Learning 3D Faces from Photo-Realistic Facial Synthesis Supplementary Materials

Ruizhe Wang * †1, Chih-Fan Chen *†1, Hao Peng†1, Xudong Liu †1, 2, and Xin Li ‡2

[1]Oben, Inc
[2]West Virginia University

Section 3.2. Photo-Realistic Rendering

Figure 1 demonstrates the 30 different manually created lighting conditions that are used to render photo-realistic facial images. Figure 2 shows final high-quality rendering samples with V-ray as well as those with Unity, which is used for ablation study in Figure 7 (main paper). Note that, for both rendering methods, we randomized the head pose, environment map, lighting condition, and the field of view (FOV) to mimic the selfie in the real world.

Section 3.4. Refinement and Texture Transfer

Landmarks Localization To achieve higher landmark localization accuracy, we have developed a coarse-to-fine approach. First, we predict all facial landmarks from the detected facial bounding box. Then, given the initial landmarks, we crop the eye, nose, and mouth areas for the second stage fine-scale landmark localization. Figure 3 shows our landmark mark-up as well as the bounding boxes used for the fine scale landmark localization stage. We have used a regression forest based approach [6] as the base landmark predictor and we train 4 landmark predictors in total, (i.e., for overall face, eye, nose and mouth.)

Non-rigid Deformation We propose to utilize facial landmarks detected in a coarse-to-fine fashion and formulate non-rigid deformation as an optimization problem that jointly optimizes over camera intrinsic, camera extrinsic, facial expression and a per-vertex displacement field.

Problem Formulation To handle facial expressions, we transfer the expression blendshape model in FaceWarehouse [2] to the same head topology with artist's assistance as $\{\mathbf{B}_1, \mathbf{B}_2, ..., \mathbf{B}_M\}$. In addition, we introduce a per-vertex correction field $\delta\mathbf{P}$ to cover out of space non-rigid deformation. Finally, a 3D face is reconstructed as $\mathbf{P}_F = \mathbf{P} + \sum_{i=1}^{M} \boldsymbol{\beta}_i \mathbf{B}_i + \delta\mathbf{P}$. Camera extrinsic $\mathbf{T}$ transforms the face from its canonical reference coordinate system to the camera coordinate system. It has a 3-DoF vector $\mathbf{t}$ for translation and a 3-DoF quaternion representation $\mathbf{q}$ for rotation. Camera intrinsic $\mathbf{K}$ projects the 3D model to the image plane. During the optimization, we have found that using a scale factor f_s to update the intrinsic matrix by $K = \begin{pmatrix} f_s f & 0 & c_x \\ 0 & f_s f & c_y \\ 0 & 0 & 1 \end{pmatrix}$ leads to the best numerical stability. Here $[f, c_x, c_y]$ are all initialized from the size of the input image as $c_x = \frac{im_h}{2}$, $c_y = \frac{im_w}{2}$, and $f = max(im_h, im_w)$. Putting things together, we can represent the overall parameterized vector by $\mathbf{p} = [\boldsymbol{\beta}, \delta\mathbf{P}, \mathbf{t}, \mathbf{q}, f_s]$.

Landmark Term We employ a global-to-local method for facial landmark localization. For global inference, we first detect the standard 68 facial landmarks, and use this initial estimation to crop local areas including eyes, nose, and mouth (i.e., a total of 4 cropped images). Then we perform fine-scale local inference on the cropped images (Please see the supplementary material for more details). The landmark localization approach produces a set of facial landmarks $\mathbf{L}$ where $\mathbf{L}_i = [L_i^x, L_i^y]$. We propose to minimize the distance between the predicted landmarks on the 3D model and the detected landmarks,

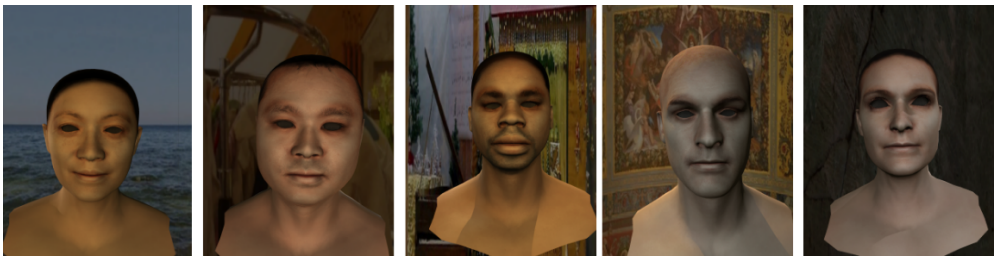
$$E_l = \frac{100}{W_{eye}} \sum_{i=1}^{K} \|\mathcal{K}(\mathcal{T}(\mathcal{S}_{\mathcal{M}}(\mathbf{P}_F, \mathbf{m}_i), \mathbf{T}), \mathbf{K}) - \mathbf{L}^i\|^2, \quad (1)$$

where $\mathcal{S}_{\mathcal{M}}(\mathbf{P}, \mathbf{m}_i)$ samples a 3D vertex from $\mathbf{P}$ given a production-ready and sparse triangulation $\mathcal{M}$ on barycentric coordinates $\mathbf{m}_i$, $\mathcal{K}(\cdot,\cdot)$ and $\mathcal{T}(\cdot,\cdot)$ are perspective projection and rigid transformation operators respectively, W_{eye} is the distance between two outermost eye landmarks and $\frac{100}{W_{eye}}$ is used to normalize the eye distance to 100. We pre-select $\mathbf{m}$ on $\mathcal{M}$ and follow the sliding scheme [1] to update the barycentric coordinates of the 17 facial contour

*equal contribution
†{ruizhe, chihfan, hpeng, xudong}@oben.com
‡xin.li@mail.wvu.edu

Figure 1: Different lighting conditions for photo-realistic rendering augmentation

(a) V-ray rendering samples



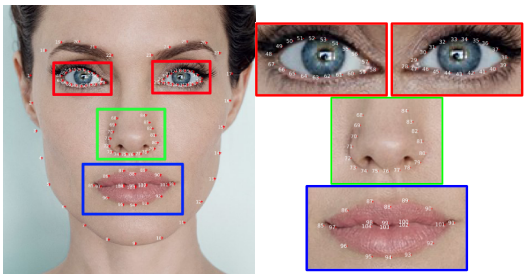
(b) Unity rendering samples

Figure 2: The synthetic facial images from (a) Maya V-ray and (b) Unity

landmarks at each iteration.

Corrective Field Regularization: To enforce a smooth and small per-vertex corrective field, we combine the following two losses,

$$E_c = \|\mathcal{L}(\mathbf{P}_F, \mathbf{M}) - \mathcal{L}(\mathbf{P} + \sum_{i=1}^{M} \boldsymbol{\beta}_i^{t-1} \mathbf{B}_i, \mathbf{M})\|^2 + \lambda_\delta \|\delta \mathbf{P}\|^2. \quad (2)$$



(a) Input Image (b) Landmark Detection of each parts

Figure 3: Our landmark mark-up consists of 104 points, (i.e., face contour (1-17), eye brows (18-27), left eye (28-47), right eye (48-67), nose (68-84) and mouth (85-104)). (a) Coarse detection of all landmarks and corresponding bounding boxes for fine scale detection. (b) Separate fine-scale detection result of local areas.

The first loss is used to regularize a smooth deformation by maintaining the Laplacian operator $\mathcal{L}$ on the deformed mesh (please refer to [9] for more details). $\boldsymbol{\beta}^{t-i}$ indicates the estimated facial expression blendshape weights from the last iteration and is a fixed value. The second loss is used to enforce a small corrective field and λ_δ is used to balance the two terms.

Other Regularization Terms: We further regularize on facial expression, focal length scale factor, and rotation component of camera extrinsic as follows,

$$E_r = \sum_{i=1}^{M} \frac{\boldsymbol{\beta}_i^2}{\boldsymbol{\sigma}_i^2} + \lambda_f \log^2(f) + \lambda_q \|\mathbf{q}\|^2, \tag{3}$$

where $\boldsymbol{\sigma}$ is the vector of eigenvalues of the facial expression covariance matrix obtained via PCA. λ_f and λ_q are regularization parameters.

Summary: Our total loss function is given by

$$E = E_l + \omega_c E_c + \omega_r E_r, \tag{4}$$

where ω_c and ω_r are used to balance relative importance of the three terms. E is optimized by Gauss-Newton approach over parameters $\mathbf{p}^t$ for a total of N iterations. For the initial parameter vector $\mathbf{p}^0$, $\boldsymbol{\beta}^0$ and $\delta\mathbf{P}^0$ are initialized as all-0 vectors, $\mathbf{t}^0$ and $\mathbf{q}^0$ are estimated from the EPnP approach [7], and f_s^0 is initialized to be 1.

Implementation Details We use a total of $N = 5$ iterations. When minimizing Equation (4), we use $\omega_c = 25$ and $\omega_r = 10$. In Equation (2), we set $\lambda_\delta = 4$, and in Equation (3) we set $\lambda_f = 5$ and $\lambda_q = 5$.

Section 4.4. More Qualitative Results

In this section, we provide more comparison results that cannot be included in the paper due to page limits. For GanFit [4], we have requested them to run the reconstructed results of our test data. Thus, we are only able to show the qualitative comparison with GanFit in our test database. For those images/selfies in the other database, we have compared our results with those papers whose codes are available online including RingNet [8], PRNet [3], Extreme3D [12] and 3DMM-CNN [11].

More Qualitative Results of Our Data In Figure 5, we provided the qualitative results of each categories. The first and second columns are the input image and the ground truth. Instead of showing the cropped mesh, we decided to show the whole models for each method in Figure 5. It is worth noting that our reconstructed full head model is ready to be deployed for different applications.

Qualitative ESRC and JUN-Validate Due to the paper limitation, we are not able to show the qualitative result of ESRC and JUN-validate Dataset. Thus, we provide the visual comparison in Figure 6. The similar results confirmed that proposed method can correctly replicate the 3D models from single selfies with much lower polygon, as we claimed in the paper.

More Qualitative Results of MoFA In Figure 7, we have requested the results from MoFA [10] for side-by-side comparisons. Although the quality of reconstructed models are not as good as the results from other database due to the image resolution, large head pose variation, occlusion such as hair and glasses, our model is still considerably better than other methods.

More Celebrity-In-the-Wild Results In Figure 8 - 9, we present the results of several celebrities and compare our method not only for geometry but also in appearance. Note that by projecting the selfie to a high-resolution UV texture, our reconstructed models has photo-realistic appearance while 3DMM-CNN [11] and PRNet [3] used vertex color results in limited texture reapplication.

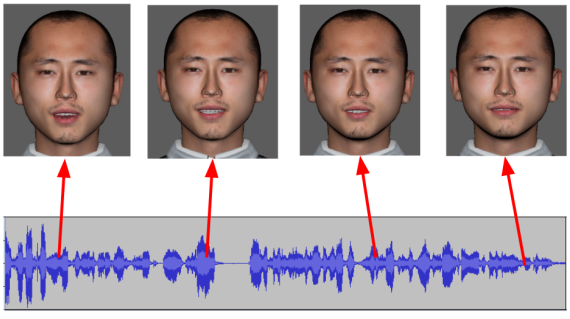


Figure 4: Audio driven lip syncing on our production ready head model

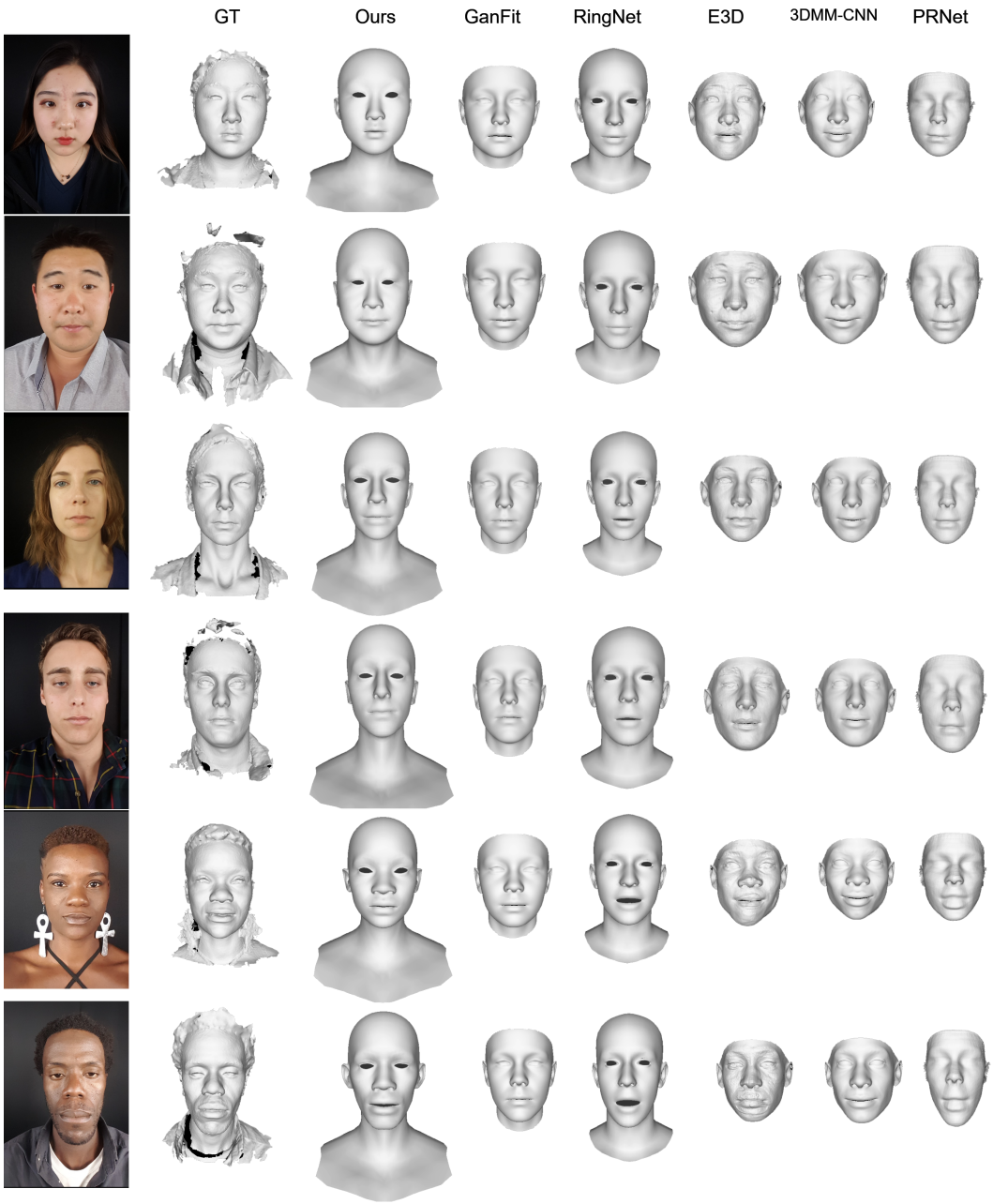


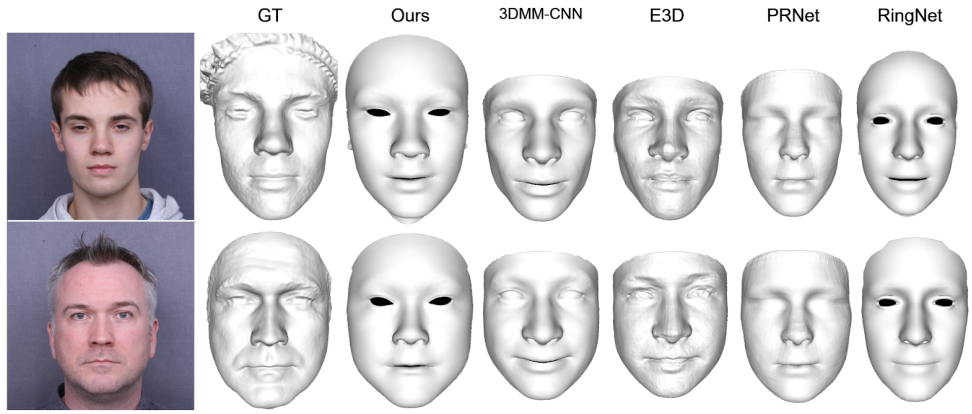
Figure 5: Qualitative results on our test dataset. From left to right, input image, ground truth, our method, GanFit [4], RingNet [8], E3D [12], 3DMM-CNN [11], and PRnet [3].

Application - Audio-driven Avatar Animation

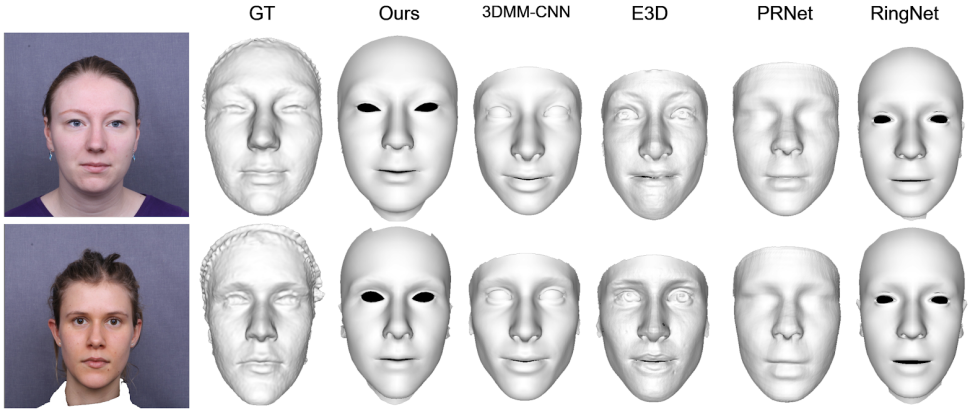
Our automatically generated head model is ready for different applications. Here we demonstrate a case of automatic lip syncing driven by a raw waveform audio input as shown in Figure 4. For data collection and deep neural network structure, we adopt a similar pipeline as that of [5] to drive the reconstructed model. All the animation blendshapes are transferred to our generic topology. Please refer to our video for more details.

References

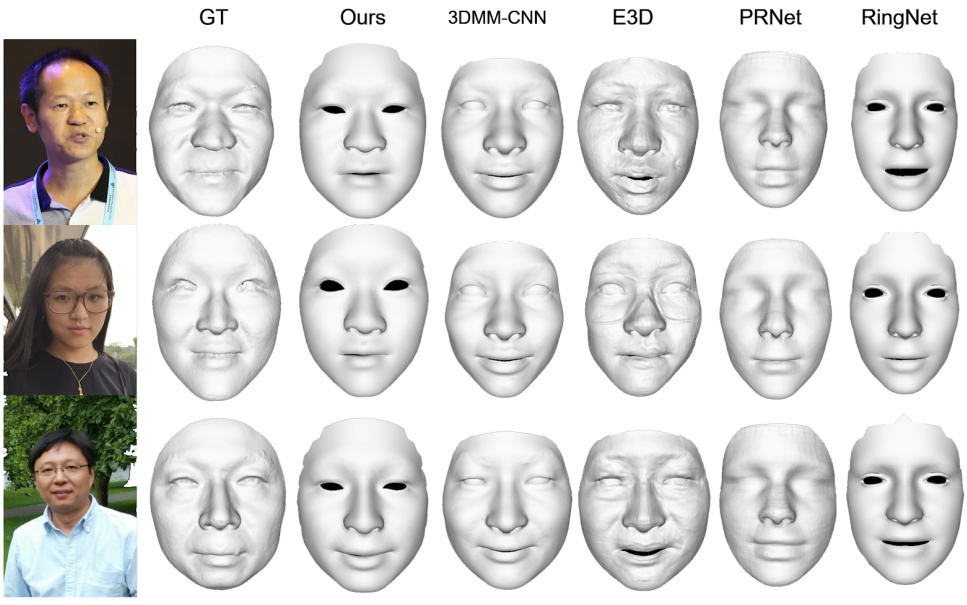
[1] C. Cao, Q. Hou, and K. Zhou. Displaced dynamic expression regression for real-time facial tracking and animation. *ACM Transactions on graphics (TOG)*, 33(4):43, 2014. 1



(a) ESRC Male

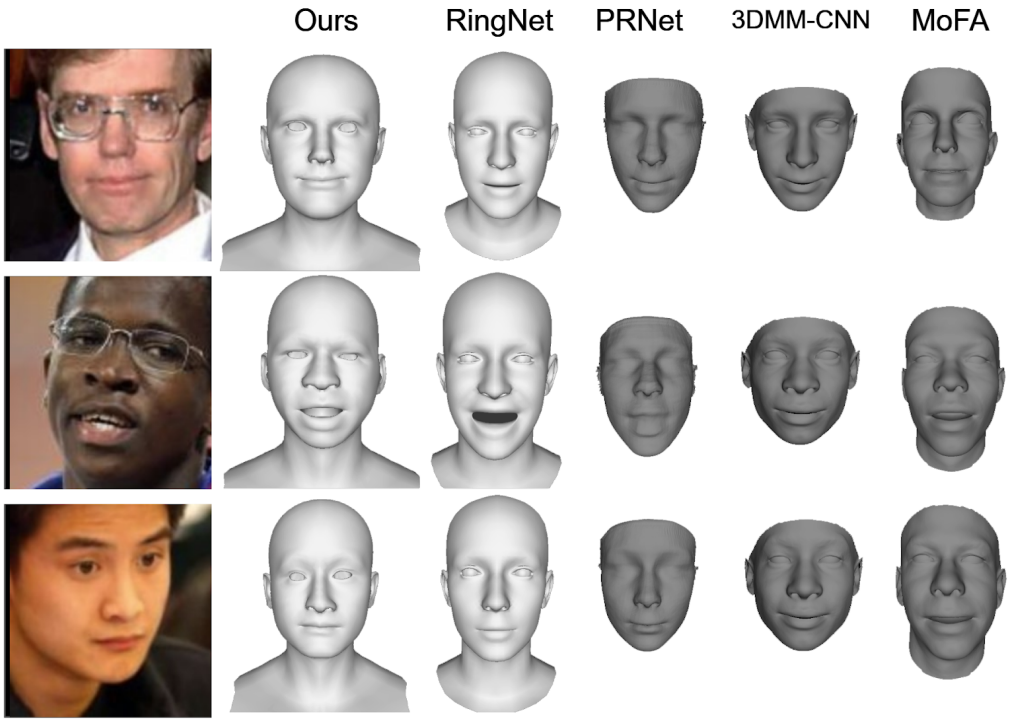


(b) ESRC Female

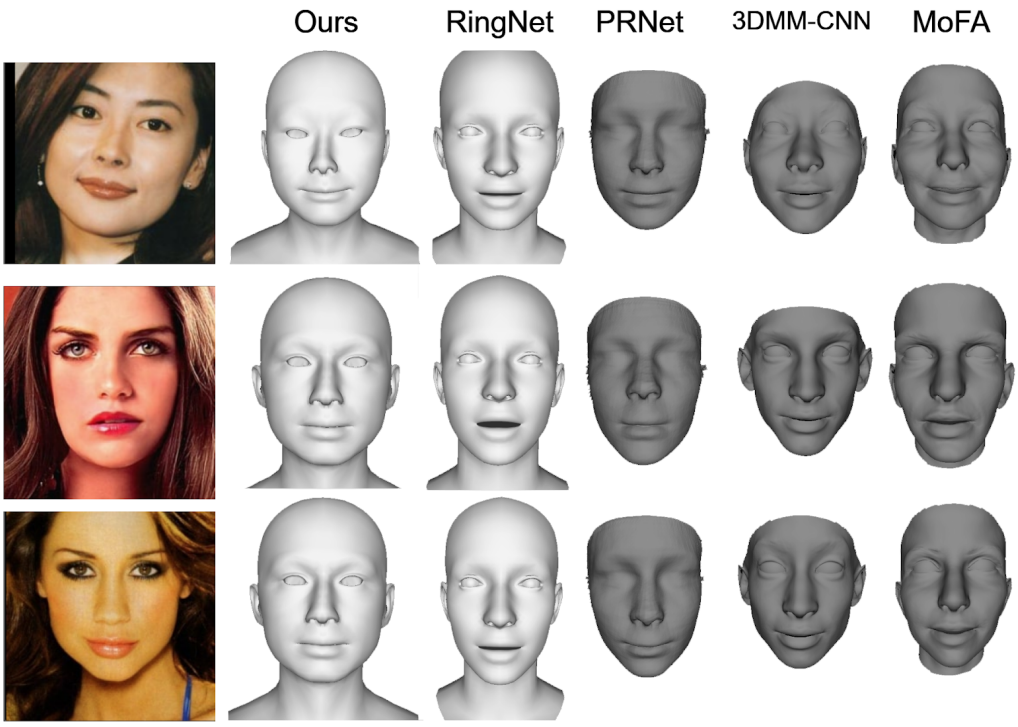


(c) JUN-Validation Database

Figure 6: Qualitative results on the ESRC and JUN-Validation datasets. From left to right, input image, ground truth, our method, 3DMM-CNN[11], E3D [12], PRnet [3], RingNet [8].



(a) MoFA Male



(b) MoFa Female

Figure 7: Qualitative results of MoFa dataset. From left to right, input image, our method, RingNet [8], PRnet [3], 3DMM-CNN[11], and MoFA [10].

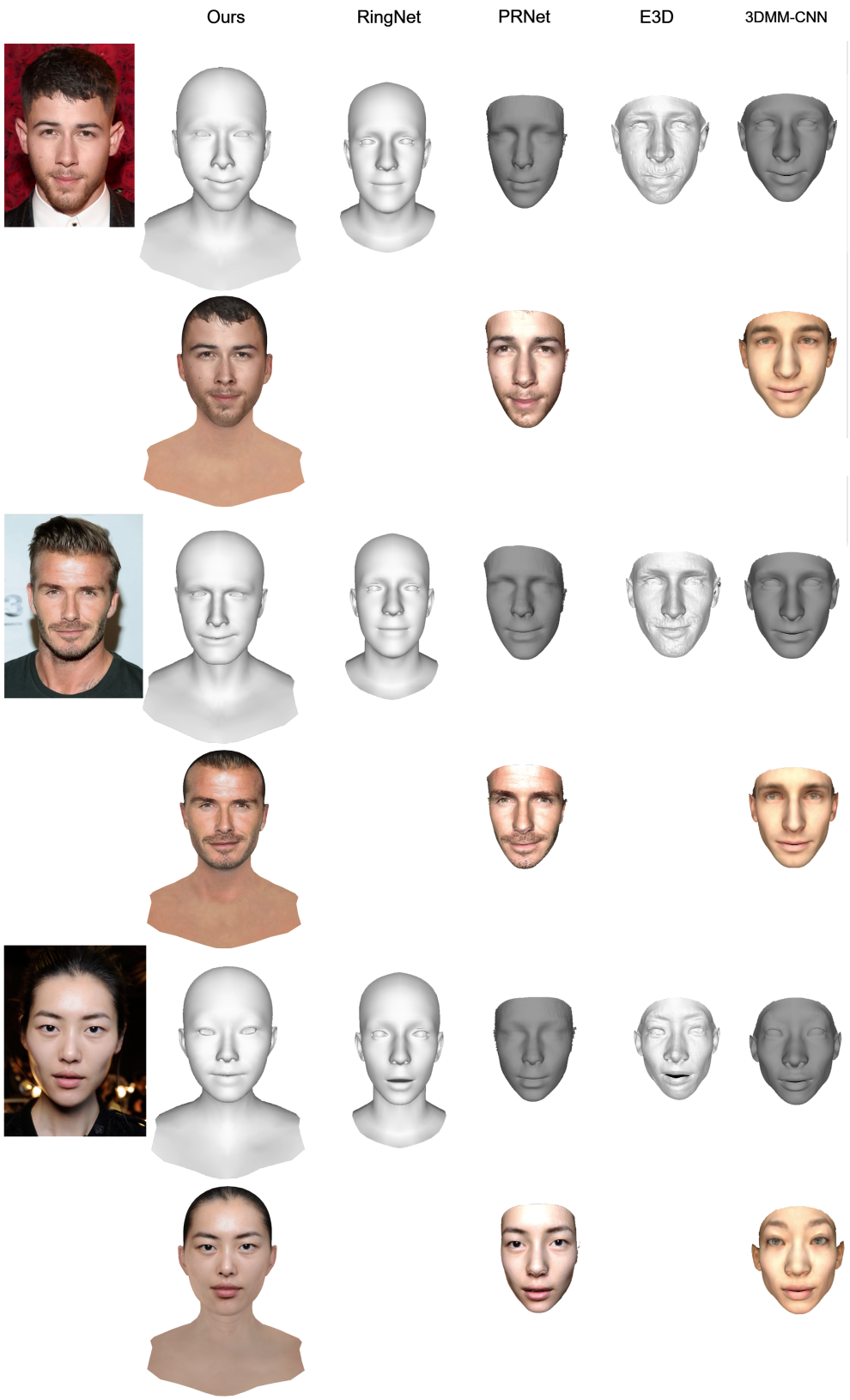


Figure 8: Qualitative results of our method compare to RingNet [8], PRnet [3], E3D [12], and 3DMM-CNN [11].

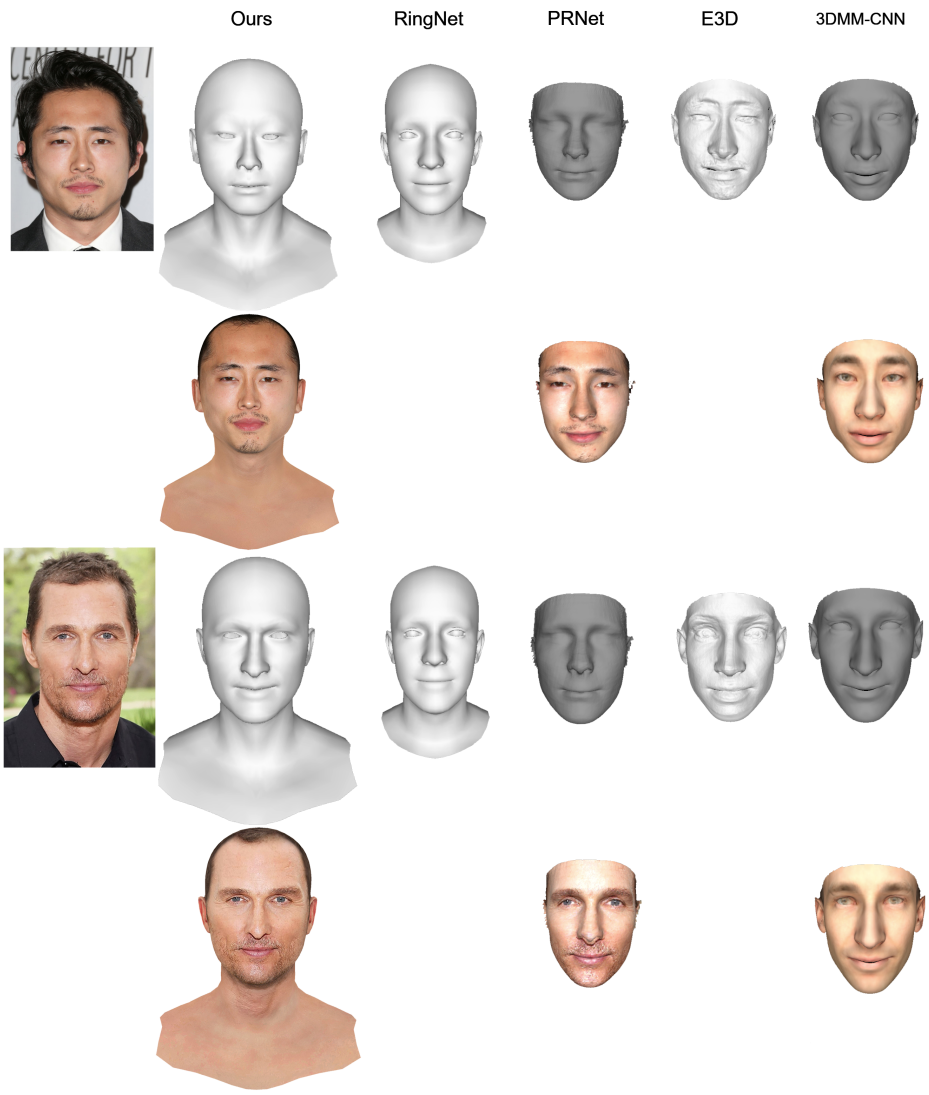


Figure 9: Qualitative results of our method compare to RingNet [8], PRnet [3], E3D [12], and 3DMM-CNN [11].

[2] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, March 2014. 1

[3] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *ECCV*, 2018. 3, 4, 5, 6, 7, 8

[4] B. Gecer, S. Ploumpis, I. Kotsia, and S. Zafeiriou. Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1155–1164, 2019. 3, 4

[5] T. Karras, T. Aila, S. Laine, A. Herva, and J. Lehtinen. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Trans. Graph.*, 36(4):94:1–94:12, July 2017. 4

[6] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1867–1874, 2014. 1

[7] V. Lepetit, F. Moreno-Noguer, and P. Fua. Epnp: An accurate o (n) solution to the pnp problem. *International journal of computer vision*, 81(2):155, 2009. 3

[8] S. Sanyal, T. Bolkart, H. Feng, and M. J. Black. Learning to regress 3d face shape and expression from an image without 3d supervision, 2019. 3, 4, 5, 6, 7, 8

[9] O. Sorkine, D. Cohen-Or, Y. Lipman, M. Alexa, C. Rössl, and H.-P. Seidel. Laplacian surface editing. In *Proceedings of the 2004 Eurographics/ACM SIGGRAPH symposium on Geometry processing*, pages 175–184. ACM, 2004. 3

[10] A. Tewari, M. Zollhöfer, H. Kim, P. Garrido, F. Bernard, P. Pérez, and C. Theobalt. Mofa: Model-based deep convo-

lutional face autoencoder for unsupervised monocular reconstruction. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3735–3744, 2017. 3, 6

[11] A. Tuan Tran, T. Hassner, I. Masi, and G. Medioni. 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*, pages 5163–5172, 2017. 3, 4, 5, 6, 7, 8

[12] A. Tuan Tran, T. Hassner, I. Masi, E. Paz, Y. Nirkin, and G. Medioni. Extreme 3d face reconstruction: Seeing through occlusions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3935–3944, 2018. 3, 4, 5, 7, 8