ISSN: 2688-8645

Weakly Supervised 3D Face Reconstruction with Joint Spatial and Frequency Domain Information

Haojie Diao[1, *], Xingguo Jiang[1, 2]

[1]School of Automation and Information Engineering, Sichuan University of Science and Engineering, Yibin 644000, China

[2]Artificial Intelligence Key Laboratory of Sichuan Province, Sichuan University of Science and Engineering, Yibin 644000, China

*2207896191@qq.com

Abstract

3D face reconstruction is an important research direction in computer vision, and its goal is to recover a 3D face model from a single face picture. In the absence of real 3D face data, how to reconstruct a 3D face with a high degree of realism has become a hot research topic in recent years. Existing reconstruction algorithms usually rely on 3D labels generated from a large number of 2D face images as training data, however, inaccurate data will seriously affect the reconstruction quality. For this reason, the paper proposes a joint spatial-frequency domain decoupled weak supervision to achieve 3D face reconstruction, the main idea of which is to construct a multi-level loss function by using the weakly supervised information extracted from the spatial domain, and separating the frequency-domain information between the input and the rendered image in the frequency domain, and minimizing the difference between the two by difference computation. The method combines deep learning with 3D deformable models to reconstruct 3D models with high quality texture and shape from only a single face image. Quantitative experiments on the AFLW2000-3D and MICC Florence datasets show that the normalized average error in the small pose interval is as low as 2.42%, and the face reconstruction accuracy in the outdoor scene is 0.98 ± 0.22 mm. Qualitative experiments on the MoFa-test, MICA datasets show that when faced with reconstruction with different poses, lighting, and expressions , our method outperforms other state-of-the-art reconstruction methods.

Keywords

Deep Learning; 3D Deformable Model; 3D Face Reconstruction; Weakly Supervised Method; Discrete Fourier Transform.

1. Introduction

3D face reconstruction has been a popular research direction in computer vision, with wide applications in face attribute editing [1], 3D gaming [2], virtual reality [3], and other fields. In the early days, practitioners often needed to use expensive equipment such as laser scanners in order to obtain accurate 3D face data, as opposed to recovering a 3D face from a single face image, which requires no expensive equipment and has a wider range of application scenarios. The traditional 3D face reconstruction techniques are mainly divided into two kinds, one is to use the lighting, texture and other information of the face image to establish a mathematical model [4], and infer the 3D structure of the face through mathematical analysis, the method often exists in the problem of large computa-tional cost and poor reconstruction robustness. Another method is based on the 3D Mor-phable Model (3DMM) [5] for iterative fitting to complete, the method through the 3D face data set for principal component analysis, the

ISSN: 2688-8645

establishment of low latitude linear space to characterize the 3D face, by adjusting the texture and shape and other parameters can be obtained by the corresponding face model, however, the method in the iterative process is easy to fall into the However, this method tends to fall into the local optimal solution in the iterative process, resulting in poor accuracy of the reconstructed 3D face. With the rapid development of deep learning, researchers have begun to use deep neural networks to predict the 3DMM coefficients of 2D face images, however, the first problem to be faced by using deep learning for 3D face reconstruction is the insufficient 3D face data and poor reconstruction accuracy.

In order to solve the problem of insufficient 3D face data, Richardson et al. [6] innova-tively introduced convolutional neural networks into 3D face reconstruction and input them as a set of training pairs to train the network. Tran et al. [7] proposed a method to es-timate the 3DMM parameters from a dataset, transforming the complex reconstruction task into a task of regressing the 3DMM coefficients using a neural network, which ena-bles supervised reconstruction by producing a 3D labeled dataset. The supervised recon-struction based method requires high accuracy of the synthetic labels. In order not to rely too much on 3D data, researchers have started to explore self-supervised or unsupervised methods to reconstruct 3D faces. Tewari et al. [8] proposed a method to learn 3DMM pa-rameters from unlabeled 2D datasets by regressing texture, shape, illumination, and other parameters in the face image through an autoencoder without any 3D labeled supervisory signals during the training process. Chen et al. [9] transformed face images into a specific rendering style by CGAN and reconstructed the rendered 2D images in 3D based on 3DMM and used a new reprojection loss function to constrain the network, which realized unsupervised 3D face reconstruction by loss function design.

In order to enhance the realism of the reconstructed face shape and texture, the re-searchers designed different layers of loss functions to constrain the characteristics of the reconstructed face model. Tewari et al. [8] trained an autoencoder to regress 3DMM coeffi-cients against texture parameters, however, the method only uses a pixel-based loss func-tion between the input image and the rendered image, and the network is prone to confuse the differences between the variables of interest. Genova et al. [10] proposed an identity loss function based on the former approach, which utilizes pre-trained face recognition networks to constrain the difference between the input and the cosine distance between the input and the output. Deng et al. [11] used a skin detector to apply different weights to the pixel points to train the network to focus on the skin color region in a weakly super-vised manner. Lin et al. [12] used a face segmentation network to pre-process the input image to generate a global mask that contains only the face in order to address the effect of occlusion on the 3D face, and trained the reconstruction and segmentation networks sim-ultaneously in a weakly supervised manner. In addition, some 3D face reconstruction methods [13] [14] use UV mapping to map the texture onto the 3D face surface, which can be stretched or compressed in the face of gesture face reconstruction, resulting in texture distortion.

In summary, existing weakly supervised reconstruction methods tend to focus on constraining the features in the spatial domain when designing the loss function, and pay less attention to the edge high-frequency features, such as simply calculating the locally weighted contour shapes and the key point losses of each part, which leads to the diffi-culty of efficiently constraining the model to reconstruct the details. To alleviate this situa-tion, this paper designs a multi-level loss function based on spatial domain information features that can learn multiple weakly supervised information from a large number of unlabeled 2D images to constrain the model learning. Further, this paper innovatively in-troduces the frequency-domain loss [15] to the 3D face reconstruction task, and designs a frequency-domain loss function to analyze the error between different frequency compo-nents. In order to verify the effectiveness of the

ISSN: 2688-8645

method, the reconstruction is visualized on MoFA-test [8] with the previous method, and the paper's method shows significant im-provement in both shape and texture reconstruction..

2. Key Technologies

2.1. 3D Face Generation

(1) Three-dimensional deformable model (3DMM)

In the 3D deformable model [19], the face space is assumed to be a linear space, and any 3D face can be linearly combined by other 3D faces, and the dimensions of the shape, expression, and texture parameters contained in the newly generated 3D face after dimensionality reduction decomposition using principal component analysis (PCA) are respectively $\alpha \in \mathbb{R}^{80}$, $\beta \in \mathbb{R}^{64}$, $\delta \in \mathbb{R}^{80}$. The formula for its generation is as follows:

$$\begin{aligned} S &= S(\alpha,\beta) = \overline{S} + B_{id}\alpha + B_{\exp}\beta \\ T &= T(\delta) = \overline{T} + B_t\delta \end{aligned} \tag{1}$$

where $\overline{S}$ and $\overline{T}$ are the average face shape and texture; B_{id}, $B_{\exp}$ and B_t are the principal component bases of the shape, expression and texture of the face, and , α , β and δ represent the 3DMM coefficients of shape, expression and texture, respectively.

(2) Camera model

After obtaining the 3D face model, it is necessary to constrain the generation of the shape by projecting it onto the 2D image using the perspective projection camera model with the following formula:

$$\begin{aligned} V' &= R \times \left(\overline{S} + B_{\mathrm{id}}\partial + B_{\exp}\beta\right) + t \\ V_p &= \mathbf{Pr} \times V' \end{aligned} \tag{2}$$

where $R \in SO(3)$ is the rotation matrix, t is the translation vector of the 3D face in space, $\mathbf{Pr} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$ is the projection matrix, V' is the set of points in the 3D space after rotation and translation, and V_p is the set of points of the 2D face after projection.

(3) Illumination

The face is assumed to be a Lambertian reflection model and the scene illumination is estimated based on the spherical harmonic function. The surface normal vector V_i of the face is calculated from the 3D face shape n_i, plus the 3D face texture t_i, and each light reflection texture of the 3D face is given by equation (3):

$$C(n_i, t_i | \ \gamma) = t_i \times \sum_{b=1}^{B^2} \gamma_b \Phi_b(n_i) \tag{3}$$

where the function $\Phi_b : \mathbb{R}^3 \to \mathbb{R}$ is the spherical harmonic basis function and γ_b is the parameter of the corresponding spherical harmonic basis function, with reference to the work of Deng et al. [11], setting B = 3 and assuming a white light illumination $\gamma \in \mathbb{R}^9$.

(4) Differentiable rendering

In order to supervise the pixel loss between the input image and the reconstructed face, it is usually necessary to render the reconstructed 3D face model onto a 2D plane. The specific steps are as follows: (a)Obtain the initial parameters for targeting by neural network (b) Define a differentiable rendering function that satisfies the principle of gradient descent (c) Define an objective function to guide the optimization process (d) The rendering function takes the parameters of shape, camera, material, and lighting as inputs and outputs the rendered 2D

ISSN: 2688-8645

image (e) The network updates the scene parameters according to the calculated gradient parameters.

2.2. Face Segmentation Algorithm

Since occluders such as glasses, hats, hair, etc. are prone to appear in the face data used for training, resulting in texture distortion in the corresponding occluded regions af-ter reconstruction, preprocessing operations on the training data are needed to increase the robustness of the model to occluders. Since the traditional Bayesian-based trained skin mask does not distinguish skin color from occluders well [11] , the semantic segmenta-tion network BiseNet [18] is trained on the CelebAMask-HQ dataset [17], and the finely discretized face segmentation mask enables the network to be constrained by different weights according to different regions when performing the loss computation.

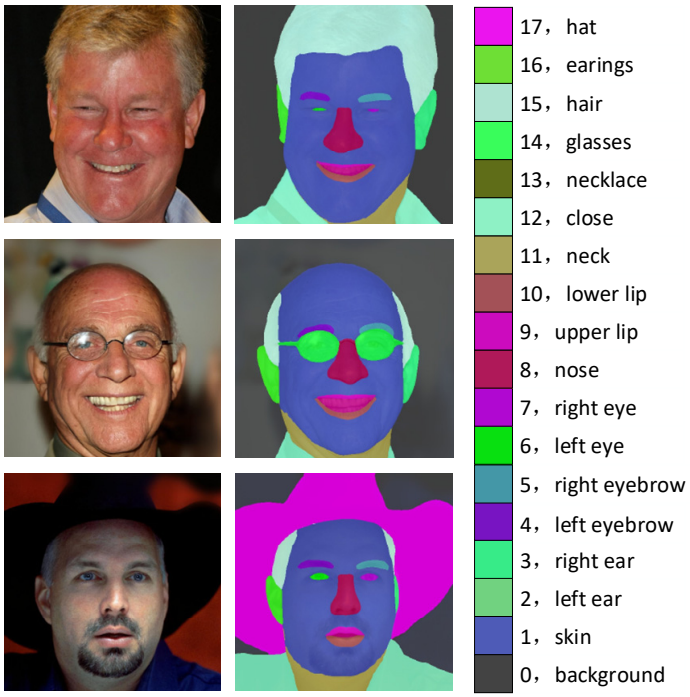

Figure 1. Examples of face segmentation algorithm

2.3. Overview of Our Approach

The method flow of this paper is shown in Fig. 2, firstly, given a 2D face image with-out any 3D labels, the face segmentation algorithm in section 2.2 is used to compute the segmentation mask, which can be utilized to remove the influence of occlusions in a better way, and at the same time, different proportions of weights are given to constrain the face features, and furthermore, 68 landmarks are computed by using the landmark detection algorithm to serve as the weak supervisory information. Subsequently, the feature coefficients of a single face, such as texture, shape, and pose, are regressed by a feature encoder, and then the 3D face is reconstructed by 3DMM decoding. In order to minimize the gap between the reconstructed face and the input face, the reconstructed 3D face is rendered to the 2D level using a differentiable renderer given the camera parameters and illumination parameters, and the input image and the rendered image are converted to the frequency domain range by the discrete Fourier transform, and the frequency magnitude and the phase are used to separately constrain the intensity of the different frequency components based on the multi-level loss function in the spatial domain. Based on the multi-level loss function in the spatial domain, the frequency amplitude and phase are used to constrain the intensity of different frequency components and different edge texture features, which not only retains the high-frequency information better, but also

ISSN: 2688-8645

measures the image similarity in the spatial and frequency domains to realize more accurate reconstruction.

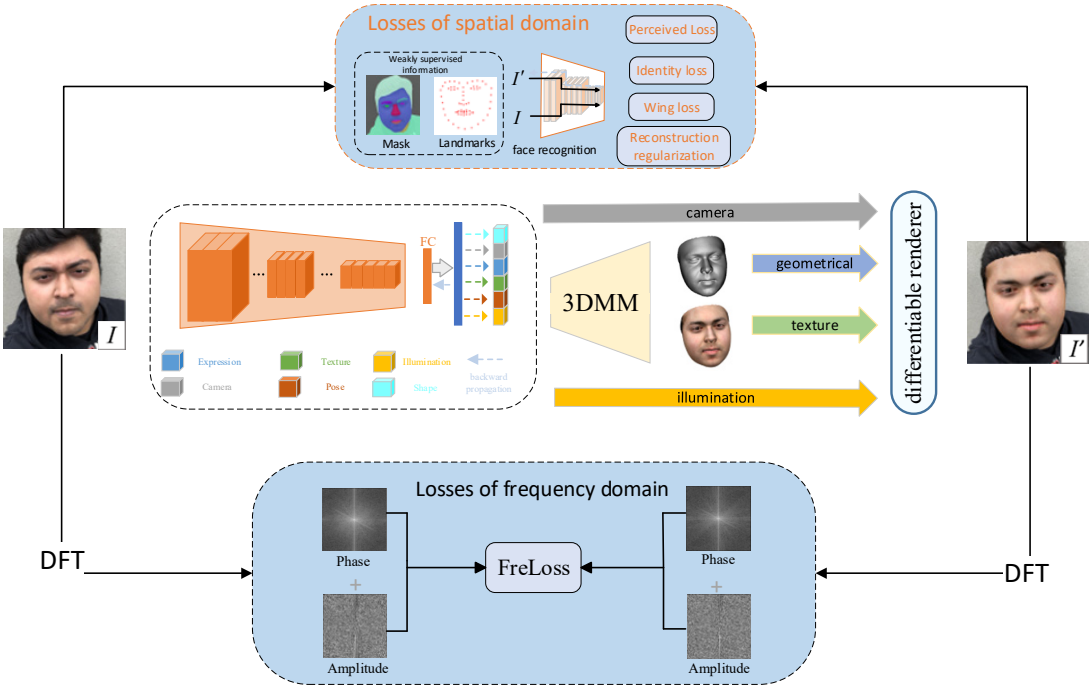

Figure 2. Framework diagram of 3D face reconstruction

3. Loss Function

3.1. Loss Function based on Spatial Domain

(1) Perceived loss of low-level

In order to calculate the pixel difference between the input image and the rendered image, a semantic segmentation network is used to generate a face mask with different weights for each part so that the network is able to constrain the pixel loss between the two images when performing the loss calculation. Its loss is defined as follows:

$$L_{photo} = \frac{\sum_{(i,j)\in R} P_{(i,j)} W_{(i,j)} \left\| I_{(i,j)} - I'_{(i,j)} \right\|_2}{\sum_{(i,j)\in R} P_{(i,j)}} \tag{4}$$

Where $P_{(i,j)} = \begin{cases} 1, & \text{Pixels in face area} \\ 0, & \text{Pixels not in face area} \end{cases}$ is the weight of face and background segmentation and $W_{(i,j)}$ is the weight of each region of the face, where the mask weights at the eyes, mouth, nose, and skin are 1.5:1.2:1.5:1. The use of weight masks allows the network to focus only on the face region, which improves the robustness of the network to occlusions to some extent.

(2) Landmark loss of mid-level

Face landmarks contain the main contours of each region of the face and can effectively allow the network to constrain the face pose. In this paper, we use a landmark detection algorithm to detect 68 landmarks $\{l_n\}$ for all faces in the dataset and use them as a weakly supervised signal. During the training process, the 3D landmarks $\{l'_n\}$ of the reconstructed model are projected onto a 2D image to compute its wing loss [22]:

ISSN: 2688-8645

$$WingLoss = \begin{cases} w\ln\left(1+\frac{|\Delta L|}{\epsilon}\right), & \text{if } |\Delta L| < w \\ |\Delta L| - c, & \text{otherwise} \end{cases} \tag{5}$$

where $\Delta L = l_n' - l_n$ represent the error between landmarks, w is the range threshold between linearity and nonlinearity of the wing loss curve, and ϵ is the curvature that constrains the nonlinear region, which is set according to the reference [22], with w=10 and e=2 , to avoid problems such as gradient explosion.

(3) Identity loss of deep-level

Using only low dimensional pixel loss leads to a local minimum problem, so the use of high dimensional features of the image is considered to avoid this problem. The pre-trained face recognition network is first used to regress the 512-dimensional feature vectors of the input image and the rendered image, and then the cosine similarity of the two images is minimized by equation (6):

$$L_{per} = 1 - \frac{f(I)f\left(I^{'}\right)}{\| f(I)\| \left\| f\left(I^{'}\right)\right\|} \tag{6}$$

where $f(\cdot)$ denotes the feature vector in the face recognition network.

(4) Reconstruction regularization

In order to prevent problems such as shape and texture degradation of 3D faces during reconstruction, regularization is used to constrain the 3DMM parameters. It is defined as follows:

$$L_{\text{reg}} = \omega_{id} \| \alpha\|^2 + \omega_{\text{exp}} \| \beta\|^2 + \omega_{shape} \| \delta\|^2 + \omega_{light} \| \gamma\|^2 \tag{7}$$

where the weights of the parameters are set respectively: $\omega_{id} = 10^{-4}$, $\omega_{\text{exp}} = 0.8\times10^{-4}$, $\omega_{shape} = 10^{-1}$, $\omega_{light} = 10^{-2}$

3.2. Loss Function based on Frequency Domain

Frequency domain loss is a loss function that measures the quality of 2D image reconstruction, where the amplitude spectrum represents the amplitude of each frequency component in the image and the phase spectrum is the phase of each frequency component. Since this paper is to render the reconstructed 3D face model back to the 2D image so the loss function can be established by calculating the difference between the amplitude spectrum and the phase spectrum after the discrete Fourier transform (DFT). Its calculation formula is as follows:

$$F(u,v) = \sum_{x=0}^{M-1}\sum_{y=0}^{N-1} f(x,y) e^{-i2\pi\left(\frac{ux}{M}+\frac{vy}{N}\right)} \tag{8}$$

$$e^{-i2\pi\left(\frac{ux}{M}+\frac{vy}{N}\right)} = \cos 2\pi\left(\frac{ux}{M}+\frac{vy}{N}\right) - i\sin 2\pi\left(\frac{ux}{M}+\frac{vy}{N}\right) \tag{9}$$

where the size of the input face image is $M\times N$, (x,y) is the pixel coordinate in the spatial domain of the image, (u,v) is the coordinate in the frequency domain, e and i represent Eulerian and imaginary units respectively. Let $a(u,v)$ and $b(u,v)$ represent the real and imaginary parts of $F(u,v)$ respectively, then the expressions for amplitude and phase are as follows:

ISSN: 2688-8645

$$|F(u,v)|=\sqrt{a^2(u,v)+b^2(u,v)} \tag{10}$$

$$\theta(u,v)=\arctan\frac{b(u,v)}{a(u,v)} \tag{11}$$

(1) Loss of amplitude

In order to be able to introduce a loss function in the network in the frequency domain, and this loss function needs to satisfy the principles of differentiability and support for gradient descent, the difference between the input image and the rendered image in the frequency domain is measured by using the frequency distance, mapping each frequency value to a two-dimensional coordinate and represented by a Euclidean vector as shown in Figure 3.

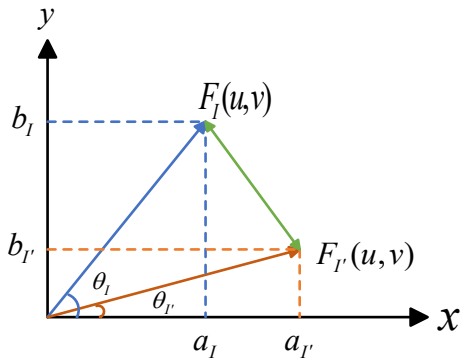

Figure3. Framework diagram of 3D face reconstruction

Its frequency domain distance is defined as follows:

$$d\left(\vec{r_I},\vec{r_{I'}}\right)=\left\|\vec{r_I}-\vec{r_{I'}}\right\|_2^2=\left|F_I(u,v)-F_{I'}(u,v)\right|^2 \tag{12}$$

where the frequency values of the input image and the rendered image at the frequency coordinates (u,v) are $F_I(u,v)=a_I+b_Ii$, $F_{I'}(u,v)=a_{I'}+b_{I'}i$. In practice, it is necessary to orthogonalize each frequency, i.e., divide it by $\sqrt{MN}$ to ensure the smoothness of the gradient, and the amplitude loss can be obtained by calculating the L2 between F_I and $F_{I'}$:

$$FALoss=\frac{1}{MN}\sum_{u=0}^{M-1}\sum_{v=0}^{N-1}\left|F_I(u,v)-F_{I'}(u,v)\right|^2 \tag{13}$$

(2) Loss of phase

The magnitude spectrum responds to the intensity distribution of the image, while the phase spectrum contains the edge details and overall texture information of the image, using the phase angle difference 0 as part of the loss function can be very helpful for the model to learn the difference between the two images. The definition is as follows:

$$FPLoss=\frac{1}{MN}\sum_{u=0}^{M}\sum_{v=0}^{N}\left|\theta_I(u,v)-\theta_{I'}(u,v)\right| \tag{14}$$

Since the frequency loss focuses on the texture and structural features of the face image, while the magnitude loss mainly focuses on the edges, contrast and other features of the image, adding them together can be a more comprehensive consideration of the features of the image, the total loss in the frequency domain used in this paper is as follows, where a and b in Eq. 15, where α_1 are α_2 taken to be 0.5 and 0.5, respectively.

$$FreLoss=\alpha_1 FALoss+\alpha_2 FPLoss \tag{15}$$

ISSN: 2688-8645

4. Experimental Analysis and Result Discussion

4.1. Dataset and Experimental Setup

To train the model, one million images were selected as training data from two publicly available face datasets, VGGFace2 [23] and CelebA [24], the input image size was 224×224, the encoder was ResNet50, the batch size was set to 16, the learning rate was $1e-4$, the deep learning framework used was Pytorch, and the optimization iterator was Adam. the spatial domain multi-level total loss function of each loss weight is $w_{photo}=3.0$, $w_{wing}=2.4\times10^{-3}$, $w_{per}=0.24$, $w_{reg}=4\times10^{-4}$. The preprocessed landmarks information along with the segmentation mask is fed as a weakly supervised signal to the network for training.

4.2. Qualitative Analysis

In order to be able to visually compare the quality of reconstructed faces, several im-ages were selected for 3D face reconstruction on MoFA-test and visualized against the al-gorithms of Genova et al. [10], Gecer et al. [25], Tran et al. [6], Li et al. [26], and Tewari et al. [8], the results of which are shown in Figure 4. Some of the results of previous work are derived from references [10][25].As can be seen in Fig. 4, the 3D face obtained by our method outperforms other methods in terms of facial expression as well as the shape of the five senses, and better reflects the identity of the face, and the reconstruction of the shape benefits from the complementation of the frequency-domain information, which allows the network to accurately recover the expression features. The reconstructed faces of other methods have no big difference when facing faces with different lighting and skin color, and cannot reflect the appearance characteristics.

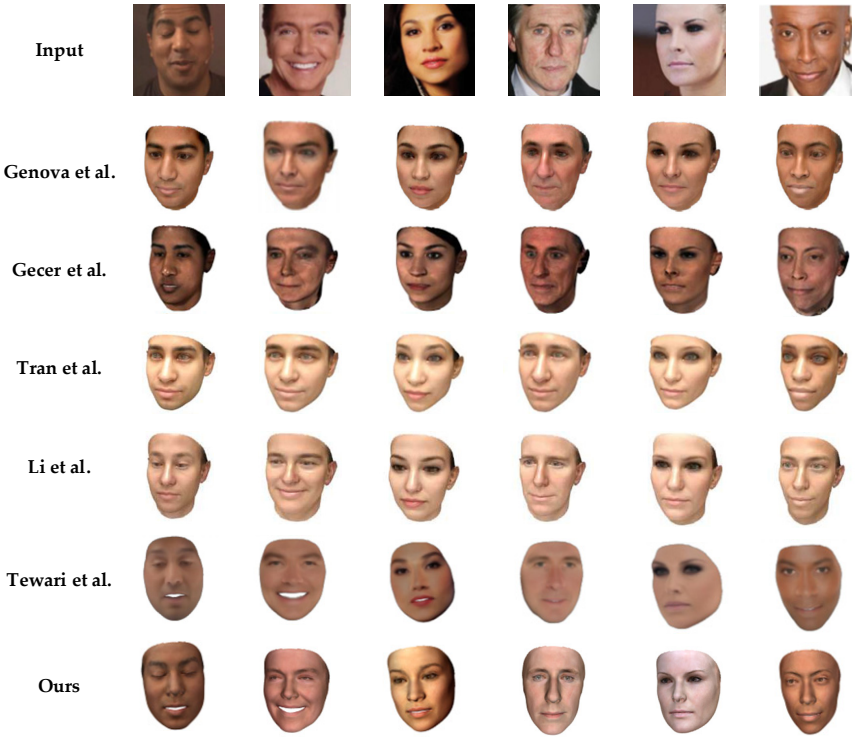

Figure 4. Visualization comparison with previous methods on the MoFA-test dataset.

In order to verify the robustness of the model to occluded images, several face images with poses within the interval of were selected from the AFLW2000-3D [29] dataset. And qualitatively analyzed with 3DDFA-V2 [13] and DECA [14], the results are shown in Fig. 5.

ISSN: 2688-8645

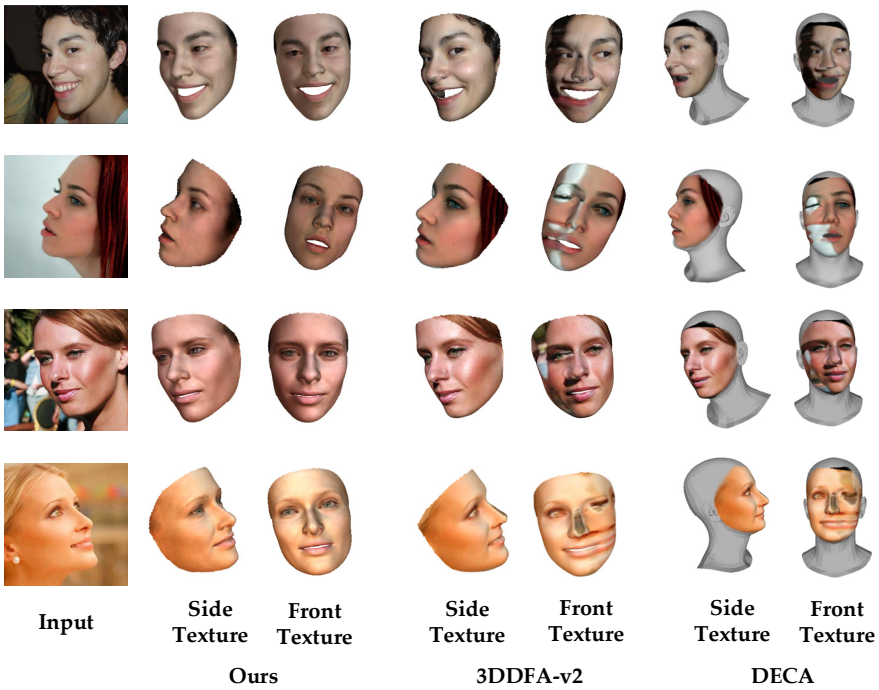

Figure 5. Visualization comparison with previous methods on the MoFA-test dataset.

As can be seen from Fig. 5, the other two methods directly sample from the input image to obtain the texture mapping, and for the reconstruction of occluded faces, the texture information of the occluded parts cannot be extracted from the input 2D face image, resulting in texture distortion. Thanks to the face segmentation algorithm in this paper and the weight constraints on the face region, a 3D face with real identity features can be reconstructed when facing the reconstruction of occluded images.

In order to verify the generalization of the model, multiple pictures of the same person with different states (illumination, expression, etc.) are reconstructed, and the results are shown in Fig. 6, which shows that for the same identity under the condition of different states, a 3D face with consistency can still be reconstructed.

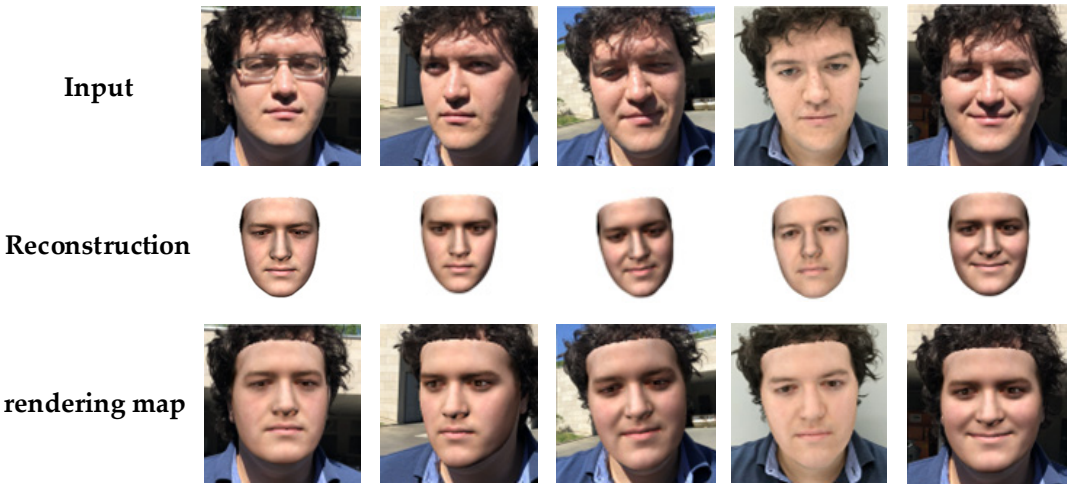

Figure 6. Visualization results for the same person in different states.

4.3. Quantitative Analysis

Consistent with references [27][28], the normalized mean error (NME) was used on the AFLW2000-3D dataset to evaluate the accuracy of the reconstruction and to compare it with other methods for quantitative analysis. It is defined as follows:

ISSN: 2688-8645

$$NME = \frac{1}{N}\sum_{k=1}^{N}\frac{\|x_k - y_k\|_2}{d} \tag{16}$$

where x_k and y_k represent the coordinates of the key points of the ground-truth and reconstructed face, d is the size of the input face box, and the smaller the NME value represents the smaller gap with the ground-truth.

As listed in Table 1, the dataset is divided into $[0° \sim 30°]$, $[30° \sim 60°]$, $[60° \sim 90°]$ according to its pose labels, and the average NME values in these three intervals are given. In the table, our method achieves the optimal NME values in the small pose interval, and the optimal one in the large pose interval is MGCNet, but the method adopts multi-view images as input, which is robust to faces with large poses. In addition, SADRNet is trained using data with real 3D labels.

Table 1. Normalized Mean Error (NME, %) of Different Methods on the AFLW2000-3D Dataset

Methods	$[0° \sim 30°]$	$[30° \sim 60°]$	$[60° \sim 90°]$	**Average↓**
3DDFA [29]	3.78	4.54	7.93	5.42
DeFA [30]	-	-	-	4.50
PRNet [27]	2.75	3.51	4.61	3.62
3DDFA_V2 [13]	2.63	3.44	4.45	3.51
MGCNet [31]	2.63	**3.12**	**3.76**	3.45
SADRNet [28]	2.66	3.30	4.42	**3.2**
Ours	**2.42**	3.25	4.78	3.48

4.4. Ablation Experiment

In order to verify the effectiveness of each loss function, the effectiveness of each loss function is verified on the AFLW2000-3D dataset. The ablation experiments of each loss function are shown in Table 3, and the average NME values of the proposed method in this paper on the dataset gradually show a decreasing trend under the gradual superposition of the main four losses. Other weakly supervised methods usually use the loss functions as the first three items in the table, and on this basis, after adding the frequency domain loss proposed in this paper, the average NME value of the dataset is 3.011%.

Table 2. Results of ablation experiments for each loss function.

L_{photo}	L_{wing}	L_{per}	L_{fre}	**NME/%**
✓				3.227
✓	✓			3.143
✓	✓	✓		3.057
✓	✓	✓	✓	3.011

To verify the importance of frequency loss, we train the model with or without frequency loss and perform visualization experiments. As shown in Fig. 7, the frequency loss can help the network to learn some tiny features and get closer to the original image in the synthesis of eyes and textures.

ISSN: 2688-8645

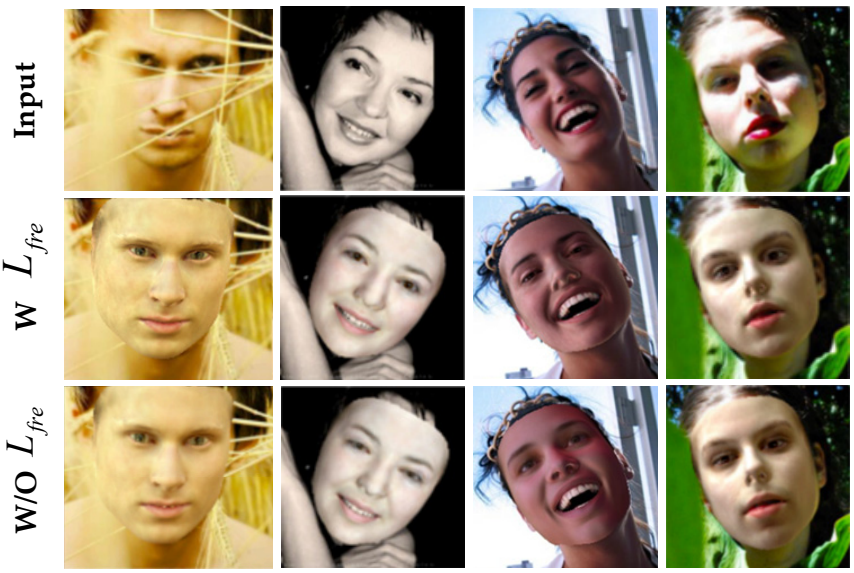

Figure 7. Comparison chart of frequency loss ablation experiments.

5. Conclusion

Existing face reconstruction algorithms focus on using the information in the spatial domain to establish the loss calculation, which lacks the attention to the detailed infor-mation in the frequency domain. In this paper, we propose a method that combines the spatial domain loss and the information domain loss, which does not require any real 3D face annotation, and can reconstruct a 3D face with real emotions based on a single face image. The main idea is to use the loss of the landmarks, the loss of the identity, and other the main idea is that on the basis of spatial domain loss, such as key point loss, identity loss, etc., the advanced face segmentation algorithm is used to constrain the different weight loss of each region of the face, so as to improve the robustness of the model to the occluded image. Further, on the basis of the spatial domain loss, the Fourier transform is used to transform the image into the frequency domain to construct the frequency domain loss, and the combination of the two is used to construct a weakly supervised 3D face re-construction algorithm, and the reconstructed 3D face is enhanced in terms of both shape and texture. However, the proposed method is still limited to the linear space of 3DMM, and the reconstructed 3D face is a bit smooth, which can't emphasize the detailed features, and can't reconstruct the neck, hair and other parts of the face, and the next research work should consider how to reconstruct the whole head region instead of a single face region.

Acknowledgments

This work was supported in part by the Scientific Research Foundation of Sichuan University of Science and Engineering under Grant 2019RC12.

References

[1] Z Abate, Andrea F., et al. "2D and 3D face recognition: A survey." Pattern recognition letters 28.14 (2007): 1885-1906.

[2] Diao, Haojie, et al. "3D Face Reconstruction Based on a Single Image: A Review." IEEE Access (2024).

[3] JingTing W, et al. Review of Single-Image 3D Face Reconstruction Methods. Computer Engineering and Applications 2023,59(17):1-21.

[4] Yue, W, et al. 3D Face Shape and Texture Reconstruction Based on Weakly Supervised Learning. Computer Systems & Applications, 2020, 29(11):183-189

[5] Blanz, Volker, and Thomas Vetter. "A morphable model for the synthesis of 3D faces." Seminal Graphics Papers: Pushing the Boundaries, Volume 2023. 157-164.

[6] Richardson, Elad, Matan Sela, and Ron Kimmel. "3D face reconstruction by learning from synthetic data." 2016 fourth international conference on 3D vision (3DV). IEEE, 2016.

[7] Tuan Tran, Anh, et al. "Regressing robust and discriminative 3D morphable models with a very deep neural network." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[8] Tewari, Ayush, et al. "Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction." Proceedings of the IEEE international conference on computer vision workshops. 2017.

[9] Chen Z, Wang Y, Guan T, et al. Transformer-based 3d face reconstruction with end-to-end shape-preserved domain transfer[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(12): 8383-8393.

[10] Genova K, Cole F, Maschinot A, et al. Unsupervised training for 3d morphable model regression[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 8377-8386.

[11] Deng Y, Yang J, Xu S, et al. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. 2019: 0-0.

[12] Lin J, Yuan Y, Shao T, et al. Towards high-fidelity 3d face reconstruction from in-the-wild images using graph convolutional networks[C]//Proceedings of the ieee/cvf conference on computer vision and pattern recognition. 2020: 5891-5900.

[13] Guo J, Zhu X, Yang Y, et al. Towards fast, accurate and stable 3d dense face alignment[C]//European Conference on Computer Vision. Cham: Springer International Publishing, 2020: 152-168.

[14] Feng Y, Feng H, Black M J, et al. Learning an animatable detailed 3D face model from in-the-wild images[J]. ACM Transactions on Graphics (ToG), 2021, 40(4): 1-13.

[15] Jiang L, Dai B, Wu W, et al. Focal frequency loss for image reconstruction and synthesis [C]// Proceedings of the IEEE/CVF international conference on computer vision. 2021: 13919-13929.

[16] Huang G B, Mattar M, Berg T, et al. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments[C]//Workshop on faces in'Real-Life'Images: detection, alignment, and recognition. 2008.

[17] Lee C H, Liu Z, Wu L, et al. Maskgan: Towards diverse and interactive facial image manipulation [C]// Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 5549-5558.

[18] Yu C, Wang J, Peng C, et al. Bisenet: Bilateral segmentation network for real-time semantic segmentation[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 325-341.

[19] Paysan P, Knothe R, Amberg B, et al. A 3D face model for pose and illumination invariant face recognition[C]//2009 sixth IEEE international conference on advanced video and signal based surveillance. IEEE, 2009: 296-301.

[20] Cao C, Weng Y, Zhou S, et al. Facewarehouse: A 3d facial expression database for visual computing[J]. IEEE Transactions on Visualization and Computer Graphics, 2013, 20(3): 413-425.

[21] Ravi N, Reizenstein J, Novotny D, et al. Accelerating 3d deep learning with pytorch3d[J]. arXiv preprint arXiv:2007.08501, 2020.

[22] Feng Z H, Kittler J, Awais M, et al. Wing loss for robust facial landmark localisation with convolutional neural networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 2235-2245.

[23] Cao Q, Shen L, Xie W, et al. Vggface2: A dataset for recognising faces across pose and age[C]//2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018). IEEE, 2018: 67-74.

ISSN: 2688-8645

[24] Liu Z, Luo P, Wang X, et al. Deep learning face attributes in the wild[C]//Proceedings of the IEEE international conference on computer vision. 2015: 3730-3738.

[25] Gecer B, Ploumpis S, Kotsia I, et al. Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 1155-1164.

[26] Li C, Morel-Forster A, Vetter T, et al. Robust model-based face reconstruction through weakly-supervised outlier segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 372-381.

[27] Liu F, Zeng D, Zhao Q, et al. Joint face alignment and 3d face reconstruction[C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14. Springer International Publishing, 2016: 545-560.

[28] Ruan Z, Zou C, Wu L, et al. Sadrnet: Self-aligned dual face regression networks for robust 3d dense face alignment and reconstruction[J]. IEEE Transactions on Image Processing, 2021, 30: 5793-5806.

[29] Zhu X, Liu X, Lei Z, et al. Face alignment in full pose range: A 3d total solution[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 41(1): 78-92.

[30] Liu Y, Jourabloo A, Liu X. Learning deep models for face anti-spoofing: Binary or auxiliary supervision [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 389-398.

[31] Shang J, Shen T, Li S, et al. Self-supervised monocular 3d face reconstruction by occlusion-aware multi-view geometry consistency[C]//European Conference on Computer Vision. Cham: Springer International Publishing, 2020: 53-70.

[32] Booth J, Antonakos E, Ploumpis S, et al. 3d face morphable models" in-the-wild"[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 48-57.

[33] Bagdanov A D, Del Bimbo A, Masi I. The florence 2d/3d hybrid face dataset[C]//Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding. 2011: 79-80.