From 2D Images to 3D Model: Weakly Supervised Multi-View Face Reconstruction with Deep Fusion

Weiguang Zhao*[1], Chaolong Yang*[1], Jianan Ye*[1], Yuyao Yan[1], Xi Yang[1], Kaizhu Huang†[2]

[1] Department of Intelligent Science, Xi'an Jiaotong-Liverpool University
[2] Department of Electrical and Computer Engineering, Duke Kunshan University
{weiguang.zhao, chaolong.yang, jianan.ye}20@student.xjtlu.edu.cn, joshuayyy@gmail.com, xi.yang01@xjtlu.edu.cn, kaizhu.huang@dukekunshan.edu.cn

Abstract

We consider the problem of Multi-view 3D Face Reconstruction (MVR) with weakly supervised learning that leverages a limited number of 2D face images (e.g. 3) to generate a high-quality 3D face model with very light annotation. Despite their encouraging performance, present MVR methods simply concatenate multi-view image features and pay less attention to critical areas (e.g. eye, brow, nose and mouth). To this end, we propose a novel model called Deep Fusion MVR (DF-MVR) and design a multi-view encoding to single decoding framework with skip connections, able to extract, integrate, and compensate deep features with attention from multi-view images. In addition, we develop a multi-view face parse network to learn, identify, and emphasize the critical common face area. Finally, though our model is trained with a few 2D images, it can reconstruct an accurate 3D model even if one single 2D image is input. We conduct extensive experiments to evaluate various multi-view 3D face reconstruction methods. Our proposed model attains superior performance, leading to 11.4% RMSE improvement over the existing best weakly supervised MVRs. Source codes are available in the supplementary materials.

Introduction

Reconstructing 3D shapes of human faces from 2D images is a challenging yet essential task for numerous applications such as virtual reality, augmented reality, and facial animations. 3D Morphable Model (3DMM) (Blanz and Vetter 1999) is the pioneer in converting the 3D face model to a parameter representation. Recently, adopting convolutional neural networks (CNN) to extract 2D image information to predict 3DMM coefficients has become the mainstream method of face reconstruction. The supervised CNN-based methods (Dou and Kakadiaris 2018; Feng et al. 2018; Guo et al. 2018) need a large number of 3D face meshes or point clouds corresponding to 2D pictures as groundtruth, which is time and/or manpower consuming.

To alleviate the need for 3D face meshes or point clouds data, recent efforts have shifted to weakly supervised and self-supervised methods (Tewari et al. 2017; Tran et al.2018; Deng et al. 2019; Shang et al. 2020). Most of these methods used landmarks and differentiable rendering for training. (Tewari et al. 2017) exploited the difference between each pixel of the original image and the rendered image as training loss. (Deng et al. 2019) attempted to combine pixelwise photometric difference and the skin probability mask to calculate training loss.

All the above weakly supervised methods only exploit one single image for construction, which usually fails to estimate facial depth appropriately. For instance, the single-view reconstruction method (Richardson et al. 2017; Tewari et al. 2018; Tran et al. 2018) cannot fully explain the geometric difference of facial features, such as the height of the mouth and eye sockets. Such limitation can however be resolved by the geometric constraints contained in a few face images of different views, or multi-view images. Surprisingly, rare studies have been made on weakly supervised multi-view 3D face reconstruction tasks. To our best knowledge, Deep3DFace (Deng et al. 2019) and MGCNet (Shang et al. 2020) are the only methods currently available that utilize multi-view information from a single subject for weakly supervised reconstruction. Specifically, (Deng et al. 2019) scored each multi-view image using CNN and then selected the highest scoring image to regress shape parameters; (Shang et al. 2020) designed the consistency map based on multi-view consistency and calculated pixelwise photometric difference for the consistency map. Unfortunately, these two methods are limited because they simply concatenate multi-view image features and do not consider deep fusion of multi-view images features, nor do they pay attention to critical areas (e.g. eye, brow, nose and mouth) which may impact the reconstruction quality the most.

To cope with these drawbacks, we propose a novel end-to-end weakly supervised multi-view 3D face reconstruction network which learns to fuse deep representations and identify critical areas. First, as multi-view images all represent the same face, we develop an encoding-decoding network (Tri-Unet) with attention to extract features and deeply fuse them into one feature map. As shown in Fig. 3, multiple encoders are used to extract features from multi-view images, and one single decoder is engaged to fuse these features in deep. In order to compensate for the possible loss caused by sampling, skip connections with attention are introduced.

*These authors contributed equally.
†Corresponding author.

Second, we develop a multi-view face parse network to learn, identify, and emphasize the critical common face area. The novel face parse network is able to learn the face mask which not only acts as input features to help Tri-Unet encode/decode common area of multi-view images for better deep fusion, but also plays the role of a weight map to calculate the pixelwise photometric loss between rendered images and original images. Since pixelwise photometric loss pays more attention to the difference of RGB, we also add the mask loss to narrow the size of facial features (e.g. eye, brow, nose and mouth) between 3D and 2D faces. Finally, we import RedNet (Li et al. 2021) instead of ResNet (He et al. 2016), which is typically utilized in face reconstruction networks. RedNet is a residual network based on involution (Li et al. 2021), which more flexibly extracts channel features than traditional convolution. Combining pixelwise photometric loss, mask loss, and landmark loss, we design a novel weakly supervised training framework that is able to fuse deep features comprehensively and pay attention to critical face features specially.

The contributions of our work are as follows:

- We design a novel weakly supervised encoding-decoding framework (Tri-Unet) for deep fusion of multi-view features, which has rarely been studied in the literature.
- We develop a face mask mechanism to identify common areas in multi-view images and encourage the 3D face reconstruction network to pay more attention to critical areas (e.g. eye, brow, nose and mouth).
- Compared with traditional convolution, involution (Li et al. 2021) is spatial-specific and able to obtain features on the channel, which means it can better process deep fusion features. We are the first to apply it to face reconstruction tasks.
- On the empirical side, our novel framework attains the superior performance, leading to 11.4% RMSE improvement over the existing best weakly supervised MVRs.

Related Work

3D Morphable Model

3D Morphable Model (3DMM) is a statistical model of 3D facial shape and texture which performed principal component analysis (PCA) on the face mesh training set (Blanz and Vetter 1999). Subsequently, (Paysan et al. 2009) released a generative 3D shape and texture model, the Basel face model (BFM), and demonstrated its application to several face recognition tasks. (Booth et al. 2018) has further expanded 3DMM to build models for specific ages, genders or ethnic groups. The current multi-view reconstruction methods mostly use BFM. For a fair comparison, we also exploit BFM to represent 3D faces in our model.

Single-view Methods

Most single-view face reconstruction methods take CNN as the deep learning network to predict 3DMM coefficients. For example, (Zhu et al. 2016) deployed CNN to predict 3DMM coefficients and achieved encouraging results. (Tuan Tran et al. 2017) designed a robust method, based on CNN and ResNet101 (He et al. 2016), to regress 3DMM shape and texture coefficients directly from an input photo without annotation of landmarks. (Dou, Shah, and Kakadiaris 2017) concatenated the last two pooling layers of CNN to create a Fusion CNN branch for predicting the expression base individually. It also generated synthetic rendered face images with predicted 3D scans. However, these methods all require 3D mesh files as ground-truth, which greatly hinders their practical applications due to the shortfall of available annotated training data containing 3D shapes.

To cope with this issue, recent research focus has been put on weakly supervised and self-supervised methods. (Tewari et al. 2017; Genova et al. 2018) proposed model can be trained without 3D labels by adopting differentiable rendering for calculating the pixel difference between the rendered image and the original image. (Sengupta et al. 2018) designed an end-to-end learning framework for accurately decomposing an unconstrained human face image into shape, reflectance and illuminance. (Lin et al. 2020) used a similar method to predict 3D shapes while further added GAN to generate more detailed texture information.

Multi-view Methods

Surprisingly, there are few multi-view 3D face reconstruction methods based on machine learning in the literature. (Dou and Kakadiaris 2018) proposed to use Recurrent Neural Network (RNN) to fuse identity-related features extracted from deep convolutional neural network (DCNN) to produce more discriminative reconstructions, but their approach does not exploit multi-view geometric constraints. (Wu et al. 2019) added multi-view geometric constraints and introduced the optical flow loss to improve the reconstruction accuracy. In the feature extraction of multiple images, they only concatenated the deep features. Both methods require ground-truth of 3DMM (Dou and Kakadiaris 2018; Wu et al. 2019), which is hardly available practically.

(Deng et al. 2019) applied weakly supervised learning to multi-image training. They designed two CNN models for predicting 3DMM coefficients and scoring each image. The image with high confidence was used to regress shape coefficients, and the rest images will be used to regress coefficients such as expression and texture. (Shang et al. 2020) adopted the concept of geometry consistency to design pixel and depth consistency loss. They established dense pixel correspondences across multi-view input images and introduced the covisible maps to account for the self-occlusion. This method strengthened the attention to the common area of multiple images, but pays less attention to the local features of the face and the global features of multiple images. Our method employs the face parsing network to label the facial features of the face from multiple perspectives, which can not only focus on the common area of multiple perspectives, but also divide the common area in more detail.

Main Methodology

Overview

We first provide an overview of our proposed framework, which is shown in Fig. 1. We decide to exploit three multi-

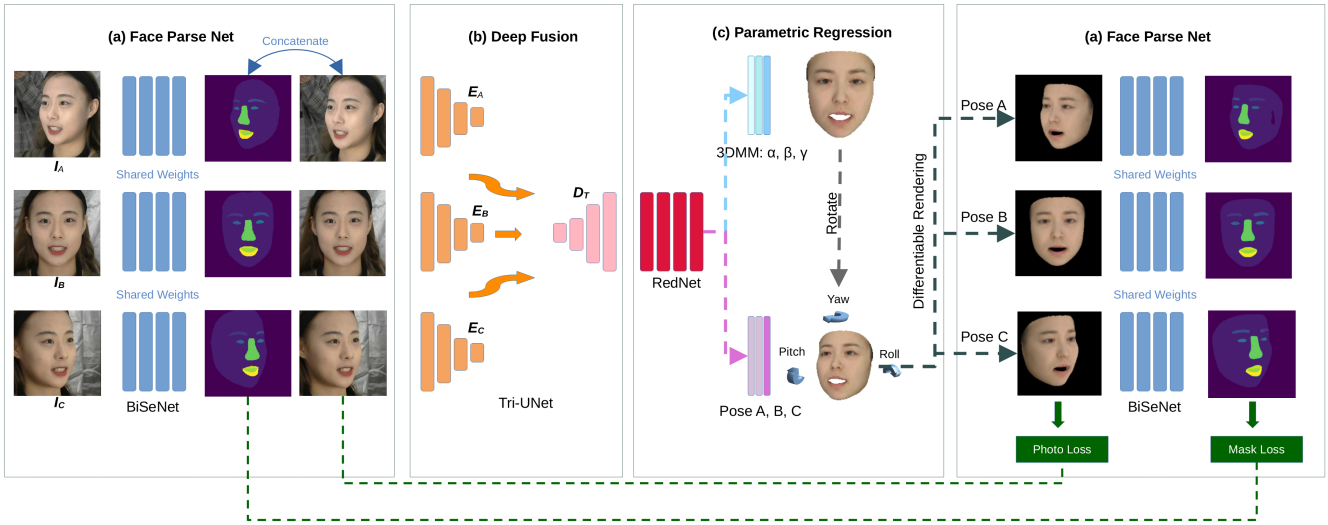

Figure 1: Overview of DF-MVR

view images of a subject for generating a corresponding 3D face and introduce the *face parse network (a)* to process these three images separately to generate unified standard face masks. An *encoding-decoding network (b)* is designed to fuse the features of multi-view images in deep by sharing a decoder with an attention mechanism to obtain information from the encoder. Moreover, RedNet (Li et al. 2021) is used as *parametric regression (c)* to regress 3DMM and pose coefficients. The reconstructed 3D face is reoriented utilising the pose coefficients and then rendered back to 2D. The photo loss between the re-rendered 2D image and the input image at the target view is calculated while the masks are exploited as the weight map to enhance the back propagation of the facial features. In this section, we will provide details on each components as below.

Face Parse Net

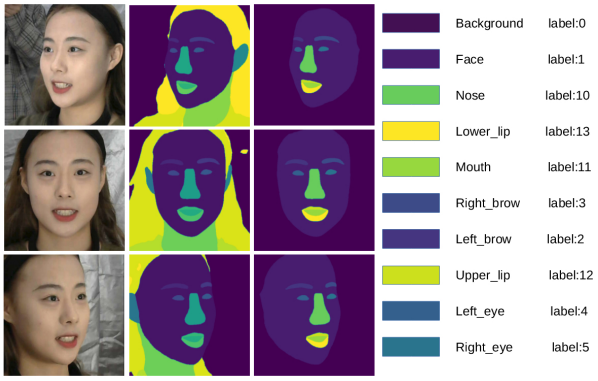

Figure 2: Face Mask Annotation. In column order: original images, preliminary segmentation, face mask, legend.

We introduce the face parse network based on BiSeNet (Yu et al. 2018) to perform preliminary analysis of the input image and identify the elements of the image. The generated face mask has only one layer of channel. For example, if the size of the input image is $224 \times 224 \times 3$, the size of the face mask will be $224 \times 224 \times 1$. In order to better highlight the face, excessive elements of face masks such as hair and neck, will be removed, and the following parts will be kept: face, nose, lower lip, upper lip, left brow, right brow, left eye, right eye and mouth. The reserved parts are marked with different numbers in order to distinguish facial features. On one hand, the face masks are concatenated with the original images to help the network understand the common area of the multi-view image. On the other hand, the face masks serve as weight map to calculate the photo loss and mask loss for training.

Deep Fusion

The existing multi-view face reconstruction networks all deployed CNN or VGG (Simonyan and Zisserman 2014) as the feature extractor. These networks concatenated the multi-graph features in the fully connected layer, which cannot perform feature interaction well. In addition, the previous networks mostly adopted shared weights or one backbone to process multi-view images, making it difficult for the network to pay attention to the unique information of each view. Differently, we design a novel feature fusion network, Tri-UNet, to extract features of multi-view images inspired by attention Unet (Oktay et al. 2018).

We denote the three-view input images as $\mathbf{I}_A$, $\mathbf{I}_B$, and $\mathbf{I}_C$, representing the three perspectives of left, front and right. Since the information and focus of each view are different, we set up three encoders to extract the features from three views respectively. Corresponding to the input images, these three encoders are represented by $\mathbf{E}_A$, $\mathbf{E}_B$, and $\mathbf{E}_C$. The weights of the three encoders are not shared. Encoders are mainly composed of double convolution and maximum pooling. At the end of encoders, the deep features of $\mathbf{I}_A$, $\mathbf{I}_B$, $\mathbf{I}_C$ will be concatenated as $\mathbf{F}_D$. Considering that $\mathbf{I}_A$, $\mathbf{I}_B$, and

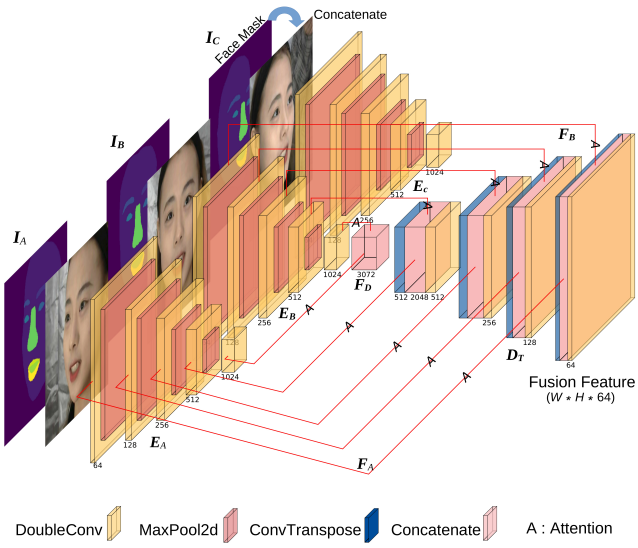

Figure 3: Tri-UNet. For conciseness, we do not draw the skip connection of the $\mathbf{E}_C$, which is similar to the $\mathbf{E}_A$.

$\mathbf{I}_C$ actually describe the same object, we only set up a shared decoder for better fusing features as well as emphasizing the common features. The decoder is mainly composed of ConvTranspose, convolution, concatenate and skip connection. We adopt the attention mechanism to extract the feature $\mathbf{F}_A$, $\mathbf{F}_B$, and $\mathbf{F}_C$ from $\mathbf{E}_A$, $\mathbf{E}_B$, and $\mathbf{E}_C$ to enrich the information in the $\mathbf{F}_D$ decoding process. Finally, the fusion feature size we retain is $224 \times 224 \times 64$, in the case where the image size is $224 \times 224 \times 3$.

Parametric Regression

We adopt RedNet50 to process the fusion features and regress parameters. RedNet replaces traditional convolution with involution on the ResNet architecture. The inter-channel redundancy within the convolution filter stands out in many deep neural networks, casting the large flexibility of convolution kernels w.r.t different channels into doubt. Compared with traditional convolution, involution is spatial-specific and able to obtain features on the channel. Therefore, we choose RedNet to perform parameter regression, and ablation experiments also verify its effectiveness.

3DMM Parameter regressed in this work include identification, expression, and texture parameters. The 3D face shape $\mathbf{S}$ and the texture $\mathbf{T}$ can be represented as:

$$\begin{aligned} \mathbf{S} &= \mathbf{S}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \overline{\mathbf{S}} + \mathbf{B}_{id}\boldsymbol{\alpha} + \mathbf{B}_{exp}\boldsymbol{\beta}, \\ \mathbf{T} &= \mathbf{T}(\boldsymbol{\gamma}) = \overline{\mathbf{T}} + \mathbf{B}_{t}\boldsymbol{\gamma}, \end{aligned} \tag{1}$$

where $\overline{\mathbf{S}}$ and $\overline{\mathbf{T}}$ are the average face shape and texture. $\mathbf{B}_{id}$, $\mathbf{B}_{exp}$, $\mathbf{B}_t$ are the PCA bases of identity, expression, and texture respectively. $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$ are the parameter vectors that the network needs to regress ($\alpha, \beta \in R^{80}$ and $\gamma \in R^{64}$). By adjusting these three vectors, the shape, expression and texture of the 3D face can be changed. In order to compare with MGCNet (Shang et al. 2020) and Deep3DFac (Deng et al. 2019), we use the same face model. BFM (Paysan et al. 2009) was adopted for $\overline{\mathbf{S}}, \mathbf{B}_{id}, \overline{\mathbf{T}}$, and $\mathbf{B}_t$. $\mathbf{B}_{exp}$ is built by (Guo et al. 2018) based on Facewarehouse (Cao et al. 2013).

Pose Parameters are used to adjust the angle and position of the 3D face in the camera coordinate system. We exploit the differentiable perspective rendering (Ravi et al. 2020) to render the 3D face back to 2D. When the camera coordinates are fixed, we could change the size and angle of the rendered 2D face by adjusting the position of the 3D face in the camera coordinate system. And the position of the 3D face in the camera coordinate system can be determined by predicting the rotation angle and translation in each axis. In order to enhance the geometric constraints of the multi-view reconstruction, we respectively predict the pose of the 3D faces in the multi-view, instead of only predicting the pose of one perspective to render 2D images.

Texture Sampling

The texture of 3D face is also an important part of 3D face reconstruction. However, the texture base contained in the 3DMM model is limited. As shown in Fig. 4, 3DMM fails to represent the colors of lipstick, beard, etc. Therefore, we develop the method of sampling from the original image to restore the texture information of the 3D face. The 3D face generated by the prediction is projected to the 2D image through the camera coordinates. Since the 3D face is composed of point clouds, every point can be projected into a 2D image. The point projected to 2D takes the average of the four neighborhood pixel values as its own texture information. In this way, the complete 3D face texture information can be obtained.

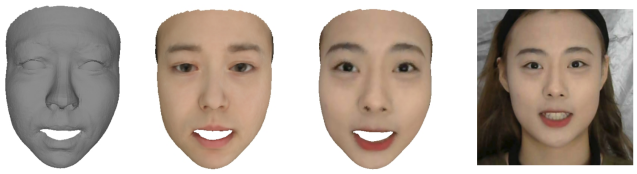

Figure 4: Texture Comparison

In this framework, two training schemes are proposed: weakly supervised and self-monitored training. Whether or not 3D landmarks are utilized is the criterion for distinguishing between the two schemes. As a weakly supervised training method, our model needs to work with slight 3D annotations as labels. On the other hand, if 3D landmarks are not introduced to calculate the loss, our model will not require any 3D labels and only require multi-view images for training. Both the schemes have been verified and compared in the following sections.

Weakly Supervised Training

In order to alleviate the strong need for the labeled data, we design a weakly supervised method for training. First, we render the predicted 3D face model back to 2D and compare the rendered image with the original image pixel by pixel. Then, the rendered 2D images are fed into the face

parse network to generate rendered face masks. According to the consistency principle, the rendered face masks should be consistent with the original face masks. Therefore, the L2 distance is treated as a mask loss. Finally, the landmark loss and regularization loss are introduced to shape 3D face and suppress the generation of distorted faces.

Photo Loss

Photo loss is often used in weakly supervised face reconstruction tasks (Thies et al. 2016; Tewari et al. 2018; Deng et al. 2019; Shang et al. 2020). Distinct with the traditional methods, we impose a weight for each pixel according to the facial features. The weight map is learned by the face mask $\mathcal{M}$ of the original image I. In order to enhance the robustness of the weight map, we dilate $\mathcal{M}$ with 20 pixel as $\mathcal{M}_d$, shown in Fig. 5. The multi-view photo loss can be expressed as:

$$L_p = \frac{1}{\mathcal{V}} \sum_{v=1}^{V} \frac{\sum_{i \in \mathcal{P}^v} \mathcal{M}_{di}^v \cdot \left\| I_i^v - I_i^{v'} \right\|_2}{\sum_{i \in \mathcal{P}^v} \mathcal{M}_{di}^v}, \tag{2}$$

where $\mathcal{V}$ is the number of the reconstructed views. $\mathcal{V}$ is 3 in the proposed model. $\mathcal{P}^v$ is the area where the rendered image $I^{v\prime}$ and the original image I^v intersect in the current view. i denotes pixel index, and $\| \cdot \|_2$ denotes the L2 norm.

Figure 5: Weight Map

Mask Loss

Photo loss focuses on the pixel difference between two pictures. It is difficult to constrain the size of the facial feature area in the two pictures. For example, the nose color is very similar to that of the cheeks, thereby leading to difficulties for the photo loss to notice the boundary line between them. For this reason, we introduce mask loss to narrow the facial features of the input image and the rendered image. The division and labeling of the facial features are shown in Fig. 2. We dialate the facemask with 20 pixels to enhance the robustness of the weight map. Then the dilated image is divided into three levels to be the weight map. In weight map, facial features are marked as 254, the rest of the facial area is marked as 128, and the background is marked as 32, as shown in Fig. 5. Similar to photo loss, we can calculate multi-view the mask loss:

$$L_m = \frac{1}{\mathcal{V}} \sum_{v=1}^{V} \frac{\sum_{i \in \mathcal{P}^v} \left\| \mathcal{M}_i^v - \mathcal{M}_i^{v'} \right\|_2}{\sum_{i \in \mathcal{P}^v} \mathcal{M}_i^v}. \tag{3}$$

Landmark Loss

We also adopt 2D landmarks and 3D landmarks for weakly supervised training. We use 3D face alignment method (Bulat and Tzimiropoulos 2017) to generate 68 landmarks $\{l_n\}$ as the groundtruth. Then the corresponding points in the predicted 3D face point cloud are projected to 2D as predicted 2D landmarks $\{l'_n\}$. Then the multi-view 2D landmark loss can be calculated:

$$L_{l\_2d} = \frac{1}{NV} \sum_{v=1}^{V} \sum_{n=1}^{N} \omega_n \left\| l_n^v - l_n^{v'} \right\|_2, \tag{4}$$

where ω_n is the weight for each landmark. We set the weight to 20 only for the nose and inner mouth landmarks, and to 1 else.

2D landmarks are still insufficient for the reconstruction of 3D face shapes. In order to obtain better reconstruction effect, we select 101 3D landmarks $\{q'_n\}$ to impose a weak constraint on the shape of the 3D face. According to the 3DMM index, 101 predicted landmarks $\{q_n\}$ can be found. Then, we select 7 points $\{a'_n\}$ and $\{a_n\}$ in $\{q'_n\}$ and $\{q_n\}$ respectively as alignment points to calculate the alignment parameters of $\{q'_n\}$ and $\{q_n\}$. The alignment parameters include: scale $\mathbf{s}$, rotation $\mathbf{R}$ and translation $\mathbf{t}$. These parameters can be obtained by the following optimization equation (Tam et al. 2012; Sanyal et al. 2019):

$$Optim(\mathbf{s}, \mathbf{R}, \mathbf{t}) = \min_{\mathbf{s},\mathbf{R},\mathbf{t}} \sum_i \left\| \mathbf{a}'_i - \mathbf{s}\left(\mathbf{R} \cdot \mathbf{a}_i + \mathbf{t}\right) \right\|_2. \tag{5}$$

After the optimal $\mathbf{s}$, $\mathbf{R}$ and $\mathbf{t}$ are obtained, the predicted 101 landmarks $\{q_n\}$ can be converted to the space of $\{q'_n\}$ as $\{q_{nt}\} = s\left(\mathbf{R} \cdot \mathbf{q}_n + \mathbf{t}\right)$.

Then the multi-view 3D landmark loss can be calculated:

$$L_{l\_3d} = \frac{1}{N} \sum_{n=1}^{N} \left\| q_{nt} - q'_n \right\|_2. \tag{6}$$

In summary, the landmark loss can be expressed as:

$$L_l = \omega_{2d} L_{l\_2d} + \omega_{3d} L_{l\_3d}, \tag{7}$$

where ω_{2d} and ω_{3d} represent respectively the weight of 2D landmark loss and 3D landmark loss. In this work, we set them to 0.02 and 1 as tuned empirically.

Regularization Loss

To suppress the generation of distorted faces, we add the regularization loss which is commonly-used in face reconstruction task (Thies et al. 2016; Tewari et al. 2018; Deng et al. 2019; Shang et al. 2020):

$$L_{\text{reg}} = \omega_\alpha \|\boldsymbol{\alpha}\|^2 + \omega_\beta \|\boldsymbol{\beta}\|^2 + \omega_\gamma \|\boldsymbol{\gamma}\|^2, \tag{8}$$

where $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$ are 3DMM parameter vectors that the network predicted. ω_α, ω_β and ω_γ are the weights for 3DMM parameter vectors. Following Deep3DFace (Deng et al. 2019), we set them to 1, 0.8 and 0.017 with fine tuning.

Overall Loss

The overall loss required by our end-to-end net for weakly supervised training can be represented as:

$$L_{all} = \omega_p L_p + \omega_m L_m + \omega_l L_l + \omega_{reg} L_{reg},$$

where ω_p, ω_m, ω_l, ω_{reg} are the weights for photo loss, mask loss, landmark loss and regularization loss. Following Deep3DFace, we set $\omega_{reg} = 3.0 \times 10^{-4}$. Since ω_{2d} and ω_{3d} has been determined, we just fix $\omega_l = 1$ to adjust ω_p and ω_m by sensitivity analysis. Then, we set $\omega_p = 4$ and $\omega_m = 3$ as empirically obtained in sensitivity analysis.

Experiment

Setup

Dataset. Pixel-Face (Lyu et al. 2020) is a large-scale and high-resolution MVR dataset, which contains 855 subjects ranging in age from 18 to 80 years old. Each subject has 7 or 23 samples of different expressions. Pixel-Face has 3D mesh file of each sample as groundtruth but not 3DMM parameters or angle of multi-view images. Hence, it is suitable for weakly supervised or unsupervised training for MVR. In the experiment, the train test split was set to 0.8.

Unfortunately, there are rare other datasets available for comparisons in this paper. E.g., though MICC and AFLW2000-3D are more commonly used in 3D face reconstruction, neither can meet our multi-view setting: AFLW2000-3D is mostly adopted for single image testing, and the MICC dataset provides data in the form of videos, which means that its expression in each view may change. To this end, we test only on the Pixed-Face dataset.

Network. Our network is shown in Fig. 1 and described in the methodology section. Based on the pre-trained BiseNet (Yu et al. 2018) with frozen weights, the face parse network is located in the beginning and end of the network. In the scenario of MVR, we design a fusion network consisting of three different encoders to emphasize more diverse features. A lightweight RedNet50 (Li et al. 2021) is designed as the parameter regression network, since the fusion network has already extracted sufficient information.

Evaluation Metric. Following the previous works, RMSE (mm) (Wu et al. 2019; Deng et al. 2019; Shang et al. 2020) is used to compute point-to-plane L2 distance between predict 3D scans and groundtruth 3D scans. Concretely, the front face area is cropped for metrics calculation instead of using a complete BFM model (Sanyal et al. 2019; Deng et al. 2019; Shang et al. 2020). Before calculating point-to-plane L2 distance, the predicted 3D scans need to be registered with ground-truth 3D scans. Also, we used the ICP registration method (Li et al. 2017) the same as (Deng et al. 2019).

Comparison to SOTAs

We compare our method with the existing weakly supervised MVRs. The parameterized results of the comparison are shown in Table 1. As observed, our proposed model attains the superior performance, leading to 11.4% RMSE improvement over the existing best weakly supervised MVRs. Since (Shang et al. 2020) and (Deng et al. 2019) did not use 3D landmarks, to be fair, we also provide the results of our model without using 3D landmarks for comparison. Our model (without 3D landmarks) shows a 7.2% improvement compared to the existing methods with even highest stability according to the standard deviation.

Method	Dataset	Mean	Std
(Shang et al. 2020)	Pixel-Face	1.8877	0.4378
(Deng et al. 2019)	Pixel-Face	1.6641	0.3690
Ours (w/o L_{l_3d})	Pixel-Face	1.5436	**0.2128**
Ours	Pixel-Face	**1.4738**	0.3059

Table 1: Comparison of RMSE (mm)

More specifically, only the two methods can be found in the literature related to multi-view weakly supervised 3D face reconstruction, both of which are used as the comparison methods in this paper. (Shang et al. 2020) uses multiple images for training, and then a single image for testing. We select the best results among the three images for display. (Deng et al. 2019) does not release their source codes of its scoring network. We use their codes to train/test on Pixel-Face. The visual comparison is shown in the first 3 rows of Fig. 6 given 3-view faces. It is evident that our predicted model is more accurate, especially in terms of facial depth estimation in the facial features. In addition, our model can better learn human facial expressions, such as closing eyes and pursing lips. Finally, the last three rows in Fig. 6 indicate that our model can still outperform the other SOTAs even if one single face is input. More analysis can be seen in the supplemental.

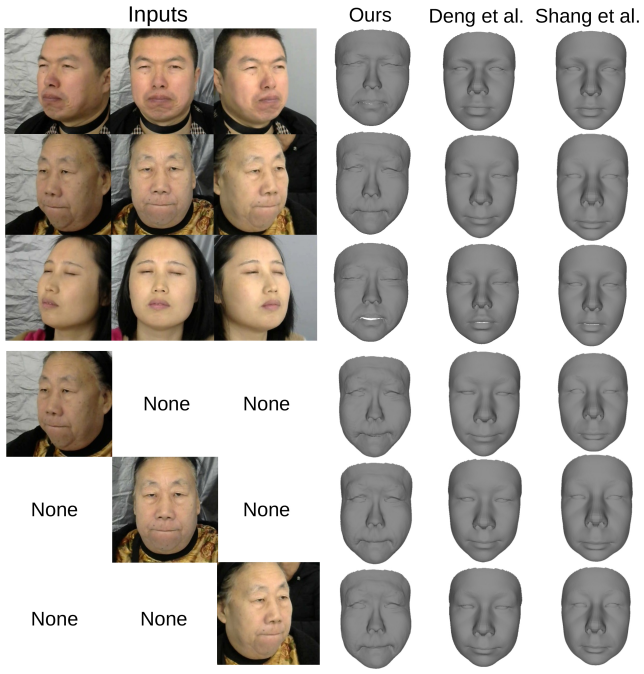

Figure 6: Comparisons to SOTA Methods

Detailed Comparison. A closer comparison is shown in

Fig. 7. Because 3D landmarks will improve the reconstruction of facial features, for fairness, we also report the results (without 3D landmarks) for comparison, which can better reflect the effect of face mask mechanism on facial feature adjustment. In the first sample, our model can predict the expression of pursed lips. The upper lip of our model is almost invisible, compared to the other models. In the second sample, the eyebrows and eyes of our model appear more similar to those of the original image.

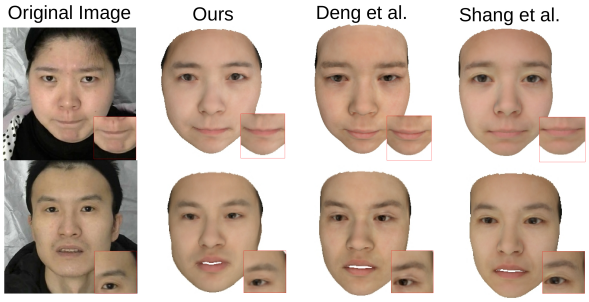

Figure 7: Detailed Comparison

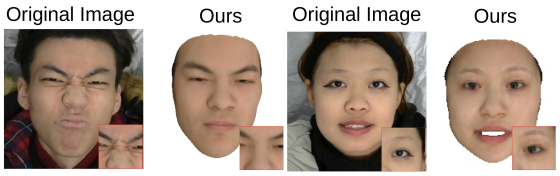

Figure 8: Some Challenging Cases

Limitations. While our model achieves high accuracy, it also has some limitations. Three multi-view images as input make the model less flexible for some fixed scenes. Second, our model is based on 3DMM which has finite vector base ($\mathbf{B}_{id}$, $\mathbf{B}_{exp}$ and $\mathbf{B}_t$). To this end, our model cannot reconstruct wrinkles, beards, eye shadow, etc., as shown in Fig. 8. We will focus on solving these two obstacles in the future.

Ablation Study

In order to verify the effectiveness of Tri-Unet and the mask mechanism we designed, we perform more ablation experiments as shown in Table 2. The mean and standard deviation of RMSE are again used as the evaluation metric. First, from v1, v2, v7, it can be found that the multi-view feature fusion network we designed is superior to traditional CNN and VGG in this task. Then, the results of v3 and v7 hint that the multi-layer feature interaction in the feature extraction stage is better than the direct concatenation of features at the end. To be fair, we set the number of layers of RedNet and ResNet to 50. Through the RMSE of v4 and v7, it is clear that RedNet performs better than ResNet in this task. For v5, we not only remove the mask loss but also the face mask $\mathbf{I}_A$, $\mathbf{I}_B$ and $\mathbf{I}_C$, which is concatenated to the original image. By comparing v5 and v7, we can see that the face mask mechanism promotes the network to generate a higher-precision model. Finally, we remove L_{lan_3d}, which means that our model can be trained with only three multi-view images without any 3D label (as denoted as v6). The result also shows that our model is accurate and stable.

Ours	Backbone	/	L_m	L_{l_3d}	Mean	Std
v1	VGG	/	Y	Y	1.4942	0.2808
v2	CNN	RedNet	Y	Y	1.5415	0.3120
v3	Unet	RedNet	Y	Y	1.5207	0.3063
v4	Tri-Unet	ResNet	Y	Y	1.5802	0.3294
v5	Tri-Unet	RedNet	N	Y	1.5332	0.3244
v6	Tri-Unet	RedNet	Y	N	1.5436	**0.2128**
v7	Tri-Unet	RedNet	Y	Y	**1.4738**	0.3059

Table 2: Ablation Study

As shown in Fig. 9, we selected 3 representative samples from the verification set for visualization. The first sample is an elderly person with one eye open and one eye closed. From the results, our model can predict her skin color and expression with smaller error. Due to the limitations of the 3DMM shape vector base, her wrinkles cannot be refined. The other two samples are angry young women and calm middle-aged man.

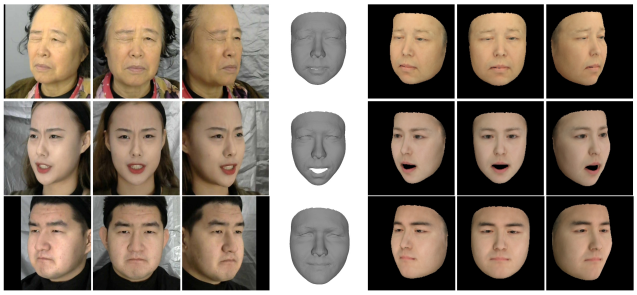

Figure 9: DF-MVR Result (v7) with 3DMM Texture.

Sensitivity Analysis

We conduct sensitivity analysis to examine if the coefficients of Photo loss and Mask loss have impact on the model performance. In order to ensure the accuracy of the model, we performed a parameter sensitivity analysis on ω_p and ω_m. As shown in Fig. 10, we first fix other parameters and only change ω_p. When ω_p is between 4 and 5, the model can obtain higher accuracy. Then, we fix ω_p at 4 and only change ω_m. When ω_m is near 3, the model can obtain higher accuracy. In this way, we set ω_p and ω_m to 4 and 3 respectively.

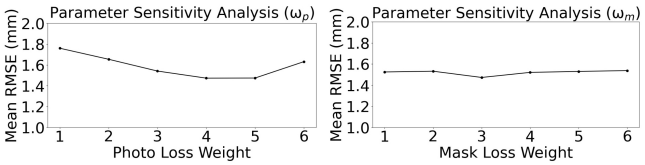

Figure 10: Sensitivity Analysis

Conclusion

In this paper, we design a novel end-to-end weakly supervised multi-view 3D face reconstruction network that exploits multi-view encoding to a single decoding framework with skip connections, able to extract, integrate, and compensate deep features with attention. In addition, we develop a multi-view face parse network to learn, identify, and emphasize the critical common face area. Combining pixelwise photometric loss, mask loss, and landmark loss, we complete the weakly supervised training. Extensive experiments verify the effectiveness of our model. Our further research will focus on deploying multi-view images for training and only using a single image to reconstruct 3D faces.

Reference

Blanz, V.; and Vetter, T. 1999. A morphable model for the synthesis of 3D faces. In *Proceedings of the Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, 187–194.

Booth, J.; Roussos, A.; Ponniah, A.; Dunaway, D.; and Zafeiriou, S. 2018. Large scale 3d morphable models. *International Journal of Computer Vision (IJCV)*, 126(2): 233–254.

Bulat, A.; and Tzimiropoulos, G. 2017. How far are we from solving the 2D & 3D Face Alignment problem? (and a dataset of 230,000 3D facial landmarks). In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Cao, C.; Weng, Y.; Zhou, S.; Tong, Y.; and Zhou, K. 2013. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics (T-VCG)*, 20(3): 413–425.

Deng, Y.; Yang, J.; Xu, S.; Chen, D.; Jia, Y.; and Tong, X. 2019. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 0–0.

Dou, P.; and Kakadiaris, I. A. 2018. Multi-view 3D face reconstruction with deep recurrent neural networks. *Image and Vision Computing*, 80: 80–91.

Dou, P.; Shah, S. K.; and Kakadiaris, I. A. 2017. End-to-end 3D face reconstruction with deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5908–5917.

Feng, Y.; Wu, F.; Shao, X.; Wang, Y.; and Zhou, X. 2018. Joint 3d face reconstruction and dense alignment with position map regression network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 534–551.

Genova, K.; Cole, F.; Maschinot, A.; Sarna, A.; Vlasic, D.; and Freeman, W. T. 2018. Unsupervised training for 3d morphable model regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 8377–8386.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Guo, Y.; Cai, J.; Jiang, B.; Zheng, J.; et al. 2018. Cnn-based real-time dense face reconstruction with inverse-rendered photo-realistic face images. *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, 41(6): 1294–1307.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Li, D.; Hu, J.; Wang, C.; Li, X.; She, Q.; Zhu, L.; Zhang, T.; and Chen, Q. 2021. Involution: Inverting the inherence of convolution for visual recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 12321–12330.

Lin, J.; Yuan, Y.; Shao, T.; and Zhou, K. 2020. Towards high-fidelity 3D face reconstruction from in-the-wild images using graph convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5891–5900.

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 (TOG)*, 36(6).

Lyu, J.; Li, X.; Zhu, X.; and Cheng, C. 2020. Pixel-Face: A Large-Scale, High-Resolution Benchmark for 3D Face Reconstruction. *ArXiv Preprint ArXiv:2008.12444*.

Oktay, O.; Schlemper, J.; Folgoc, L. L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N. Y.; Kainz, B.; et al. 2018. Attention u-net: Learning where to look for the pancreas. *ArXiv Preprint ArXiv:1804.03999*.

Paysan, P.; Knothe, R.; Amberg, B.; Romdhani, S.; and Vetter, T. 2009. A 3D face model for pose and illumination invariant face recognition. In *Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 296–301.

Ravi, N.; Reizenstein, J.; Novotny, D.; Gordon, T.; Lo, W.-Y.; Johnson, J.; and Gkioxari, G. 2020. Accelerating 3D Deep Learning with PyTorch3D. *ArXiv Preprint ArXiv:2007.08501*.

Richardson, E.; Sela, M.; Or-El, R.; and Kimmel, R. 2017. Learning detailed face reconstruction from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1259–1268.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 234–241. Springer.

Sanyal, S.; Bolkart, T.; Feng, H.; and Black, M. J. 2019. Learning to regress 3D face shape and expression from an image without 3D supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 7763–7772.

Sengupta, S.; Kanazawa, A.; Castillo, C. D.; and Jacobs, D. W. 2018. Sfsnet: Learning shape, reflectance and illuminance of facesin the wild'. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6296–6305.

Shang, J.; Shen, T.; Li, S.; Zhou, L.; Zhen, M.; Fang, T.; and Quan, L. 2020. Self-supervised monocular 3d face reconstruction by occlusion-aware multi-view geometry consistency. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 53–70. Springer.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *ArXiv Preprint ArXiv:1409.1556*.

Tam, G. K.; Cheng, Z.-Q.; Lai, Y.-K.; Langbein, F. C.; Liu, Y.; Marshall, D.; Martin, R. R.; Sun, X.-F.; and Rosin, P. L. 2012. Registration of 3D point clouds and meshes: A survey from rigid to nonrigid. *IEEE Transactions on Visualization and Computer Graphics (T-VCG)*, 19(7): 1199–1217.

Tewari, A.; Zollhöfer, M.; Garrido, P.; Bernard, F.; Kim, H.; Pérez, P.; and Theobalt, C. 2018. Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2549–2559.

Tewari, A.; Zollhofer, M.; Kim, H.; Garrido, P.; Bernard, F.; Perez, P.; and Theobalt, C. 2017. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, 1274–1283.

Thies, J.; Zollhofer, M.; Stamminger, M.; Theobalt, C.; and Nießner, M. 2016. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2387–2395.

Tran, A. T.; Hassner, T.; Masi, I.; Paz, E.; Nirkin, Y.; and Medioni, G. G. 2018. Extreme 3D Face Reconstruction: Seeing Through Occlusions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3935–3944.

Tran, L.; and Liu, X. 2018. Nonlinear 3d face morphable model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 7346–7355.

Tuan Tran, A.; Hassner, T.; Masi, I.; and Medioni, G. 2017. Regressing robust and discriminative 3D morphable models with a very deep neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5163–5172.

Wu, F.; Bao, L.; Chen, Y.; Ling, Y.; Song, Y.; Li, S.; Ngan, K. N.; and Liu, W. 2019. Mvf-net: Multi-view 3d face morphable model regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 959–968.

Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; and Sang, N. 2018. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, 325–341.

Zhu, X.; Lei, Z.; Liu, X.; Shi, H.; and Li, S. Z. 2016. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 146–155.

From 2D Images to 3D Model: Weakly Supervised Multi-View Face Reconstruction with Deep Fusion

Extended Experiment

Single-view Reconstruction. The single-view reconstruction method only requires one image to generate the 3D face. From the practical viewpoint, it is more flexible though it may be inferior to multi-view methods in terms of accuracy. Our method can also be adapted in the single view scenario. More specifically, during the training process, we only change the input, without changing other parts. As shown in Fig. 1, the original input has been changed to four different forms, according to the probability: $\boldsymbol{P}'_{\mathrm{a}}$, $\boldsymbol{P}_{\mathrm{l}}$, $\boldsymbol{P}_{\mathrm{f}}$,$\boldsymbol{P}_{\mathrm{r}}$. The input of multi-view images still needs to be dominant to preserve accuracy, so we set its probability to 2/3, and the other inputs equally distribute with the probability of 1/3.

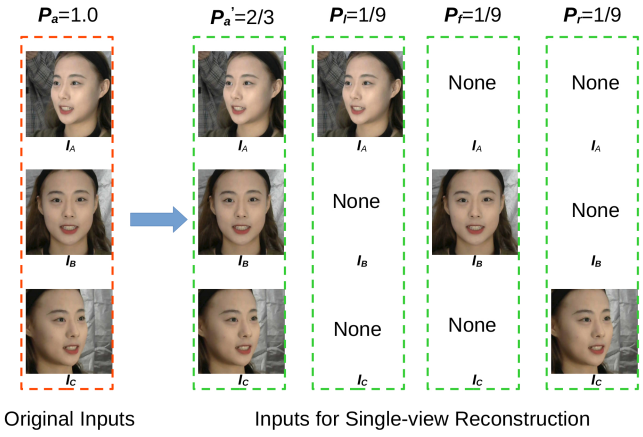

Figure 1: Inputs for Single-view Reconstruction. 'None' means a black image.

The parameterized results of the comparison are shown in Table 1. As observed, our proposed model also attains the superior performance. In the case of single image testing, our model is more effective than (Deng et al. 2019) and (Shang et al. 2020). More specifically, (Shang et al. 2020) adopted multiple images for training, and then a single image for testing, which is the same as our model in the case of sing-view. (Deng et al. 2019) does not release their source codes of their scoring network. We use their codes to train/test on Pixel-Face. Visual comparisons are shown in Fig. 2.

Method	Dataset	Mean	Std
(Shang et al. 2020)	Pixel-Face	1.8877	0.4378
(Deng et al. 2019)	Pixel-Face	1.6641	0.3690
Ours (single-view)	Pixel-Face	1.5437	0.3088
Ours	Pixel-Face	**1.4738**	**0.3059**

Table 1: Comparison of RMSE (mm)

It can also be seen from Fig. 2 that our model is more sensitive to depth changes. The mouths and cheeks reconstructed by our model are more accurate. On the other hand, the three 3D faces reconstructed from multi-view in (Shang et al. 2020) have greater differences, while the three 3D faces reconstructed by our method from multi-view appear more similar.

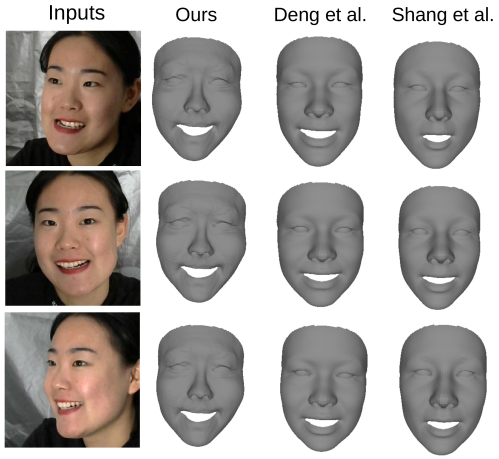

Figure 2: Comparisons to SOTA Methods. We only input one image to the network.

DF-MVR Results

In this section, we provide more visualization results with 3DMM texture.

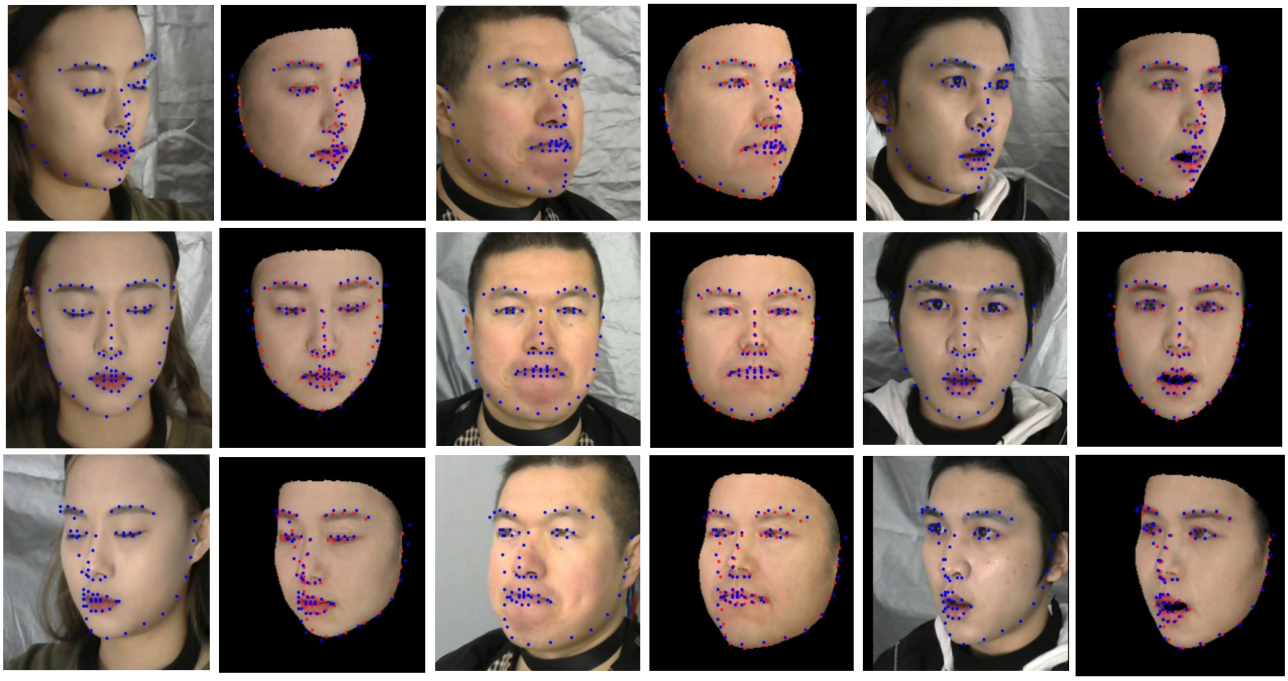

Figure 3: DF-MVR Results with 3DMM Texture. The blue dots are the groundtruth of the 2D lanmarks, and the red dots represent the projection of the corresponding 3D points which we predict.

Figure 4: DF-MVR Results with 3DMM Texture (w/o L_{l_3d})

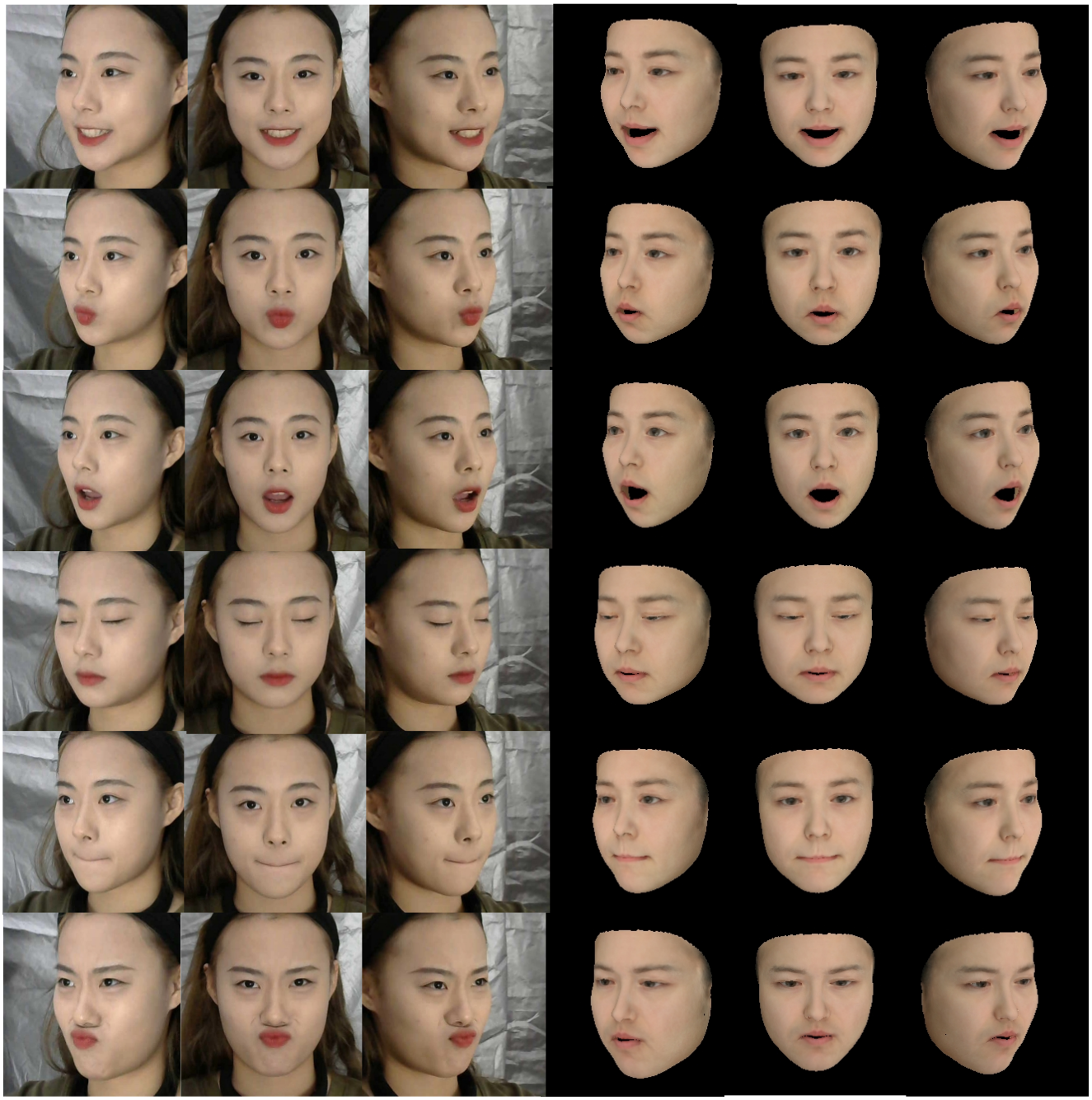

Figure 5: DF-MVR Results with 3DMM Texture. We provide multiple expression reconstruction results of the same person to show the effectiveness of our model