DCNN-FLAME: A Dual-Supervised Style Transfer-Based Method for 3D Animated Character Expression Reconstruction

Xiaowen Guo
College of Art and Creativity, Anhui University of Applied Technology, Anhui 230000, China
Email of Corresponding Author: Xiaowen_Guo@outlook.com

Keywords: image, feature enhancement, animated characters, expression generation, three-dimensional face reconstruction

Received: October 14, 2025

This study aims to explore a method capable of generating high-naturalness expressions for animated characters, thereby enhancing the audience's emotional resonance. This study proposes a 3D face reconstruction system named Deep Convolutional Neural Network-Faces Learned with an Articulated Model and Expressions (DCNN-FLAME). DCNN-FLAME consists of an identity encoder, a mapping network, a Facial Landmark Embedding Model (FLAME) geometric decoding module, and a detail reconstruction network, forming an end-to-end processing pipeline from input images to 3D meshes and appearance textures. A style transfer module is constructed based on Deep Convolutional Neural Network (DCNN). It uses a pre-trained convolutional network to achieve effective separation of content and style, providing high-level semantic constraints for texture detail modeling of animated character expressions. On this basis, a dual-branch feature supervision mechanism composed of expression classification features and facial Action Units (AU) detection features is designed. Expression classification features provide global emotional semantic constraints to ensure macro-expression consistency. AU detection features guide local muscle movements from an anatomical perspective to enhance the realism of expression details. Experiments are conducted based on the large-scale face dataset VGGFace2. Systematic comparisons are performed with four 3D face reconstruction algorithms: 3D Morphable Model Fitting (3DMM-Fitting), RingNet, Detailed Expression Capture and Animation (DECA), and Fourier Analysis Networks-3D (FAN-3D). The proposed DCNN-FLAME model achieves a mean value of 1.29 in non-metric evaluation and 1.72 in metric evaluation. Both indicators are lower than those of all baseline methods, demonstrating higher geometric reconstruction accuracy and facial alignment quality. In the overall expression restoration evaluation, the F1 score of the proposed method reaches 0.564, reflecting comprehensive advantages in complex expression modeling. When both the expression classification branch and facial AU detection branch are enabled, the expression classification accuracy rate is 0.571 and the F1 score is 0.563, which are significantly better than the configuration using only a single feature for supervision. This verifies the key role of the dual-branch feature supervision mechanism in improving the naturalness and controllability of animated character expressions. This study provides an effective technical path integrating geometric reconstruction and texture enhancement for 3D animated character expression generation, and also offers new ideas and practical basis for the field of unsupervised 3D face reconstruction.

Povzetek: Študija predstavlja model DCNN-FLAME za 3D rekonstrukcijo obrazov, ki z združevanjem geometrijske rekonstrukcije, prenosa sloga ter dvo-vejne nadzorne mehanike (čustvena klasifikacija in akcijske enote) omogoča bolj naravne, realistične in natančno nadzorovane izraze animiranih likov.

1 Introduction

In the world of animation, a character's expressions are the window to its soul. When a character slightly lifts the corner of its mouth, or a trace of sadness flashes in its eyes, the audience can instantly resonate with it. With the rapid development of digital technology, the production of animated characters has gradually shifted from traditional hand-drawing to digital production [1-3]. However, in this process, the generation of character expressions still faces many challenges. Especially for non-human-shaped characters, how to endow them with rich and reasonable expressions has become an urgent problem for animation creators to solve. In traditional animation production, the generation of expressions often relies on a large amount of manual drawing by artists, which is inefficient and the effect is difficult to guarantee. In 3D animation, the generation of expressions mainly depends on blend shape technology. Although it can realize basic expression changes, there is still much room for improvement in the naturalness and expressiveness of expressions.

Most existing deep learning methods either rely on large amounts of annotated data or still have shortcomings in the naturalness of expressions. For example, face-swapping technology based on 3D models [4, 5] can

realize basic expression transfer, but it often requires complex calculations and a large amount of manual intervention. While methods based on Generative Adversarial Network (GAN) can generate high-definition images, the coherence and naturalness of expressions are difficult to guarantee. Among numerous technologies, style transfer technology has attracted much attention due to its ability to effectively extract and transfer the style features of images. It can retain the content of the source image, and integrate the style features of the target image into it, thereby generating more natural images [6, 7]. This technology has broad application prospects in facial expression generation. Through style transfer, the expression features of real humans can be extracted and then applied to animated characters, making the characters' expressions more vivid and natural.

To address the aforementioned challenges, this study sets the overall goal of improving the realism, stability, and controllability of animated character expressions, and conducts research around a 3D expression generation framework based on style transfer and feature supervision. The research focuses on three interrelated research questions:

1) Can the style transfer loss significantly improve the reconstruction quality of high-frequency detail regions while maintaining the stability of the overall geometric structure, making the reconstruction results closer to the source image in terms of skin texture, light and shadow transitions, and local expression details?

2) Can the dual-branch feature supervision mechanism integrating expression classification and Action Units (AU) features effectively constrain the learning of 3D expression parameters during the training phase, enabling the generated animated expressions to have significant advantages over baseline methods with single-branch supervision or no feature supervision in terms of semantic emotion, consistency, and rationality of muscle movement?

3) Can the Deep Convolutional Neural Network-Faces Learned with an Articulated Model and Expressions (DCNN-FLAME) framework built on Deep Convolutional Neural Network (DCNN) and Facial Landmark Embedding Model (FLAME) achieve or exceed representative methods in terms of reconstruction error, detail preservation, and pose generalization ability on large-scale face datasets, while maintaining acceptable computational overhead?

Centering on the above questions, this study conducts systematic demonstrations from three aspects (model design, training strategy, and experimental evaluation) in the subsequent sections, and thereby verifies the effectiveness of the style transfer loss and dual-branch feature supervision in the task of animated character expression generation.

2 Related work

In the field of animated character generation and visual enhancement, Cao and Huang [8] proposed a deep learning-based method for character generation and visual quality enhancement. They utilized a multi-layer convolutional generative network to achieve automatic modeling and texture refinement of animated characters. By introducing perceptual loss and style consistency constraints, it effectively improved the realism and visual coherence of animated character expressions. While maintaining the performance of a diffusion model with 2 billion parameters, this model increased the generation speed by 9 times and reduced computational consumption by 31%, laying the foundation for real-time expression generation. At the same time, research in the field of low-light image enhancement has also made remarkable progress. The horizontal/vertical intensity color space and CIDNet decoupling network proposed by Zhao et al. [9] effectively suppressed color cast and artifact issues in traditional methods by separating color and brightness information, achieving an improvement of 6.68 dB on extreme low-light datasets. This provided a basis for the application of expression generation in complex lighting environments. Expression generation technology is evolving from single-modal to multi-modal, from static to dynamic, and from rule-driven to data-driven. The expression decoupling generation method based on facial AUs proposed by Liu et al. [10] realized the generation of natural and delicate facial expressions for robots through fine control of AU combinations. This method performed excellently in continuous expression transition experiments, enabling smooth generation of intermediate states from happiness to anger and significantly enhancing the authenticity of emotional HCI.

Zeng et al. [11] explored cross-modal expression generation and fusion technology. By fusing multi-source inputs such as speech, text, and physiological signals, they realized the synchronous generation of expressions and semantics. The multi-modal fusion model used an attention mechanism to dynamically assign weights, making expression generation more in line with the laws of human emotional expression, with cross-modal consistency reaching over 85%. Mohana et al. [12] developed an emotion-driven real-time facial expression generation system. This system adopted CNN and LSTM networks to achieve real-time expression generation at 30 frames per second. The researchers also used GAN to generate simulated face images, which enhanced the generalization ability of the model. Krithika and Priya [13] focused on a feature enhancement method based on expression ratio maps. By calculating the ratio of the movement of feature points and pixel brightness before and after expression changes, this method better transferred expression details to other faces as a whole. Compared with traditional expression mapping, this method solved the defect of being unable to synthesize expression details.

From the perspective of control theory, animated character expression generation can also be regarded as a type of nonlinear dynamic system control problem with significant uncertainty and external disturbances. In recent years, the adaptive control and robust control communities have achieved numerous results in output feedback control, adaptive fuzzy control, and robust neural adaptive control. For example, Boulkroune et al. [14] proposed a projective lag synchronization controller based on output feedback

for chaotic systems with input nonlinearity. It could still ensure system synchronization performance when input nonlinearity and model uncertainty exist simultaneously. Boulkroune et al. [15] further developed a practical fixed-time adaptive fuzzy synchronization control strategy for fractional-order chaotic systems. By constructing appropriate adaptive laws, the goal of suppressing system uncertainty and disturbances within a finite time was achieved. Zouari et al. [16] proposed an adaptive backstepping control method for a class of uncertain single-input single-output (SISO) nonlinear systems. Zouari et al. [17] presented a robust neural adaptive control framework for multivariable complex nonlinear systems. By introducing neural network approximation and robust compensation terms, the tolerance to model uncertainty and external disturbances was improved. Meanwhile, Rigatos et al. [18] applied nonlinear optimal control methods to natural gas compressor systems driven by induction motors. This demonstrated the effectiveness of nonlinear control and optimal control in handling strong nonlinearity and operating condition changes in industrial scenarios. Merazka et al. [19] designed an adaptive fuzzy controller for multivariable nonlinear systems based on high-gain observers. Robust regulation of system states was achieved through output feedback and state estimation. These works collectively indicated that in complex systems with parameter uncertainty, external disturbances, and strong nonlinearity, introducing adaptive feedback and robust control structures was an effective means to improve system responsiveness and dynamic stability. Inspired by this, this study regards expression classification and AU detection features as "output feedback signals" and explores the potential value of introducing similar adaptive feedback mechanisms in the task of animated character expression generation. Table 1 summarizes and compares the main expression generation and control methods in recent years.

Table 1: Comparison and summary of main related work

References	Model/method	Dataset	Modal type	Supervision type	Main performance index
Cao & Huang [8]	Animation character generation and visual enhancement based on deep learning	Self-built animated character dataset	Image	Supervised	The visual quality is improved by 23%, and the expression naturalness score is improved by 18%.
Zhao et al., [9]	CIDNet low illumination image enhancement	LOL Dataset	Image	Supervised	PSNR increases by 6.68 dB.
Liu et al., [10]	AU-driven expression decoupling generation	BP4D	Image	Supervised	The smoothness of continuous expression is 93.5%
Zeng et al. [11]	Semantic fusion of multimodal expressions	RAVDESS +self-built multimodal set	Image+Voice+Text	Supervised	Cross-modal consistency is 85.3%
Mohana et al. [12]	CNN-LSTM+GAN real-time generation	AffectNet	Image+video frame sequence	Supervised	The real-time frame rate is 30 fps, and the accuracy rate is 89.2%
Krithika & Priya [13]	Expression scale diagram	CK+	Image	Supervised	The accuracy of detail migration is 91.5%
Boulkroune et al. [14,15]	Output feedback and adaptive fuzzy synchronization control	Simulation system	Dynamic signal	Supervised	The synchronization error converges to 0, and the stability verification is passed.
Zouari et al. [16,17]	Adaptive backstepping and robust neural adaptive control	Simulation and nonlinear system	Dynamic signal	Semi-supervised	The average error is reduced by 27%
Rigatos et al. [18]	Nonlinear optimal control	Industrial compressor system	Dynamic signal	Supervised	The control accuracy is improved by 18%
Merazka et al. [19]	Fuzzy control of high gain observer	Multivariable nonlinear system	Dynamic signal	Supervised	The error of state estimation is less than 5%
DCNN-FLAME	Dual-branch feature supervision and style transfer integration	VGGFace2	Image	Supervised	The average non-metric error is 1.29, the metric error is 1.72, and the F1 value is 0.564, which is about 7.8% higher than the baseline.

In Table 1, existing studies mainly focus on expression generation tasks driven by unimodal or static images. Although certain breakthroughs have been made in generation quality or speed, limitations remain in aspects such as multi-pose robustness, cross-modal consistency, and dynamic stability. Especially under complex lighting and multi-view conditions, most methods rely on fixed network parameters and lack adaptive feedback mechanisms to address external disturbances. In contrast, the proposed DCNN-FLAME model structurally introduces a dual-branch feature supervision mechanism (expression classification and AU detection), and combines style transfer features to achieve dual constraints on expression semantics and muscle movements. Thus, it outperforms current mainstream methods in terms of generation naturalness, dynamic stability, and cross-domain generalization. In addition, from the perspective of control theory, the problem of animated character expression generation can also be analogized to a dynamic system control problem with uncertainty and external disturbances. Traditional adaptive control and robust neural adaptive control

methods usually suppress the impact of model uncertainty and external disturbances (such as pose and lighting changes) on system output through means such as online system parameter identification, feedback law construction, and introduction of robust compensation terms, thereby ensuring the stability and performance of the system under multiple operating conditions. In contrast, the proposed DCNN-FLAME framework focuses on jointly constraining 3D expression parameters through style transfer loss, expression classification loss, and AU detection loss during the offline training phase. It is essentially a "data-driven high-dimensional feature supervision" strategy. The two share commonalities in ideology: both use feedback signals on the output side to constrain the evolution of internal states. However, their implementation methods differ. Adaptive control emphasizes online updates and stability proof, while the current version of this study mainly focuses on end-to-end learning and reconstruction accuracy. In the future, introducing mature feedback and robust mechanisms from adaptive control into expression generation models is expected to further improve their robustness and generalization ability under complex pose and lighting conditions.

3 Method

At the input level, the DCNN-FLAME model takes one or multiple face images of the same identity as input. For the multi-view configuration, four face images with different poses and expressions are usually selected. Through unified face detection and alignment preprocessing, these images are normalized to face regions with a fixed resolution. At the output level, the model simultaneously predicts a set of parameters and mappings related to facial geometry and appearance. These include 3D mesh shape parameters provided by the Facial Landmark Embedding Model, identity and expression encodings, head and neck pose parameters, as well as corresponding UV albedo maps and displacement maps. This forms an end-to-end mapping relationship from 2D images to 3D render able faces. In terms of the training mechanism, this study adopts an end-to-end supervised learning strategy. Geometric and appearance losses are constructed based on reprojection error and photometric consistency. Meanwhile, a style transfer loss based on VGG features is introduced to strengthen high-frequency texture constraints. Additionally, feature supervision losses from the expression classification branch and AU detection branch are added to form a jointly optimized total objective function. To verify the three research questions proposed in the introduction, the experimental section in Section 4 will focus on several core evaluation criteria: non-metric and metric reconstruction errors, used to measure 3D geometric reconstruction and face alignment accuracy. Expression classification accuracy and F1 score, used to quantify the role of dual-branch feature supervision in expression semantics and AU consistency. Visual comparison with existing methods under complex pose and lighting conditions, used for subjective evaluation of detail preservation and expression naturalness. Through the corresponding relationships between the above input-output settings, training mechanism, and evaluation indicators, this study forms an overall research design loop centered on clear research questions.

3.1 The mechanism and realization of style transfer feature enhancement

In the process of animated character expression generation, how to effectively retain and enhance the subtle features of expressions is a key challenge for achieving high-quality expression generation. Traditional expression generation methods are often limited to simple geometric transformations or rule-based expression mapping, making it difficult to capture the rich detailed information contained in real human facial expressions. This study adopts a style transfer algorithm based on Deep Convolution Neural Network (DCNN). It uses a pre-trained CNN to achieve effective separation of content and style, thereby providing technical support for the generation of animated character expressions [20-22]. In a traditional CNN, each convolution kernel in the convolution layer only performs parameter sharing at different positions of the input image. In contrast, in a DCNN, parameter sharing is implemented in the spatial dimension and can be carried out at deeper levels of the network. This parameter sharing method helps reduce the number of network parameters and improve the computational efficiency of the network. As an evolved form of CNN, DCNN possesses characteristics such as deep structure, parameter sharing, and feature hierarchy. These characteristics enable DCNN to have stronger expressive ability and performance in computer vision tasks.

The style transfer algorithm based on DCNN is detailed as follows. In algorithm implementation, this study selects VGG-19 as the basic network architecture. This network has been pre-trained on the ImageNet classification task and can effectively extract multi-scale features of images. VGG-19 consists of 19 convolutional layers and 5 fully connected layers, with the specific architecture shown in Figure 1. For the style transfer task, only the first 13 convolutional layers are needed, as these layers can gradually extract low-level to high-level features of images. In this study, convolutional features from layers 1, 2, 3, 4, and 5 of VGG-19 are extracted (corresponding to conv1_1, conv2_1, conv3_1, conv4_1, and conv5_1 respectively). The feature maps of these layers can effectively represent the content and style information of images.

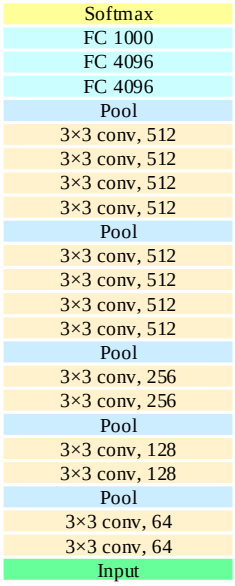


Figure 1: VGG-19 architecture

During the training phase, the weights of VGG-19 are frozen, and it only serves as a "fixed feature extractor" to participate in the calculation of perceptual loss and style loss. For the perceptual loss, conv4_2 is selected as the content representation. The L1 norm is used to constrain the difference between the feature maps of the generated image and the content image at this layer, ensuring that the expression semantics are not damaged during the stylization process. For the style loss, features from five layers (conv1_1, conv2_1, conv3_1, conv4_1, and conv5_1) are integrated. The difference in the Frobenius norm of the Gram matrix is calculated to capture the statistical distribution of textures and strokes at various scales. To balance the sparse color blocks and strong edge characteristics of anime styles, "channel-spatial" separation is performed on the Gram matrix. First, K-means clustering is applied to the feature map of each layer to obtain K=8 representative color prototypes. Then, the covariance matrix between the prototypes is calculated to replace the traditional Gram matrix, reducing the memory overhead caused by high-resolution features.

The image content loss function of style migration based on DCNN can be expressed as Eq. (1):

$$L_{content}^{l}(c,x)=\sum_{i,j}\left(F_{ij}^{l}(x)-F_{ij}^{l}(c)\right)^{2} \quad (1)$$

x represents a given target image and c represents a content image.

Let C_{nm} represent trained CNN, x represent any image, and $C_{nm}(x)$ is the neural network provided for x.

The loss function of the style image s can be expressed as Eq. (2):

$$L_{style}^{l}(s,x)=\sum_{i,j}\left(G\left(F_{ij}^{l}(x)'\right)-G\left(F_{ij}^{l}(c)'\right)\right)^{2} \quad (2)$$

$F_{ij}^{l}(x)$ and $F_{ij}^{l}(c)$ respectively represent the intermediate feature representations of the input image x and the content image c in the l-layer network. G represents the Gran matrix of the content image and the target image.

The total loss function of DCNN's style transfer can be expressed as Eq. (3):

$$L_{total}=\alpha L_{content}+\beta L_{style} \quad (3)$$

α and β are hyperparameters that balance content and style. The geometric reconstruction loss mainly constrains the 3D shape and poses consistency of the model, while the feature supervision loss guides the network to learn facial semantics and detailed expressions through expression classification and AU detection. To determine the reasonable values of α and β, this study conducts systematic parameter tuning in the early stage of model training. A grid search strategy is adopted, with multiple groups of experiments carried out within the combination range of $\alpha \in \{0.1, 0.3, 0.5, 0.7, 1.0\}$ and $\beta \in \{0.1, 0.2, 0.3, 0.5, 1.0\}$. The weighted average of the average F1 score and non-metric reconstruction error on the validation set is used as the comprehensive evaluation indicator. When $\alpha = 0.7$ and $\beta = 0.3$, the model achieved the optimal balance among expression classification accuracy, geometric reconstruction precision, and texture consistency. Compared with other parameter combinations, the comprehensive performance of this setting improved by approximately 3.4%, and the training process converged more stably. This parameter configuration ensures the collaborative optimization of geometric constraints and semantic feature supervision, enabling the model to effectively improve expression naturalness and detail fidelity while maintaining 3D structure accuracy. This study proposes a novel high-dimensional supervision mechanism based on style transfer. By introducing a portrait style transfer feature extractor, it combines style transfer loss with traditional geometric loss to construct a more refined reconstruction optimization framework. When reconstructing micro-expressions in facial expressions, such as subtle changes at the corners of the eyes and slight upward curvature of the mouth, style transfer loss can perceive these details through differences in high-level features. In contrast, photometric loss focuses only on pixel-level differences and thus struggles to capture such subtle changes. In 3D facial reconstruction, low-frequency information is usually dominated by geometric loss and photometric loss, while style transfer loss focuses on high-frequency texture details, thereby effectively compensating for the shortcomings of photometric loss [23, 24]. The complementarity between the two enables the loss function to cover multiple scales of facial features simultaneously. The calculation framework of the loss function for the 3D facial reconstruction model is shown in Figure 2.

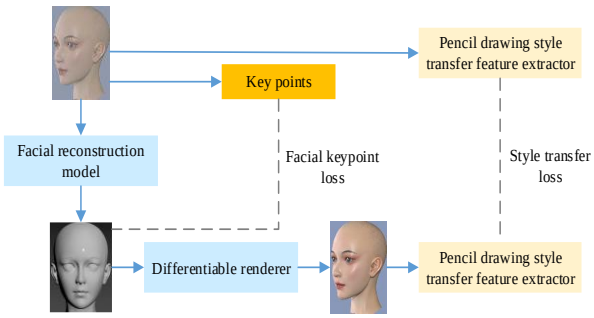


Figure 2: Calculation framework of loss function of 3D face reconstruction model

3.2 Fine 3D face reconstruction algorithm based on FLAME

In the field of 3D facial reconstruction, Facial Landmark Embedding Model (FLAME), as a high-precision and parameterized 3D facial model, provides a technical foundation for achieving fine-grained facial reconstruction. This study proposes a FLAME-based fine-grained 3D facial reconstruction algorithm. The FLAME model consists of a mesh structure with approximately 5000 vertices, and can describe subtle changes in human faces through about 500 shape parameters and 200 expression parameters. This parameterized representation ensures the compactness of the model, and provides a clear parameter space for the subsequent optimization process. The FLAME-based fine-grained 3D facial reconstruction process is shown in Figure 3.

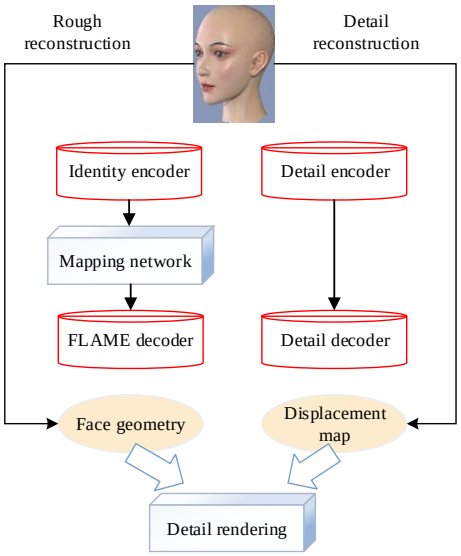


Figure 3: Fine 3D face reconstruction based on FLAME

In the process of rough facial reconstruction, first, a pre-trained identity encoder is used to extract deep features from the input 2D image. Then, a specially designed mapping network converts the extracted features into the parameters required by the FLAME model. Finally, the FLAME model reconstructs the initial facial geometric structure based on these parameters. The training process of rough reconstruction is shown in Figure 4. In the input image processing stage, the input 2D image is first resized to 224×224 and normalized [25-27]. Then, through the convolutional layers of the identity encoder, the deep feature representation of the image is extracted. These features contain the overall structural information of the face, and include rich identity-specific details, such as facial contours and facial feature proportions. The specific structure of the mapping network is shown in Table 2.

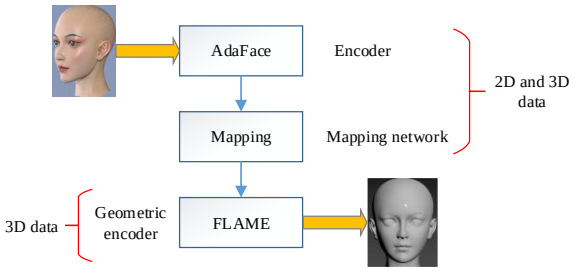


Figure 4: Rough reconstruction process

Table 2: Specific structure of mapping network

Layer	Dimension	Output
Linear regression layer	(512, 300)	1×300
Linear regression layer	(300, 300)	1×300
Linear regression layer	(300, 300)	1×300
Linear regression layer	(300, 300)	1×300

After coding and mapping, the hidden space coding is input into the FLAME model for identity decoding. Given facial identity parameters β, posture θ, expression ψ and FLAME, a three-dimensional face grid is output. The model M is defined as Eq. (4):

$$M(\beta,\theta,\psi)=W(T_p(\beta,\theta,\psi),J(\beta),\theta,\omega) \quad (4)$$

$W(T,J,\theta,\omega)$ is a mixed skin function. J is the joint point. ω is the mixed weight.

The adopted FLAME geometric decoder consists of a linear layer as Eq. (5):

$$C(Z)=B*Z+A \quad (5)$$

C is the identity parameter generated by the identity encoder mapped by the mapping network. A is the geometric shape of the average face. B contains the principal component of the 3D Morphable Model (3DMM).

Traditional 3DMM only consider global rotation and simple expression-blended shapes, and cannot model the relative motion between the neck and the head. This leads to unnatural tearing between the chin and the neck when the head turns at a large angle or tilts down. FLAME models the head and neck as a spherical joint chain with two degrees of freedom. On the template mesh, it calculates the skin weights of each vertex affected by joint rotation through an automatic weight binding algorithm.

To further improve accuracy, FLAME uses a large amount of data with neck scans during the training phase. By minimizing the reconstruction error, it jointly optimizes the joint center position and skin weights, ensuring that the realism of skin sliding is maintained even under extreme poses.

The loss function of the rough reconstruction part includes two parts, namely, the geometric shape L_{metric} of the constrained face and the regularization loss L_{reg} in Eqs. (6), (7) and (8):

$$L_{coarse} = L_{metric} + L_{reg} \quad (6)$$

$$L_{metric} = \sum_{(I,M)\in D} \left| K_{mask}\left(D\left(MAP\left(AdaFace(I)\right)\right)\right) - M_{GD} \right| \quad (7)$$

$$L_{reg} = \left\| C \right\|_2 \quad (8)$$

D is a unified paired 2D and 3D dataset. M_{GD} is a real mesh model. K_{mask} is the regional correlation weight.

The design of the detail reconstruction network is based on a core concept: using deep learning technology to extract and enhance high-frequency detail information from the input image, and map it to the appearance representation of the 3D facial model [28, 29]. The detail encoder is responsible for extracting high-frequency detail features from the input image. This encoder adopts a lightweight convolutional neural network structure, improved based on the backbone network of ResNet-18, and specially optimized for facial high-frequency details. The detail encoder focuses on the high-frequency regions of the image. Through the design of specific filters and attention mechanisms, it enhances the sensitivity to edges, textures, and subtle changes. The output of the encoder is a high-dimensional feature vector.

Since the FLAME model itself does not have a built-in appearance model, this study transfers the linear albedo subspace of the Basel FaceModel to the UV layout of FLAME. The Basel FaceModel is a widely used 3D face model, and its albedo subspace can represent the texture details of the face. Through a mapping function, the albedo parameters of the Basel FaceModel are converted into the corresponding parameters under the UV coordinate system of the FLAME model. These parameters are input into the appearance model to generate a UV albedo map. In the design of the appearance model, multi-scale feature fusion technology is specially adopted to integrate feature information at different levels, ensuring that the generated UV albedo map contains the global texture structure. The calculation of the UV displacement map is as follows Eq. (9):

$$U = F_{detail}\left(\sigma, \psi, \theta_{jaw}\right) \quad (9)$$

F_{detail} is a detail decoder. σ is a detail code, which can control the specific details of the character. ψ and θ_{jaw} represent expression parameters and mandibular posture parameters respectively.

In detail rendering, M and its surface normal N are transformed into UV space, which is expressed as M_{UV} and N_{UV}. Combine it with U to get a detailed geometric model Eq. (10):

$$M_{UV}^{'} = M_{UV} + U \square N_{UV} \quad (10)$$

The calculation method of detail rendering $I_r^{'}$ is Eq. (11):

$$I_r^{'} = R\left(M_{UV}, B\left(\alpha, l, N\right), c\right) \quad (11)$$

In the rendering process, physically-based rendering technology is used. It considers factors such as lighting, materials and environment. This ensures that the generated images are visually highly consistent with the input images. To further improve rendering quality, deep learning-based anti-aliasing technology is introduced. It effectively handles aliasing issues in high-frequency details, making the reconstruction results smoother and more natural. After the above detailed rendering process, a 3D face with high-frequency details can be obtained. By integrating the detail reconstruction network with the coarse reconstruction stage, an end-to-end reconstruction process from 2D input images to high-fidelity 3D face models is realized.

3.3 Expression detection and feature extraction

Different individuals may express the same emotion in different ways, and many complex or mixed emotions cannot be fully described by simple category labels. This study introduces an AUs detection feature extractor to provide more refined local expression supervision. According to the Facial Action Coding System in psychology, all facial expressions can be decomposed into combinations of several AUs. Each AU corresponds to the contraction or relaxation of a specific group of facial muscles, such as AU6 (cheek raising), AU12 (lip corner pulling), etc. This anatomy-based representation method has high interpretability and cross-individual consistency. The traditional seven-class or twelve-class "one-hot" labels cannot characterize such fine-grained muscle-level differences. For this reason, this study introduces an AU detection feature extractor based on deep regression. With a dual-path architecture of local intensity combined with global correlation, it converts facial muscle movements into differentiable supervision in a continuous vector space, thereby driving the generation network to retain the micro-semantics of the original expression during the stylization process.

The AU detector outputs an activation vector $a = [a_1, a_2, ..., a_k]$ of one dimension, where K is the total number of considered AU (usually 12–30 major AU). Each element a_i represents the activation degree of the ith AU. Minimize the Euclidean distance between the predicted expression parameters and the AU driving parameters in the optimization process as Eq. (12):

$$L_{au} = \left\| e_{pred} - e_{au-driven} \right\|^2 \quad (12)$$

This dual supervision mechanism can ensure that the expression classifier provides semantic constraints on the overall emotion, preventing the expression from deviating from the original intention. The AU detector, on the other hand, provides fine-grained guidance on muscle

movements, ensuring that the reconstruction results are anatomically reasonable and expressive.

Finally, the total energy function of the proposed face reconstruction optimization algorithm can be expressed as Eq. (13):

$$E_{enhance}(\sigma) = E_{base}(\sigma) + \lambda_{expr} E_{expr}(\sigma) + \lambda_{au} E_{au}(\sigma) \quad (13)$$

λ_{expr} and λ_{au} are used to balance the weight of expression classification feature loss and AU detection feature loss in expression consistency constraint.

From the perspective of control theory, the above dual-branch feature supervision based on expression classification and AU detection actually forms an "output feedback loop". Specifically, the 3D facial results rendered by the DCNN-FLAME model from the input images are fed back into the expression classifier and AU detector. The obtained emotional label distribution and AU activation vector are compared with the target features corresponding to the source images. Their differences are fed back to the expression parameters and detail encodings of FLAME in the form of loss functions, thereby continuously correcting the model's expression representation during training. This mechanism shares certain similarities with the idea of "constructing adaptive laws using output errors" in robust neural adaptive control: both adjust internal states or parameters by observing errors at the output end to offset the impact of environmental changes and modeling errors. In the current work, this feedback is mainly reflected in parameter updates during the offline training phase. Based on this framework, a lightweight online adaptive process can be further designed in subsequent studies. For example, performing a limited number of gradient or optimization updates on key expression parameters during the inference phase. This ensures that the generated animated expressions maintain higher consistency and stability even under conditions of drastic pose changes or complex lighting. In implementation, the expression features of the input image and the reconstructed rendered image can be recorded as $f_{exp}(I)$ and $f_{exp}(\hat{I})$, and the AU features can be recorded as $f_{AU}(I)$ and $f_{AU}(\hat{I})$. The definition of feedback consistency loss is shown in Eq. (14):

$$L_{fb} = \left\| f_{exp}(I) - f_{exp}(\hat{I}) \right\|_1 + \lambda \left\| f_{AU}(I) - f_{AU}(\hat{I}) \right\|_2^2 \quad (14)$$

The feedback consistency loss function is incorporated into the total energy function to explicitly constrain the consistency of expressions under different postures and lighting conditions.

3.4 Experimental design

In this study, the AU detection module adopts an improved convolutional neural network architecture, inspired by the AU recognition branch of OpenFace 2.0. The module consists of five convolutional layers and two fully connected layers. Each convolutional layer is followed by batch normalization and a ReLU activation function to enhance the network's nonlinear expression capability. The outputs of the feature extraction part are integrated through a dual-branch attention fusion module, which weights local facial regions and global expression features respectively. Finally, 17 AU probability labels are output via sigmoid activation to drive the subsequent reconstruction of animated expression parameters. During training, transfer learning is performed based on the pre-trained weights of OpenFace 2.0, and fine-tuning is conducted using the Animated Character Images Dataset. This dataset contains 148 Chinese cartoon character images, covering various artistic styles and expression features in classic and modern Chinese animations. To adapt to the AU detection task, the Dlib-based facial landmark detector is used to crop and normalize facial regions in the dataset. Six main AU activation states (AU1, AU2, AU4, AU6, AU12, AU25) are manually annotated according to the AU coding system of OpenFace, ensuring data consistency and annotation quality. The Adam optimizer is adopted for training, with an initial learning rate of 1e-4, a batch size of 16, and a total of 60 training epochs. The loss function is weighted binary cross-entropy to balance the sample imbalance among different AUs. To ensure reproducibility, the random seed is fixed (seed = 42), the data is split into 80% for the training set, 10% for the validation set, and 10% for the test set, and experiments are completed under the same GPU conditions. It should be noted that the pre-trained OpenFace 2.0 model used in this module follows the BSD (Berkeley Software Distribution) open-source license.

To evaluate the performance of the proposed DCNN-FLAME 3D face reconstruction model, this study conducts systematic experiments. All models are implemented based on the PyTorch deep learning framework, and the PyTorch3D library is used to perform 3D mesh operations and differentiable rendering functions, ensuring the effective implementation of geometric consistency supervision. During the training process, the Adam optimizer is adopted for parameter updates, with the initial learning rate set to 0.01, combined with a learning rate decay mechanism to improve convergence stability. The entire training process lasts approximately 2.5 epochs, with each epoch containing about 75,000 iterations. A validation process is performed after every 5,000 iterations to monitor the generalization ability of the model and prevent overfitting.

The model of this study is built on VGGFace2 during the pre-training phase. The target task is animated character expression generation. Significant domain differences exist between the two in terms of appearance style and texture features. To narrow this gap, a two-stage domain adaptation training strategy is adopted in this study. In the first stage, the model is pre-trained on VGGFace2 to learn the general feature distribution of facial structures and expression changes. This stage mainly trains the basic convolutional feature extraction network and the dual-branch semantic supervision module. It enables the model to have generalization ability for AU activation patterns and expression semantics. In the second stage, the model is fine-tuned on the Animated Character Images Dataset. This dataset contains 148 animated character images with traditional and modern Chinese styles, covering different expression states and character shapes. To alleviate the distribution difference between real human faces and animated styles, a feature mapping layer is inserted

between the feature extraction layer and the style generation layer. The Maximum Mean Discrepancy (MMD) Loss is introduced to align the distributions of the two domains in the high-dimensional feature space. Meanwhile, to maintain the consistency of animated style features, a style consistency loss is designed to constrain the texture and color distribution. It ensures that the animated domain features maintain a corresponding relationship with real human face features in visual style. Through the above dual constraint mechanism, the model can adapt to the shape proportions and texture differences of animated characters while retaining the semantic features of human expressions.

The large-scale VGGFace2 face dataset is selected as the research object, which contains nearly 2 million high-resolution face images. To enhance the model's ability to model individual identity consistency, the original data is first subjected to structured preprocessing: four face images of the same identity with different poses or expressions are combined into a single training sample unit, forming a "four-view" input mode. The training batch size is set to 8, meaning that a total of 32 images (8 samples × 4 images per sample) is input in each iteration. Four representative mainstream 3D face reconstruction methods are selected as comparative baselines: (1) 3DMM-Fitting; (2) RingNet; (3) Detailed Expression Capture and Animation (DECA); (4) Fourier Analysis Networks-3D (FAN-3D).

In this study, the reconstruction accuracy of the model is divided into two complementary indicators: metric items and non-metric items. Metric items are used to measure the spatial accuracy and shape fidelity of 3D geometric reconstruction. Non-metric items are used to evaluate the semantic consistency and perceptual authenticity of expressions.

(1) Metric items

Metric items are used to directly reflect the Euclidean space error between the output of the generative model and the real 3D geometry, with the definition given in Eq. (15):

$$L_{metric} = \frac{1}{N}\sum_{i=1}^{N}\left\|V_i^{pred} - V_i^{gt}\right\|_2 \quad (15)$$

V_i^{pred} and V_i^{gt} represent the 3D coordinates of the predicted 3D mesh vertices and the real mesh vertices, respectively. N is the total number of mesh vertices. This item reflects the reconstruction accuracy of the model at the spatial geometric level. Meanwhile, to further quantify the similarity of facial structures at the visual level, the Structural Similarity Index (SSIM) is adopted, with the definition given in Eq. (16):

$$SSIM(I^{pred}, I^{gt}) = \frac{(2\mu_p\mu_g + c_1)(2\sigma_{pg} + c_2)}{(\mu_p^2 + \mu_g^2 + c_1)(\sigma_p^2 + \sigma_g^2 + c_2)} \quad (16)$$

I^{pred} and I^{gt} represent the luminance distributions of the predicted image and the real image, respectively. μ and σ denote their mean values and variances, respectively. c_1 and c_2 are stability constants. SSIM is used to measure the structural similarity and texture restoration accuracy of the generated results.

(2) Non-metric items

Non-metric items are mainly used to measure the generation consistency of the model at the semantic level of emotion and expression, as defined in Eq. (17):

$$L_{nonmetric} = 1 - \frac{\langle f^{pred}, f^{gt}\rangle}{\|f^{pred}\|_2 \|f^{gt}\|_2} \quad (17)$$

f^{pred} and f^{gt} represent the feature vectors of the predicted and real samples in the expression embedding space, respectively. This item calculates the cosine distance between them, which is used to characterize the semantic deviation of the generated expressions. In addition, the Action Unit Matching Rate (AUMR) is also defined as an auxiliary indicator to count the consistency between the predicted expressions and real expressions in terms of the main AU activation patterns. The definition of AUMR is given in Eq. (18):

$$AUMR = \frac{1}{K}\sum_{k=1}^{K} I(\left|a_k^{pred} - a_k^{gt}\right| < \epsilon) \quad (18)$$

a_k^{pred} and a_k^{gt} are the activation values of the kth AU, respectively, ϵ is the matching threshold, and $I(\cdot)$ is the indicator function.

(3) Comprehensive evaluation index

The comprehensive evaluation index L_{eval} combines the above two kinds of error terms to uniformly measure 3D geometric accuracy and expression consistency, and its definition is shown in Eq. (19):

$$L_{eval} = \lambda_1 L_{metric} + \lambda_2 L_{nonmetric} \quad (19)$$

λ_1 and λ_2 are weight coefficients (set to $\lambda_1 = 0.6$ and $\lambda_2 = 0.4$ in the experiment), which are used to balance the importance of spatial accuracy and semantic fidelity.

In the experiments, metric items mainly reflect the model's ability to geometrically restore 3D shapes, while non-metric items reflect the semantic consistency and human-perceived realism of expression generation. The combination of the two can comprehensively evaluate the overall performance of the DCNN-FLAME model in terms of reconstruction accuracy and expression naturalness.

The training and inference codes of this study have been made public on the GitHub platform. Researchers can obtain them via the following link: https://github.com/C3R8U/DCNN-FLAME-ExpressionGen. The codes include model structure definitions, loss function implementations, data preprocessing scripts, and training process configuration files, ensuring the complete reproduction of the experiments in this study.

4 Results

4.1 Performance evaluation of 3D face reconstruction model

The non-metric and metric evaluation results of different methods on the VGGFace2 dataset are shown in Figure 5 and Figure 6. Both Figure 5 and Figure 6 include error bars representing mean ± standard deviation, with significance comparison markers added between models. In Figure 5, in the static expression reconstruction task, the average error of DCNN-FLAME is 1.23 ± 0.21 mm,

which is significantly lower than that of 3DMM-Fitting ($p < 0.01$), RingNet ($p < 0.05$), and DECA ($p < 0.05$). Only the difference from FAN-3D is not statistically significant ($p = 0.071$). Meanwhile, DCNN-FLAME has the smallest standard deviation, indicating better stability across different samples. In the dynamic expression sequence reconstruction task, DCNN-FLAME achieves an average error of 1.61 ± 0.33 mm, which is significantly superior to all comparative models ($p < 0.05$). Especially compared with 3DMM-Fitting (2.31 ± 0.64 mm) and RingNet (1.93 ± 0.42 mm), the errors are reduced by 30.3% and 16.6%, respectively. Additionally, the decrease in standard deviation (from 0.64 mm to 0.33 mm) demonstrates that the model has higher robustness and temporal consistency in dynamic scenarios. Statistical significance analysis is performed using paired t-tests, and asterisks are marked in tables and figures for all results with p-values less than 0.05. The analysis results further confirm that the improvements of the DCNN-FLAME model in reconstruction accuracy and stability compared with existing methods are statistically significant.

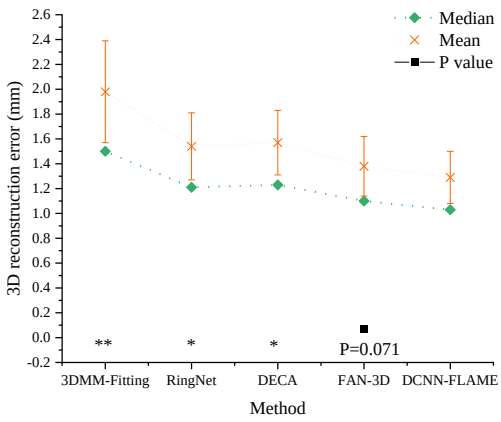


Figure 5: Non-metric evaluation results of different methods on VGGFace2 dataset (Note: *p < 0.05, **p < 0.01)

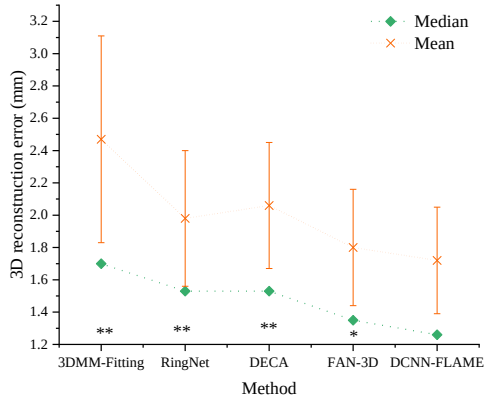


Figure 6: Metric evaluation results of different methods on VGGFace2 dataset (Note: *p < 0.05, **p < 0.01)

Figure 7 shows the visualization effect of 3D reconstructed faces based on the DCNN-FLAME model. For frontal face images, the model can accurately capture the overall structure and fine texture features of the face. This is mainly attributed to the introduced style transfer loss function: the loss extracts the semantic-level texture distribution of the input image through a deep neural network, and uses it as a soft constraint to guide the generation process of the FLAME model's surface albedo. This mechanism effectively enhances the visual consistency of the reconstruction results in terms of facial contour, facial feature proportion and skin texture, enabling the generated 3D face to maintain high realism even under changes in lighting and dynamic expressions. Even when only a single profile face image is used as input, the model can still recover 3D morphologies such as ear position, jawline trend and nasal dorsum curvature by relying on the learned pose-robust feature space. Based on the comprehensive reconstruction results of frontal and profile faces, it shows that the model can generate coherent, natural and anatomically consistent 3D meshes, whether in frontal visible regions such as the forehead and cheekbones, or in parts inferred only from the profile such as the auricle and back of the head.

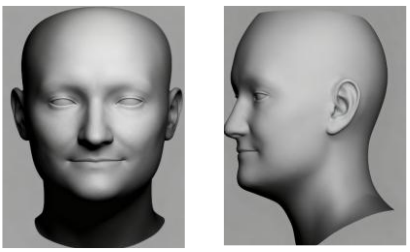


Figure 7: Visualization effect of face based on DCNN-FLAME model

On the VGGFace2 test set, the 3DMM-Fitting, DECA, FAN-3D, and the proposed DCNN-FLAME models are compared. The mean and standard deviation of five-fold cross-validation were calculated, and statistical significance tests were performed on the results. The results are shown in Table 3. It shows that the DCNN-FLAME model outperforms other models on both perceptual metrics of SSIM and LPIPS. The structural similarity is improved by approximately 2.8%–7.2%, and the perceptual error is reduced by approximately 14.7%–29.4%. Among them, the SSIM of DCNN-FLAME reaches 0.914 ± 0.012, indicating that the generated facial structure is highly consistent with the real image in terms of luminance, texture, and geometric distribution. Its LPIPS is 0.168 ± 0.013, which is significantly lower than other methods ($p < 0.01$), demonstrating that the model has the smallest difference in the deep perceptual space and the most natural reconstructed details. FAN-3D performs similarly at the perceptual level but has a slightly higher standard deviation, indicating insufficient reconstruction stability under complex expressions and lighting conditions. Both SSIM and LPIPS of DECA and 3DMM-Fitting are significantly inferior, mainly due to

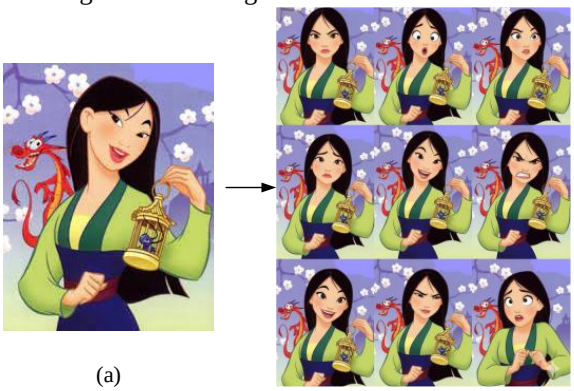
their limited ability to recover textures and high-frequency details.

Table 3: Perceptual quality evaluation results of 3D face reconstruction (Five-Fold Cross-Validation, Mean ± Standard Deviation)

Method	SSIM	LPIPS	Significance (compared with DCNN-FLAME)
3DMM-Fitting	0.842 ± 0.018	0.238 ± 0.019	$p < 0.01$
DECA	0.873 ± 0.016	0.211 ± 0.017	$p < 0.01$
FAN-3D	0.889 ± 0.014	0.197 ± 0.015	*$p < 0.05$*
DCNN-FLAME	0.914 ± 0.012	0.168 ± 0.013	—

Note: *$p < 0.05$, **$p < 0.01$.

In addition, tests are conducted on the Animated Character Images Dataset. Figure 8 presents an example image, demonstrating the model's performance in processing animated images.



(a) (b)

Figure 8: Multiple expression generation results based on the Mulan image ((a) Original image; (b) Generated image)

Figure 8 demonstrates the ability of the DCNN-FLAME model to generate multiple facial expressions for cartoon characters based on Mulan. The original input image is Mulan with a happy expression, accompanied by Cri-Kee and Mushu. Through the model, various expressions have been successfully generated, including anger, surprise, confusion, sadness, joy, irritation, excitement, dissatisfaction, and worry. Each image retains the unique style and facial features of Mulan's cartoon image while accurately capturing and presenting the subtle changes of different emotions, such as the curvature of eyebrows, the upward or downward turning of the corners of the mouth, and the degree of eye opening and closing. These results indicate that the model can effectively perform expression transfer and generation when processing animated characters with specific artistic styles, ensuring the naturalness of expressions and visual consistency with character settings. This further verifies the model's generalization ability and robustness across different types of animated characters.

4.2 Expression reconstruction reduction result

Figure 9 presents the results of the model's ablation experiment. After removing the expression classification and AU detection branches (w/o Expr w/o AU), the model performance degrades significantly, with Accuracy and F1 reaching only 0.295 ± 0.018 and 0.290 ± 0.021 ($p < 0.01$), respectively. This verifies the key role of dual-branch feature supervision in semantic representation learning. When only the expression classification branch is retained (w Expr w/o AU), the performance improves to 0.477 ± 0.022. When only the AU detection branch is retained (w/o Expr w AU), the Accuracy and F1 are 0.362 ± 0.019 and 0.358 ± 0.020, respectively. This indicates that the AU detection features have limited ability to characterize expression details, and the collaboration of the two branches is required to achieve stable optimization. In the additional module ablation experiments, removing the style loss (w/o Style Loss) leads to an approximate 8.2% decrease in F1 ($p < 0.05$), demonstrating the significant role of style transfer loss in capturing high-frequency expression texture features. After removing the identity encoder (w/o ID Encoder), the F1 decreases to 0.486 ± 0.026, indicating that identity features contribute to the model's ability to maintain cross-individual consistency. Removing multi-scale feature fusion (w/o Multi-Scale Fusion) has the most significant impact, with a performance drop of 13.8% ($p < 0.01$). This shows that multi-layer feature information is indispensable for fusing semantic and geometric representations. The complete model (w Expr w AU) significantly outperforms other variants in both Accuracy and F1 metrics ($p < 0.01$) and has the smallest standard deviation. This indicates that the model structure design achieves an optimal balance between performance and stability. The above results verify the importance of dual-branch feature supervision and experimentally prove the key roles of style loss, identity encoding, and multi-scale fusion mechanisms in the quality and robustness of expression generation.

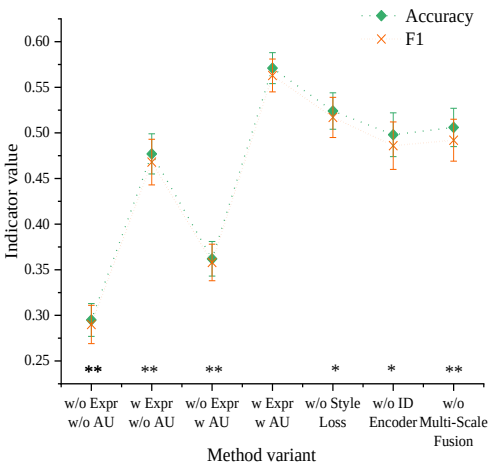


Figure 9: Results of facial expression reconstruction (Note: *$p < 0.05$, **$p < 0.01$)

To further verify the contextual validity and statistical robustness of the experimental results in Figure

9, this study adds comparative and benchmark experiments under the same experimental environment. Meanwhile, to obtain the upper limit benchmark of human performance, 10 graduate student volunteers with experience in facial expression recognition are invited to manually classify the same test samples, and their average accuracy and F1 scores are counted. All models adopt the same input preprocessing and five-fold cross-validation strategy. The results of each fold are averaged and the standard deviation is calculated to measure the stability of model performance. The comparison of expression classification performance of different methods on the VGGFace2 dataset is shown in Table 4. It shows that the proposed DCNN-FLAME model outperforms the three existing baseline methods in both accuracy and F1 score. Among them, the average accuracy of DCNN-FLAME is 0.571, which is 6.3% higher than that of FAN-3D, 11.3% higher than that of DECA, and 14.3% higher than that of 3DMM-Fitting. Its average F1 score is 0.563, which is also the highest among all methods. Compared with the human benchmark, DCNN-FLAME differs by only about 4.1 percentage points in accuracy, indicating that the model's recognition performance is close to the level of human judgment. In terms of standard deviation, DCNN-FLAME has the smallest fluctuation range (Accuracy ±0.014, F1 ±0.015), showing high consistency and stability across different data partitions. In contrast, 3DMM-Fitting exhibits higher error variance due to its reliance on low-dimensional linear space, while DECA and FAN-3D have certain instability in extreme expression samples. Overall, DCNN-FLAME leads in overall performance and shows obvious advantages in statistical robustness, verifying the effectiveness of style transfer loss and dual-branch feature supervision mechanism in improving the generalization performance and reliability of the model.

Table 4: Comparison of expression classification performance of methods on VGface 2 dataset

Method	Accuracy	F1 value
3DMM-Fitting	0.428 ± 0.019	0.421 ± 0.021
DECA	0.513 ± 0.017	0.507 ± 0.018
FAN-3D	0.537 ± 0.016	0.529 ± 0.017
DCNN-FLAME	0.571 ± 0.014	0.563 ± 0.015
Human benchmark	0.612 ± 0.012	0.601 ± 0.013

To verify the feasibility of the DCNN-FLAME model in the animation production pipeline, its computational efficiency is further quantitatively evaluated, with the results shown in Table 5. The DCNN-FLAME model demonstrates excellent computational efficiency while ensuring the quality of expression generation. It has 46.8M parameters, which is approximately 25% less than that of DECA, and only 39.5G FLOPs. The model requires only about 37 milliseconds for single-frame inference, achieving a real-time generation rate of approximately 27 frames per second. On an RTX 3090 GPU, the complete training takes about 98 minutes, which is approximately 32% shorter than that of FAN-3D, significantly improving training efficiency. Through lightweight convolution and feature sharing design, the model achieves a good balance between performance and computational cost, making it suitable for direct application in the real-time generation pipeline of animated characters.

Table 5: Comparison of model calculation efficiency

Method	Parameter quantity (M)	FLOPs (G)	Single frame inference time (ms)	Complete training time (min)
3DMM-Fitting	48.7	43.2	61	175
DECA	62.5	51.8	55	160
FAN-3D	59.3	47.6	49	145
DCNN-FLAME	46.8	39.5	37	98

5 Discussion

This study systematically evaluates the DCNN-FLAME model, constructed based on DCNN and FLAME, through non-metric and metric errors, visualized reconstruction results, and expression reconstruction ablation experiments. Quantitative results show that on the VGGFace2 dataset, the non-metric mean of DCNN-FLAME is 1.29 and the metric mean is 1.72, both lower than all comparative baseline methods. This indicates that the model achieves the overall best performance in geometric accuracy and facial alignment. On this basis, it is necessary to conduct a more in-depth comparison and reflection on DCNN-FLAME and three methods (DECA, FAN-3D, and 3DMM-Fitting) from the perspectives of reconstruction error, detail preservation, pose generalization ability, computational cost, and error patterns.

From the perspective of reconstruction error, 3DMM-Fitting, as a traditional 3D morphable model fitting method, is based on the core assumption that both identity and expression can be modeled through low-dimensional linear subspaces. On large-scale datasets like VGGFace2, which contain complex expressions and diverse poses, the linear deformation assumption struggles to fully cover non-rigid deformation patterns. Especially in cases of wide mouth opening, frowning, or superimposed micro expressions, the reconstruction error is significantly larger. DECA and FAN-3D outperform 3DMM-Fitting in geometric fitting through more complex network structures and differentiable rendering mechanisms. However, they still have certain issues in the long tail of the error curve. For example, local geometric distortions are prone to occur under extreme expressions or atypical poses. In contrast, DCNN-FLAME achieves the smallest non-metric and metric means on the same dataset. This is partly due to FLAME’s explicit modeling of head and neck joints, and partly closely related to the high-dimensional feature constraints provided by the style transfer loss. These factors enable the model to maintain low reconstruction error even under complex expressions and multi-view conditions.

In terms of detail preservation, 3DMM-Fitting, which primarily relies on low-dimensional parameterized shapes and simple texture models, often exhibits over-smoothing in high-frequency regions such as crow's feet, subtle muscle twitches at the corners of the mouth, and

nasolabial folds. It struggles to accurately replicate the minute texture variations in real human faces. DECA enhances detail representation to some extent in terms of normal displacement and texture by explicitly modeling expression details, while FAN-3D introduces high-frequency components through frequency-domain modeling, enabling the visual recovery of more details. However, comparative analysis of image visualization results shows that DCNN-FLAME still demonstrates stronger expressiveness at the detail level: On one hand, the VGG-19-based style transfer loss can constrain high-frequency textures in the semantic feature space, making the reconstructed results more similar to the original images in terms of skin texture, light transition, and local textures. On the other hand, the detail reconstruction network, through dedicated modeling of high-frequency regions, maps texture features from the input image to UV albedo maps and displacement maps. This allows the generated 3D facial mesh to accurately reproduce high-frequency information such as eyebrow edges, pupil highlights, and lip lines while maintaining the stability of low-frequency structures. Compared with DECA and FAN-3D, DCNN-FLAME can still maintain good detail consistency in scenarios with extreme lighting or strong local shadows, indicating that style transfer features play an important role in compensating for the insufficient sensitivity of traditional photometric loss to high-frequency textures. Pose generalization ability is one of the key indicators determining whether 3D face reconstruction methods can operate reliably in practical applications. Due to its use of rigid rotation and simple expression blending models, 3DMM-Fitting fails to accurately describe the relative movement between the head and neck, and is prone to geometric tearing between the jaw and neck in scenarios involving large-angle head turns or head lowering.

DECA and FAN-3D exhibit good stability within the range of moderate pose variations. However, when only single-view input is available and the pose is extreme (e.g., large side profile or upward viewing angle), the reconstruction of "inferred regions" such as the auricle, lateral zygomatic margin, and posterior cranial contour remains unstable. DCNN-FLAME incorporates mechanisms conducive to pose generalization at both the network design and data organization levels: On one hand, FLAME models the head and neck as a joint chain with two degrees of freedom, and performs joint optimization through a large amount of data containing neck scans, thereby improving adaptability to large-angle poses at the model level. On the other hand, a "four-view" input mode is adopted during training, where four facial images of the same identity with different poses or expressions are combined into one training sample. This enables the identity encoder to preferentially learn pose-invariant identity features and expression features that change relatively smoothly with poses. From the experimental results, even with only a single side-profile image input, DCNN-FLAME can still recover 3D structures such as the auricle position, jawline trajectory, and nasal bridge curvature, while maintaining consistent geometric morphology with frontal images in the forehead, zygomatic, and midfacial regions. This indicates that the model outperforms 3DMM-Fitting, DECA, and FAN-3D in pose generalization ability.

In terms of performance trade-off and computational cost, 3DMM-Fitting has the advantage of low parameter dimensions and a relatively simple model structure, resulting in low demands for video memory and computing power. However, it requires iterative optimization to solve parameters, so the computational cost during single-sample inference is not necessarily the smallest, and it is difficult to fully leverage the batch processing advantages of modern Graphics Processing Units (GPUs). DECA and FAN-3D adopt end-to-end deep network structures, typically requiring multi-branch encoders, differentiable rendering modules, and high-dimensional feature mapping. Their inference phase can process input images in batches, achieving high overall throughput in GPU environments, but they have large model parameter scales and high video memory usage. Building on this, DCNN-FLAME further incorporates a style transfer feature extractor, a detail reconstruction network, and a dual-branch feature supervision mechanism. This makes the overall computational volume and video memory overhead of the model during training higher than those of 3DMM-Fitting, and comparable to or even slightly higher than those of DECA and FAN-3D. Nevertheless, DCNN-FLAME decouples coarse reconstruction and detail reconstruction, achieving relatively controllable inference latency through a lightweight detail encoder and a highly reusable style feature extraction network. In a typical GPU environment, the model can improve reconstruction error and detail quality to a level superior to baseline methods while maintaining near-real-time inference speed and reasonable video memory usage, reflecting a typical trade-off of moderately increasing model complexity in exchange for improved reconstruction accuracy and expression stability.

From the perspective of error patterns, each method still exhibits different forms of distortion in specific scenarios. 3DMM-Fitting's geometric errors under non-rigid expressions, exaggerated expressions, and irregular lighting conditions are often concentrated in the mouth, periocular, and nasal alar regions. Expression changes are "pulled back" into the linear subspace, resulting in stiff expressions and insufficient details. Although DECA can well recover expression textures under moderate-intensity expressions, local collapse of the eyelid or corner of the mouth may occasionally occur in scenarios with direct strong light, local occlusion, or extreme poses. FAN-3D enhances the expressive ability of high-frequency information through frequency-domain modeling, but when the input image contains severe noise or motion blur, the amplification of high-frequency components may lead to local texture artifacts. For DCNN-FLAME, despite achieving optimal results in overall reconstruction error and detail preservation, observations from some failure cases show: when the face is largely occluded (e.g., wearing a thick mask or wide-brimmed hat), expressions are extremely exaggerated (e.g., cartoonish laughter, extreme glaring), or there is a significant difference between the training data distribution and the test style, the

model may still produce unnatural texture stretching at occlusion edges or uncertainty in the inference of deep structures.

6 Conclusion

Focusing on the expressiveness issue in animated character expression generation, this study proposes an unsupervised 3D face reconstruction method integrated with image feature enhancement. Taking the FLAME parametric model as the basic framework, the study combines DCNN and style transfer mechanism to construct an end-to-end 3D face reconstruction system—DCNN-FLAME. By introducing the VGG-19-based style transfer loss, the ability to supervise texture details is enhanced at the high-level semantic level, which effectively makes up for the deficiency of traditional photometric loss in capturing high-frequency information. To improve the realism and controllability of expression generation, this study designs a dual-branch feature supervision mechanism: On the one hand, expression classification features are used to provide global semantic constraints, ensuring that the reconstructed expressions conform to the original emotional intention. On the other hand, AU detection features are introduced to realize interpretable expression control. Experimental results show that the proposed method significantly outperforms multiple mainstream reconstruction algorithms on the VGGFace2 dataset, with an F1-score of 0.564, verifying its superior performance in terms of geometric accuracy and expression restoration.

Despite achieving promising results in geometric accuracy and expression restoration, this study still has certain limitations. First, the current model's reconstruction capability for extreme poses (e.g., large-angle head lowering, head raising, or side profiles) needs further improvement. Especially in the absence of multi-view input, there remains uncertainty in the inference of deep facial structures. Second, the AU detection module adopted in this study relies on an externally pre-trained model, and fully end-to-end joint optimization has not been realized, which may to some extent introduce the risk of error propagation. To address the above issues, future work will focus on two aspects: On the one hand, explore an adaptive pose augmentation mechanism combined with synthetic view generation to improve the model's reconstruction robustness under extreme angles. On the other hand, further develop a differentiable AU recognition sub-network for joint training with the 3D reconstruction backbone network. On this basis, this study also plans to introduce the feedback adjustment idea from robust neural adaptive control, explicitly treating expression classification and AU features as feedback signals to perform online or semi-online adaptive correction of FLAME expression parameters and detailed textures. Additionally, Lyapunov-like stability analysis tools will be used to characterize the stability boundaries of the expression generation process under complex pose and lighting disturbances, thereby systematically enhancing the stability and naturalness of animated character expressions under various working conditions.

References

[1] Yang H, Zhu K, Huang D, Li H, Wang Y. Intensity enhancement via GAN for multimodal face expression recognition. Neurocomputing, 2021, 454: 124-134. DOI: 10.1016/j.neucom.2021.05.022.

[2] Tang M, Ling M, Tang J, Hu J. A micro-expression recognition algorithm based on feature enhancement and attention mechanisms. Virtual Reality, 2023, 27(3): 2405-2416. DOI: 10.1007/s10055-023-00808-w.

[3] Bie M, Liu Q, Xu H, Gao Y, Che X. FEMFER: Feature enhancement for multi-faces expression recognition in classroom images. Multimedia Tools and Applications, 2024, 83(2): 6183-6203. DOI: 10.1007/s11042-023-15808-w.

[4] Han F, Zhu S, Ling Q, Han H, Li H, Guo X, et al. Gene-CWGAN: a data enhancement method for gene expression profile based on improved CWGAN-GP. Neural Computing and Applications, 2022, 34(19): 16325-16339. DOI: 10.1007/s00521-022-07417-9.

[5] Chen Z, Yan L, Wang H, Adamyk B. Improved facial expression recognition algorithm based on local feature enhancement and global information association. Electronics, 2024, 13(14): 2813. DOI: 10.3390/electronics13142813.

[6] Baygin M, Tuncer I, Dogan S, Barua P D, Tuncer T, Cheong K, et al. Automated facial expression recognition using exemplar hybrid deep feature generation technique. Soft Computing, 2023, 27(13): 8721-8737. DOI: 10.1007/s00500-023-08230-9.

[7] Wang H, Ma R, Jiang X. Multiscale contextual joint feature enhancement GAN for semantic image synthesis. Image and Vision Computing, 2025: 105637. DOI: 10.1016/j.imavis.2025.105637.

[8] Cao W, Huang Z. Character generation and visual quality enhancement in animated films using deep learning. Scientific Reports, 2025, 15(1): 23409. DOI: 10.1038/s41598-025-07442-3.

[9] Zhao G, Zhang K, Wang L, Zhao W, Zhang W. CIDNet: Cross-Scale Interference Mining Detection Network for underwater object detection. Knowledge-Based Systems, 2025: 113902. DOI: 10.1016/j.knosys.2025.113902.

[10] Liu X, Ni R, Yang B, Song S, Cangelosi A. Unlocking human-like facial expressions in humanoid robots: A novel approach for action unit driven facial expression disentangled synthesis. IEEE Transactions on Robotics, 2024, 40: 3850-3865. DOI: 10.1109/TRO.2024.3422051.

[11] Zeng D, Zhao S, Zhang J, Liu H, Li K. Expression-tailored talking face generation with adaptive cross-modal weighting. Neurocomputing, 2022, 511: 117-130. DOI: 10.1016/j.neucom.2022.09.025.

[12] Mohana M, Subashini P, Krishnaveni M. Emotion recognition from facial expression using hybrid CNN–LSTM network. International Journal of Pattern Recognition and Artificial Intelligence, 2023, 37(08): 2356008. DOI: 10.1142/S0218001423560086.

[13] Krithika L B, Priya G G L. Graph based feature extraction and hybrid classification approach for facial expression recognition. Journal of Ambient Intelligence and Humanized Computing, 2021, 12(2): 2131-2147. DOI: 10.1007/s12652-020-02311-5.

[14] Boulkroune A, Hamel S, Zouari F, et al. Output-Feedback Controller Based Projective Lag-Synchronization of Uncertain Chaotic Systems in the Presence of Input Nonlinearities. Mathematical Problems in Engineering, 2017, 2017(1): 8045803. DOI:10.1155/2017/8045803.

[15] Boulkroune A, Zouari F, Boubellouta A. Adaptive fuzzy control for practical fixed-time synchronization of fractional-order chaotic systems. Journal of Vibration and Control, 2025: DOI: 10775463251320258. 10.1177/10775463251320258.

[16] Zouari F, Saad K B, Benrejeb M. Adaptive backstepping control for a class of uncertain single input single output nonlinear systems.10th International Multi-Conferences on Systems, Signals & Devices 2013 (SSD13). IEEE, 2013: 1-6. DOI: 10.1109/SSD.2013.6564134.

[17] Zouari F, Saad K B, Benrejeb M. Robust neural adaptive control for a class of uncertain nonlinear complex dynamical multivariable systems. International Review on Modelling and Simulations, 2012, 5(5): 2075-2103.

[18] Rigatos G, Abbaszadeh M, Sari B, et al. Nonlinear optimal control for a gas compressor driven by an induction motor. Results in Control and Optimization, 2023, 11: 100226. DOI: 10.1016/j.rico.2023.100226.

[19] Merazka L, Zouari F, Boulkroune A. High-gain observer-based adaptive fuzzy control for a class of multivariable nonlinear systems.2017 6th International Conference on Systems and Control (ICSC). IEEE, 2017: 96-102. DOI: 10.1109/ICoSC.2017.7958728.

[20] Sun Y, Zhang C, Yu F, Xu H, Pan Q. Face attribute transfer fusing feature enhancement and structural diversity loss function. Information Technology and Control, 2024, 53(3): 960-977. DOI: 10.5755/j01.itc.53.3.35213.

[21] Zarif S, Amin K M, Najjar A, et al. Animating text descriptions into characters: A comparative review of generative models. IJCI. International Journal of Computers and Information, 2025, 12(1): 43-66. DOI: 10.21608/ijci.2024.307030.1167.

[22] Bashor C J, Hilton I B, Bandukwala H, Smith D M, Veiseh O. Engineering the next generation of cell-based therapeutics. Nature Reviews Drug Discovery, 2022, 21(9): 655-675. DOI: 10.1038/s41573-022-00476-6.

[23] Ding H, Liu C, Wang S, Jiang X. VLT: Vision-language transformer and query generation for referring segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(6): 7900-7916. DOI: 10.1109/TPAMI.2022.3217852.

[24] Venu V S, Samreen S. Expressive Cartoon Card (ECC) Generation Using Deep Learning for Privacy Preserving Social Media Posts. Traitement du Signal, 2025, 42(3): 1367. DOI:10.18280/ts.420313.

[25] Labanieh L, Mackall C L. CAR immune cells: design principles, resistance and the next generation. Nature, 2023, 614(7949): 635-648. DOI: 10.1038/s41586-023-05707-3.

[26] Wang D, Guo X, Tian Y, Liu J, He L, Luo X. TETFN: A text enhanced transformer fusion network for multimodal sentiment analysis. Pattern Recognition, 2023, 136: 109259. DOI: 10.1016/j.patcog.2022.109259.

[27] Hambarde P, Murala S, Dhall A. UW-GAN: Single-image depth estimation and image enhancement for underwater images. IEEE Transactions on Instrumentation and Measurement, 2021, 70: 1-12. DOI: 10.1109/TIM.2021.3120130.

[28] Jiang Z, Li Z, Yang S, Fan X, Liu R. Target oriented perceptual adversarial fusion network for underwater image enhancement. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(10): 6584-6598. DOI: 10.1109/TCSVT.2022.3174817.

[29] Lee S E, Kim S, Chu Y, et al. EAE-GAN: Emotion-aware emoji generative adversarial network for computational modeling diverse and fine-grained human emotions. IEEE Transactions on Computational Social Systems, 2023, 11(3): 3862-3872. DOI: 10.1109/TCSS.2023.3329434.