Multi-View 3D Face Reconstruction Based on Flame

Wenzhuo Zheng, Junhao Zhao, Xiaohong Liu, Yongyang Pan, Zhenghao Gan, Haozhe Han and Ning Liu

August 2, 2023

CGI2023 manuscript No.
(will be inserted by the editor)

Multi-view 3D Face Reconstruction Based on Flame

Wenzhuo Zheng* · **Junhao Zhao*** · **Xiaohong Liu**† · **Yongyang Pan** · **Zhenghao Gan** · **Haozhe Han** · **Ning Liu**†

Abstract At present, face 3D reconstruction has broad application prospects in various fields, but the research on it is still in the development stage. In this paper, we hope to achieve better face 3D reconstruction quality by combining multi-view training framework with face parametric model Flame, propose a multi-view training and testing model **MFNet** (Multi-view Flame Network). We build a self-supervised training framework and implement constraints such as multi-view optical flow loss function and face landmark loss, and finally obtain a complete MFNet. We propose innovative implementations of multi-view optical flow loss and the covisible mask. We test our model on AFLW and facescape datasets and also take pictures of our faces to reconstruct 3D faces while simulating actual scenarios as much as possible, which achieves good results. Our work mainly addresses the problem of combining parametric models of faces with multi-view face 3D reconstruction and explores the implementation of a Flame based multi-view training and testing framework for contributing to the field of face 3D reconstruction.

Keywords 3D reconstruction · Human face · Multi-view · Parametric model

1 Introduction

3D reconstruction is a technology that uses 2D information such as images or videos to recover and reconstruct specific 3D objects or scenes [13]. Among the various 3D reconstruction techniques, human face 3D reconstruction has been a very popular research topic. Face 3D reconstruction mainly focuses on the reconstruction of human facial regions, and broadly speaking, also includes hair, ear, neck and other regions. The human face is a special 3D object that has not only more complex shape and texture features, but also strong prior constraints. This poses a great challenge to face 3D reconstruction on one hand, and on the other hand, it also provides feasible technical approaches to reconstruct the face 3D structure from 2D information, and face parametric model is one of them. The face parametric model is a statistical model based on a large number of faces, and its core idea is that faces can be matched one-to-one in the 3D feature space and can be obtained by weighted linear summation of orthogonal bases for a large number of other faces. The most widely used model is 3DMM [1,2], but it has two core problems: (1) 3DMM is in a low-dimensional space and thus the face detail characterization is weak; and (2) 3DMM only reconstruct the front face region without neck or hindbrain. Therefore, we choose Flame [11], which has better characterization of details and more complete reconstruction. FLAME has three parameters: shape, pose, and expression, which can more accurately classify faces into more dimensions. And The face reconstructed by FLAME not only includes the lateral face, the back of the head, but also the neck. However, there is not much research work on Flame so far, and there is a gap in the field of multi-view training using Flame. Our work fills this gap and makes an exploratory contribution to Flame-based multi-view training.

In the past decade, deep learning technologies based on neural networks have become a dominant trend in the field of computing and artificial intelligence. Their end-to-end training method and simple and general learn-

* Equal contribution.
† Corresponding authors.

Wenzhuo Zheng, Junhao Zhao, Xiaohong Liu, Yongyang Pan, Zhenghao Gan, Haozhe Han, Ning Liu
School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China.

This work was supported in part by the Shanghai Pujiang Program under Grant 22PJ1406800

ing paradigm have brought breakthroughs in many domains, and face 3D reconstruction is no exception. Some works [7,15] use neural networks to regress end-to-end to compute the inputs needed for face parameterization models, but are limited to single-view, while our proposed MFNet can utilize features from multiple views and fuse them to obtain more complete face information. In this paper, we use Flame as a powerful tool to reconstruct fine-grained 3D face models with low cost and only 2D RGB images.

Our main contributions are listed as follows:

- We innovatively combine multi-view training with Flame, propose a multi-view self-supervised framework and implement a complete multi-view training and testing process. Our proposed model MFNet achieve good results on both test datasets and actual captured images.
- We propose a multi-view optical flow loss for our multi-view training framework and propose a novel implementation of the technical details such as covisible mask.

2 Related Work

2.1 Parametric model

In 1999, Blanz and Vetter et al. [1,2] proposed the 3D Morphable Model (3DMM) for the human face, which is the most widely used 3D face reconstruction model. Subsequent studies related to 3DMM have been published in the next decade, either by adding coefficients to the original model, such as Pascal Paysan et al. [9] updated the expression coefficients of the 3DMM model for BFM (Basel Face Model) model in 2017, or build larger datasets, such as James Booth et al. [4] built a dataset of 9663 faces, or propose better ways to optimize the solution coefficients, such as adding deep learning ideas to the coefficient solution in recent years to achieve better results [3,19], or make nonlinear adjustments to the model, such as the nonlinear 3DMM model proposed by Luan Tran et al. [18], but none of them have departed from the original framework of 3DMM. This also leads to the fact that these changes do not solve two core problems of 3DMM:(1) The 3DMM model parameter space is a relatively low-dimensional parameter space, the texture model is too simple, the generated face model is too average, it is difficult to reconstruct detailed features such as face wrinkles, and the recovery of details such as occlusions and shadows is poor; and (2) 3DMM only models the front human face, but it does not include other parts of the human head, while our work wants to reconstruct the whole face region as complete as possible, including the ear, neck and other regions. Therefore, we choose Flame[11] as our face parametric model.

Flame was proposed by Li Tianye et al., referring to the expression of the body model SMPL[12], combining linear blend skining (LBS) and the corresponding corrected blendshape. Not many researches have been done on Flame[7,15], and they are all limited to single-view. We want to utilize the features and data from multiple perspectives, so we propose a self-supervised multi-view training framework and achieve better reconstruction results.

2.2 Multi-view reconstruction

There are many works based on face parametric models, but very few of them[16,21] are trained using multi-view data, and the only ones are based on 3DMM. MVFNet[21] is the first work that proposed the idea of multi-view parametric model training, but it is based on 3DMM and the implementation is very rough, which lead to poor results. MGCNet[16] makes some improvements on its basis, proposing novel multi-view loss functions, using multi-view training, but only using a single image for testing. It does improve the quality of the face reconstruction, but the reconstructed faces were still rough and incomplete. The field of Flame based multi-view training still remains a gap. Our work implements a multi-view training framework for face reconstruction based on Flame to provide more information with multi-view input. To the best of our knowledge, MFNet is the first work on 3D face reconstruction using multi-view training and testing framework based on the face parametric model Flame, and we find that the combination of multi-view training framework and parametric model Flame can lead to better quality of face reconstruction.

3 Method

3.1 Overall architecture

MFNet learns to regress a parameterized face model with geometric detail on Multi-PIE dataset with multiview images of one person. In order to establish multiview geometric constraints and correct optical estimation for now we assume the facial images are taken at the same time under the same lighting condition, and this assumption can be satisfied by Multi-PIE dataset.

The overall architecture for our proposed is show in Figure 1. Resnet is a highly mature technology that has performed well in numerous image recognition and classification. So we extract features from each input image by a shared weight Resnet50, and then concatenate the features together and put them into a fully connected layer to regress a set of flame parameters for the person. Because the input is a multi-view images, we cannot directly obtain the pose parameters from the fully connected layer. We separate a pose and texture feature from Resnet50 for each perspective for subsequent reconstruction work.

Next, we will introduce Flame (Sec. 3.2), Feature extraction (Sec. 3.3), Differentiable renderer (Sec. 3.4) and Loss function (Sec. 3.5).

3.2 Flame

After extracting features from the multi view images in the input batch through Resnet50 and converting them into fully connected layers, we can obtain the desired Flame model input vectors $\vec{\beta}$, pose $\vec{\theta}$, expression $\vec{\psi}$. Next, the Flame model acts as a decoder to convert these hidden layer vectors into three-dimensional facial information.These three-dimensional information mainly consists of two parts, the first is the information of each vertex, such as coordinate T_P, Normal vector N_{uv} and faces F, and the second is landmark coordinates of the face. The equation of the Flame model is as follows:

$$M(\vec{\beta},\vec{\theta},\vec{\psi}) = W(T_P(\vec{\beta},\vec{\theta},\vec{\psi}),\mathbf{J}(\vec{\beta}),\vec{\theta},\mathcal{W}) \tag{1}$$

3.3 Feature extraction

A pre-trained model is a model that was previously trained on a large dataset, typically for a large-scale task, and then saved for reuse. Pretrained models can save time and improve performance by leveraging the knowledge learned from previous tasks. Here we use a pretrained model for better feature extractoin. In order to obtain better feature information, we use a fully connected layer to fuse the features extracted by Resnet50 from three perspectives together for consideration, thereby obtaining a more accurate model.

3.4 Differentiable renderer

After getting the 3D information of the face through Flame model, we need to use 3D rendering to get the 2D image.Our shadow facial image $B(alpha, l, N_{UV})$ is calculated based on the following equation:

$$B(\alpha, l, N_{uv})_{i,j} = A(\alpha)_{i,j} \odot \sum_{k=1}^{9} l_k H_k(N_{i,j}) \tag{2}$$

In the equation 2, $A(\alpha)$ represents UV albedo map, $N_{UV}$is the normal vector of the face surface output by Flame. $B_{i,j} \in R^3$, $A_{i,j} \in R^3$, $N_{i,j} \in R^3$ represents the various attributes of pixel (i, j) in the UV coordinate system. $\odot$ represents Hadamard product.

In addition, we also need to extract texture from the original input image and obtain vertex coordinates T_P and faces F to calculate the correspondence between points in the 3D mesh and the 2D texture map U_V. Then, the texture map I'_{uv} is obtained from the original input image by using this correspondence U_V, and the missing part in the middle is supplemented by bilinear interpolation. We extract the texture of multi views and perform simple fusion to obtain I'_{uv}, which contains information from multi views. Finally, we use facial mask M_{face} to get UV texture map I_{uv}:

$$I_{uv} = M_{face} \odot I'_{uv} \tag{3}$$

Given the geometric parameters $(\vec{\beta},\vec{\theta},\vec{\psi})$, albedo α, lighting condition l, and camera parameter c of the mesh, we can render different two-dimensional face images I_r from various perspectives:

$$I_r = \mathcal{R}(M, B, c, I_{uv}) \tag{4}$$

3.5 Loss function

3.5.1 Multiview optical loss

The multi-perspective projection loss function calculates the photometric consistency loss with the input image after projecting the 3D model from different perspectives to each perspective, obtaining the 2D face under different perspectives. This calculation method is very common in other fields, but in fact it is difficult to play a supervisory role in our model.

We find that the multi-perspective projection loss function has two major drawbacks. On the one hand, because the pixel gray value difference between adjacent regions of the face is very small, even if the model estimation is slightly different, the loss function cannot accurately estimate; on the other hand, suppose that a reconstruction model has a slight deviation in the 3D angle pose, even small, but rendering to 2D image will cause large-scale misalignment, resulting in a very large loss function. This reflects an essential problem: the photometric consistency loss estimates the gray value difference between each pixel in two 2D images, but optimizing in this direction will not make our 3D model reconstruction better, because the model will mistakenly focus on the gray value difference and ignore a more important measure, that is, whether the coordinate prediction of each point of the 3D model model is

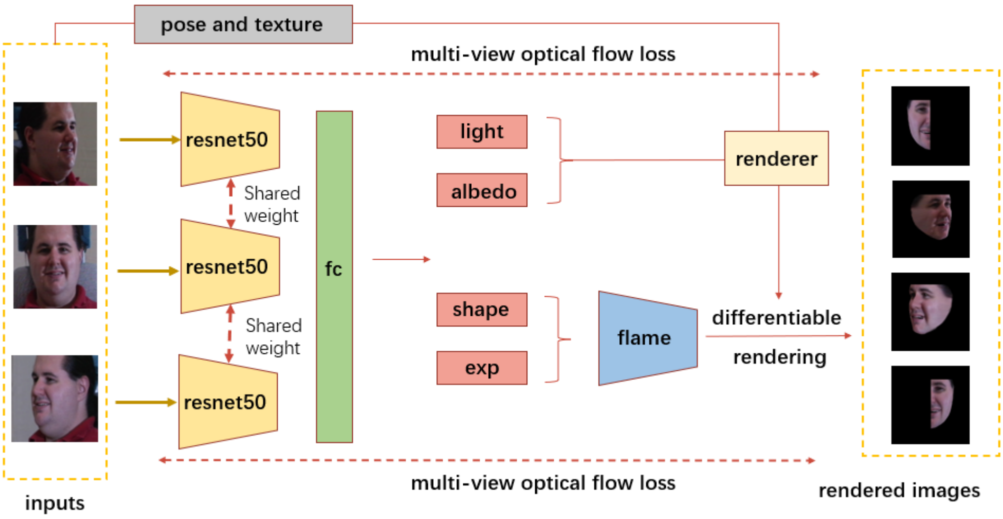

Fig. 1 Architecture of MFNet.

accurate or not. We hope to calculate the reconstruction error of the 3D model, but it is not advisable to calculate it directly because of the lack of true value of the 3D model. Therefore, we need to estimate it through the rendered 2D image, but the photometric consistency loss obviously cannot provide accurate estimation. After analysis, we think that the correct measure should be that for the same vertex in the 3D model, we find its corresponding point position in the rendered image and the corresponding point position in the true value image, and calculate the distance between their coordinates as the reconstruction error. We hope that these two points can coincide, so the distance should be as close to zero as possible. But it is very difficult to calculate this distance directly. The main difficulty lies in that it is very difficult to find the pixel point positions between two 2D images corresponding to the same vertex in the 3D model. MGCNet[16] proposed a method to estimate using camera parameters and depth information, but our model does not have depth information, and this estimation method is also difficult to implement. After research, we choose to use optical flow method to estimate this.

Optical flow[22] in computer vision mainly refers to the movement of objects in the image, specifically the change of pixel points corresponding to the same object in two images, which is represented by a two-dimensional vector. In our model, we use a dense optical flow estimator to estimate the change distance of two faces, and use the sum of squares of all changes as the loss function for supervision. Here we show the extraction effect of optical flow (Figure 2). The first two images are the original and rendered images, and the third image is the estimated optical flow. It can be seen that the optical flow estimator can estimate the difference between face images well, and the effect basically meets the expected goal.

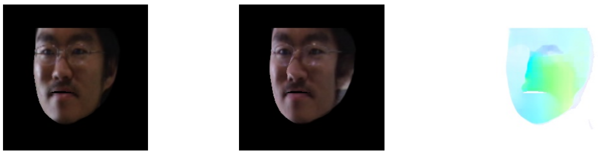

Fig. 2 Optical flow estimation.The first column is the input of the original image, the second column is the rendered image, and the last column is the forward optical flow.

Due to the occlusion of the face, the area of the right face will not appear in the left perspective facial photo. So when we calculate the optical flow loss between the reconstructed left face and the front face of the original image, we should mask the area of the right face and only calculate the area where the two perspectives are common. So here we propose a new implementation of the covisible mask. For the input face image, we first generate a projected two-dimensional face mask MF according to the position of the three-dimensional face model by some more complex mathematical operations during the differentiable rendering. This face mask can better extract the face part in the two-dimensional image, but we still need to extract the covisible mask for the common visible regions of different viewpoints. At this time, we will use the extracted face keypoints. For a three-dimensional face model, we take its eyebrow keypoints, nose keypoints, cheek keypoints and chin keypoints as the boundaries of the common visible region, and obtain the upper, lower, left and right boundaries of the bounding box MB, such as the left eyebrow, left cheek, left chin and left side of the nose. The bounding box MB composed of keypoints and the face mask MF can be combined to obtain a better covisible mask MC. We formulate it as follows: given the common visible region $MB_{a,b}$ between viewpoint a and viewpoint b, and the face mask MF_b of viewpoint b, we can obtain the covisible mask $MC_{a,b}$ between viewpoint b and viewpoint a:

$$MC_{a,b} = MB_{a,b} \odot MF_b \tag{5}$$

Here we show the effect of the covisible mask that we re-implemented (Fig. 3). The first column is the original image, the second column is the face with the face mask MF added, and the third column is the face with the covisible mask MC added. It can be seen that when we calculate the face mask, in order to reduce the estimation error of the optical flow for the uninterested region, we also mask the complex regions such as the mouth, so that the covisible mask basically achieves our expected goal.

Therefore, given the image I_b and the rendered image $I_{a\to b}$, the optical flow estimator $\mathbf{F}$, the covisible mask $MC_{a,b}$, we can calculate the multi-view optical flow loss function $L_{multiop}$:

$$L_{multiop}(I_b, I_{a\to b}) = |\mathbf{F}(MC_{a,b} \odot I_b, MC_{a,b} \odot I_{a\to b})| \tag{6}$$

3.5.2 Single View Keypoint Loss

The original multiview keypoint loss was overly focused on achieving multi-view consistency, but neglected the constraint of the face itself. Therefore, we abandoned excessive multi-view constraints and implemented a single-view keypoint loss function. Similar to the multi-view keypoint loss function, we also projected the 3D face keypoints to the 2D image, but instead of projecting the keypoints from one view to another, we re-projected them back to the original view and compared them with the original image. We hope that this can provide stronger constraints for the model and prevent it from ignoring the constraints of the face itself.

We formalize it as follows: for view a, we compute the error between the ground truth keypoints k_a of view a and the 2D projected keypoints $k_{a\to a}$ obtained by re-projecting the 3D keypoints of the model M_a generated from view a back to view a:

$$L_{singlelmk}(k_a, k_{a\to a}) = \sum_{i \in MF_a} \|k_a(i) - k_{a\to a}(i)\|_1 \tag{7}$$

3.5.3 Eye keypoint loss

Since the eye area of the face is relatively complex, we implemented an eye keypoint loss function to achieve better face reconstruction results. We compute the relative offset between the keypoints $k_a(i)$ and $k_a(j)$ of the upper and lower eyelids on a certain view a, and measure the difference between their offset and the offset between the corresponding re-projected keypoints $k_{a\to a}(i)$ and $k_{a\to a}(j)$ of the 3D model, as the eye keypoint function:

$$L_{eye}(k_a, k_{a\to a}) = \sum_{(i,j)\in E_a} \|k_a(i) - k_a(j) - (k_{a\to a}(i) - k_{a\to a}(j))\|_1 \tag{8}$$

Where E_a denotes the set of upper and lower eyelid keypoints of view a. The resulting L_{eye} will focus more on penalizing the error of the relative offset between the eyelid keypoints, while $L_{singlelmk}$ is the error calculation for the keypoints of the whole face. Compared to equation 7, it is more robust to the misalignment problem between the projected 2D face and the original image, while if only absolute distance is used to measure the loss function, the reconstructed face image will show some abnormal facial shapes, which is reflected in the ablation experiment.

3.5.4 Lip keypoint loss

Since the lip area of the face is also complex, we implemented a similar reconstruction error as the eye keypoint loss function, computing the relative offset between the keypoints $k_a(i)$ and $k_a(j)$ of the upper and lower lips on a certain view a, and measuring the difference between their offset and the offset between the corresponding re-projected keypoints $k_{a\to a}(i)$ and $k_{a\to a}(j)$ of the 3D model, as the lip keypoint function:

$$L_{lip}(k_a, k_{a\to a}) = \sum_{(i,j)\in P_a} \|k_a(i) - k_a(j) - (k_{a\to a}(i) - k_{a\to a}(j))\|_1 \tag{9}$$

Where P_a denotes the set of upper and lower lip keypoints of view a. The resulting L_{lip} will focus more on penalizing the error of the relative offset between the upper and lower lip keypoints.

3.5.5 Regularized loss

We need to regularize some vectors to prevent overfitting, including shape vector $\vec{\beta}$ regularization, expres-

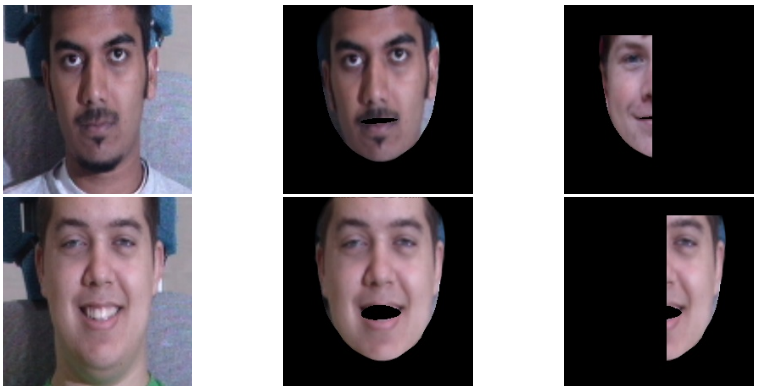

Fig. 3 Covisible mask.The last column is the rendering of our covisible masked images. The dark regions of last column are excluded using 3D landmarks on nose tip and eyebrows

sion vector $\vec{\psi}$ regularization and albedo α regularization:

$$L_{reg} = \|\vec{\beta}\|_2 + \|\vec{\psi}\|_2 + \|\alpha\|_2 \tag{10}$$

3.5.6 Total loss

The total loss function is shown below:

$$\mathrm{L}_{total} = \lambda_1 L_{multiop} + \lambda_2 L_{singlelmk} + \lambda_3 L_{eye} + \lambda_4 L_{lip} + \lambda_5 L_{reg} \tag{11}$$

3.6 Testing

First, we found in the experiment that if we replace the pose parameter $\vec{\theta}$ of the Flame model during reconstruction, we can achieve the effect of projecting to different views more easily, which is equivalent to the pose parameter $\vec{\theta}$ acting as the camera parameter, and we can use the DECA pre-trained model to extract the pose parameter for the input image, thus omitting the process of camera calibration for the input image, which is more suitable for daily scenarios and reduces the reconstruction cost, in line with our research purpose. Second, we added three view information fusion functions in the feature extraction module of MFNet, which can provide more view information for the model.

4 Experiments

In this section, we first introduce our implementation details for conducting the experiments, including the datasets and evaluation metrics(Sec. 4.1). Then we make qualitative and quantitative comparisons to other 3D face reconstruction methods(Sec. 4.2 and Sec. 4.3). Finally, we demonstrate the effectiveness of the proposed method with extensive ablation studies in Sec. 4.4.

4.1 Implementation Details

Training Datasets Our training is performed on Multi-PIE dataset, which contains over 750,000 images recorded from 337 subjects using 15 cameras in different directions 963 under various lighting conditions. We take frontal-view, images as anchors with the number 05 and randomly select side-view images (left and right) to form a three view triplet which is the input of our model and also images for reconstruction. Note that whether an image is in frontal, left, or right view can be determined by the provided camera ID, so we can easily select those images we want. In this way, we take 36k training triplets.

Evaluation Datasets We mainly perform quantitative and qualitative evaluations on the facescape benchmark containing in-the-wild and in-the-lab data. 14 recent methods are evaluated on the dimensions of camera pose and focal length, which provides a comprehensive evaluation.

The benchmark of FS-Wild consists of 400 face images of 400 synthesized subjects. The data are uniformly divided into 4 sets according to the angle between camera orientation and face orien tation ($0° - 5°$,$5° - 30°$,$30° - 60°$,$60° - 90°$),with a reference 3D face model per subject. The images consist of indoor and outdoor images, neutral expression and expressive face images, and varying viewing angles ranging from frontal view to side view

The benchmark of FS-Lab consists of the 20 detailed 3D face models, which are randomly selected from the unpublished testing set of FaceScape. These subjects' age ranges from 17 to 63, with an average age of 38.7. Centering on the head and starting from the front, Everyone's picture has 11 different camera locations. So we can choose three views from FS-Lab as the input for our nulti-view model.

Pretrained Model After conducting a literature review, we chose DECA (Detailed Expression Capture and Animation)[7]. Its E_{coarse} model is used as a pretrained model, and we import part of its parameters into our Resnet50, and fine-tune it in the subsequent training.

Optical flow extractor We use RAFT[17] to extract optical flow. The RAFT model extracts pixel-level features and establishes multi-scale four dimensional correlation information, and iteratively updates the estimated optical flow field through four dimensional information lookup.

Evaluation Metrics In the single-view quantitative evaluations on FS-Wild dataset, we follow the evaluation metrics on FS-Wild dataset, which compute Chamfer Distance(CD), Mean Normal Error(MNE), and Complete Rate(CR). Among them, CD (Chanmfer Distance) refers to the chamfer distance, which represents the difference between the model we reconstructed and the real 3D facial model. MNE (Normal Mean Error) refers to the average error, which is the sum of the distances between the normalized predicted values of facial feature points and the true values. CR (Completeness Rate) refers to the completeness of the predicted model, calculated as the proportion of the number of points in both the real and predicted models to the total number of points in the real model. In the multi-view quantitative evaluations on FS-Lab dataset, we follow the evaluation metrics on FS-Lab dataset, but take three view of FS-Lab dataset as the input of our model.

Hyper-Parameters Setting In actual training, we set the hyperparameters in equation (11) to $\lambda_1 = 1$, $\lambda_2 = 1$, $\lambda_3 = 1$, $\lambda_4 = 0.5$, $\lambda_5 = 1e - 04$. learning rate $= 1e - 3$. Train epochs on multi-PIE are 10.

4.2 Qualitative Results

We first present our reconstruction results, as shown in Figure 4.

We obtained the reconstructed model based on three inputs, and then projected it to a certain perspective. We can see that MFNet's reconstructed facial model performs well in various perspectives, achieving our expected goals. Next, we compared the reconstruction effects of DECA and MFNet. We used DECA and our model to reconstruct 2000 images from AFLW2000-3D respectively. Some of them are shown in Figure 5.

Through observation, it can be found that DECA has problems in predicting facial edges in certain situations, which can cause the predicted 3D model of the face to be narrow, such as in the first row of images; When there is occlusion on the face, such as in the second row of images, the reconstruction results of DECA may have some deviation, while MFNet can model more accurately due to the involvement of multiple perspectives; When the face has a certain angle and there is a visual deviation in the judgment of the face shape, as shown in the third line of the left image, DECA also makes a mistake in the judgment of the character's face shape; In addition, for the reconstruction of some fa-

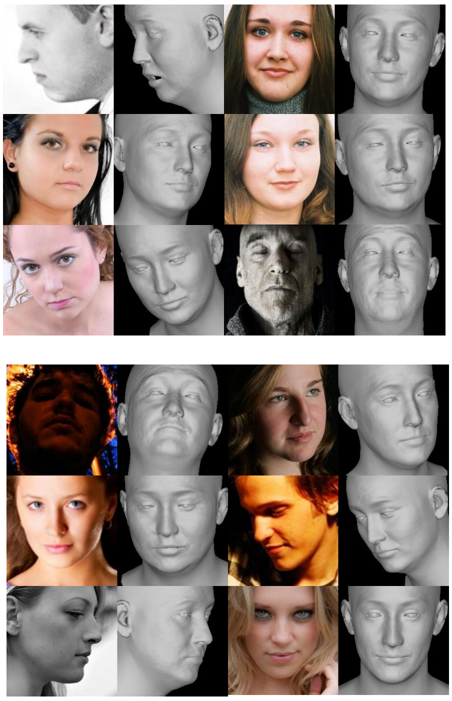

Fig. 4 MFNet reconstruction.The first and third columns represent a certain perspective of the three input images, while the second and fourth columns represent the MFNet reconstruction effect.It can be seen that both the shape and angle of the face have been well reconstructed in MFNet.

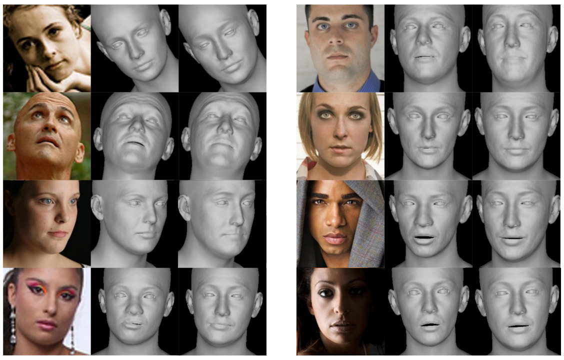

Fig. 5 Qualitative experiment of DECA and MFNet. The first column represents a certain perspective of the input image, the second column represents the DECA reconstruction effect, and the third column represents the MFNet reconstruction image.It can be seen that sometimes the lips reconstructed by DECA cannot be closed, but MFNet can reconstruct the lips very well. And DECA cannot reconstruct a chubby face shape well, but MFNet does it well.

cial details such as eyes, eyebrows, lips, etc., MFNet is more suitable for reconstruction compared to DECA, as shown in the third line of the right image. Therefore, through comparison, it can be found that due to the training from multiple perspectives, MFNet performs better than DECA in most cases.

Finally, we also set up three-viewed cameras on site to take images of the people around us, obtaining multi-view images that are close to the real environment. We tested the reconstruction effect of MFNet on these images and added texture, as shown in the Figure 5.

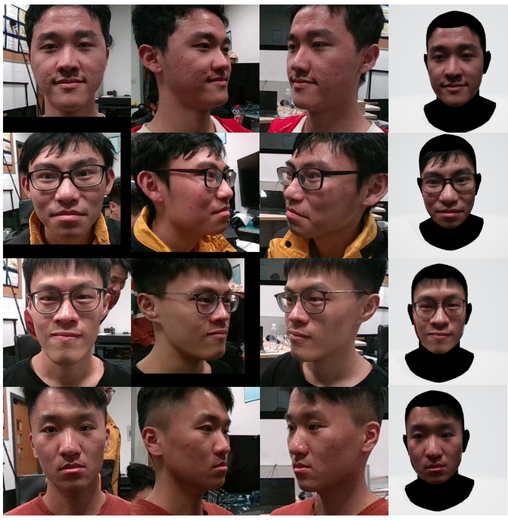

Fig. 6 MFNet reconstruction of shot images.The first three columns are facial photos from the three input perspectives. The last column is the reconstructed face of MFNet after texture rendering.

4.3 Quantitative Results

At present, the number of benchmarks suitable for multi view Iterative reconstruction test of parametric models is small, and basically only 3DMM type face truth values are provided, so it is difficult to find a benchmark suitable for Flame model comparison. Therefore, in order to conduct a broader comparison, we test our model on a single view Iterative reconstruction test set and compare it with other algorithms. Since the benchmark is a single view Iterative reconstruction of face, and MFNet is designed to serve multiple views, we will repeat the test image three times as the input of MFNet's left, middle and right views. Due to the original intention of designing MFNet for multi view input methods, this testing method inevitably reduces the reconstruction effect of MFNet. However, compared with the quantitative indicators of single view models, it can also reflect the progress of MFNet in side face reconstruction to some extent.

Fisrt, we used the FS-Wild test set to test the effectiveness of MFNet and compared it with other single view reconstruction algorithms. Table 1 shows the performance of various single view reconstruction algorithms from small to large poses. The top row represents the head deflection angle, and each column is best represented in bold.

Next, we use images in the FS-Lab dataset of the same person from three different perspectives as the input to MFNet, and randomly selected images from one perspective as the input for other single-view models. After obtaining the 3D model for facial reconstruction, we followed the evaluation metrics on FS-Lab dataset to calculate the three test metrics. The specific results are shown in Table 2.

It can be seen that on the facescape lab dataset, when MFNet was tested with a complete multi-view input, its various indicators showed significant improvement compared to DECA and also other single-view models, indicating that our multi-view training gives MFNet better reconstruction ability and achieves our expected goals.

4.4 Ablation Study

In this section, we conduct an ablation study on the mentioned loss function. In the ablation experiment, we remove one Loss function, keep other Loss function unchanged, and train the same epochs on the same training set. Testing is performed on the fasescape-wild dataset.The results are shown in Table 3. Finally, we reconstructed each ablation model on the alfw dataset. The effectiveness of Loss function can be directly reflected by the reconstruction of ablation model.And the reconstructions are shown in Figure 7 From Table 3, we have some findings in our loss function:

- After removing the multi view optical flow loss, the reconstruction effect of the front face or near the front face decreases, while the reconstruction effect of the side face is similar to MFNet, indicating that multi view optical flow loss can better consider the information of multiple views comprehensively.
- After the regularized Loss function is removed, the model reconstruction effect is also greatly reduced, indicating that the regularized Loss function has achieved good results in preventing over fitting and other functions.

	0-5			5-30			30-60			60-90		
methods	CD	MNE	CR	CD	MNE	CR	CD	MNE	CR	CD	MNE	CR
extreme3dface[20]	5.02	0.16	0.62	5.512	0.18	0.56	7.91	0.20	0.40	25.3	0.26	0.27
PRNet[8]	2.61	0.12	0.83	3.11	0.11	0.83	4.25	0.11	0.78	3.88	0.14	0.75
Deep3DFaceRec[6]	2.30	**0.07**	0.83	**2.50**	**0.07**	0.83	3.56	0.08	0.77	6.81	0.14	0.62
RingNet[15]	2.40	0.08	**0.99**	2.99	0.09	**0.99**	4.78	0.10	0.98	10.7	0.18	0.97
DFDN[24]	3.67	0.09	0.87	3.27	0.09	0.86	7.29	0.12	0.84	27.4	0.30	0.57
DF2Net[24]	2.92	0.12	0.57	4.21	0.13	0.56	6.54	0.15	0.46	19.7	0.30	0.30
UDL[5]	**2.27**	0.09	0.69	2.59	0.09	0.68	3.45	0.10	0.64	6.32	0.17	0.49
facescape_opti[23]	2.81	0.09	0.84	3.17	0.09	0.82	4.08	0.10	0.78	6.57	0.16	0.67
facescape_deep[23]	2.70	0.08	0.87	3.69	0.09	0.86	4.22	0.09	0.85	9.09	0.15	0.70
MGCNet[16]	2.97	**0.07**	0.84	2.94	**0.07**	0.85	**2.78**	**0.07**	0.81	4.20	**0.09**	0.74
3DDFA_V2[10]	2.49	**0.07**	0.86	2.66	**0.07**	0.86	3.17	**0.07**	0.83	**3.67**	**0.09**	0.79
SADRNet[14]	6.60	0.11	0.90	6.87	0.11	0.89	6.39	0.10	0.84	8.62	0.16	0.82
LAP[25]	4.19	0.11	0.94	4.47	0.12	0.93	6.15	0.14	0.87	13.7	0.20	0.68
DECA[7]	2.88	0.08	**0.99**	2.64	**0.07**	**0.99**	2.88	0.08	**0.99**	4.83	0.11	**0.99**
MFNet	3.98	0.11	**0.99**	4.07	0.11	**0.99**	3.60	0.10	**0.99**	5.25	0.12	**0.99**

Table 1 Comparison with other single-view methods.Among them, CD (Chanmfer Distance) refers to the chamfer distance, which represents the difference between the model we reconstructed and the real 3D facial model. MNE (Normal Mean Error) refers to the average error, which is the sum of the distances between the normalized predicted values of facial feature points and the true values. CR (Completeness Rate) refers to the completeness of the predicted model, calculated as the proportion of the number of points in both the real and predicted models to the total number of points in the real model.

Table 2 comparison of MFNet and other single-view models. Through the table, we can find that MFNet outperforms many single-view models in various indicators, indicating that the information provided by multiple views has helped in facial reconstruction

	facescape-lab		
methods	CD	MNE	CR
DECA[7]	5.25	0.16	0.97
LAP[25]	9.76	0.20	0.85
SADRNet[14]	7.21	0.18	0.89
DFDN[24]	14.10	0.32	0.93
Deep3DFaceRec[6]	5.28	0.15	0.80
extreme3dface[20]	15.38	0.26	0.66
PRNet[8]	4.97	0.15	0.85
facescape_opti[23]	5.14	0.16	0.76
DF2Net[24]	7.39	0.17	0.67
MFNet	**4.89**	**0.14**	**0.99**

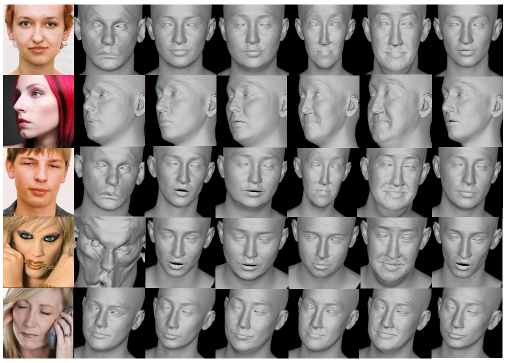

Fig. 7 Ablation study of loss function.From left to right are the images with reg, lip, lmk, eye, multiop removed respectively, and the last column is the reconstruction of MNFet.

- After removing the Loss function of single view face key points, the Loss function of eye key points, and the Loss function of lip key points, the reconstruction performance of the model has declined significantly, especially the Loss function of eye key points, which shows that the eye and lip areas are complex, and it is necessary to provide a good Loss function for supervision, and the Loss function of key points of complete face is also very useful, It can provide the model with correct facial prior knowledge. We can see that after removing constraints such as eyes and lips, the reconstruction effect of the front face decreases the most significantly, which is also in line with our analysis, because the structure of the eyes and nose in the front face is the most complete, accounting for a large proportion, and the performance degradation is most significant after the lack of constraints.

In general, the ablation experiment of Loss function shows that the performance of the model has declined to varying degrees after the removal of some Loss function, which shows that the Loss function we designed and the implementation method are reasonable.

5 Conclusion

Our main contribution is to innovatively combine the face parametric model Flame with a multi-view training and testing framework, and propose a multi-view face 3D reconstruction model MFNet based on Flame. We firstly analyzes the research significance and

Table 3 Ablation study of loss function.

		0-5			5-30			30-60			60-90	
methods	CD	MNE	CR	CD	MNE	CR	CD	MNE	CR	CD	MNE	CR
- multiop	4.29	0.12	0.98	4.43	0.12	**0.99**	3.62	**0.09**	**0.99**	**5.12**	**0.12**	**0.99**
- singlelmk	6.54	0.14	**0.99**	5.85	0.13	**0.99**	12.2	0.18	0.97	38.6	0.25	0.93
- eye	140	0.33	**0.99**	423	0.38	0.98	61.8	0.24	0.96	5.91	0.14	**0.99**
- lip	6.95	0.13	**0.99**	11.2	0.15	0.98	13.7	0.17	0.94	13.6	0.18	0.95
- reg	23.3	0.19	**0.99**	32.3	0.19	**0.99**	7.39	0.12	**0.99**	8.75	0.16	**0.99**
MFNet	**3.98**	**0.11**	0.98	**4.06**	**0.11**	0.98	**3.60**	0.10	**0.99**	5.25	**0.12**	**0.99**

research status of face 3D reconstruction. However, the face 3D reconstruction technology at home and abroad is still in the development stage, and there are problems such as hardware cost and product quality are difficult to be satisfied. We propose our model MFNet, which is implemented in a multi-view training and testing framework and achieves excellent face 3D reconstruction results by using multiple RGB images with simple acquisition difficulty.

Although our model MFNet achieves the expected results on multi-view 3D reconstruction, our work still has many problems, such as our quantitative experiments are not sufficient, our results are not very good when performing quantitative analysis on single-view datasets, and the reconstruction results have many flaws, these shortcomings give us room for improvement. We believe that a broader comparison of MFNet on multi-view test sets is needed to validate our performance, while using larger training datasets, more complex deep neural networks, and more diverse loss functions can bring better results to MFNet.

References

1. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: Proceedings of the 26th annual conference on Computer graphics and interactive techniques, pp. 187–194 (1999)
2. Blanz, V., Vetter, T.: Face recognition based on fitting a 3d morphable model. IEEE Transactions on pattern analysis and machine intelligence **25**(9), 1063–1074 (2003)
3. Booth, J., Antonakos, E., Ploumpis, S., Trigeorgis, G., Panagakis, Y., Zafeiriou, S.: 3d face morphable models" in-the-wild". In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 48–57 (2017)
4. Booth, J., Roussos, A., Zafeiriou, S., Ponniah, A., Dunaway, D.: A 3d morphable model learnt from 10,000 faces. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5543–5552 (2016)
5. Chen, Y., Wu, F., Wang, Z., Song, Y., Ling, Y., Bao, L.: Self-supervised learning of detailed 3d face reconstruction. IEEE Transactions on Image Processing **29**, 8696–8705 (2020)
6. Deng, Y., Yang, J., Xu, S., Chen, D., Jia, Y., Tong, X.: Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pp. 0–0 (2019)
7. Feng, Y., Feng, H., Black, M.J., Bolkart, T.: Learning an animatable detailed 3d face model from in-the-wild images. ACM Transactions on Graphics (ToG) **40**(4), 1–13 (2021)
8. Feng, Y., Wu, F., Shao, X., Wang, Y., Zhou, X.: Joint 3d face reconstruction and dense alignment with position map regression network. In: Proceedings of the European conference on computer vision (ECCV), pp. 534–551 (2018)
9. Gerig, T., Morel-Forster, A., Blumer, C., Egger, B., Luthi, M., Schönborn, S., Vetter, T.: Morphable face models-an open framework. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pp. 75–82. IEEE (2018)
10. Guo, J., Zhu, X., Yang, Y., Yang, F., Lei, Z., Li, S.Z.: Towards fast, accurate and stable 3d dense face alignment. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX, pp. 152–168. Springer (2020)
11. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4d scans. ACM Trans. Graph. **36**(6), 194–1 (2017)
12. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. ACM transactions on graphics (TOG) **34**(6), 1–16 (2015)
13. Roberts, L.G.: Machine perception of three-dimensional solids. Ph.D. thesis, Massachusetts Institute of Technology (1963)
14. Ruan, Z., Zou, C., Wu, L., Wu, G., Wang, L.: Sadrnet: Self-aligned dual face regression networks for robust 3d dense face alignment and reconstruction. IEEE Transactions on Image Processing **30**, 5793–5806 (2021)
15. Sanyal, S., Bolkart, T., Feng, H., Black, M.J.: Learning to regress 3d face shape and expression from an image without 3d supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7763–7772 (2019)
16. Shang, J., Shen, T., Li, S., Zhou, L., Zhen, M., Fang, T., Quan, L.: Self-supervised monocular 3d face reconstruction by occlusion-aware multi-view geometry consistency. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV, pp. 53–70. Springer (2020)
17. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16, pp. 402–419. Springer (2020)
18. Tran, L., Liu, X.: Nonlinear 3d face morphable model. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7346–7355 (2018)

19. Tran, L., Liu, X.: On learning 3d face morphable model from in-the-wild images. IEEE transactions on pattern analysis and machine intelligence **43**(1), 157–171 (2019)
20. Trn, A.T., Hassner, T., Masi, I., Paz, E., Nirkin, Y., Medioni, G.: Extreme 3d face reconstruction: Seeing through occlusions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3935–3944 (2018)
21. Wu, F., Bao, L., Chen, Y., Ling, Y., Song, Y., Li, S., Ngan, K.N., Liu, W.: Mvf-net: Multi-view 3d face morphable model regression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 959–968 (2019)
22. Wu, G., Liu, X., Luo, K., Liu, X., Zheng, Q., Liu, S., Jiang, X., Zhai, G., Wang, W.: Accflow: Backward accumulation for long-range optical flow. International Conference on Computer Vision (2023)
23. Yang, H., Zhu, H., Wang, Y., Huang, M., Shen, Q., Yang, R., Cao, X.: Facescape: A large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
24. Zeng, X., Peng, X., Qiao, Y.: Df2net: A dense-fine-finer network for detailed 3d face reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2315–2324 (2019)
25. Zhang, Z., Ge, Y., Chen, R., Tai, Y., Yan, Y., Yang, J., Wang, C., Li, J., Huang, F.: Learning to aggregate and personalize 3d face from in-the-wild photo collection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14,214–14,224 (2021)