Detailed 3D Face Reconstruction in Full Pose Range

Anonymous submission

In this supplement, we will provide additional more experimental details.

Related Work

Coarse 3D Face Reconstruction

In work based on the BFM models, 3DDFA combined 3DMM, cascaded regression and CNN to align and reconstruct 3D face. In this process, alignment and reconstruction are done together, and 3DDFA fits the entire 3D dense face model to the face picture instead of the sparse feature point shape. Different from 3DDFA, MOFA proposes a framework for self-supervised 3D face reconstruction. Its input is a picture and 2D key points (optional), and its encoder extracts the pose, shape, expression, facial reflection, and illumination of the picture, and uses these parameters to reconstruct a 3D face. During training, the reconstructed 3D face is projected back to 2D, and a loss is computed with the input image to optimize the reconstruction process. Since most 3D face reconstruction methods use synthetic data or 3D labels, the quality of the data will affect the accuracy of the reconstruction. To alleviate this problem, 19qingghua proposes a hybrid loss function for weak supervision, which improves the reconstruction accuracy by integrating pixel loss and perceptual loss. In work based on the FLAME model, DECA decouples camera, albedo, light, shape, pose, and expression parameters for the input image, and loads them into the FLAME model to obtain a 3D face. Similar to MOFA, DECA finally projects the 3D face into a 2D image for self-supervised training, and DECA also has 2D keypoint labels for weakly supervised training. In order to obtain more realistic 3D face, MICA employs 3D labels for supervised training. However, these methods need expensive large-scale training set to achieve good results, otherwise, their reconstruction accuracy will be difficult to guarantee.

Small Pose Facial Details Reconstruction

Sela17 estimates the high-frequency part of the texture by subtracting the low-pass filtered part from the luminance values of the input image, which could be obtained by convolving the texture with a spatially varying Gaussian kernel in the image domain. Then Sela17 uses the high-frequency part of the texture to reconstruct the details of the face, but in practical applications, the details of the face reconstructed by Sela17 are rough and accompanied by a large number of noise points. Extreme 3D uses the details calculated by the SFS method as labels, trains a network to estimate the bump map (displacement map) of the input image and reconstructs it as facial details. The disadvantage of Extreme 3D is that the facial details it reconstructs are not real and it could not reconstruct small facial details. The UV displacement map of chen20 is obtained by the difference between the input image and the coarse face rendering image, and the input image is used as the self-supervised training label. The detail reconstruction effect of chen20 is more realistic, but some subtle facial details are missed. DECA does not estimate the UV displacement map from the input image, but uses a generator to generate it. DECA could not realistically restore the facial details of the input image. Unlike methods using UV displacement maps, based on self-supervised decomposition of diffuse normals and specular normals, fg23 uses SFS to obtain detailed facial geometry and approximate diffuse albedo, diffuse shadows, and specular shadows. However, the details reconstructed by fg23 are sparse, and there are obvious noise points. Overall, the common problem of existing methods is that it is difficult to capture subtle details in human faces.

Objective Function

Coarse 3D Face Reconstruction The total objective function is denoted as:

$$\mathcal{L} = w_1\mathcal{L}_p + w_2\mathcal{L}_{lm} + w_3\mathcal{L}_{id} + w_4\mathcal{R}_{param} \tag{1}$$

Specifically, the every objective function is denoted as:

$$\mathcal{L}_p = \frac{1}{|\mathcal{M}|}\sum_{(i,j)\in\mathcal{M}} \left\| I_{i,j} - I^R_{i,j} \right\|_2, \tag{2}$$

where M are the pixels of the visible region on the I^R, while i and j are their positions. $\mathcal{L}_p$ is obtained by averaging the $L_{2,1}$ distance of all visible pixels.

$$\mathcal{L}_{lm} = \frac{1}{N}\sum_{i=1}^{N} \left\| p_i - p_i^R \right\|_2^2 \tag{3}$$

where p_i represents the i-th landmark position in the input image, p_i^R is the corresponding i-th landmark position in the rendered face image, and N = 68 is the number of landmarks.

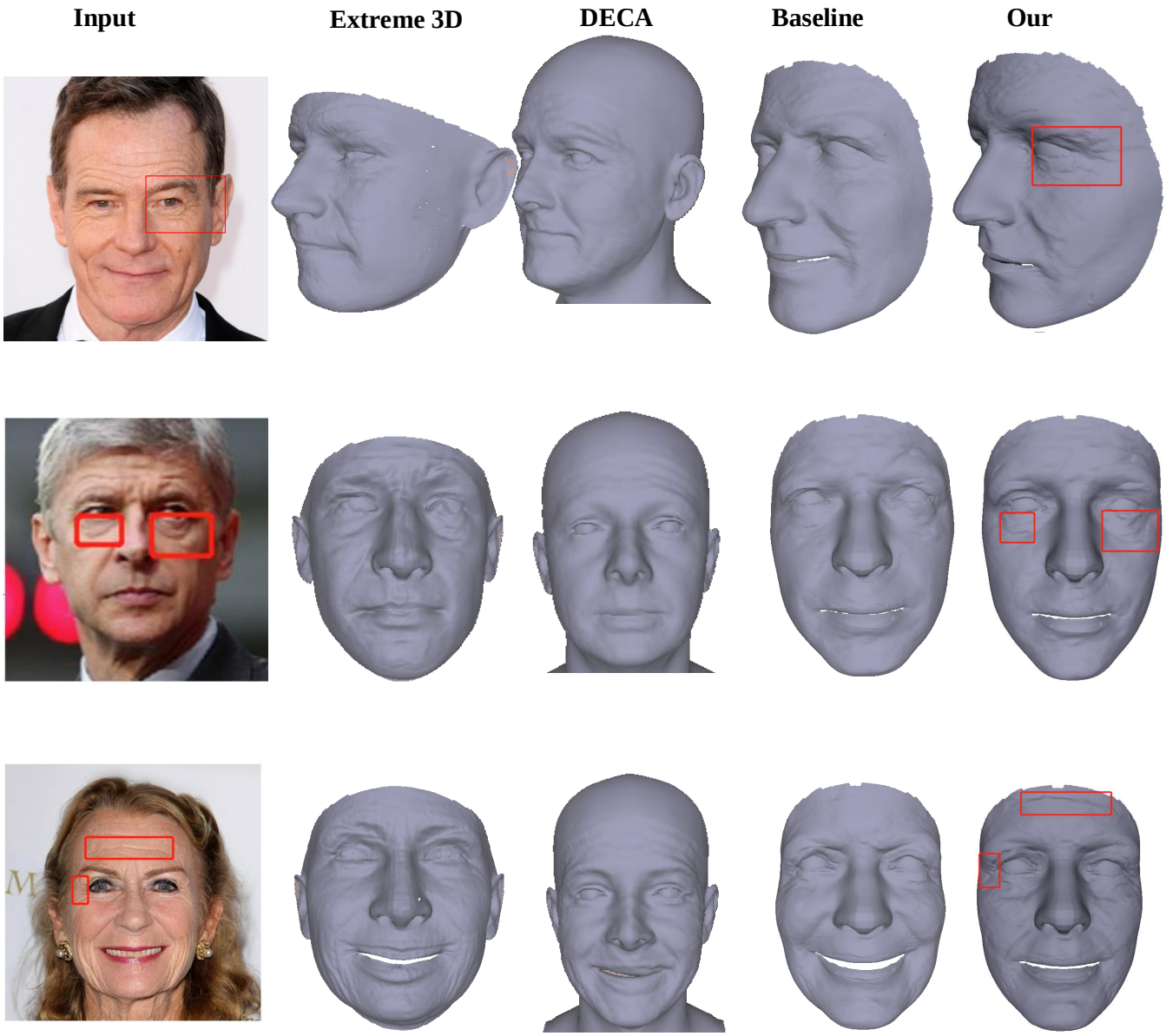

Figure 1: Detailed 3D face reconstruction results of the existing and our methods.

$\mathcal{L}_{lm}$ could constrain the network's learning of pose and expression parameters.

$$\mathcal{L}_{id} = \left\| \phi(I) - \phi\left(I^R\right) \right\|_2^2 \tag{4}$$

where ϕ represents the feature extraction process of the VGG-16 network.

$$\mathcal{R}_{param} = \omega_s \left\| x_{shape} \right\|^2 + \omega_e \left\| x_{exp} \right\|^2 \tag{5}$$

where ω_s and ω_e are weighting parameters.

Detailed 3D Face Reconstruction The total objective function is denoted as:

$$\mathcal{L} = a_1 \mathcal{L}_p + a_2 \mathcal{L}_s + a_3 \mathcal{R}_{disp} \tag{6}$$

Specifically, $\mathcal{L}_p$ is the same as the $\mathcal{L}_p$ in Coarse 3D Face Reconstruction section, $\mathcal{L}_s$ is introduced in the body of the paper, and $\mathcal{R}_{disp}$ is denoted as:

$$\mathcal{L}_{disp} = \sum_i w_{dn} \|\Delta n(i)\|^2 + w_{dz} \|\Delta z(i)\|^2 \tag{7}$$

This regularization term could alleviate the drastic depth value variation in the estimation of the displacement depth map, thereby reducing the distortion of the reconstructed 3D face.

Implementation Details

We first train the Baseline method for 20,0000 steps on the small training set, and then employ its VGG-16 as the network of our coarse 3D face reconstruction. The number of training steps for the coarse 3D face reconstruction is 250 steps, and the batch size is 6. In the Self-augment mechanism, the rendering number of large pose faces is set to 26, while the pose parameter yaw yaw_i and its displacement tx_i are set to $\begin{cases} yaw_i = \delta(\pi/180) \cdot [10 \cdot (6 + i/4)] \\ tx_i = \delta(160/9) \cdot (6 + i/4) \end{cases}$, pitch $pitch_i$ and its displacement ty_i are set to $\begin{cases} pitch_i = 0 \\ ty_i = 0 \end{cases}$, roll $roll_i$ and its displacement sth_i are set to $\begin{cases} roll_i = 0 \\ sth_i = 0 \end{cases}$, where δ= $\begin{cases} -1, i \in [0, 12] \\ 1, i \in [13, 25] \end{cases}$ and i represents the serial number of the pose parameter. In the training of the small pose detailed 3D face reconstruction, the t of RI2ITN is 2, the number of training steps is 20,000. In the training of the large pose detailed 3D face reconstruction, we load the RI2ITN trained by the small pose detailed 3D face reconstruction, and continue to train for 5000 steps. We write the codes with TensorFlow 1.15 and train on a Tesla V100 and a GTX 2080Ti for coarse and detailed face reconstruction, respectively. The optimizer and all hyperparameter settings such as the objective function weights w_1, w_2, w_3, w_4, ω_s, ω_e, a_1, a_2, a_3, b_1, and b_2 are all consistent with the open source version of the Baseline. The code of Baseline is available from https://github.com/cyj907/unsupervised-detail-layer. The Baseline mentioned in the experimental part of this paper was trained on the small-scale dataset using its open source version. The epoch of DECA trained through small-scale dataset is set to 10 times to published version, and other all hyperparameter settings are the same to the published version. The code of DECA is available from https://deca.is.tue.mpg.de/.

Datasets

Training Datasets CelebFaces Attributes Dataset (CelebA) is a large-scale face attributes dataset with more than 200K celebrity images. We use 1,000 images from the training set of the CelebA dataset.

Evaluation Datasets

- MICC Florence. In MICC, videos are taken on 53 subjects under Indoor Cooperative, PTZ Indoor, and PTZ Outdoor conditions. The ground truth 3D scans are provided for 52 out of the 53 people. We select the Indoor Cooperative condition which has the clearest face, and crop a face picture with a small pose and a face picture with a large pose for each available subject from the clear video frame as the evaluation dataset.
- NOW Validation. In this dataset, there are 20 subjects in total, each subject has face pictures from 5 or 6 poses and a ground truth 3D scan. For each subject, we select a face picture with small pose and a face picture with large pose as the evaluation dataset under the two poses respectively.

Experiment

Qualitative evaluation

We show more results of details reconstruction in Figure 1. It could be seen that our method captures and reconstructs finer wrinkles, which are often lost by other methods.