Learning a 3D Morphable Face Reflectance Model from Low-cost Data
Supplementary Document

Yuxuan Han[1] Zhibo Wang[2] Feng Xu[1]

[1]School of Software and BNRist, Tsinghua University [2]SenseTime Research

A. More Implementation Details

A.1. Multi-PIE Dataset Preprocessing

We select 9 viewpoints (09_0, 08_0, 13_0, 14_0, 05_1, 05_0, 04_1, 19_0, and 20_0) and 11 flashes (03, 04, 05, 06, 07, 08, 09, 10, 11, 14, and 18) for reflectance parameter estimation. Please refer to [11] for the detailed configuration of the viewpoints and flashes. We develop a model-based method to reconstruct the camera parameters and the BFM09 [14] geometry coefficients for each identity. According to the Multi-PIE dataset [11], each selected viewpoint has one selected flash attached to it[1]. Hence, we approximate the flash position as the camera position.

We use the room-light images [11] for reconstruction. Specifically, we first adopt a CNN-based single-view face reconstruction method [5] to obtain the BFM09 coefficients, illumination coefficients, and head pose for each room-light image of a given identity. Then, we apply an offline optimization using the same loss function as [5] to improve the reconstruction accuracy. During the offline optimization, each room-light image shares the same BFM09 coefficients since they are the multi-view images of the given identity, and we initialize them as the average of the coefficients of all views predicted by the face reconstruction CNN. Similar to [5], we use the perspective camera model with a reasonable predefined focal length to represent the 3D-2D projection. After reconstruction, we can compute the camera parameters from the head pose $\mathbf{R}$ and $\mathbf{t}$ for each viewpoint:

$$\mathbf{R}_{cam} = \mathbf{R}^{\mathrm{T}}, \quad \mathbf{t}_{cam} = -\mathbf{R}^{\mathrm{T}} \cdot \mathbf{t} \tag{1}$$

Here, $\mathbf{R}_{cam}$ and $\mathbf{t}_{cam}$ are the camera rotation and translation in the BFM09 canonical space, respectively. We repeat the steps above for all the identities in the Multi-PIE dataset.

Before reflectance parameter estimation, we obtain the OLAT image by removing the effect of the room light in the flash image. Specifically, we subtract the room-light image from the flash image in linear space with a reasonable mapping function[2]:

$$I_{OLAT} = (I_{flash})^{1.2} - (I_{roomlit})^{1.2} \tag{2}$$

Here, I_{OLAT} is the OLAT image in linear space, I_{flash} and $I_{roomlit}$ are the flash and room-light image provided by the Multi-PIE dataset, respectively. We then estimate the reflectance parameters from I_{OLAT} and build our morphable face reflectance model in linear space. To synthesis a face image in nonlinear space, we convert the shading s to pixel color c using the inverse mapping:

$$c = s^{\frac{1}{1.2}} \tag{3}$$

Demographics Our initial morphable face reflectance model is built from a total of 128 manually selected individuals from the Multi-PIE dataset. We release the ID of the selected individuals in our *code repository*.

Feasibility of reflectance parameter estimation The RGB diffuse color and 3 linear combination weights are the only unknowns in our reflectance representation. Theoretically, the ambiguity can be solved with 6 independent equations. We have 99 light-view direction pairs (the combination of 9 viewpoints and 11 light directions) in total, and if considering visibility, most of the vertices have 50+ light-view direction pairs. Different light-view direction pairs give independent equations. Thus, it's feasible to estimate the BRDF parameters theoretically.

Practically, the light-view direction pairs which are not hitting the lobe of the BRDF would lead to a low activation value, and thus solving the reflectance parameters from these equations are highly ill-posed. In our setup, we find that the ill-posed scenario only happens on very few face vertices on the side face or with normal directions going down like nares. For most of the face vertices, our setup can provide enough well-conditioned equations with the corresponding light-view direction pairs hitting the lobe. Thus, it's feasible to estimate the BRDF parameters practically.

[1]We use the viewpoints 08_1 and 19_1 to solve the position of the flashes 14 and 18. However, we do not use the images captured by 08_1 and 19_1 since there exists apparent color inconsistency between these two viewpoints and the other selected 9 viewpoints.

[2]We empirically find that performing image differencing in linear space leads to better reflectance parameter estimation than in non-linear space.

A.2. Model Finetuning

Recall that in model finetuning, the learnable parameters are the morphable model parameters, including the mean $\bar{R}$ and bases $\mathrm{M_R}$, and face reconstruction network parameters θ. We optimize them with the combination of a reconstruction loss $\mathcal{L}_{rec}$ and a regularization loss $\mathcal{L}_{reg}$:

$$\underset{\bar{R},\mathrm{M_R},\theta}{\arg\min}\, \mathcal{L}_{rec} + \mathcal{L}_{reg} \tag{4}$$

$\mathcal{L}_{rec}$ is the combination of a L1 term $\mathcal{L}_{l1}$ and a perceptual term $\mathcal{L}_{per}$:

$$\mathcal{L}_{rec} = \omega_{l1} \cdot \mathcal{L}_{l1} + \omega_{per} \cdot \mathcal{L}_{per}, \text{where} \tag{5}$$

$$\mathcal{L}_{l1} = M_{skin} \cdot ||\hat{I} - I||_1 \tag{6}$$

$$\mathcal{L}_{per} = 1 - \langle \phi_{feat}(\hat{I}), \phi_{feat}(I) \rangle \tag{7}$$

Here, M_{skin} is the mask indicated skin region, obtained by an off-the-shelf face parsing method [24]; $\langle \cdot, \cdot \rangle$ is the inner product operation; ϕ_{feat} is a pretrained FaceNet architecture [17] for feature extraction. Note that we directly compute the reconstruction loss $\mathcal{L}_{rec}$ in the linear space. Although ϕ_{feat} is trained on images in the nonlinear space, we empirically find that it can still provide a reasonable supervision signal if the input image is in the linear space.

In our regularization loss $\mathcal{L}_{reg}$, we first adopt $\mathcal{L}_{coef}$ to constrain the predicted PCA coefficients β and γ:

$$\mathcal{L}_{coef} = \sum_{i=1}^{N_R} (\frac{\beta_i}{\sigma_{\beta_i}})^2 + \sum_{i=1}^{N_L} (\frac{\gamma_i}{\sigma_{\gamma_i}})^2 \tag{8}$$

Here, σ_β and σ_γ are the standard deviations of the initial morphable face reflectance model and the lighting PCA model, respectively. Then, to constrain the updating of our morphable reflectance model, we design $\mathcal{L}_{upd}$ as:

$$\mathcal{L}_{upd} = ||\bar{R} - \bar{R}_0||_1 + ||\mathrm{M_R} - \mathrm{M_{R_0}}||_1 \tag{9}$$

Here, $\bar{R}_0$ and $\mathrm{M_{R_0}}$ are the mean and bases of our initial morphable face reflectance model built from the Multi-PIE dataset. To resolve the color ambiguity between albedo and lighting, we involve $\mathcal{L}_{light}$ to encourage monochromatic environment lighting as [4]:

$$\mathcal{L}_{light} = ||l - l_{mean}||_2^2 \tag{10}$$

Here, l is the retrieved 8-th order SH coefficients; l_{mean} is the mean of l over the color channel dimension, representing the monochromatic counterpart of l. Thus, our regularization loss $\mathcal{L}_{reg}$ can be written as:

$$\mathcal{L}_{reg} = \omega_{coef} \cdot \mathcal{L}_{coef} + \omega_{upd} \cdot \mathcal{L}_{upd} + \omega_{light} \cdot \mathcal{L}_{light} \tag{11}$$

In our experiments, we set $\omega_{l1}, \omega_{per}, \omega_{coef}, \omega_{upd}, \omega_{light}$ to 2, 0.1, 0.001, 10, 10, respectively.

Table 1. Quantative face geometry reconstruction error on the validation set of the NoW challenge.

	Median (mm) ↓	mean (mm) ↓	std (mm) ↓
BFM09	1.44	2.06	2.51
Ours	1.51	2.15	2.61

B. More Results

B.1. Model Visualization

In Figure 2 and Figure 3, we visualize our model by showing random samples drawn from it before and after fine-tuning, respectively. The images are rendered in non-linear space with a white frontal point light.

B.2. Face Reconstruction

More Reconstruction Results We show more face reconstruction results on in-the-wild face images in Figure 1, including diverse ethics groups and challenging cases with facial occlusions and makeups. We multiply the linear combination weights (columns 3, 4 5 in Figure 1) by 3 for better visualization.

Thanks to the model-finetuning process, our method is robust to handle diverse input images and predicts plausible reflectance attributes. However, it has the same limitation as previous in-the-wild face reconstruction methods [5,19,20]: *i)* the global skin tone can not be disentangled from the illumination due to the scale ambiguity between lighting and reflectance (row 5), and *ii)* shadow cast by external geometry (hat in row 9) bakes into the reflectance channels.

Evaluation on Geometry Reconstruction Although our goal is not to better reconstruct face shape from images, we compare our method and BFM09 [14] on the validation set of the NoW challenge [16] to help the readers better understand our model. Note that both methods use the same BFM09 geometry model; we do not compare to AlbedoMM since AlbedoMM [18] is built on top of the BFM17 [10] geometry model.

In this experiment, we adopt a similar network architecture as [5] by simply modifying the number of neurons of the last fully-connect layer of $E_\theta(\cdot)$ from $N_R + N_L + 3$ to $N_S + N_E + N_P + N_R + N_L + 3$ to predict the shape and expression coefficients and the head pose. We use the first 80 and 64 bases of the BFM09 shape and expression morphable model, respectively; thus, $N_S = 80$ and $N_E = 64$. For the head pose, we use the Euler angle to represent rotation and a 3D vector to represent translation; thus, $N_P = 6$. To train the network for geometry reconstruction, we involve a landmark loss term akin to previous works [5,16,21,22]:

$$\mathcal{L}_{ldm} = \sum_{n=1}^{68} ||\hat{q_n} - q_n||_2^2 \tag{12}$$

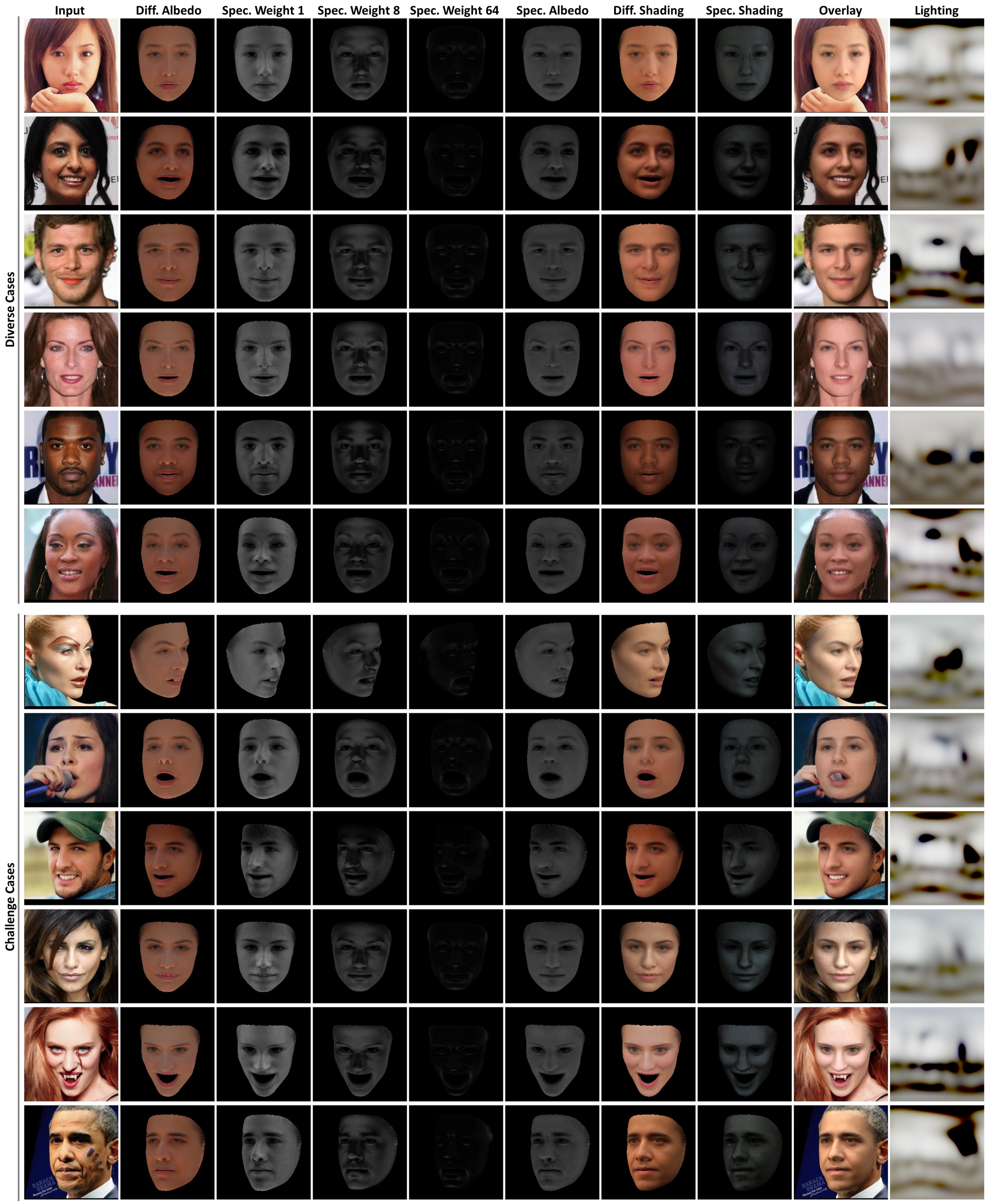

Figure 1. Face reconstruction results on diverse in-the-wild face images.

Figure 2. 60 random samples drawn from our initial morphable face reflectance model (***before model finetuning***). Rendered in nonlinear sRGB space with a white frontal point light.

Here, q_n are the 2D landmarks obtained from an off-the-shelf landmark detector [1]; $\hat{q_n}$ are the 2D projection of the 3D landmarks defined on the reconstructed shape. In addition, we modify $\mathcal{L}_{coef}$ to add constraints on the shape and expression coefficients:

$$\mathcal{L}_{coef} = \sum_{i=1}^{N_S}(\frac{\alpha_i}{\sigma_{\alpha_i}})^2 + \sum_{i=1}^{N_E}(\frac{\delta_i}{\sigma_{\delta_i}})^2 + \sum_{i=1}^{N_R}(\frac{\beta_i}{\sigma_{\beta_i}})^2 + \sum_{i=1}^{N_L}(\frac{\gamma_i}{\sigma_{\gamma_i}})^2 \quad (13)$$

Here, $\alpha \in \mathbb{R}^{N_S}$ and $\delta \in \mathbb{R}^{N_E}$ are the predicted shape and expression coefficients, respectively; σ_α and σ_δ are the standard deviations of the shape and expression morphable model, respectively. Our full loss functions for geometry reconstruction can be written as:

$$\begin{aligned}\mathcal{L} &= \omega_{l1} \cdot \mathcal{L}_{l1} + \omega_{per} \cdot \mathcal{L}_{per} \\ &+ \omega_{coef} \cdot \mathcal{L}_{coef} + \omega_{light} \cdot \mathcal{L}_{light} + \omega_{ldm} \cdot \mathcal{L}_{ldm} \quad (14)\end{aligned}$$

In the geometry reconstruction experiments, we set ω_{l1}, ω_{per}, ω_{coef}, ω_{light}, ω_{ldm} to 2, 0.2, 0.001, 10, 0.002, respectively. We train the geometry reconstruction network on the FFHQ [12] dataset for 20 epochs.

As shown in Table 1, our method just obtains similar quantitative results compared to the BFM09 under the same CNN-based face geometry reconstruction pipeline. However, we believe that our model has the potential to achieve better geometry reconstruction results with the advance of lighting estimation and differentiable ray tracer.

B.3. Face Relighting and OLAT Rendering

See our *project page* for the video results.

C. Limitations and Discussions

Our method still has several limitations. We adopt the Lambertian BRDF to represent diffuse reflectance. Thus, we cannot model the subsurface scattering effect. Integrating a more complicated reflectance representation [23] into our morphable face reflectance model to improve face rendering realism is an interesting direction.

Our model cannot well represent the specularities around the eyes. We try a straightforward way by adding more mirror-like specular terms in our reflectance representation but find it does not work. We attribute this to the following two reasons: *i)* the reconstructed geometry is inaccurate around eyes during inverse rendering, and *ii)* our BRDF reflectance representation cannot well model the complex properties of eyes (e.g. refraction).

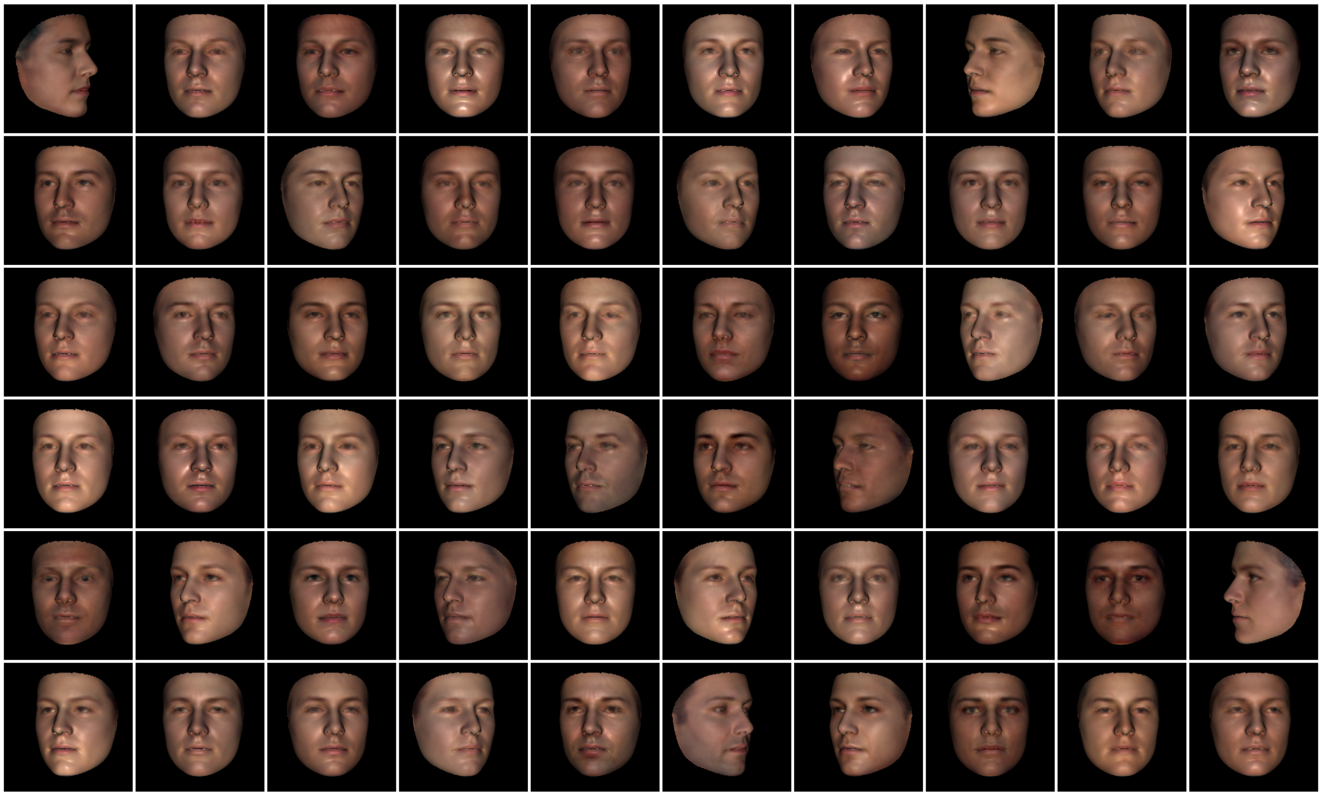

Figure 3. 60 random samples drawn from our final morphable face reflectance model (***after model finetuning***). Rendered in nonlinear sRGB space with a white frontal point light.

During model finetuning, we use a differentiable rasterizer with an efficient local shading technique to render the reconstructed image, without considering global illumination effects like self-shadowing, considering that the illumination is soft, and the self-shadows are insignificant in most in-the-wild images. We believe that using a differentiable ray tracer [13] would slightly improve the current results as demonstrated in existing works [6–8]. Moreover, leveraging a multi-view in-the-wild face image dataset [2] or video dataset [3] could improve the face reconstruction results, as demonstrated by the previous works [9, 19]. We leave these as our future works.

In addition, there is an inevitably global scale between the reflectance parameters in our model and the ground truth since the low-cost data does not provide lighting information [15].

References

[1] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *International Conference on Computer Vision*, 2017. 4

[2] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018. 5

[3] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018. 5

[4] Yu Deng. Deep3dfacereconstruction, 2019. https://github.com/microsoft/Deep3DFaceReconstruction. 2

[5] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 1, 2

[6] Abdallah Dib, Junghyun Ahn, Cedric Thebault, Philippe-Henri Gosselin, and Louis Chevallier. S2f2: Self-supervised high fidelity face reconstruction from monocular image. *arXiv preprint arXiv:2203.07732*, 2022. 5

[7] Abdallah Dib, Gaurav Bharaj, Junghyun Ahn, Cédric Thébault, Philippe Gosselin, Marco Romeo, and Louis Chevallier. Practical face reconstruction via differentiable ray tracing. In *Computer Graphics Forum*, volume 40, pages 153–164. Wiley Online Library, 2021. 5

[8] Abdallah Dib, Cédric Thébault, Junghyun Ahn, Philippe-Henri Gosselin, Christian Theobalt, and Louis Chevallier. Towards high fidelity monocular face reconstruction with rich reflectance using self-supervised learning and ray tracing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12819–12829, 2021. 5

[9] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (ToG)*, 40(4):1–13, 2021. 5

[10] Thomas Gerig, Andreas Morel-Forster, Clemens Blumer, Bernhard Egger, Marcel Luthi, Sandro Schönborn, and Thomas Vetter. Morphable face models-an open framework. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 75–82. IEEE, 2018. 2

[11] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and vision computing*, 28(5):807–813, 2010. 1

[12] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4396–4405, 2018. 4

[13] Tzu-Mao Li, Miika Aittala, Frédo Durand, and Jaakko Lehtinen. Differentiable monte carlo ray tracing through edge sampling. *ACM Trans. Graph. (Proc. SIGGRAPH Asia)*, 37(6):222:1–222:11, 2018. 5

[14] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 296–301, 2009. 1, 2

[15] Ravi Ramamoorthi and Pat Hanrahan. A signal-processing framework for inverse rendering. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 117–128, 2001. 5

[16] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael Black. Learning to regress 3D face shape and expression from an image without 3D supervision. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 7763–7772, June 2019. 2

[17] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015. 2

[18] William AP Smith, Alassane Seck, Hannah Dee, Bernard Tiddeman, Joshua B Tenenbaum, and Bernhard Egger. A morphable face albedo model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5011–5020, 2020. 2

[19] Ayush Tewari, Florian Bernard, Pablo Garrido, Gaurav Bharaj, Mohamed Elgharib, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. Fml: Face model learning from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10812–10822, 2019. 2, 5

[20] Ayush Tewari, Hans-Peter Seidel, Mohamed Elgharib, Christian Theobalt, et al. Learning complete 3d morphable face models from images and videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3361–3371, 2021. 2

[21] Ayush Tewari, Michael Zollhöfer, Pablo Garrido, Florian Bernard, Hyeongwoo Kim, Patrick Pérez, and Christian Theobalt. 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*, pages 2549–2559, 2018. 2

[22] Ayush Tewari, Michael Zollhofer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian Theobalt. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1274–1283, 2017. 2

[23] Tim Weyrich, Wojciech Matusik, Hanspeter Pfister, Bernd Bickel, Craig Donner, Chien Tu, Janet McAndless, Jinho Lee, Addy Ngan, Henrik Wann Jensen, et al. Analysis of human faces using a measurement-based skin reflectance model. *ACM Transactions on Graphics (ToG)*, 25(3):1013–1024, 2006. 4

[24] Changqian Yu, Changxin Gao, Jingbo Wang, Gang Yu, Chunhua Shen, and Nong Sang. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *International Journal of Computer Vision*, 129(11):3051–3068, 2021. 2