CIT. Journal of Computing and Information Technology, Vol. 33, No. 2, June 2025, 91–107
doi: 10.20532/cit.2025.1005965

3D Facial Modeling Based on Multi-scale Feature Fusion and Lighting Robustness

Hongyan Zhu

Sichuan Vocational and Technical College, Suining, Sichuan, China

Facial 3D modeling technology is widely used and has become an important research direction in the fields of artificial intelligence and computer vision. However, the modeling accuracy and robustness of existing technologies in dealing with weak texture areas and complex lighting conditions are insufficient, which limits their practical application in production. Therefore, a facial 3D modeling method based on multi-scale feature fusion and lighting robustness optimization was proposed, and a multi-scale dense feature network and lighting robustness feature fusion network were constructed. The experimental outcomes indicated that the method exhibited excellent performance on the dataset. Among them, structural similarity reached 0.954, and the average absolute error was the lowest at 0.63 mm. Under dynamic lighting conditions, the feature consistency reached 0.941, and the point cloud error was reduced to 0.85 mm. In addition, tests in security and virtual reality scenarios showed that after using this method, the accuracy increased to 92.8%, the peak signal-to-noise ratio reached 33.0 dB, and the model running efficiency improved to 36 frames per second, verifying the practicality and reliability of the method. The research provides new ideas for developing stable, efficient, and practical facial 3D modeling methods, which is expected to promote the widespread application of related technologies in complex environments.

ACM CCS (2012) Classification: Computing methodologies → Computer graphics → Shape modeling → Shape analysis

Keywords: 3D facial modeling, multiple scale, feature fusion, dynamic lighting, robustness

1. Introduction

With the acceleration of human society towards digitization and intelligence, facial 3D modeling technology has become one of the core technologies that has attracted much attention due to its wide application in security monitoring, medical diagnosis, virtual reality and other fields [1–2]. Facial 3D modeling restores the 3D form of objects from 2D images, providing richer and more accurate biometric information for various production practices, and can enhance the technological level of various industries [3]. For example, in security, 3D facial modeling can significantly improve the accuracy and robustness of identity authentication systems under complex lighting conditions and exhibits stronger anti-interference ability compared to traditional 2D recognition [4]. In the medical industry, 3D modeling technology provides precise and intuitive auxiliary support for simulation design and postoperative evaluation of plastic surgery [5]. In the fields of virtual reality and entertainment, high-precision 3D facial models have laid the technological foundation for creating virtual images and providing users with immersive experiences [6]. However, current technology still faces challenges such as low recognition accuracy in weak texture areas, multiple lighting conditions, and difficulty in recognition in dynamic scenes. Previous studies have utilized binocular stereo vision and deep learning techniques to optimize the 3D modeling of faces, but the matching accuracy is insufficient under smooth areas and lighting

conditions, making it difficult to meet practical production needs [7]. Meanwhile, although some deep learning algorithms can achieve matching under single frame conditions, their performance in handling multi-scale features and enhancing lighting robustness is limited [8]. In addition, many studies focus on the design of theoretical algorithms, with limited validation of their adaptation in actual production, which hinders the widespread application of technology in complex environments. In view of this, a facial 3D modeling method combining multi-scale feature fusion and lighting robustness optimization is proposed. It adopts Multi-scale Dense Feature Network (MSDF-Net) and Illumination-Robust Feature Fusion Network (IRFF-Net), and improves modeling accuracy and adaptability to complex lighting conditions by introducing multi-scale feature pooling and group correlation matching strategies. The aim of this research is to create a reliable, efficient, and practical facial 3D modeling method that meets the diverse application needs of medical, security, and entertainment fields.

2. Literature Review

Facial 3D modeling, as a key technology that can provide rich biometric information, is of great significance in supporting identity verification, expression analysis, and interactive applications. As a result, numerous scholars have undertaken extensive studies on facial 3D modeling, constantly exploring more efficient and accurate algorithms and technologies. D'Ettorre *et al.* proposed a method of using a smartphone application that supports TrueDepth system for facial scanning, which addresses the issues of large size, high price, and complex operation of 3D facial scanning devices. This approach reduced device costs, improved portability, and enhanced operational flexibility [9]. Mehta *et al.* proposed a method based on a 3D dense connected self-attention neural network to accurately evaluate the participation of students in online education. This method identified the emotional state and participation of students by analyzing their facial expressions through 3D facial modeling, thereby achieving efficient monitoring and evaluation of their learning status [10]. Florkow *et al.* proposed techniques such as short echo time acquisition and post-processing to generate synthetic computed tomography scans to address the problem of low hard tissue signals, thereby optimizing high contrast imaging and 3D modeling of hard tissues, providing support for improving the modeling accuracy of facial bone structures [11]. Luo *et al.* proposed a multimodal perception analysis method to address the issue of real-time assessment of students' learning interests. By using head pose estimation, facial expression recognition, and interactive data fusion, a 3D learning interest model was constructed to optimize interest assessment capabilities, providing support for the application of 3D facial modeling in educational settings [12].

In addition, Chen *et al.* raised an end-to-end reconstruction method that combines cross domain face synthesis conditional generative adversarial networks and grid transformers to address the problem of limited performance in single image face 3D reconstruction due to the lack of 3D annotated data. This optimized the ability to construct face 3D models based on real, artificial synthesis, and field image training [13]. Sun *et al.* proposed a network based on 3D facial feature reconstruction and learning to address the issue of facial expression recognition being affected by posture, lighting, and occlusion in outdoor environments. By reconstructing 2D frontal facial data and 3D facial geometric features, the network integrated appearance path and geometric path features, thereby improving the accuracy and robustness of facial expression recognition in complex scenes [14]. Munir *et al.* proposed a method combining convolutional neural networks and deep learning frameworks to optimize the accuracy and robustness of 3D modeling of faces and facial hair in complex scenes, addressing the difficulty of 3D modeling and reconstruction of facial details caused by interference factors in a single image. This provided strong support for achieving high-quality 3D modeling of faces [15]. Zhou *et al.* proposed a network based on attention and data augmentation to tackle the issue of insufficient feature recognition in unconstrained environments for facial 3D modeling in single view images. By adaptively recalibrating attention weights to improve feature recognition, the accuracy and generalization ability of facial 3D modeling

were optimized, providing an effective solution for 3D modeling in complex scenes [16].

In summary, existing research has achieved notable advancements in the area of facial 3D modeling. However, current research still has shortcomings in dealing with complex lighting conditions and weak texture areas, thus they cannot fully meet the requirements of facial 3D modeling in regard to accuracy, robustness, and real-time performance. Therefore, the study proposes a facial 3D modeling method based on multi-scale feature fusion and lighting robustness. The novelty of the research consists in improving the feature extraction ability of weak texture regions through multi-scale feature pooling technology, and introducing group correlation matching strategy to enhance robustness under complex lighting conditions. In addition, the adaptability of the model in different scenarios is optimized by combining data diversity expansion.

3. Research Methodology

This section provides a detailed explanation of the proposed facial 3D modeling method based on multi-scale feature fusion and lighting robustness. The research method is divided into two core networks, among which MSDF-Net achieves multi-scale feature extraction and fusion, improving the modeling accuracy of weak texture regions. The adaptability of the IRFF-Net optimization model to complex lighting conditions further enhances the robustness and accuracy of the modeling.

3.1. MSDF-Net Based on Multi-scale Feature Extraction and Fusion

In the research of facial 3D modeling, complex lighting conditions and weak texture areas have always been the main challenges affecting modeling accuracy. These issues not only limit the accurate restoration of facial details by the model but also reduce its applicability in practical scenarios [17–18]. Therefore, to address the shortcomings of traditional methods in feature extraction and lighting adaptability, a facial 3D modeling method combining multi-scale feature fusion and lighting robustness optimization has been proposed, based on two core networks, MSDF-Net and IRFF-Net. The core goal of MSDF-Net is to enhance the feature extraction capability for weak texture regions through multi-scale feature fusion, thereby improving the overall accuracy of face modeling. The workflow of MSDF-Net based on multi-scale feature extraction and fusion is shown in Figure 1.

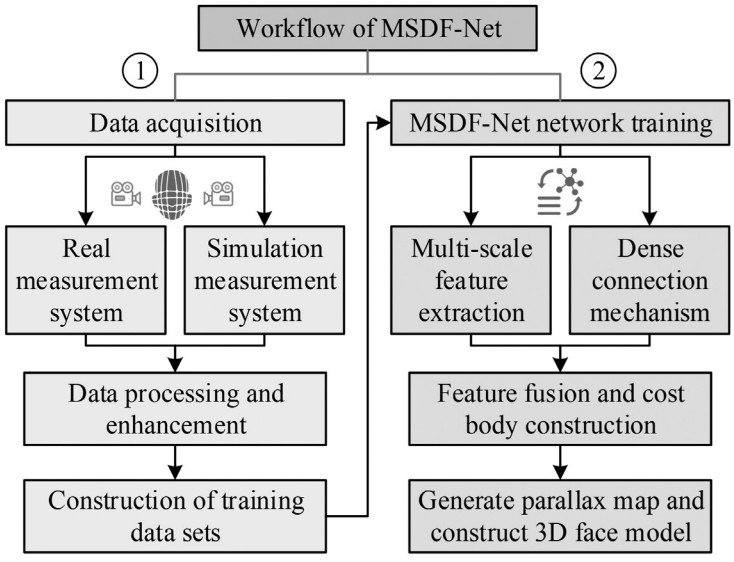


Figure 1. Schematic diagram of the workflow for MSDF-Net.

As shown in Figure 1, the entire workflow is divided into two parts: data acquisition and network training. During the data acquisition phase, left and right view images are obtained through real measurement systems and simulation measurement systems. The real measurement system utilizes cameras and projectors to capture high-quality input data, while the simulation measurement system generates diverse synthetic data through virtual simulation. It is combined with noise enhancement techniques to expand the adaptability and robustness of the dataset, ultimately constructing a high-quality training dataset that includes left and right views and ground truth values. In the network training phase, the left and right view data are input into the MSDF-Net network, and local and global information is captured through a multi-scale feature extraction module. The dense connection mechanism is used to efficiently fuse features of different scales and optimize the construction of cost bodies. After generating the disparity map through the network, the high-precision 3D model of the target is further output through the 3D reconstruction module to improve the performance of facial 3D modeling in weak texture areas and complex lighting conditions. The specific details of data acquisition are shown in Figure 2.

As shown in Figure 2, the data acquisition process includes four steps: hardware setup, pattern projection, image acquisition, and preliminary data processing. Firstly, in the hardware setup phase, the fixed angles, focal lengths, and projection ranges of the left and right cameras and projectors are adjusted to ensure complete coverage of the projection pattern and seamless connection of the field of view. In the pattern projection stage, the projector sequentially projects high-precision generated speckle and stripe patterns onto the surface of the target object, enhancing the feature expression of weak texture areas, and capturing the geometric structure of the object through stripe phase changes [19]. In the image acquisition stage, the left and right cameras work synchronously with the projector to record the reflection images of the target surface at a high frame rate, generating the original image sequence of the left and right views [20]. Finally, in the preliminary data processing stage, the left and right views are denoised, aligned, and cropped to extract speckle and stripe details from the projected patterns, providing support for constructing high-precision ground truth values. The MSDF-Net network structure is shown in Figure 3.

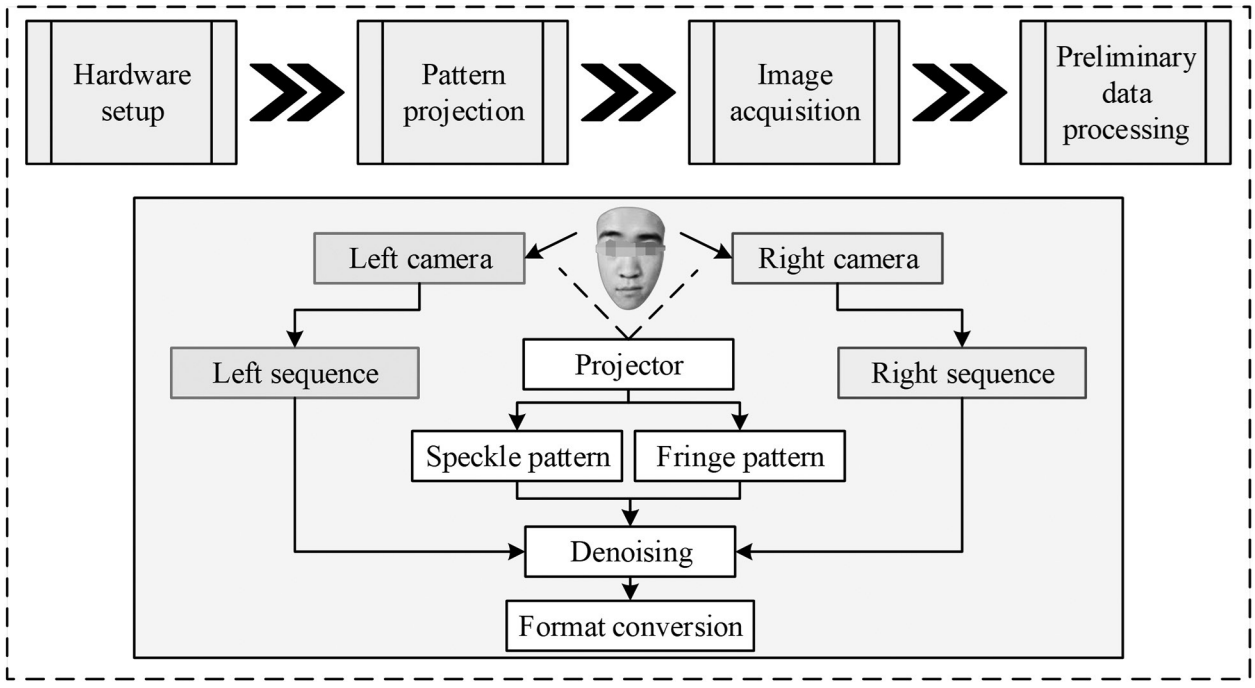


Figure 2. Detailed process of data acquisition.

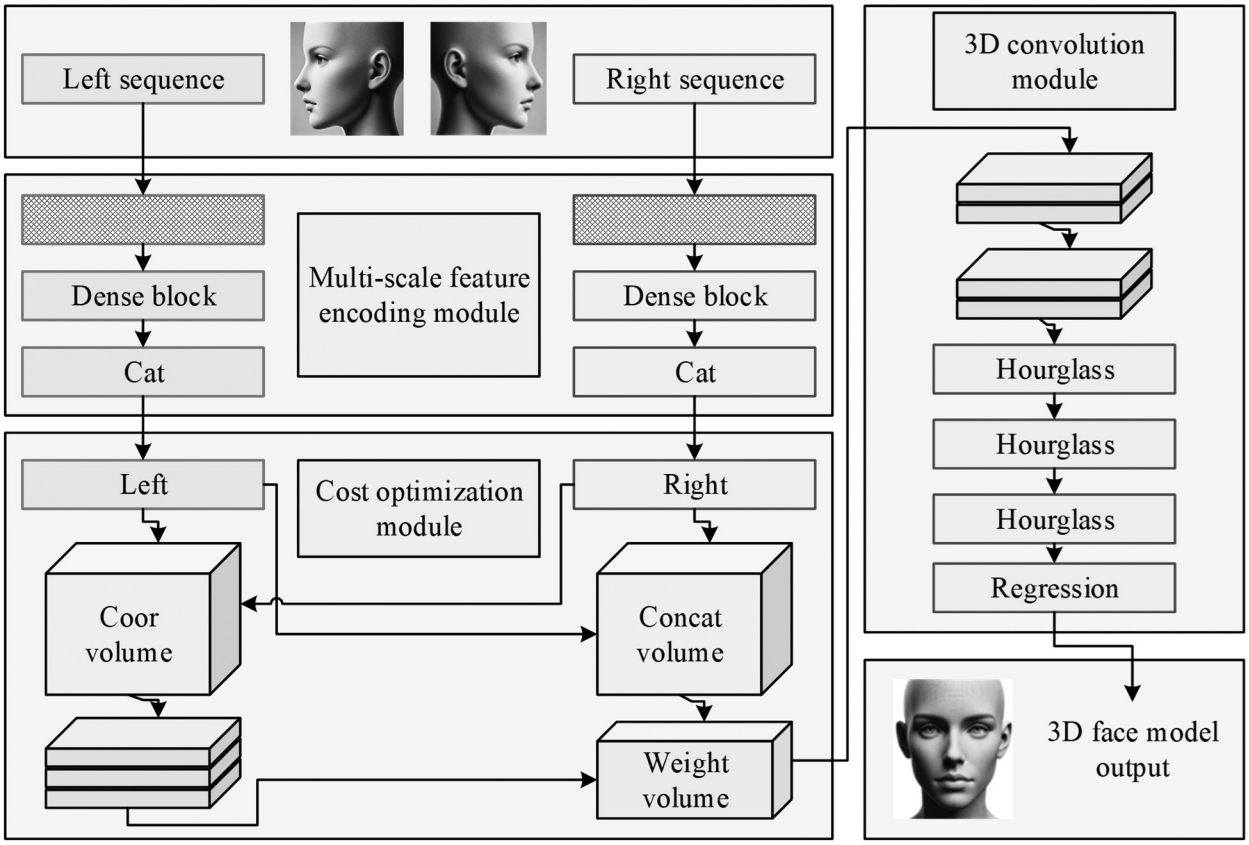


Figure 3. Schematic diagram of MSDF-Net network structure.

In Figure 3, the MSDF-Net network consists of a multi-scale feature encoding module, a cost optimization module, and a three-dimensional convolution module. MSDF-Net first generates multi-scale feature representations through a multi-scale feature encoding module, integrating local details and global contextual information. Subsequently, in the cost optimization module, the left and right view features are matched, and the feature representation is further optimized by concatenating cost bodies and weight cost bodies. Finally, MSDF-Net utilizes 3D convolution modules to refine the cost volume and generate high-precision disparity maps for subsequent 3D point cloud reconstruction. For left and right input images $I_L(x, y)$ and $I_R(x, y)$, features are extracted layer by layer through dense blocks, and each layer's feature map is represented as equation (1).

$$F_i^L = \sigma\left(W_i * F_{i-1}^L + b_i + \sum_{j=1}^{i-1} g_{ij} F_j^L\right), i = 1, 2, ..., n \quad (1)$$

In equation (1), F_i^L and F_{i-1}^L respectively represent the left view features of layer i and layer $i-1$, with the formula for extracting right view features being similar. W_i and b_i respectively represent the weight matrix and bias of the i-th layer convolution, g_{ij} represents the feature reuse weight in dense connections, used to control the contribution of the previous j layer features to the current layer, and n represents the total number of layers in the dense block [21]. The extracted multi-scale features are upsampled and fused, as shown in equation (2).

$$\begin{cases} \widetilde{F}_L = \sum_{s=1}^{S} u_s \cdot \text{Upsample}\left(F_L^s\right) \\ \widetilde{F}_R = \sum_{s=1}^{S} u_s \cdot \text{Upsample}\left(F_R^s\right) \end{cases} \quad (2)$$

In equation (2), $\widetilde{F}_L$ and $\widetilde{F}_R$ respectively represent the fused left and right view features, and F_L^s and F_R^s respectively represent the features of the left and right views at the s-th scale, S represents the number of multi-scale features,

u_s represents the weight of each scale, and *Upsample* represents the upsampling operation. The left and right view features are matched through a cost optimization module. The calculation of the initial matching cost body is shown in equation (3).

$$C_{raw}(x,\ y,\ d) = \frac{\left\| \widetilde{F}_L(x,y) - \widetilde{F}_R(x-d,y) \right\|^2}{\left\| \widetilde{F}_L(x,y) \right\|_2 \cdot \left\| \widetilde{F}_L(x-d,y) \right\|_2} \quad (3)$$

In equation (3), $C_{raw}(x, y, d)$ represents the preliminary matching cost volume, which is the similarity of the left and right view features under disparity d. To further optimize the cost body, the study introduces a concatenated cost body $V_{\text{concat}}(x, y, d)$, as shown in equation (4).

$$\begin{gathered} V_{\text{concat}}(x,y,d) = \\ W_{cat} * \text{Concat}\left(\widetilde{F}_L(x,y), \widetilde{F}_R(x-d,y) \right) + b_{cat} \end{gathered} \quad (4)$$

In equation (4), W_{cat} and b_{cat} are respectively the weights and bias terms of the concatenated convolution. The weight cost body $V_{weight}(x, y, d)$ is used to dynamically adjust the matching result of the cost body, as shown in equation (5).

$$\begin{gathered} V_{weight}(x,y,d) = \\ \text{Sigmoid}\left(W_w * V_{\text{concat}}(x,y,d) + b_w \right) \end{gathered} \quad (5)$$

In equation (5), *Sigmoid* is used to restrict within the range of [0, 1]. W_w and b_w represent convolution kernels and biases, respectively. The final cost body $V_{final}(x, y, d)$ is represented as shown in equation (6).

$$V_{final}(x, y, d) = C_{raw}(x, y, d) \cdot V_{weight}(x, y, d) \quad (6)$$

Furthermore, $V_{final}(x, y, d)$ is input into the 3D convolution module and the cost volume will be optimized through the Hourglass network. After three-dimensional convolution and deconvolution, multi-stage disparity prediction results are generated, as shown in equation (7).

$$\widehat{D}_k(x,y) = \arg\min_d \widehat{C}_k(x,y,d),\ k = 1,\ 2,\ ...,\ K \quad (7)$$

In equation (7), $\widehat{D}_k(x,y)$ and $\widehat{C}_k(x,y,d)$ respectively represent the disparity prediction and cost volume output of the k-th stage, and K indicates the total number of stages. The final disparity map $D(x, y)$ is generated through weighted fusion of multi-stage outputs, as shown in equation (8).

$$D(x,y) = \sum_{k=1}^{K} \alpha_k \cdot \widehat{D}_k(x,y) \quad (8)$$

In equation (8), α_k represents the weighting coefficient of the k-th stage.

3.2. IRFF-Net for Light Robustness Optimization

MSDF-Net can improve the accuracy of 3D facial modeling, but under complex lighting conditions, due to changes in light source intensity, orientation, and distribution, the features of the left and right views may exhibit inconsistency, such as local overexposure, loss of information in shaded areas, *etc.*, which increases the difficulty of feature matching [22–23]. Therefore, further research is needed to design IRFF-Net to enhance the adaptability of facial 3D modeling to complex lighting conditions. IRFF-Net optimizes the feature matching process through group correlation matching and multi-level block feature fusion strategies, and improves the robustness of the cost volume, thereby achieving more accurate facial 3D modeling in dynamic lighting scenes. The network structure of IRFF-Net is shown in Figure 4.

As shown in Figure 4, the difference between the network structure of IRFF-Net and MSDF-Net lies in the group correlation matching module and the multi-level block feature fusion module. Specifically, IRFF-Net still includes a multi-scale feature encoding module for capturing local details and global contextual information. On this basis, the group correlation matching module groups the extracted left and right view features, enhances the robustness of feature matching through intra group correlation calculation, and reduces the interference of lighting changes on the matching results. Subsequently, the multi-level block feature fusion module divides the features into blocks of different scales, optimizes the matching results by combining the concatenation cost body and the weight cost body, and fuses the feature representations of multi-level blocks through a weighted strategy. Finally, the optimized cost volume is input into the 3D convolution module to generate high-precision disparity maps, providing

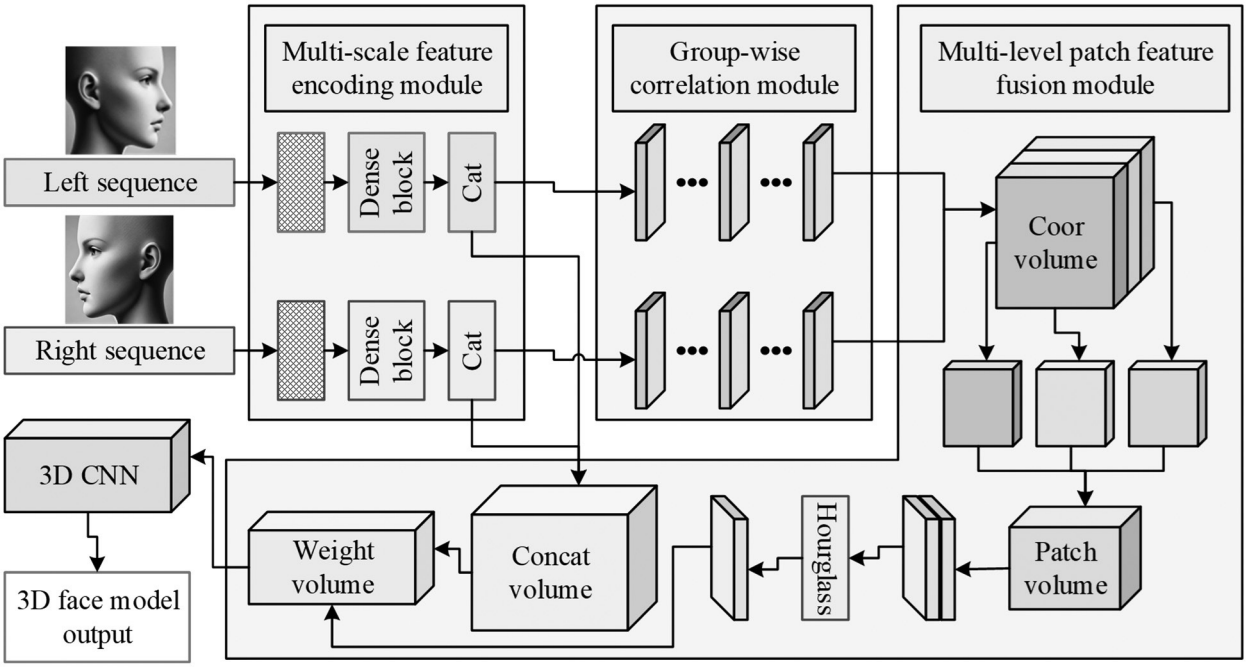


Figure 4. Schematic diagram of IRFF-Net network structure.

stronger robustness and adaptability for 3D facial modeling under complex lighting conditions.

In the group correlation matching module, to raise the robustness of left and right view feature matching under complex lighting conditions, a method of grouping processing and correlation calculation is studied. The multi-scale feature maps extracted from the left and right views are divided into channels to form several groups. Each group of features is aligned according to disparity, and matching proxy values are generated through intra group correlation calculation [24]. The goal of correlation calculation is to measure the degree of matching between left and right views under different disparities, and to reduce the interference of lighting changes on the matching outcomes. The schematic diagram of intra group correlation calculation is shown in Figure 5.

From Figure 5, the multi-channel feature maps of the left and right views are represented as $F_L(x, y)$ and $F_R(x, y)$, respectively. The dimension of each feature map is $C \times H \times W$, where C is the number of channels. H and W are the height and width of the feature map. During the feature alignment process, the right view feature map is offset in the width direction based on disparity $D = d$ to form an aligned feature map. Subsequently, the correlation within each group g of the left and right feature maps is calculated under a specific disparity, as shown in equation (9).

$$C_{group}(x,y,g,d) = \frac{\sum_{k\in g} F_L^k(x,y)\cdot F_R^k(x-d,y)}{\sqrt{\sum_{k\in g}\left(F_L^k(x,y)\right)^2 \cdot \sum_{k\in g}\left(F_R^k(x-d,y)\right)^2}} \quad (9)$$

In equation (9), $C_{group}(x, y, g, d)$ represents the correlation value of the group g under a specific disparity d, $F_L^k(x, y)$ and $F_R^k(x - d, y)$ are the feature values of the left and right view channels, respectively. The denominator of equation (9) is normalized to eliminate the scale influence of the feature values. After the group related matching module generated the preliminary group related cost body, the study further optimized the cost body using a multi-level block feature fusion module. The study introduces multi-level dilated convolution technology into the multi-level block feature fusion module, dynamically expanding the receptive field range by adjusting the expansion rate of the convolution kernel, thus balancing local details and overall contextual information in lighting changing scenes. Multi-level dilated convolution is shown in Figure 6.

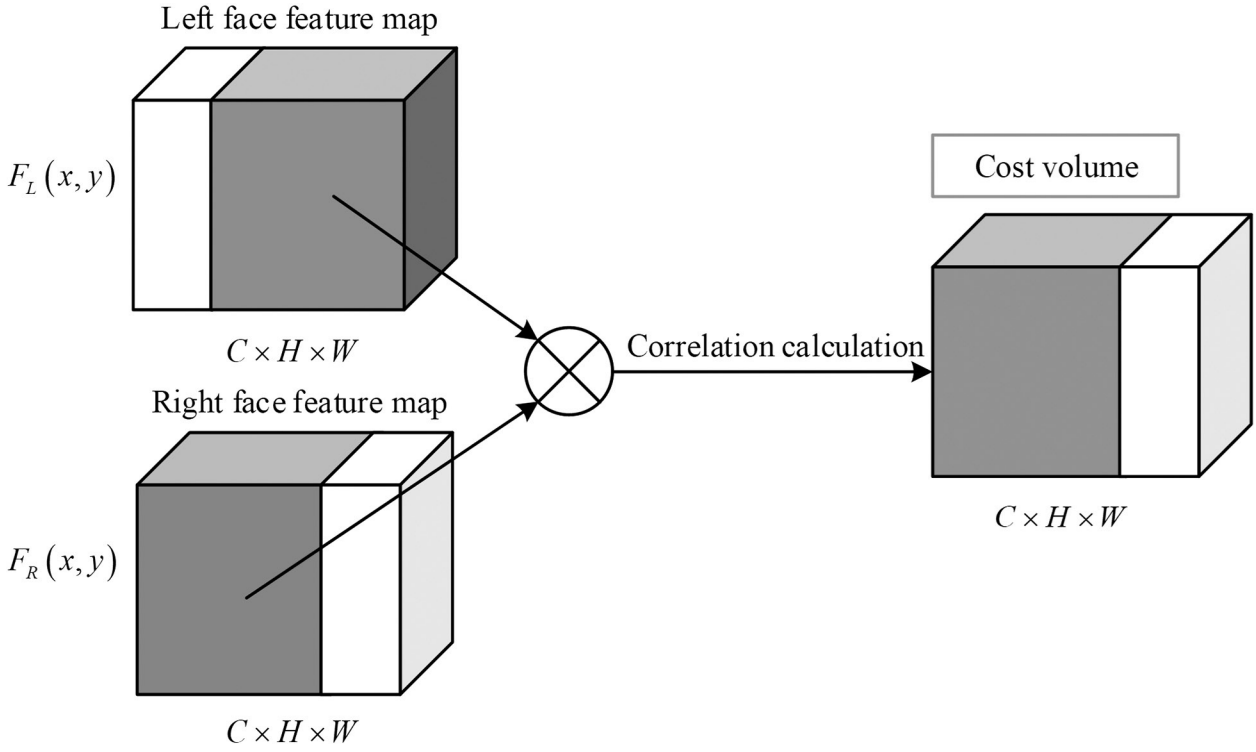


Figure 5. Correlation calculation diagram.

As shown in Figure 6, the multi-level dilated convolution consists of three convolution operations with different dilation rates, including the basic convolution with dilation rate 1, the medium dilation convolution with dilation rate 2, and the large dilation convolution with dilation rate 3. The features extracted by each convolution operation can be represented as shown in equation (10).

$$F_i^{(p)}(x, y) = \sigma(W_i^{(p)} * F_{\text{input}}(x, y) + b_i^{(p)}), \quad p = \{1, 2, 3\} \tag{10}$$

In equation (10), $F_{\text{input}}(x, y)$ represents the feature map of the input block, $F_i^{(p)}(x, y)$ represents

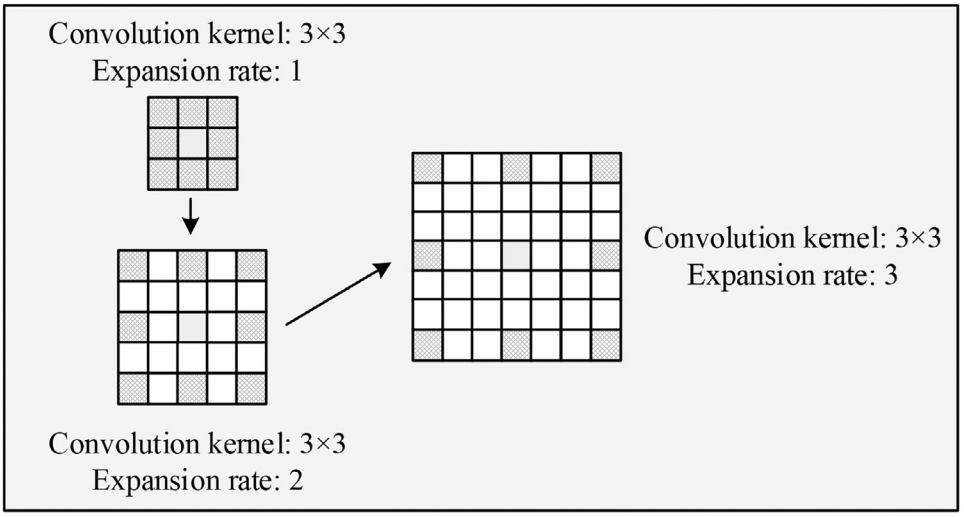


Figure 6. Schematic diagram of multi-level cavity convolution.

the features extracted by the dilated convolution with an expansion rate of p, $W_i^{(p)}$ and $b_i^{(p)}$ represent the corresponding convolution kernel weights and biases, respectively. Hollow convolutions with different dilation rates can capture feature information of different scales within a block [25–26]. After fusing the features extracted from various levels of dilated convolutions, it is possible to generate multi-scale feature representations within the block. Therefore, through the normalization processing of the group correlation matching module and the multi-scale dilated convolution of the multi-level block feature fusion module, IRFF-Net can effectively reduce the interference of lighting changes on the matching of left and right view features, achieving more stable and accurate facial 3D modeling in dynamic lighting scenes.

4. Results and Discussion

To verify the effectiveness and superiority of the proposed facial 3D modeling method, a large number of experiments were conducted using the BU3D-FE dataset and 3DFAW dataset as data sources. Among them, the BU3D-FE dataset can provide high-quality 3D facial scanning data containing different expression changes, which helps evaluate the performance of the method in modeling expression and weak texture areas. The 3DFAW dataset can provide rich real-world facial images and 3D annotation information, including samples with complex lighting conditions and dynamic pose changes, to verify the lighting robustness and adaptability of the method in practical scenarios. The specific configuration and parameter design for the experiment are shown in Table 1.

Table 1. Experimental environment configuration and network parameter design.

/	**Category**	**Configuration**	/	**Category**	**Parameter**
Hardware configuration	CPU	Intel Core i9-12900K	Network parameters	Learning rate	0.001
	GPU	NVIDIA RTX 3090		Optimizer	Adam
	Memory	128 GB DDR4		Batch size	32
	Storage	2 TB SSD		Training Epochs	100
Software configuration	Operating system	Ubuntu 20.04		Weight initialization	Xavier Initialization
	Deep learning Framework	PyTorch 1.13		Number of Multi-scale features	4
	Programming language	Python 3.9		Dilation rates for atrous convolution	[1, 2, 3]
	Other libraries	NumPy, OpenCV, Matplotlib		Weighted cost volume fusion weights	Dynamic adjustment

As can be seen from Table 1, the Adam optimizer and Xavier initialization were used to optimize the model. The learning rate was initially set at 0.001, and the cosine annealing strategy was used for dynamic adjustment. In the training process, PSNR and feature consistency are used as the model selection criteria, and the model with the best performance on the verification set is selected for testing. In addition, the cavity convolution expansion rate is set to [1, 2, 3], which is determined by preliminary experiments, aiming at balancing local details with global receptive field. The number of multi-scale features is set to 4, and the modeling accuracy is improved by dynamic weighted fusion of each scale feature. The cost-body fusion module introduces an adaptive weight mechanism to enhance the matching robustness under different parallax levels. The optimization strategy not only ensures the modeling accuracy, but also effectively controls the model complexity and training cost.

On the basis of Table 1, the study first verified the performance of MSDF-Net. Semi-Global Matching (SGM), Pyramid Stereo Matching Network (PSMNet), and Residual Regression Network (RRNet) were selected as comparison methods, and the experiment outcomes are shown in Figure 7.

As shown in Figure 7 (a), in the comparison of Structural Similarity Index Measure (SSIM), MSDF-Net consistently scored higher than other comparison methods. In the high visual difference range of 80–128 pixels, SSIM reached 0.949 to 0.954, which was better than SGM's 0.861 to 0.870 and PSMNet's 0.902 to 0.911. As shown in Figure 7 (b), in the Mean Absolute Error (MAE) metric, MSDF-Net exhibited the lowest reconstruction error across all disparity ranges. Within the high disparity range of 80–128 pixels, MAE decreased from 0.72 mm to 0.63 mm, demonstrating stronger modeling accuracy and error convergence ability. Other methods such as PSMNet and RRNet had errors of 0.86 mm and 0.73 mm, respectively, within the disparity range of 112–128 pixels. Therefore, MSDF-Net's multi-scale feature extraction and fusion strategy could better capture the structural information of weak texture regions. Furthermore, robustness testing was conducted under varying lighting conditions, and the Photometric Stereo Reconstruction (PSR) method was selected as a supplementary approach for comparison with IRFF-Net. The results are shown in Figure 8.

According to Figure 8 (a), in strong light scenes, the average point cloud error of IRFF-Net was the lowest, being only 0.85 mm, which

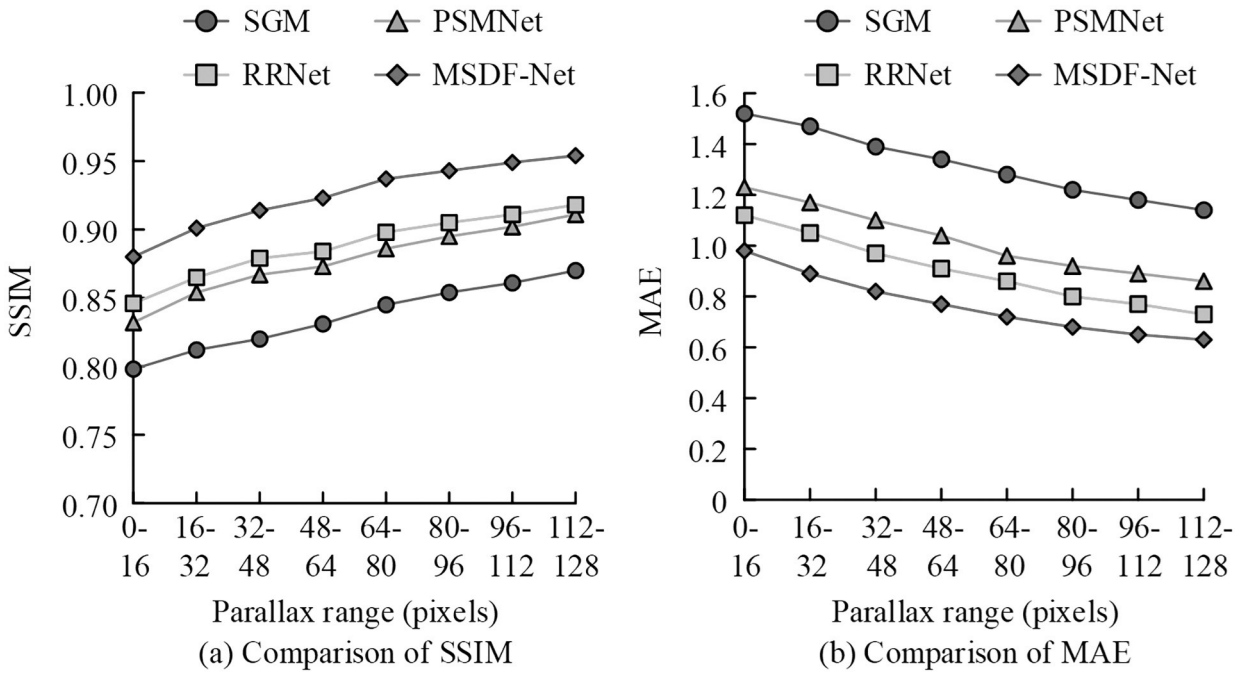


Figure 7. Performance verification results of MSDF-Net.

was 13% lower than the suboptimal method RRNet's 0.98 mm. The feature consistency of IRFF-Net reached 0.941, which performed the best among all methods. As shown in Figure 8 (b), the performance of various methods generally decreased in low light scenes, but IRFF-Net still had the best performance with an error of only 0.92 mm and a feature consistency of 0.924. As shown in Figure 8 (c), in the shadow changing scene, the error of IRFF-Net was 1.01mm, which was still lower than all comparison methods, and the feature consistency was 0.908, which was higher than other methods. From this, IRFF-Net consistently exhibited the lowest MSE and the highest feature consistency in three lighting scenarios, demonstrating the advantages of its group correlation matching and multi-level block feature fusion strategies in dealing with complex lighting conditions. On this basis, ablation experiments were conducted, and the results are shown in Table 2.

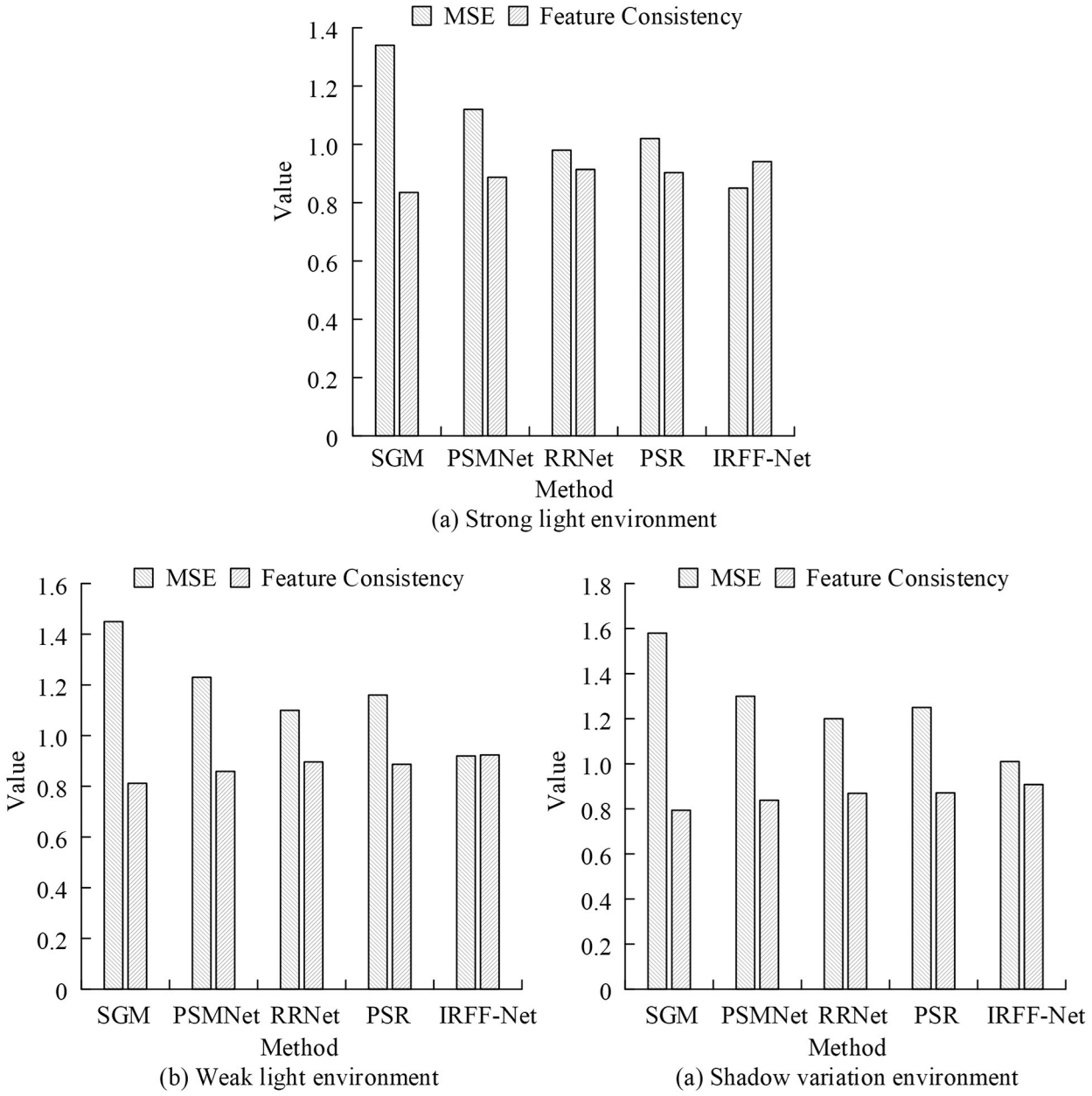


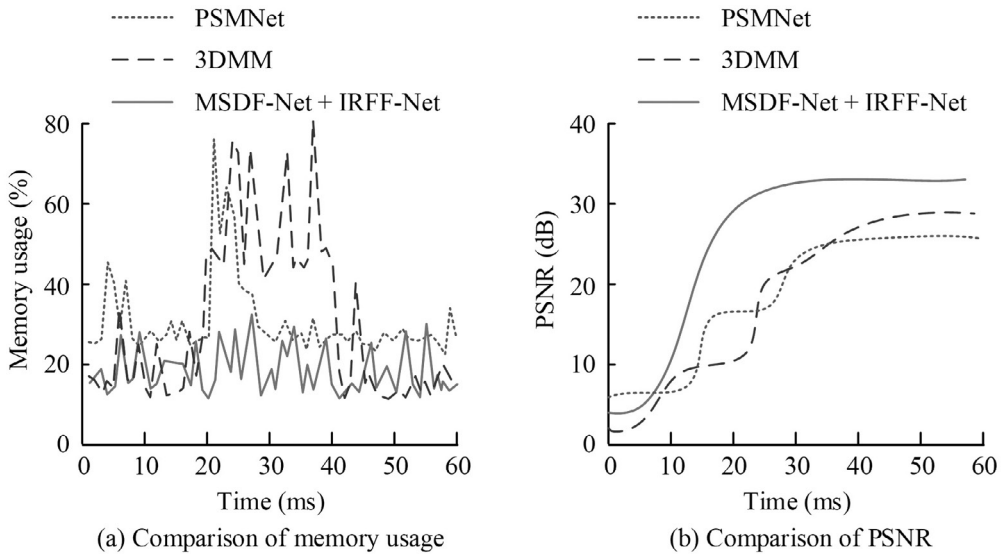
Figure 8. Robustness test under illumination variation.

According to Table 2, the complete network MSDF-Net+IRFF-Net performed the best in terms of Peak Signal to Noise Ratio (PSNR), MSE, and feature consistency, with a PSNR of 32.1 dB and a minimum MSE of 0.72 mm. The feature consistency and weak texture region feature extraction accuracy in the complete network reached 0.941 and 91.3%, respectively. In contrast, after removing modules from MSDF-Net or IRFF-Net, PSNR and feature consistency both decreased, while MSE significantly increased, indicating the crucial role that these modules played in feature extraction and matching. Although the complete network slightly increased runtime, its performance improvement fully demonstrated the irreplaceability of each module. Overall, the synergistic effect of each module and strategy was the key to achieving high-precision, low error, and strong robustness in 3D modeling. Furthermore, a comprehensive performance comparison of end-to-end networks was conducted, and PSMNet based on binocular vision and 3D Morphable Model (3DMM) based on monocular reconstruction were selected as the comparison methods among the current mainstream 3D modeling methods. The results are shown in Figure 9.

According to Figure 9 (a), the overall memory usage of PSMNet fluctuated between 30% and 40% throughout the entire 60-second running process. At 20–30 ms, the memory usage soared to 78%. The memory usage of 3DMM exceeded 50% within 20–40 ms. The memory usage of the complete network MSDF-Net+IRFF-Net proposed in the study remained between 15% and 35%. According to Figure 9 (b), the PSNR of PSMNet and 3DMM during operation ultimately reached 25.8 and 28.3, respectively, while the PSNR of the proposed complete network MSDF-Net+IRFF-Net reached 32.1. From this, the proposed multi-scale feature fusion and lighting robustness optimization strategy demonstrated superiority in improving model-

Table 2. Results of ablation experiment.

Configuration	PSNR (dB)	MSE(mm)	Runtime (ms)	Feature Consistency	Feature extraction accuracy in weak-texture regions (%)
Complete network (MSDF-Net + IRFF-Net)	32.1	0.72	85	0.941	91.3
Only MSDF-Net	30.4	0.91	78	0.885	89.7
Only IRFF-Net	30.9	0.87	82	0.903	85.2
Without cost Optimization module (MSDF-Net only)	29.6	0.95	72	0.862	81.8
Without group-wise correlation module (IRFF-Net only)	30.2	0.90	78	0.879	84.7
Without multi-level patch feature fusion module (IRFF-Net only)	30.1	0.92	79	0.881	83.6

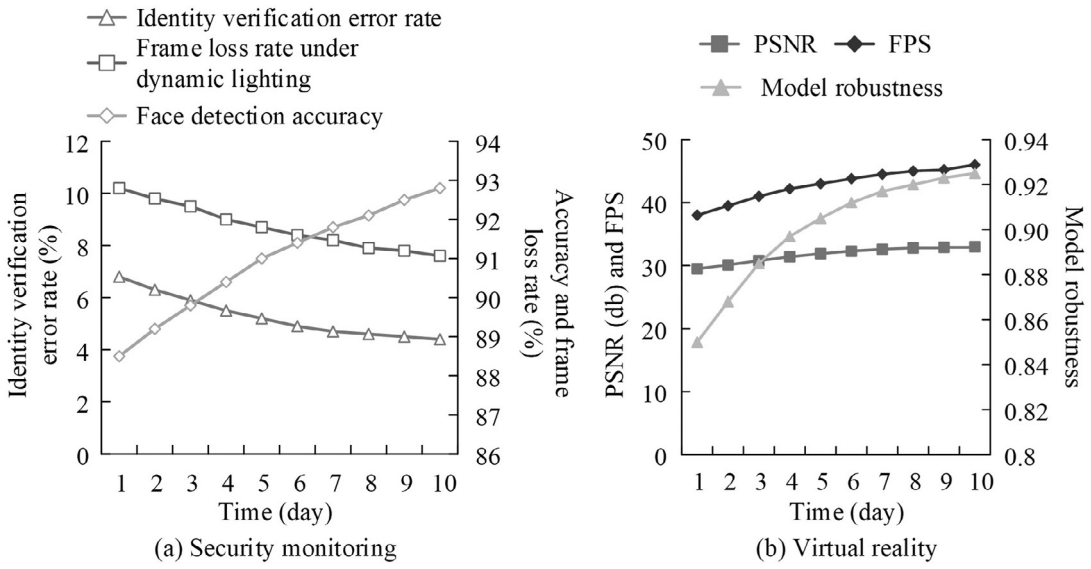


(a) Comparison of memory usage (b) Comparison of PSNR

Figure 9. Overall performance comparison of the end-to-end network.

ing performance and resource efficiency. Then, the applicability of the proposed method in two practical scenarios, security monitoring and virtual reality face modeling, was verified through research. The results are shown in Figure 10.

According to Figure 10 (a), in security monitoring, from the first day to the tenth day of the experiment, the accuracy steadily increased from 88.5% to 92.8%, the error rate decreased from 6.8% to 4.4%, and the frame loss rate under dynamic lighting decreased from 10.2% to 7.6%. This verified the adaptive optimization ability and stability and reliability of the method in longterm use. According to Figure 10 (b), in the virtual reality scene, the PSNR increased from 29.5 dB on day 1 to 33.0 dB on day 10, the model robustness increased from 85.0% to 92.5%, and the operating efficiency increased from 28FPS to 36FPS. This indicated that the method could adaptively adjust network parameters and handle dynamic changes in complex scenes, providing reliable technical support for virtual reality applications.



(a) Security monitoring (b) Virtual reality

Figure 10. Comparison of applicability in real-world scenarios.

Finally, in order to further verify the performance of the proposed method in practical applications, two typical scenarios are designed again, namely, medical-assisted modeling and virtual reality dynamic interaction. Among them, the medical-assisted modeling simulates the 3D facial reconstruction before plastic surgery in medical treatment, and there are many smooth areas and shadow occlusion in the scene. Virtual reality dynamic interactive simulation of real-time face driving and expression capture in VR. Stereo Transformer (STTR), a new stereo matching method based on Transformer architecture is selected for comparison. The experimental results are shown in Table 3.

As can be seen from Table 3, the proposed method shows excellent comprehensive performance in both medically assisted modeling and virtual reality interaction scenarios. In the medical scenario, the proposed method achieved the minimum modeling error of 0.68 mm, a peak signal-to-noise ratio of 32.7 dB in image quality, a frame rate of 33 frames per second, 134G FLOPs, and 27% memory consumption, which ensured high modeling accuracy and strong operating efficiency. In the virtual reality interactive scene, the method in this paper also maintains the lead. The modeling error is 0.72 mm, the PSNR reaches 33.0 dB, the frame rate is increased to 36 frames per second, and the memory consumption is further reduced to 24%. In contrast, although STTR is based on Transformer structure, due to high computational complexity, its accuracy and real-time performance in the two scenarios are not as good as the proposed method. PSMNet and 3DMM are also weak in terms of accuracy, speed, and resource consumption. In general, the proposed method achieves a good balance between precision, real-time and resource efficiency, and shows strong practicability and application potential.

Table 3. Performance Comparison of Different Methods in Real-World Scenarios.

Scenario	Method	MAE (mm)	PSNR (dB)	FPS	FLOPs (G)	Parameters (M)	Memory Usage (%)
Medical Reconstruction	Proposed	0.68	32.7	33	134	28.4	27
	STTR	0.81	30.3	22	176	41.2	26
	PSMNet	0.89	29.4	25	162	35.7	40
	3DMM	1.15	28.2	29	98	15.9	51
VR Interaction	Proposed	0.72	33.0	36	134	28.6	24
	STTR	0.86	30.9	20	176	41.2	38
	PSMNet	0.94	29.5	27	162	35.7	42
	3DMM	1.21	28.0	30	98	15.9	49

5. Conclusion

A method based on multi-scale feature fusion and lighting robustness optimization was proposed to address the problem of insufficient feature extraction ability in weak texture areas and poor robustness under complex lighting conditions in facial 3D modeling. The study utilized MSDF-Net to achieve multi-scale feature extraction and fusion, enhancing the modeling capability of weak texture regions. By optimizing lighting robustness through IRFF-Net and introducing group correlation matching and multi-level block feature fusion strategies, the accuracy of feature matching under complex lighting conditions was improved. The experimental results showed that the SSIM of MSDF-Net was as high as 0.954, and the MAE was as low as 0.63mm, which was superior to the comparative methods. In the robustness test of lighting changes, IRFF-Net showed the lowest MSE and highest feature consistency in strong light, weak light, and shadow changing scenes, reaching 0.85mm and 0.941, respectively. In the ablation experiment, the complete network MSDF-Net+IRFF-Net performed the best in terms of PSNR, MSE, and feature consistency, with a PSNR of 32.1dB and a minimum MSE of 0.72mm. The end-to-end comprehensive performance comparison showed that the complete network was superior to mainstream methods in terms of PSNR, resource efficiency, and memory usage. Finally, in actual scenario testing, the accuracy of security monitoring increased from 88.5% to 92.8%, and the PSNR of virtual reality scenes increased from 29.5 dB to 33.0 dB, fully verifying the adaptive optimization capability and applicability and stability of the proposed method in complex dynamic environments.

However, there is still room for improvement in the operational efficiency of complete networks, especially in dynamic application scenarios with high real-time requirements. Future research can focus on the lightweight design of network structure to further reduce computational overhead and improve modeling speed. At the same time, in order to adapt to the diversified device deployment requirements, it is necessary to systematically evaluate the model performance on different hardware platforms (such as mobile terminals, edge computing devices, *etc.*), and explore more efficient hardware optimization schemes and model compression strategies. In addition, although the performance of this study is stable under common light changes, there may still be insufficient feature extraction under extreme light conditions (such as strong backlight, local saturation or dark regions), and a reconstruction mechanism based on illumination invariance coding or generative adversarial network (GAN) will be considered in the future to further enhance robustness. At the same time, considering that practical applications may face challenges such as occlusion, multi-pose, large expression changes, and environmental noise, the research will further introduce self-supervised learning, cross-domain adaptation and multi-modal perception technologies to improve the generalization ability and adaptability of the model in uncontrolled scenarios, and provide more reliable technical support for complex applications such as security monitoring, virtual reality, and medical modeling.

Declaration of Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This research work was not funded by any program.

Data availability

The data used in this study is proprietary and not suitable for sharing.

References

[1] F. Jin *et al.*, "Bottomup Synthesis of 8-connected Three-dimensional Covalent Organic Frameworks for Highly Efficient Ethylene/ethane Separation", *Journal of the American Chemical Society*, vol. 144, no. 12, pp. 5643–5652, 2022. http://dx.doi.org/10.1021/jacs.2c01058

[2] S. Y. Lee *et al.*, "In Vitro Three-dimensional (3D) Cell Culture Tools for Spheroid and Organoid Models", *SLAS Discovery*, vol. 28, no. 4, pp. 119–137, 2023. http://dx.doi.org/10.1016/j.slasd.2023.03.006

[3] C. Agarwal and C. Bhatnagar, "Unmasking the Potential: Evaluating Image Inpainting Techniques for Masked Face Reconstruction", *Multimedia Tools and Applications*, vol. 83, no. 1, pp. 893–918, 2024. https://doi.org/10.1007/s11042-023-15807-x

[4] B. Anđić *et al.*, "Phenomenography Study of STEM Teachers' Conceptions of Using Three-dimensional Modeling and Printing (3DMP) in Teaching", *Journal of Science Education and Technology*, vol. 32, no. 1, pp. 45–60, 2023. http://dx.doi.org/10.1007/s10956-022-10005-0

[5] S. Sharma and V. Kumar, "3D Face Reconstruction in Deep Learning Era: A Survey", *Archives of Computational Methods in Engineering*, vol. 29, no. 5, pp. 3475–3507, 2022. http://dx.doi.org/10.1007/s11831-021-09705-4

[6] C. Shen *et al.*, "Large-view 3D Color Face Reconstruction from Dingle Image Via UV Location Map and CGAN", *Journal of Computer-Aided Design & Computer Graphics*, vol. 34, no. 4, pp. 614–622, 2022. http://dx.doi.org/10.3724/SP.J.1089.2022.18959

[7] S. Rani *et al.*, "Three Dimensional Objects Recognition & Pattern Recognition Technique; Related Challenges: A Review", *Multimedia Tools and Applications*, vol. 81, no. 12, pp. 17303–17346, 2022. http://dx.doi.org/10.1007/s11042-022-12412-2

[8] S. Pal *et al.*, "Adapting a Swin Transformer for License Plate Number and Text Detection in Drone Images. Artificial Intelligence and Applications", vol. 1, no. 3, pp. 145–154, 2023. http://dx.doi.org/10.47852/bonviewAIA3202549

[9] G D 'Ettorre *et al.*, "A Comparison Between Stereophotogrammetry and Smartphone Structured Light Technology for Three-dimensional Face Scanning", *The Angle Orthodontist*, vol. 92, no. 3, pp. 358–363, 2022. http://dx.doi.org/10.2319/040921-290.1

[10] N. K. Mehta *et al.*, "Three-dimensional DenseNet Self-attention Neural Network for Automatic Detection of Student's Engagement", Applied *Intelligence*, vol. 52, no. 12, pp. 13803–13823, 2022. http://dx.doi.org/10.1007/s10489-022-03200-4

[11] M. C. Florkow *et al.*, "Magnetic Resonance Imaging Versus Computed Tomography for Three-Dimensional Bone Imaging of Musculoskeletal Pathologies: A Review", *Journal of Magnetic Resonance Imaging*, vol. 56, no. 1, pp. 11–34, 2022. http://dx.doi.org/10.1002/jmri.28067

[12] Z. Luo *et al.*, "A Three-dimensional Model of Student Interest During Learning Using Multimodal Fusion with Natural Sensing Technology", *Interactive Learning Environments*, vol. 30, no. 6, pp. 1117–1130, 2022. http://dx.doi.org/10.1080/10494820.2019.1710852

[13] Z. Chen *et al.*, "Transformer-based 3D Face Reconstruction with End-to-end Shape-preserved Domain Transfer", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 12, pp. 8383–8393, 2022. http://dx.doi.org/10.1109/TCSVT.2022.3192422

[14] N. Sun *et al.*, "3D Facial Feature Reconstruction and Learning Network for Facial Expression Recognition in the Wild", *IEEE Transactions on Cognitive and Developmental Systems*, vol. 15, no. 1, pp. 298–309, 2022. http://dx.doi.org/10.1109/TCDS.2022.3157772

[15] H. M. U. Munir and W. S. Qureshi, "Single Image 3D Beard Face Reconstruction Approaches", *International Journal of Cyber-Physical Systems (IJCPS)*, vol. 4, no. 1, pp. 1–17, 2022. http://dx.doi.org/10.4018/IJCPS.314572

[16] Z. Zhou *et al.*, "Replay Attention and Data Augmentation Network for 3D Face and Object Reconstruction", *IEEE Transactions on Biometrics, Behavior and Identity Science*, vol. 5, no. 3, pp. 308–320, 2023. http://dx.doi.org/10.1109/TBIOM.2023.3261272

[17] S. C. Medin *et al.*, "MOST-GAN: 3D Morphable StyleGAN for Disentangled Face Image Manipulation", in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, vol. 36, no. 2, pp. 1962–1971. http://dx.doi.org/10.1609/aaai.v36i2.20091

[18] M. K. Meena *et al.*, "Partially Occluded Face Reconstruction Using Graph-based Algorithm", *Journal of Electrical Engineering & Technology*, vol. 19, no. 6, pp. 3655–3664, 2024. http://dx.doi.org/10.1007/s42835-024-01995-5

[19] C Wang *et al.*, "Cross-domain and Disentangled Face Manipulation with 3D Guidance", *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 4, pp. 2053–2066, 2022. http://dx.doi.org/10.1109/TVCG.2021.3139913

[20] S. Bahroun *et al.*, "Deep 3D-LBP: CNN-based Fusion of Shape Modeling and Texture Descriptors for Accurate Face Recognition", *The Visual Computer*, vol. 39, no. 1, pp. 239–254, 2023. http://dx.doi.org/10.1007/s00371-021-02324-x

[21] E. Tretschk *et al.*, "State of the Art in Dense Monocular Non-Rigid 3D Reconstruction", *Computer Graphics Forum*, vol. 42, no. 2, pp. 485–520, 2023. http://dx.doi.org/10.1111/cgf.14774

[22] S. Van Nguyen *et al.*, "Reconstruction of 3D Digital Heritage Objects for VR and AR Applications", *Journal of Information and Telecommunication*, vol. 6, no. 3, pp. 254–269, 2022. http://dx.doi.org/10.1080/24751839.2021.2008133

[23] X. Amezua *et al.*, "Analysis of the Influence of the Facial Scanning Method on the Transfer Accuracy of a Maxillary Digital Scan to a 3D Face Scan for a Virtual Facebow Technique: An in Vitro Study", *The Journal of Prosthetic Dentistry*, vol. 128, no. 5, pp. 1024–1031, 2022. http://dx.doi.org/10.1016/j.prosdent.2021.02.007

[24] S. M. La Cava *et al.*, "3D Face Reconstruction: The Road to Forensics", *ACM Computing Surveys*, vol. 56, no. 3, pp. 1–38, 2023. http://dx.doi.org/10.1145/3625288

[25] S. Zhao *et al.*, "Fm-3dfr: Facial Manipulation-based 3D Face Reconstruction", *IEEE Transactions on Cybernetics*, vol. 54, no. 1, pp. 209–218, 2023. http://dx.doi.org/10.1109/TCYB.2023.3242368

[26] G. Srivastava and S. Bag, "Modern-day Marketing Concepts Based on Face Recognition and Neuromarketing: A Review and Future Research Directions", *Benchmarking: An International Journal*, vol. 31, no. 2, pp. 410–438, 2024. http://dx.doi.org/10.1108/BIJ-09-2022-0588

Received: January 2025
Revised: March 2025
Accepted: March 2025

Contact address:
Sichuan Vocational and Technical College
Suining
Sichuan
China
e-mail: Zxiaoxiao068@126.com

Hongyan Zhu graduated from China West Normal University with a master's degree in Computer Education and Application direction of the Educational Technology in 2024. She is currently employed at Sichuan Vocational and Technical College as a lecturer. Her reaserch interests include: computer digital media technology, 3D modeling technology.