AU-aware 3D Face Reconstruction through Personalized AU-specific Blendshape Learning Supplementary Material

Chenyi Kuang[1], Zijun Cui[1], Jeffrey O. Kephart[2], and Qiang Ji[1]

[1] Rensselaer Polytechnic Institute
{kuangc2,cuiz3,jiq}@rpi.edu
[2] IBM Thomas J. Watson Research Ctr.
kephart@us.ibm.com

To provide complementary descriptions about our experiments and results, we divide the supplementary material into the following sections: Traning Details, AU Prior from Bayesian Network, AU-specific Blendshape Results, More Qualitative Results and Method Limitation.

1 Training Details

We train our **baseline+AU label+AU prior** model with the following loss function:

$$\begin{aligned} L =& \lambda_{img} L_{img} + \lambda_{lmk} L_{lmk} + \lambda_{id} L_{id} + \lambda_G L_G + L_{sp} \\ &+ \lambda_{au-label} L_{au-label} + \lambda_{au-corr} L_{au-corr} \end{aligned} \tag{1}$$

where we set $\lambda_{img} = 1.75$,$\lambda_{lmk} = 0.002$,$\lambda_{id} = 0.1$, $\lambda_G = 0.015$, $\lambda_{sp} = [1.0, 0.75, 0.05]$, $\lambda_{au-label} = 1.0$, $\lambda_{au-prior} = 0.1$.

2 AU Prior from Bayesian Network

In our paper, we propose to utilize prior relationship between AU pairs to constrain the learning of 3D coefficients $\boldsymbol{\alpha}$. Following the procedure of Cui et al.[3] of using expression-independent inequality constraints among AUs, we pre-train a Bayesian Network(BN) to represent the general AU prior relationship among 8 AUs. The output include an adjacency matrix indicating the structure of the BN and a 256-dim probability array of the joint configuration of 8 AUs. We select a set of eight AU pairs that are available in Table.1 in our paper and the pre-trained BN, which is: $\{(1,2),(4,7),(6,12),(15,17),(2,6),(2,7),(12,15),$ $(12,17)\}$. We can compute the joint probability for each AU pair by marginalizing out the rest six nodes.

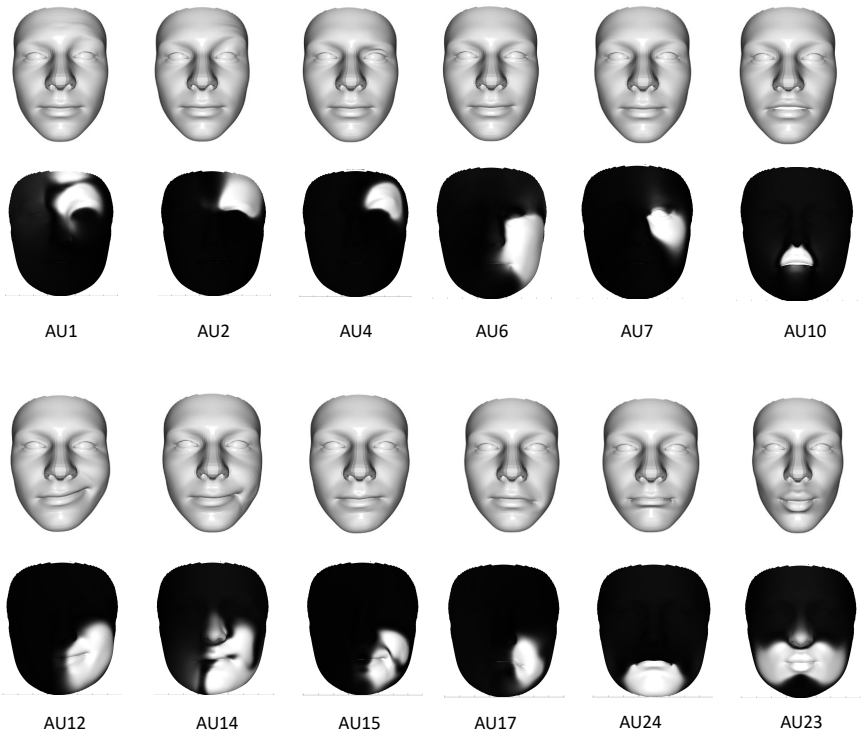

Fig. 1. Pre-computed AU mask W_k for constructing AU-aware blendshapes. **Row1 & Row2**: registered ICT shapes with BFM topology for {AU1, AU2, AU4, AU6, AU7, AU10}; **Row3 & Row4**: registered ICT shapes with BFM topology for {AU12, AU14, AU15, AU17, AU24, AU23}.

3 AU-specific Blendshape Results

For each input subject, we predict the 3D AU coefficients $\boldsymbol{\eta}$ to produce the AU-specific blendshapes follwoing Eq. 2.

$$B_{au}[k] = \sum_{m=1}^{M} \eta_{k,m} W_k \odot \boldsymbol{B}_{exp}[m], \forall k \in 1, ..., K \tag{2}$$

In Eq. 2, we use globally deformed PCA basis $\boldsymbol{B}_{exp}$ from BFM model to construct our personalized AU-aware blendshapes $\boldsymbol{B}_{au}$, which are locally deformed on certain AU region. In Eq. 2, $W_k \in [0,1]^{N\times 1}$(N is total vertex number) for k-th blendshape is defined to filter out potential deformations from irrelevant face regions. In particular, for k-th blendshape, a relevant face region consisting of a subset of vertices on 3D mesh is identified through the ICT model. We first perform a non-rigid ICP from each ICT templates to BFM topology and compute the vertex deformations between a blendshape and the mean face shape.

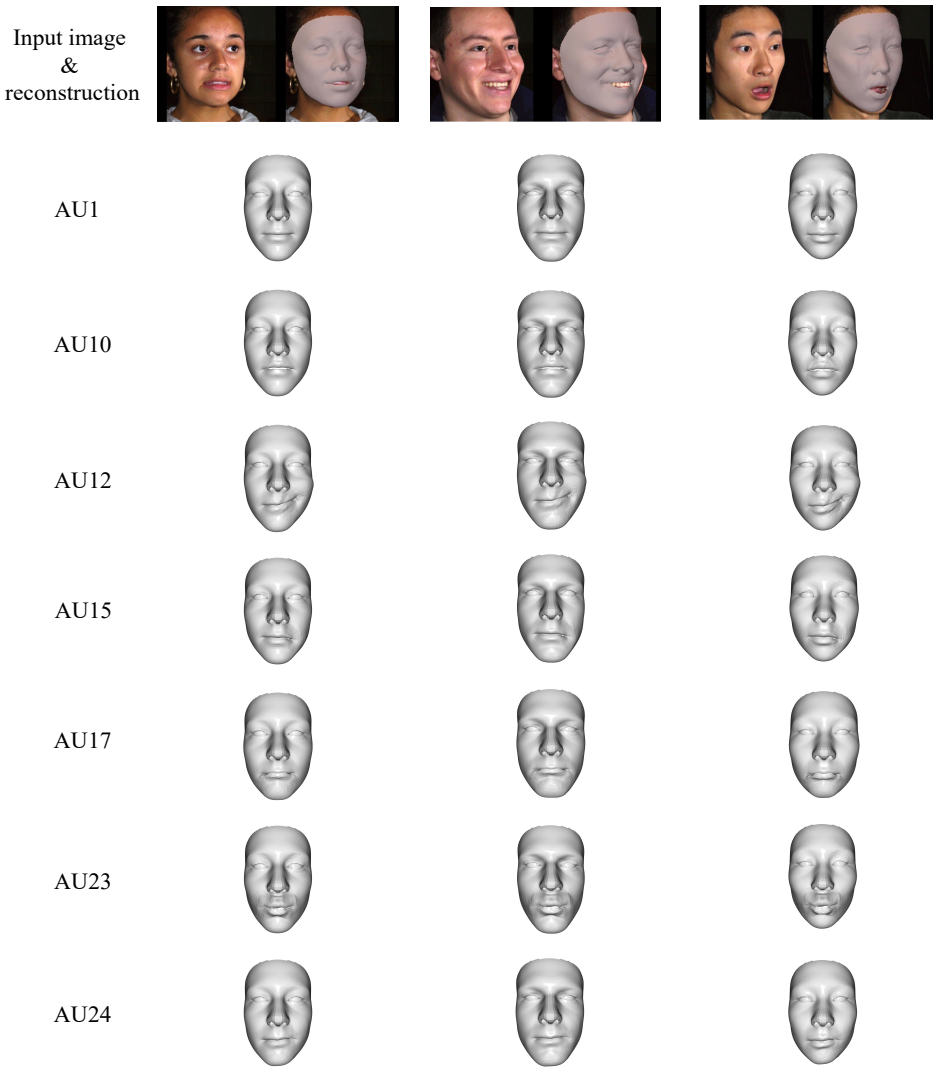

Fig. 2. AU-specific blendshape visualization of each input image.**First row**: input image and reconstruction result; **row2-row8**:AU-specific blendshapes for 7 different AUs.

Elements in W_k corresponding to identified active vertices are assigned to "1", and "0" otherwise. Then we perform a smoothing of the weights in W_k for those boundary vertices to ensure the transition between "active" vertices and "inac-

tive" vertices are smooth. In Fig. 1, we provide the visualization of W_k for 12 AUs that are involved in our paper and these pre-defined W_k are not updated during training or testing.

In Fig. 2, we provide visualization examples of personalized blendshapes for random selected testing subjects in BU3DFE [8], i.e., $S_{neu} + B_{au}[k]$ for interested AUs. Compared with original BFM[4] model, the produced blendshapes are locally deformed in the AU-related face region and contains left-blendshape and right-blendshape for most AUs. In total our model generate 49 blendshapes for constructing accurate expressions.

More discussions on person-specific parameters: we show the diversity of person specific parameters by visualizing the mean and variance of predicted facial identity parameter $\boldsymbol{\beta}$ over 100 subjects in BU3DFE in Fig. 3. On the other hand, the AU-related parameter $\boldsymbol{\alpha}$ for different subject with similar expressions will cluster. We show clustering results over 100 subjects on four expressions using the AU-related parameter $\boldsymbol{\alpha}$ in Fig. 4, by visualizing the intensity coordinates of two specific AUs. We can draw a conclusion that parameters cluster under the same expression with different subjects (e.g., happy), and vary with different subjects and expressions.

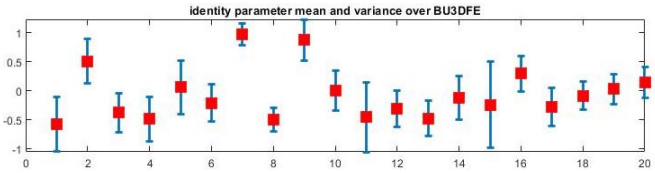

Fig. 3. Mean (red points) and variance (blue segments) of the first 20 dimensions of $\boldsymbol{\beta}$ over 100 subjects in BU3DFE.

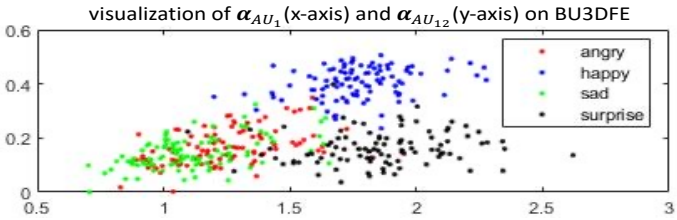

Fig. 4. Clustering using $(\boldsymbol{\alpha}_{AU_1}, \boldsymbol{\alpha}_{AU_{12}})$ for four different expressions over 100 subjects on BU3DFE.

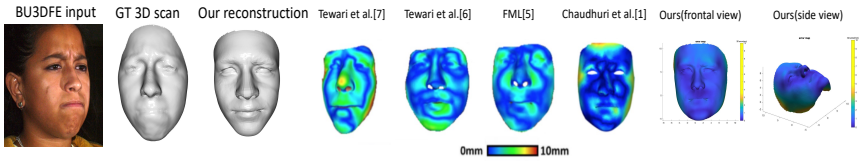

Fig. 5. 3D reconstruction error maps on one BU3DFE testing example.

4 3D Reconstruction Error on BU3DFE

Through learning personalized AU-aware blendshapes, our final model can recover subtle facial motions related to AUs in 3D space and achieves SOTA reconstruction accuracy on BU3DFE [8] dataset. In Fig. 5, we perform a quantitative comparison of 3D reconstruction error with Tewari et al. [7], Tewari et al. [6], FML [5] and Chaudhuri et al. [1](note that the color bar scaling is just used to represent different levels of errors in millimeter and our colorbar is a little bit different from the colorbar used by Chaudhuri et al. [1].)

5 More Qualitative 3D Reconstruction Results

In Fig. 7 we provide additional reconstruction results on VoxCeleb2[2]. It can be noted that our model can generate accurate 3D faces of various expressions under different conditions. Besides 3D face reconstruction, expression manipulation can also be easily achieved by our model in AU level by adjusting the value of $\boldsymbol{\alpha}$ in the reconstructed 3D face $S = \bar{B} + \boldsymbol{B}_{id}\boldsymbol{\beta} + \boldsymbol{B}_{au}\boldsymbol{\alpha}$. Compared to PCA expression basis, we can manipulate the deformation intensity for each AU while fixing other parts, as the example shown in Fig. 6.

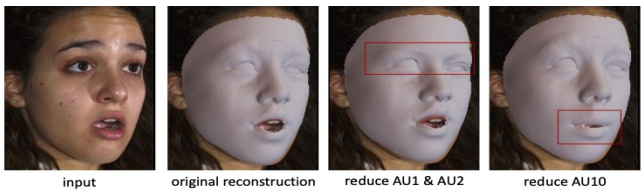

Fig. 6. Expression manipulation by adjusting $\boldsymbol{\alpha}$ value corresponding to AU1&AU2 (column 3) and AU10 (column 4).

6 Method Limitation

Our model performs joint face reconstruction and AU detection and achieves state-of-art 3D reconstruction accuracy but still have limitations in within-

dataset AU detection on BP4D compared with appearance-based methods. According to the Table. 5 in our paper, our model achieves state-of-art F1-score on AU1, AU2, AU23 but worse results on the rest AUs. It can be explained by the mechanism of the 3D blendshape model fitting. First of all, to produce accurate 3D reconstruction, we construct 49 blendshapes in total while only 12 AU labels are available for the AU label regularization loss. Besides, considering the number of nodes in the pre-trained BN, the integrated pairwise AU prior are limited within the combinations of the 8 AUs involved in the BN. Besides, as we can observe in Fig. 2, the constructed AU-specific blendshapes are not orthogonal basis and have deformation redundancy in local face regions for those positive-correlated AU groups(like AU6, AU12). At the same time, blendshapes for negative correlated AU groups(like AU12,AU15) may eliminate each other in vertex deformation. Therefore, we have to introduce a sparsity constraint on the expression coefficient $\boldsymbol{\alpha}$ to ensure the reconstruction accuracy is maintained.

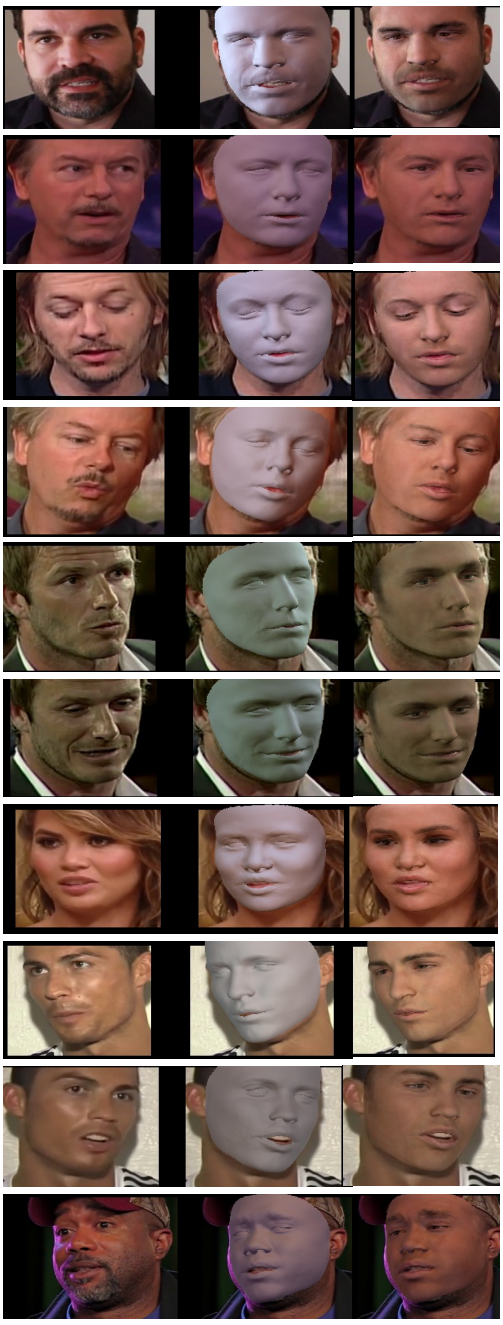

Fig. 7. More qualitative results on VoxCeleb2[2].**First column**: input images; **second column**:reconstructed shape withoout texture ; **third column**: reconstructed shape with texture.

References

1. Chaudhuri, B., Vesdapunt, N., Shapiro, L., Wang, B.: Personalized face modeling for improved face reconstruction and motion retargeting. In: European Conference on Computer Vision. pp. 142–160. Springer (2020)
2. Chung, J.S., Nagrani, A., Zisserman, A.: Voxceleb2: Deep speaker recognition. arXiv preprint arXiv:1806.05622 (2018)
3. Cui, Z., Song, T., Wang, Y., Ji, Q.: Knowledge augmented deep neural networks for joint facial expression and action unit recognition. Advances in Neural Information Processing Systems **33** (2020)
4. 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)
5. Tewari, A., Bernard, F., Garrido, P., Bharaj, G., Elgharib, M., Seidel, H.P., Pérez, P., Zollhofer, M., Theobalt, C.: Fml: Face model learning from videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10812–10822 (2019)
6. Tewari, A., Zollhöfer, M., Garrido, P., Bernard, F., Kim, H., Pérez, P., Theobalt, C.: 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. pp. 2549–2559 (2018)
7. Tewari, A., Zollhofer, M., Kim, H., Garrido, P., Bernard, F., Perez, P., Theobalt, C.: Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 1274–1283 (2017)
8. Yin, L., Wei, X., Sun, Y., Wang, J., Rosato, M.J.: A 3d facial expression database for facial behavior research. In: 7th international conference on automatic face and gesture recognition (FGR06). pp. 211–216. IEEE (2006)