Personalized Face Modeling for Improved Face Reconstruction and Motion Retargeting

Anonymous ECCV submission - **Supplementary**

Paper ID 2986

Our two main contributions in this paper are a) to model personalized expression blendshapes and dynamic albedo maps, and b) to perform joint tracking and modeling in a decoupled manner to support both reconstruction and motion retargeting. In this supplementary material, we provide more qualitative results of our approach to substantiate the improvements over baseline caused by our contributions.

1 Training Details

The architectures of the encoder and decoders of our modeling network are given in Table 1a and Table 1b. Each Conv2D and Deconv2D layer is followed by batch normalization which is then followed by ReLU activation. Our end-to-end network has a size of 240 MB and takes 15.4 ms to execute 1 image and 37.5 ms to execute 4 images on a Titan X GPU on average. The loss weights are chosen to be: $\lambda_{\text{ph}} = 200; \lambda_{\text{lm}} = 0.1; \lambda_{\text{pa}} = 50; \lambda_{\text{sd}} = 2.5; \lambda_{\text{bg}} = 1.5; \lambda_{\text{reg}} = 10^{-3}; \lambda_{\gamma} = 0.02$.

Fine-tuning the modeling network during the second stage of training ensures further decoupling between tracking and modeling. Besides, our method of obtaining the user-specific face shape and albedo from multiple frames helps in learning the static shape and albedo corrections separately from the expression-specific shape and albedo variations. As a result, our framework can produce photorealistic expression-specific deformations on a new user during testing.

2 More Qualitative Results

Fig. 1 shows 3D face reconstruction results using our method on our test data. It can be noted that our method can reconstruct faces accurately even under conditions like unusual lighting (row 6), uncommon face shapes (baby face in row 3), extreme poses (rows 4 and 7), occlusion (row 9), extreme expressions etc. However, similar to [3], our method embeds eye glasses into the albedo (row 9). We would like to point out that videos of ExpressiveFaces are captured by hand-held cameras and have resolution of 1920×1080, from which we crop faces resized to size 224×224. On the other hand, videos in Voxceleb2 [2] are scraped from Youtube and hence have a very different distribution (lower resolution) than the videos of ExpressiveFaces.

3 Video Results

The performance of our method on face videos is shown in the video attached with this document. We show three applications of our method: a) personalized reconstruction, b)

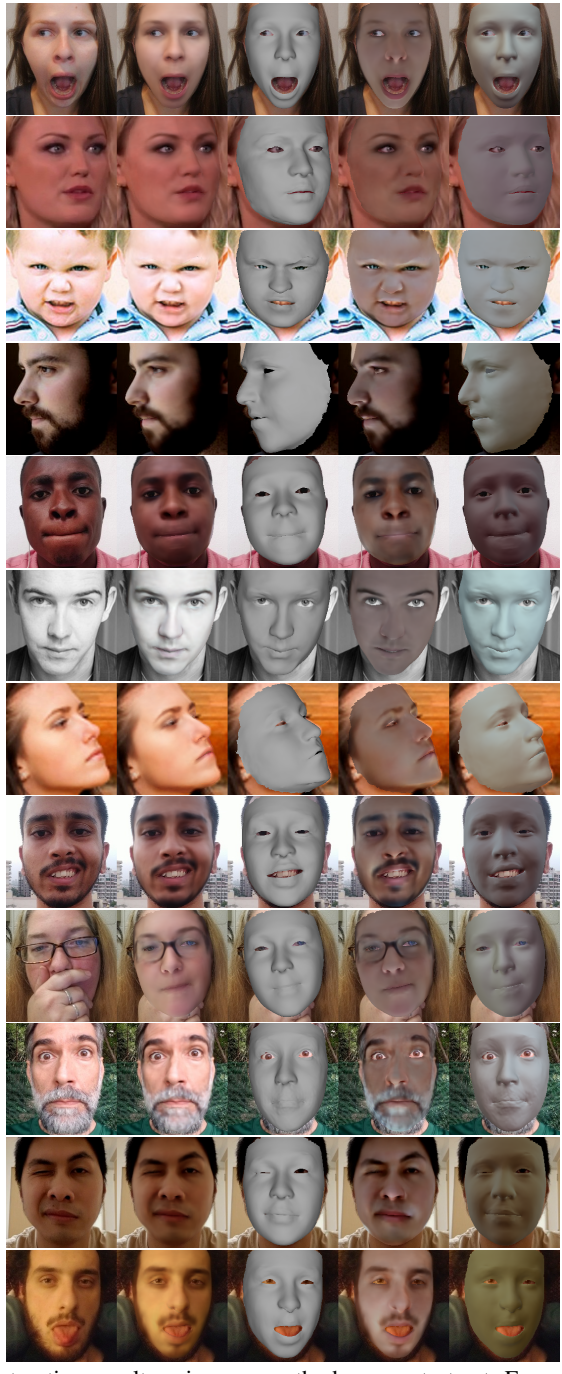

Fig. 1: Face reconstruction results using our method on our test set. From left to right: input image, overlay, shape, albedo, lighting.

Table 1: Architecture of our networks. $s\#$ refers to stride $\#$. (a) Architecture of the shared encoder of our modeling network. The outputs of the encoder for each input image in a mini-batch are average pooled to obtain a single $(7, 7, 512)$ feature that becomes the input to both the decoders. (b) Architecture of each of the decoder of our modeling network. Note that the output of the last Deconv2D layer goes into both the last 2 Conv2D layers.

(a)

Layers	Input Shape	Output Shape
Conv2D (7×7, s2)	(224,224,3)	(112,112,64)
Maxpool (3×3, s2)	(112,112,64)	(56,56,64)
Conv2D (3×3, s1)	(56,56,64)	(56,56,128)
Conv2D (3×3, s2)	(56,56,128)	(28,28,128)
Conv2D (3×3, s1)	(28,28,128)	(28,28,256)
Conv2D (3×3, s2)	(28,28,256)	(14,14,256)
Conv2D (3×3, s1)	(14,14,256)	(14,14,512)
Conv2D (3×3, s2)	(14,14,512)	(7,7,512)

(b)

Layers	Input Shape	Output Shape
Deconv2D (4×4, s2)	(7,7,512)	(16,16,512)
Deconv2D (4×4, s2)	(16,16,512)	(32,32,256)
Deconv2D (4×4, s2)	(32,32,256)	(64,64,128)
Deconv2D (4×4, s2)	(64,64,128)	(128,128,64)
Conv2D (1×1, s1)	(128,128,64)	(128,128,3)
Conv2D (1×1, s1)	(128,128,64)	(128,128,56*3)

retargeting to a different user's face model, and c) retargeting to an external 3D puppet. To process the input video, we detect the face bounding box using [1] for the first frame only. For the subsequent frames, the bounding box of each frame is obtained from the boundaries of the 2D landmarks predicted in the previous frame. This technique helps in reducing the jitter in the results due to inconsistent bounding box selection if done on a per-frame basis. However, some temporal smoothing as a post-processing step will produce better results.

References

1. Chaudhuri, B., Vesdapunt, N., Wang, B.: Joint face detection and facial motion retargeting for multiple faces. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
2. Chung, J.S., Nagrani, A., Zisserman, A.: Voxceleb2: Deep speaker recognition. In: INTERSPEECH (2018)
3. Tewari, A., Bernard, F., Garrido, P., Bharaj, G., Elgharib, M., Seidel, H., Pérez, P., Zollhöfer, M., Theobalt, C.: FML: face model learning from videos. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)