, as well as education and psychological behaviour analysis, thereby enhancing its applicability and practical value.

Declarations

The data used to support the findings of this study are all in the manuscript.

The authors declare no competing interests.

References

Chen, X. and Chen, H. (2023) 'Emotion recognition using facial expressions in an immersive virtual reality application', *Virtual Reality*, Vol. 27, No. 3, pp.1717–1732.

Debnath, T., Reza, M.M., Rahman, A. et al. (2022) 'Four-layer ConvNet to facial emotion recognition with minimal epochs and the significance of data diversity', *Scientific Reports*, Vol. 12, No. 1, p.6991.

Di, C., Peng, J., Di, Y. et al. (2021) '3D face modeling algorithm for film and television animation based on lightweight convolutional neural network', *Complexity*, Vol. 2021, No. 1, p.6752120.

Diao, H., Jiang, X., Fan, Y. et al. (2024) '3D face reconstruction based on a single image: a review', *IEEE Access*, Vol. 12, pp.59450–59473.

Esmaeili, M. and Kiani, K. (2024) 'Generating personalized facial emotions using emotional EEG signals and conditional generative adversarial networks', *Multimedia Tools and Applications*, Vol. 83, No. 12, pp.36013–36038.

Guo, M., Xu, F., Wang, S. et al. (2023) 'Synthesis, style editing, and animation of 3D cartoon face', *Tsinghua Science and Technology*, Vol. 29, No. 2, pp.506–516.

Javanmardi, A., Pagani, A. and Stricker, D. (2024) *G3FA: Geometry-Guided GAN for Face Animation*, arXiv preprint arXiv:2408.13049.

Jiang, D., Chang, J., You, L. et al. (2024) 'Audio-driven facial animation with deep learning: a survey', *Information*, Vol. 15, No. 11, p.675.

Khare, S.K., Blanes-Vidal, V., Nadimi, E.S. et al. (2024) 'Emotion recognition and artificial intelligence: a systematic review (2014–2023) and research recommendations', *Information Fusion*, February, Vol. 102, p.102019.

Kopalidis, T., Solachidis, V., Vretos, N. et al. (2024) 'Advances in facial expression recognition: a survey of methods, benchmarks, models, and datasets', *Information*, Vol. 15, No. 3, p.135.

Lei, D. and Kim, S.H. (2021) 'Design of 3D modeling face image library in multimedia film and television', *Journal of Sensors*, Vol. 2021, No. 1, p.2221893.

Li, X., Zhang, J. and Liu, Y. (2022) 'Speech driven facial animation generation based on GAN', *Displays*, September, Vol. 74, p.102260.

Mejia-Escobar, C., Cazorla, M. and Martinez-Martin, E. (2023) 'Improving facial expression recognition through data preparation and merging', *IEEE Access*, Vol. 11, pp.71339–71360.

Pourebadi, M. and Riek, L.D. (2022) 'Facial expression modeling and synthesis for patient simulator systems: past, present, and future', *ACM Transactions on Computing for Healthcare (HEALTH)*, Vol. 3, No. 2, pp.1–32.

Song, H. and Kwon, B. (2024) 'Facial animation strategies for improved emotional expression in virtual reality', *Electronics*, Vol. 13, No. 13, p.2601.

Sun, Y., Chu, W., Zhou, H. et al. (2024) 'Avi-talking: learning audio-visual instructions for expressive 3D talking face generation', *IEEE Access*, Vol. 12, pp.57288–57301.

Tu, X., Zou, Y., Zhao, J. et al. (2021) 'Image-to-video generation via 3D facial dynamics', *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 32, No. 4, pp.1805–1819.

Wang, B. and Shi, Y. (2023) 'Expression dynamic capture and 3D animation generation method based on deep learning', *Neural Computing and Applications*, Vol. 35, No. 12, pp.8797–8808.

Wang, X., Wang, Y., Li, W. et al. (2021) 'Facial expression animation by landmark guided residual module', *IEEE Transactions on Affective Computing*, Vol. 14, No. 2, pp.878–894.

Wu, Y., Deng, Y., Yang, J. et al. (2022) 'AnifaceGAN: animatable 3D-aware face image generation for video avatars', *Advances in Neural Information Processing Systems*, Vol. 35, pp.36188–36201.

Xu, P., Zhu, Y. and Cai, S. (2022) 'Innovative research on the visual performance of image two-dimensional animation film based on deep neural network', *Neural Computing and Applications*, Vol. 34, No. 4, pp.2719–2728.

Zeng, B., Liu, B., Li, H. et al. (2022) 'FNeVR: neural volume rendering for face animation', *Advances in Neural Information Processing Systems*, Vol. 35, pp.22451–22462.

Zhang, N. and Pu, B. (2024) 'Film and television animation production technology based on expression transfer and virtual digital human', *Scalable Computing: Practice and Experience*, Vol. 25, No. 6, pp.5560–5567.

Zhang, Y. and Qian, J. (2025) 'Machine learning-based expression generation technology for virtual characters in film and television art', *International Journal of Computational Intelligence Systems*, Vol. 18, No. 1, p.219.

Zhang, Y., Xu, X., Zhao, Y. et al. (2023) 'Facial prior guided micro-expression generation', *IEEE Transactions on Image Processing*, Vol. 33, pp.525–540.