Scientific Research Publishing

Open Journal of Applied Sciences, 2025, 15(12), 3987-3999
https://www.scirp.org/journal/ojapps
ISSN Online: 2165-3925
ISSN Print: 2165-3917

3D Face Reconstruction with Implicit Neural Representation and Multi-Scale Feature Fusion

Danni Peng[1*], Guoliang Wei[1#], Yuhua Ai[2]

[1]Business School, University of Shanghai for Science and Technology, Shanghai, China
[2]College of Optoelectronic Information and Computer Engineering, University of Shanghai for Science and Technology, Shanghai, China
Email: 232171036@st.usst.edu.cn, [#]guoliang.wei@usst.edu.cn, 241260098@st.usst.edu.cn

How to cite this paper: Peng, D.N., Wei, G.L. and Ai, Y.H. (2025) 3D Face Reconstruction with Implicit Neural Representation and Multi-Scale Feature Fusion. *Open Journal of Applied Sciences*, **15**, 3987-3999.
https://doi.org/10.4236/ojapps.2025.1512257

Received: November 10, 2025
Accepted: December 15, 2025
Published: December 18, 2025

Abstract

In computer vision and graphics applications, the precise representation of 3D faces is of critical importance. This paper proposes a novel 3D deformable face model that learns complex continuous spaces through implicit representation. Firstly, multi-scale context features are extracted from the input image by using a dense dilated convolution branch, capturing both global semantics and local geometric details. Then, through position encoding and gated fusion, an adaptive mapping between image features and 3D spatial coordinates is achieved. To enhance the implicit decoding capability, local expert decoders are constructed, and spatial regularization constraints are introduced to ensure the local continuity and geometric smoothness of the implicit field. Experiments show that this method performs well on the FaceScape dataset, with a chamfer distance of 0.553 and an F-score of 93.74. It also demonstrates high-fidelity details in 3D face reconstruction when compared with multiple classic algorithms.

Keywords

3D Morphable Models, Implicit Neural Representations, Dense Atrous Convolution, Spatial Regularization

1. Introduction

Three-dimensional face reconstruction (3D Face Reconstruction) [1]-[3] has become a fundamental research direction in computer vision and digital human modeling, aiming to accurately recover the 3D geometric structure and deformation at-

*First author.
#Corresponding author.

tributes of human faces from two-dimensional images. The 3D Morphable Face Model (3DMM) [4], first proposed by Blanz and Vetter, is a well-known statistical representation that provides a general framework for facial modeling. The 3DMM leverages the non-rigid iterative closest point (NICP) algorithm [5] to register a known template mesh to all training scans, and was later extended into multilinear formulations. The FLAME model [6], for instance, represents facial expressions through the combination of jaw articulation and linear expression blendshapes. Further extensions [7]-[9] introduced multilinear decompositions that encode various facial modes independently. However, due to the inherently limited representational capacity of linear models, these methods struggle to handle complex and nonlinear variations in facial geometry.

With the rapid development of deep learning, nonlinear 3DMM approaches based on neural networks have emerged. These methods [10] are capable of learning 3D facial representations directly from large-scale, unconstrained 2D datasets. Models such as DECA [11] and PRNet [12] achieved notable progress in data-driven facial feature extraction. Bagautdinov *et al.* [13] further mapped 3D meshes into 2D domains, while others applied spiral convolutions to directly learn 3DMMs from mesh data. Despite their progress, these neural approaches are built upon discrete 3D representations, which limit their ability to model complex deformations. Furthermore, because of the dimensional constraints inherent to these models, they often lack high-frequency geometric details and fail to achieve high-fidelity reconstruction.

To address these limitations, implicit representation-based methods [14]-[17] have recently demonstrated superior spatial continuity and topological flexibility, becoming a promising alternative for 3D face reconstruction. These methods implicitly represent facial surfaces by learning continuous signed distance functions (SDFs) [18] or occupancy fields [19] from low-dimensional shape embeddings of observed inputs. The continuous parameterization and unified representation enable implicit models to outperform traditional mesh- or voxel-based methods in geometric consistency and fine-grained detail reconstruction, achieving impressive results in both shape recovery and surface registration. For instance, NPHM [20] introduces a parametric head model using local implicit fields to represent identity and a global conditional implicit neural representation (INR) for expression deformation. However, its capacity to capture high-frequency facial details remains limited by the global conditioning network. H3D-Net [21] constructs an implicit head shape space from 2D inputs but cannot serve as a general model. I3DMM [22], the first implicit deformable model for human heads, achieves efficient reconstruction through disentangling global deformation and local details, yet still suffers from low-quality reconstruction in facial regions. Mildenhall *et al.* [23] employ differentiable volume rendering to learn deformable radiance fields, but their density-based geometry representation tends to introduce geometric noise. ImFace [18] combines implicit representations with 3D facial identity embeddings to enable structural priors shared across individuals; however, as it relies solely on

coordinate-based and implicit embeddings, it lacks multi-scale semantic features extracted from images and fails to effectively leverage global contextual cues for shape recovery.

In this paper, we propose a novel 3D face reconstruction framework that substantially upgrades the traditional 3DMM by integrating multi-scale image encoders and implicit neural representations (INRs). Independent INR subnetworks are designed to separately model identity and expression, while incorporating feature sampling [24] and positional encoding to enhance detail learning. Furthermore, a mixture-of-local-experts decoder, composed of multiple MLPs, adaptively partitions and models complex fine-grained facial details.

The main contributions of this work are summarized as follows:

1) Multi-scale feature extraction: We construct a multi-scale image encoder based on dense dilated convolutions, which enhances facial responses at different scales. This design enlarges the receptive field while preserving spatial resolution, thereby enabling joint modeling of global structures and local textures to capture fine-grained facial details.

2) Disentangled implicit deformation fields: The signed distance field is decomposed into two sub-fields representing identity and expression, respectively. A mixture-of-experts structure partitions the implicit field into multiple local subspaces, allowing each to learn independent implicit mappings. This alleviates overfitting and generalization issues often encountered by single-MLP representations in high-dimensional spaces.

3) Spatial regularization: A spatial regularization term is introduced to constrain the SDF prediction differences between neighboring points, ensuring that the implicit field remains spatially continuous and physically interpretable.

2. Method

We employ implicit neural representations (INRs) to learn a nonlinear 3D facial model, where the facial geometry is formulated as a conditional continuous signed

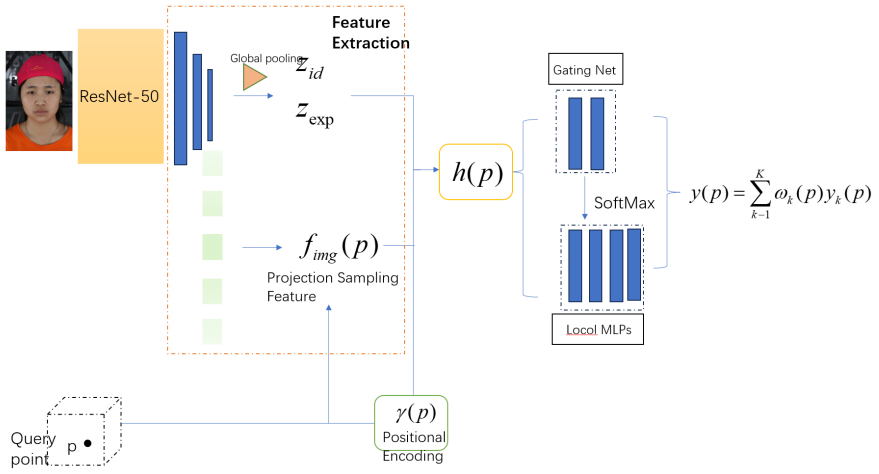

Figure 1. Algorithm flowchart.

distance function (SDF). To enhance detail fidelity and spatial stability, we incorporate multi-scale image features, implicit query embeddings, and a spatially regularized adaptive local-expert decoder. The overall framework of the proposed method is illustrated in **Figure 1**.

2.1. Multi-Scale Contextual Feature Extraction

In previous implicit representation frameworks, the conditioning of 3D query points typically relies on latent codes and sparse keypoint localization, while lacking direct guidance from 2D image features. This absence of image-level conditioning limits the recovery of high-frequency textures and geometric details, and may amplify projection ambiguities in single-image reconstruction.

To bridge this gap, we introduce a multi-scale contextual feature extraction module. Specifically, a pre-trained ResNet-50 [25] backbone is first employed to extract mid-level feature maps $F \in \mathbb{R}^{H/8\times W/8\times C}$ that capture hierarchical facial structures—such as landmark distributions and preliminary texture patterns $H/8\times H/8$—while maintaining spatial resolution across multiple scales. Before feature extraction, a Dlib-based facial alignment module [26] is applied to normalize the input image into a 256 × 256 canonical frontal view, ensuring geometric consistency across samples.

To further aggregate contextual details and enrich semantic information, we design a multi-branch dilated convolutional module with dense connections. This structure effectively reduces channel redundancy and enhances receptive field diversity for facial geometry optimization. We set the number of branches to K = 5, with a dilation rate sequence of $r=\{r_k\}_{k=1}^{K}=\{1,6,12,18,24\}$, thereby covering receptive fields ranging from local to global contexts. The outputs of all branches are concatenated along the channel dimension to form a unified multi-scale representation that encodes both fine-grained and large-scale contextual cues for subsequent implicit modeling.

The extracted image features are sampled through a weak-perspective projection, where the camera intrinsic $K \in \mathbb{R}^{3\times3}$ and extrinsic parameters $[R|t]$ are estimated from 2D facial landmarks via a least-squares optimization. Each 3D query point P is projected onto the corresponding UV coordinate $(u,v)=\pi(p;K,[R|t])$ on the image plane. To mitigate depth ambiguity and sampling artifacts, we adopt a Gaussian-weighted 3 × 3 neighborhood sampling strategy, which aggregates local contextual information around each projected location, formulated as:

$$f_{img}(p)=\sum_{(u',v')\in N(u,v)}\omega_{(u',v')}\cdot M(u',v')\in\mathbb{R}^{C} \tag{1}$$

where, $N(u,v)$ denotes the neighborhood centered at the projected point (u,v), and $\omega_{(u',v')}=\exp\left(-\frac{\|(u'-u,v'-v)\|_2^2}{2\sigma^2}\right)\Big/Z$ represents the Gaussian weighting co efficient for each sampled pixel. This sampling strategy primarily facilitates 2D -

3D alignment, providing image-guided supervision for the deformation field to enhance the consistency between image features and 3D geometry.

The multi-scale feature maps M are obtained through global average pooling across all dilated convolution branches, followed by a fusion convolution layer that aggregates contextual information from multiple receptive fields. The resulting feature maps are downsampled via bilinear interpolation to a fixed network resolution, effectively reducing computational complexity while preserving essential semantic cues.

Branches with low dilation rates focus on capturing local fine-grained details, whereas those with high dilation rates integrate global structural contours, thereby alleviating the detail loss typically caused by purely latent-code-based implicit representations. The proposed multi-scale feature extraction module thus provides rich multi-scale correlations between local and global facial regions, serving as a strong prior for implicit field learning.

Implicit Query Encoding

The core concept of implicit neural representations (INRs) is to train a neural network to approximate a continuous function f, which represents a 3D surface implicitly via its level set formulation. In this work, we employ a deep signed distance function (SDF) conditioned on latent embeddings of expression and identity to achieve a comprehensive and disentangled facial representation. For each spatial query point $x \in \mathbb{R}^3$, the network outputs a signed distance value:

$$f : \left(p, z_{exp}, z_{id}\right) \in \mathbb{R}^3 \times \mathbb{R}^{d_{exp}} \times \mathbb{R}^{d_{id}} \mapsto s \in \mathbb{R} \tag{2}$$

where, $p \in R^3$ denotes the spatial coordinate of the query point in 3D space, while z_{id} and z_{exp} represent the identity and expression deformation fields, respectively. The identity and expression deformation fields are constructed by first mapping each sampled point p into a canonical geometric space. The identity field receives only the canonical coordinates together with the identity latent code as input, and is responsible for modeling the stable and person-specific facial structure. The expression field is driven solely by the expression latent code and predicts localized displacement residuals, whose outputs are combined with the identity field's canonical geometry to obtain the final dynamic facial surface.

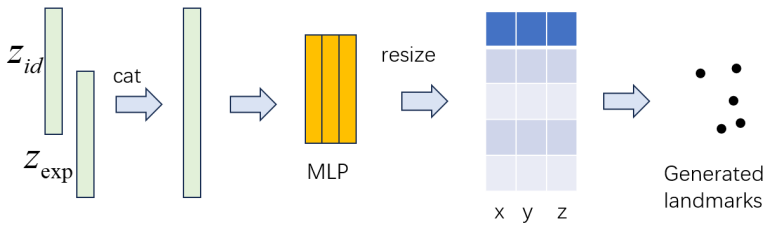

Figure 2. Landmark-Net.

The latent codes are disentangled into identity and expression components through independent multilayer perceptron (MLP) projections. Each MLP consists of three fully connected layers activated by ReLU functions, with shared parameters to en-

hance model generalization. As shown in Figure 2, a Landmark-Net module is designed to determine the spatial position of query points p within the neural blend field. By conditioning on the input identity latent code, the network predicts a set of canonical facial landmarks, denoted as $I = \eta\left(z_{exp}, z_{id}\right) \in \mathbb{R}^{68}$ and $I' = \eta'\left(z_{id}\right) \in \mathbb{R}^{68}$, which provide geometric priors for both deformation fields. Each of the three-layer MLPs within Landmark-Net guides the local semantic deformation of the identity and expression fields, ε and ζ, respectively.

In this work, we use $SE(3)$ to represent the facial shape deformation, which exhibits superior capability in modeling jaw articulation and rotation. The corresponding rotation matrix e^{ω} is formulated as:

$$e^{\omega} = I + \frac{\sin\|\omega\|}{\|\omega\|}\omega^{\wedge} + \frac{1-\cos\|\omega\|}{\|\omega\|^{2}}\left(\omega^{\wedge}\right)^{2} \tag{3}$$

All deformation fields are implemented using a shared mini-network architecture. A lightweight module conditioned on the spatial coordinates of query points is appended to the end of each mini-network, enabling adaptive local field blending to capture spatially varying deformations.

Implicit neural networks generally exhibit strong performance in representing low-frequency components but struggle to encode high-frequency geometric details. To enhance the network's capability in modeling fine-grained geometry, we apply a multi-frequency Fourier positional encoding to each spatial point p, expanding the original low-dimensional 3D coordinates into a high-dimensional periodic feature representation:

$$\gamma(p) = \left[\sin\left(2^{k}\pi p\right), \cos\left(2^{k}\pi p\right)\right]_{k=0}^{L-1} \tag{4}$$

Each coordinate component is independently encoded, ensuring that the network learns a continuous and smooth deformation field. The encoded positional features are then concatenated with the latent variables z_{id} and z_{exp} from Equation (1), forming a composite feature vector:

$$h(p) = \left[\gamma(p), f_{img}(p), z_{id}, z_{exp}\right] \tag{5}$$

The introduction of implicit representations expands the latent space and enables precise optimization during single-image fitting, while maintaining identity consistency under facial expression editing.

2.2. Adaptive Local-Expert Decoder

Traditional approaches typically employ a single MLP decoder that performs a uniform mapping for all spatial points, which makes it difficult to capture region-specific high-frequency details. To address this limitation, we propose an adaptive local-expert decoding mechanism. Specifically, the proposed model preserves the local field structures of the Mini-Nets as backbone modules and introduces multiple independent local MLPs to model different facial regions separately. The human face exhibits significant variations across different regions in terms of geometric morphology, texture complexity, and sensitivity to expression changes. For

example, areas such as the regions around the left and right eyes, the mouth, the nasal bridge, and the jaw demonstrate notably distinct geometric dynamics. Therefore, the face is partitioned into eight semantically consistent and structurally stable subregions based on these anatomical characteristics. We define $K=8$ local subregions $\{e_k\}_{k=1}^{K}$, each corresponding to a local field in the Mini-Nets Ψ_n, which facilitates fine-grained modeling of structural details specific to each region.

By incorporating region-aware specialization, facial regions are partitioned during training using landmark-based guidance on the FaceScape dataset [27], dividing the face into semantically meaningful parts such as the eyes, mouth, and cheeks. Each expert network is trained to learn localized deformations within its assigned region. A two-layer gating network G_ϕ receives the implicit feature vector $h(p)$ has input and produces soft assignment weights across the K experts. This mechanism adaptively fuses local field predictions, enhancing both the representation capability and training stability of the model while reducing the risk of overfitting in geometrically complex regions.

Finally, the outputs of the local experts are aggregated to produce the local implicit predictions $y_k(p)$, which are combined via weighted summation to yield the final signed distance field (SDF) value:

$$y(p)=\sum_{k-1}^{K}\omega_k(p)y_k(p) \tag{6}$$

2.3. Loss Function

Several loss functions are employed to train the proposed method for learning a reliable facial shape representation.

Reconstruction Loss. A basic SDF structural loss is applied to learn the implicit field:

$$\mathcal{L}_{sdf}^{i}=\lambda_1\sum_{\mathbf{p}\in\Omega_i}\left|f(\mathbf{p})-\overline{s}\right|+\lambda_2\sum_{\mathbf{p}\in\Omega_i}\left(1-\left\langle\nabla f(\mathbf{p}),\overline{\mathbf{n}}\right\rangle\right) \tag{7}$$

where $\overline{s}$ and $\overline{n}$ denote the ground-truth SDF value and field gradient, respectively, Ω_i represents the sampled space of the i-th facial scan, and λ is the loss weight coefficient.

Embedding Loss. The embedding vectors are regularized using a zero-mean Gaussian prior:

$$\mathcal{L}_{emb}=\lambda_3\left(\left\|\mathbf{z}_{exp}\right\|^2+\left\|\mathbf{z}_{id}\right\|^2\right) \tag{8}$$

Smooth Regularization Loss. Since the soft assignment of the gating network may lead to discontinuities or unstable boundaries among expert regions, we introduce a smoothness constraint to enforce local weight consistency. For each sampled neighboring point pair $\{(p_i,p_j)\}$:

$$\mathcal{L}_{smooth}=\lambda_4\mathbb{E}_{p_i,p_j\sim\mathrm{N}}\left[\left\|\alpha(p_i)-\alpha(p_j)\right\|_2^2\right] \tag{9}$$

where N denotes the neighborhood set of point p. This term encourages smooth transitions between adjacent regions, such as gradual blending from the forehead to the nose bridge.

Entropy Regularization Loss. To prevent degeneration caused by uniform expert activation in the gating network, an entropy-based regularization is introduced to promote sparse and discriminative expert selection:

$$\mathcal{L}_{ent} = -\lambda_5 \mathbb{E}_{\mathbf{p}} \left[\sum_{k=1}^{K} \alpha_k(\mathbf{p}) \log \alpha_k(\mathbf{p}) \right] \tag{10}$$

Therefore, the total loss is:

$$\mathcal{L} = \mathcal{L}_{sdf}^{i} + \mathcal{L}_{emb} + \mathcal{L}_{smooth} + \mathcal{L}_{ent} \tag{11}$$

3. Experimental Results and Analysis

3.1. Dataset

FaceScape is a large-scale, high-quality 3D face dataset consisting of 938 individuals and 20 types of facial expressions. It is one of the most comprehensive and high-precision 3D facial datasets available. Among them, data from 365 subjects are publicly released and are used in this work. Specifically, 5323 facial scans from 355 subjects with 15 expressions are used for training, while 200 scans from the remaining 10 subjects with 20 expressions are used for testing.

3.2. Experimental Details

1) Training

The proposed model is trained end-to-end using the Adam optimizer with an initial learning rate of 0.0001 for 1500 epochs. After 200 epochs, the learning rate decays by a factor of 0.95 every 10 epochs. The ResNet-50 backbone is frozen during the first 10 epochs, and the Mini-Nets component is fixed during pretraining to preserve deformation priors. Training is conducted on a single NVIDIA RTX 4090 GPU for approximately 2 days with a batch size of 72. During testing, optimizing 200 samples on a single GPU takes about 5 hours. Due to the lightweight architectures employed in all components of the proposed framework, the overall inference time is marginally reduced compared to IM-Face. Nevertheless, the model is able to preserve and reconstruct finer geometric variations, demonstrating better detail representation without sacrificing computational efficiency.

2) Data Preprocessing

Since the implicit function requires strictly aligned inputs, we adopt the pseudo-watertight mesh generation method from [18]. Facial meshes are rigidly aligned to the frontal view using landmarks and normalized to a 10 cm unit scale. Sampling is performed within a sphere centered 4 cm behind the nose tip with a radius of 10 cm, and points outside the sphere are removed. To construct a directed pseudo-watertight mesh, holes around the nose and mouth are filled, enabling distance transformation to compute SDF values.

3.3. Experimental Results

We use the proposed model to fit facial scans and compare reconstruction results against the geometric models of FLAME [6], i3DMM [22], and ImFace [18], demonstrating the superiority of our approach. The official FLAME code is used to fit full-face scans in the test set, which includes 300 identity parameters and 100 expression parameters. Since the FaceScape dataset is employed for both training and evaluation in this work, the publicly released models from the dataset are also used for comparison. For Imface, both the identity and expression embeddings are 128-dimensional, consistent with the dimensionality used in i3DMM and in the proposed method. Since the original i3DMM model is trained on only 58 subjects, we retrain it on the same dataset used in this paper for a fair comparison.

3.3.1. Quantitative Analysis

We adopt symmetric Chamfer Distance (CD) and F-score as evaluation metrics, where the F-score threshold is set to 0.001 as a strict criterion. The F-score is a comprehensive measure combining precision and recall, providing an overall assessment of model performance. A smaller Chamfer Distance indicates better reconstruction accuracy, while a higher F-score reflects improved consistency.

The results are summarized in **Table 1**, showing that our proposed method surpasses competing approaches across both metrics, clearly demonstrating its effectiveness.

Table 1. Quantitative analysis of facial reconstruction accuracy.

Methods	Dim.	Chamfer (mm)[†]	F-score@1mm[¶]
I3DMM	256	1.635	42.26
FaceScape	352	0.929	67.09
Imface	256	0.625	91.11
Ours	256	0.553	93.74

3.3.2. Qualitative Analysis

This section visualizes the reconstruction results obtained by different models, as shown in **Figure 3**, where each column corresponds to a test subject with a non-neutral facial expression.i3DMM is the first deep implicit model designed for human heads; however, when dealing with more complex scenes, it fails to capture intricate deformations and detailed features, resulting in noticeable artifacts on the reconstructed faces. FaceScape, benefiting from high-quality training scans and the inclusion of test subjects within its training set, performs well in preserving identity characteristics. Nevertheless, when handling nonlinear deformations, it tends to produce rigid facial expressions and struggles to capture fine texture details. ImFace achieves better overall performance compared to the above methods, as it incorporates both identity and expression attributes. However, it still lacks precision in reconstructing fine-grained facial structures. In contrast, our proposed method not only reconstructs faces with more accurate identity and expression representations but also maintains geometric robustness while effectively

capturing subtle and rich nonlinear facial muscle deformations—such as frowning and pouting—that contribute to more realistic and expressive results.

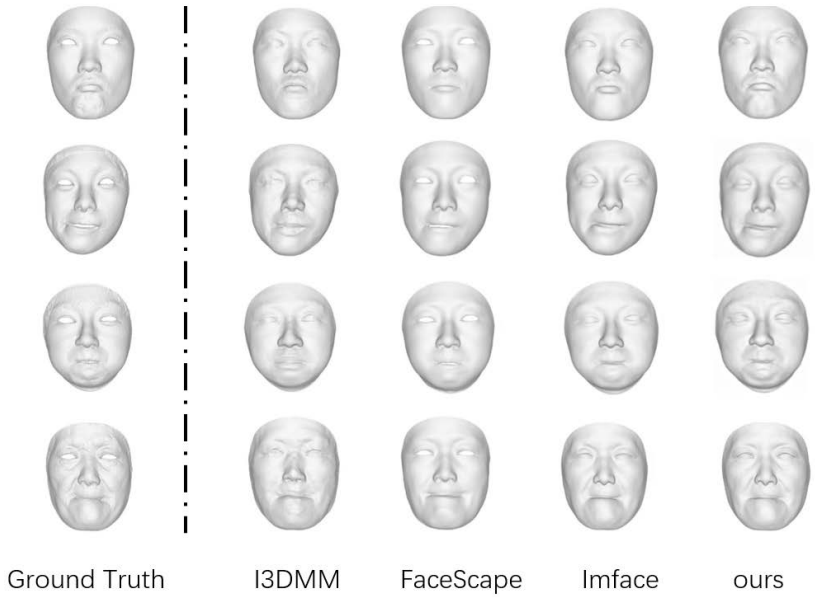

Figure 3. Comparison of results with i3DMM, FaceScape, and Imface.

3.3.3. Ablation Study

The core components of the proposed algorithm include context feature extraction, local expert decoder, and regularization losses. In this section, we conduct ablation experiments to verify the effectiveness of each key component.

Since the regularization losses are designed to prevent instability in the soft assignments of the gating network, we specifically examine cases where the local expert decoder is used alone or combined with the regularization losses.

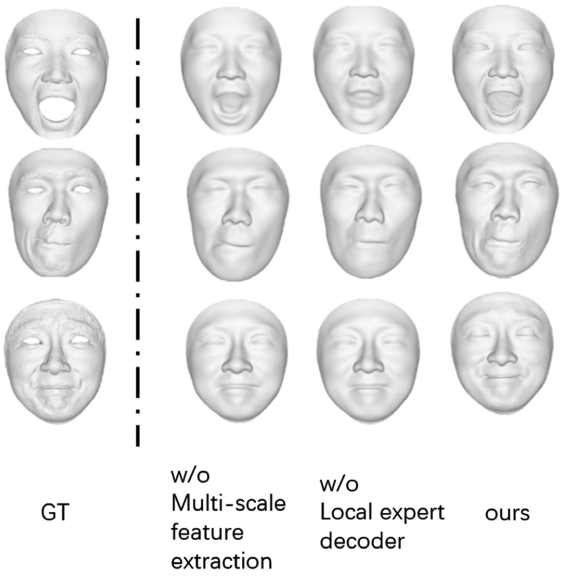

Figure 4. Ablation study.

Table 2. Presents the results of the ablation study.

Multi-Scale Feature Extraction	Local Expert Decoder	Regularization	Chamfer (mm)†	F-score@1mm¶
✓			0.602	91.82
	✓		0.619	91.59
✓	✓		0.569	92.53
	✓	✓	0.575	92.96
✓	✓	✓	0.553	93.74

From the ablation experiments in Table 2 above, it can be seen that the core components proposed in this paper are effective improvements. While adding either the feature extraction or the local expert decoder individually is effective, the combined effect of both yields better results. As shown in Figure 4, the algorithm proposed in this paper is more refined and accurate.

4. Conclusions

To enhance the fine-detail reconstruction capability of 3D face modeling, this paper proposes a novel 3D facial deformation model that substantially upgrades traditional 3DMMs by integrating context-aware encoding and decoding with Implicit Neural Representations (INRs). The improved regularization losses effectively mitigate instability caused by the soft assignment in the gating network, resulting in more refined and stable 3D face reconstructions. Experimental results, both qualitative and quantitative, demonstrate that the proposed method achieves superior performance compared to existing approaches in terms of geometric accuracy and expression fidelity.

However, the current method remains relatively limited in handling illumination and reflection variations. In future work, we plan to incorporate realistic diffuse and specular reflectance fusion to further enhance the overall robustness of 3D face reconstruction and its ability to cope with complex lighting environments.

Funding

National Natural Science Foundation of China (62273239); Shanghai "Science and Technology Innovation Action Plan" Domestic Science and Technology Cooperation Project (20015801100).

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Wang, Z., Zhu, X., Zhang, T., Wang, B. and Lei, Z. (2024) 3D Face Reconstruction with the Geometric Guidance of Facial Part Segmentation. 2024 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 16-22 June 2024, 1672-1682. https://doi.org/10.1109/cvpr52733.2024.00165

[2] Jin, H., Jian, M., Ding, D. and Yu, H. (2025) Self-Supervised Face Deocclusion via 3-D Face Reconstruction with Outlier Segmentation. *IEEE Transactions on Human-Machine Systems*, **55**, 746-755. https://doi.org/10.1109/thms.2025.3585780

[3] Dou, P., Shah, S.K. and Kakadiaris, I.A. (2017) End-to-End 3D Face Reconstruction with Deep Neural Networks. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 1503-1512. https://doi.org/10.1109/cvpr.2017.164

[4] Blanz, V. and Vetter, T. (2023) A Morphable Model for the Synthesis of 3D Faces. In: *Seminal Graphics Papers*: *Pushing the Boundaries*, *Volume* 2, ACM, 157-164. https://doi.org/10.1145/3596711.3596730

[5] Serafin, J. and Grisetti, G. (2015) NICP: Dense Normal Based Point Cloud Registration. 2015 *IEEE/RSJ International Conference on Intelligent Robots and Systems* (*IROS*), Hamburg, 28 September-2 October 2015, 742-749. https://doi.org/10.1109/iros.2015.7353455

[6] Li, T., Bolkart, T., Black, M.J., Li, H. and Romero, J. (2017) Learning a Model of Facial Shape and Expression from 4D Scans. *ACM Transactions on Graphics*, **36**, 1-17. https://doi.org/10.1145/3130800.3130813

[7] Brunton, A., Bolkart, T. and Wuhrer, S. (2014) Multilinear Wavelets: A Statistical Shape Space for Human Faces. In: *Lecture Notes in Computer Science*, Springer International Publishing, 297-312. https://doi.org/10.1007/978-3-319-10590-1_20

[8] Vlasic, D., Brand, M., Pfister, H. and Popović, J. (2005) Face Transfer with Multilinear Models. *ACM Transactions on Graphics*, **24**, 426-433. https://doi.org/10.1145/1073204.1073209

[9] Bolkart, T. and Wuhrer, S. (2015) A Groupwise Multilinear Correspondence Optimization for 3D Faces. 2015 *IEEE International Conference on Computer Vision* (*ICCV*), Santiago, 7-13 December 2015, 3604-3612. https://doi.org/10.1109/iccv.2015.411

[10] Tran, L. and Liu, X. (2018) Nonlinear 3D Face Morphable Model. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 7346-7355. https://doi.org/10.1109/cvpr.2018.00767

[11] Feng, Y., Feng, H., Black, M.J. and Bolkart, T. (2021) Learning an Animatable Detailed 3D Face Model from In-the-Wild Images. *ACM Transactions on Graphics*, **40**, 1-13. https://doi.org/10.1145/3450626.3459936

[12] Wang, Y. and Solomon, J.M. (2019) PrNet: Self-Supervised Learning for Partial-to-Partial Registration. *Advances in Neural Information Processing Systems*, **32**, 1-13.

[13] Bagautdinov, T., Wu, C., Saragih, J., Fua, P. and Sheikh, Y. (2018) Modeling Facial Geometry Using Compositional VAEs. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 3877-3886. https://doi.org/10.1109/cvpr.2018.00408

[14] Park, J.J., Florence, P., Straub, J., Newcombe, R. and Lovegrove, S. (2019) DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 165-174. https://doi.org/10.1109/cvpr.2019.00025

[15] Chiang, P.-Z., Tsai, M.-S., Tseng, H.-Y., Lai, W.-S. and Chiu, W.-C. (2022) Stylizing 3D Scene via Implicit Representation and HyperNetwork. 2022 *IEEE/CVF Winter Conference on Applications of Computer Vision* (*WACV*), Waikoloa, 3-8 January 2022, 215-224. https://doi.org/10.1109/WACV51458.2022.00029

[16] Chen, Z.Q. and Zhang, H. (2019) Learning Implicit Fields for Generative Shape Modeling. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 5932-5941.

https://doi.org/10.1109/cvpr.2019.00609

[17] Lipman, Y. (2021) Phase Transitions, Distance Functions, and Implicit Neural Representations. *Proceedings of the* 38*th International Conference on Machine Learning*, Online, 18-24 July 2021, 6702-6712.

[18] Zheng, M., Yang, H., Huang, D. and Chen, L. (2022) ImFace: A Nonlinear 3D Morphable Face Model with Implicit Neural Representations. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 18-24 June 2022, 20311-20320. https://doi.org/10.1109/cvpr52688.2022.01970

[19] Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S. and Geiger, A. (2019) Occupancy Networks: Learning 3D Reconstruction in Function Space. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 4455-4465. https://doi.org/10.1109/cvpr.2019.00459

[20] Giebenhain, S., Kirschstein, T., Georgopoulos, M., Rünz, M., Agapito, L. and Nießner, M. (2023) Learning Neural Parametric Head Models. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 17-24 June 2023, 21003-21012. https://doi.org/10.1109/cvpr52729.2023.02012

[21] Ramon, E., Triginer, G., Escur, J., Pumarola, A., Garcia, J., Giro-i-Nieto, X., *et al.* (2021) H3D-Net: Few-Shot High-Fidelity 3D Head Reconstruction. 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Montreal, 10-17 October 2021, 5600-5609. https://doi.org/10.1109/iccv48922.2021.00557

[22] Yenamandra, T., Tewari, A., Bernard, F., Seidel, H., Elgharib, M., Cremers, D., *et al.* (2021) i3DMM: Deep Implicit 3D Morphable Model of Human Heads. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 20-25 June 2021, 12798-12808. https://doi.org/10.1109/cvpr46437.2021.01261

[23] Mildenhall, B., Pratul, P.P., Tancik, M., Barron, J.T., *et al.* (2020) NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. 2020 *European Conference on Computer Vision*, Glasgow, 23-28 August 2020, 99-106.

[24] Wang, K.Z., Tan, Y.Z., Fu, Y.T., *et al.* (2025) Early Smoke Segmentation with Dense Multi-Scale Context and Hierarchical Feature Fusion Attention. https://link.cnki.net/urlid/11.2127.tp.20250225.1502.009

[25] Li, B. and Lima, D. (2021) Facial Expression Recognition via ResNet-50. *International Journal of Cognitive Computing in Engineering*, **2**, 57-64. https://doi.org/10.1016/j.ijcce.2021.02.002

[26] Zhu, X., Lei, Z., Liu, X., Shi, H. and Li, S.Z. (2016) Face Alignment across Large Poses: A 3D Solution. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 146-155. https://doi.org/10.1109/cvpr.2016.23

[27] Yang, H., Zhu, H., Wang, Y., Huang, M., Shen, Q., Yang, R., *et al.* (2020) FaceScape: A Large-Scale High Quality 3D Face Dataset and Detailed Riggable 3D Face Prediction. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 598-607. https://doi.org/10.1109/cvpr42600.2020.00068