Supplementary Material:
3DMM-RF: Convolutional Radiance Fields for 3D Face Modeling

Stathis Galanakis[1,2] Baris Gecer[2] Alexandros Lattas [1,2]
Stefanos Zafeiriou[1,2]
[1]Imperial College London {s.galanakis21,a.lattas,s.zafeiriou}@imperial.ac.uk
[2]Huawei baris.gecer@huawei.com

1. Implementation details

1.1. Training

Our codebase is based on StyleGan2[6] and more specifically, we use a Pytorch implemantationt[1]for both the synthesis and the discriminator network. We also use Pytorch3D[10] to render our synthetic face dataset.

The loss function of our generator is:

$$L = L_{gen} + L_{pht} + L_{per} + L_{ID} + L_{gen} + L_{mean} + L_{kps} \quad (1)$$

where L_{gen} is the loss of the generator, L_{pht} is the photometry loss, L_{per} perceptual loss, L_{ID} the identity loss, L_{gen} the generator loss, L_{mead} the depth loss and L_{kps} the landmark loss. We use non saturating logistic loss [5] with R_1 regularization [8] as generator loss, the perceptual similarity loss [14] as perceptual loss and $L2$-norm as photometry loss. Also, we use a facial landmark detector [1] to compare the predicted facial landmarks between the groundtruth images and the generated ones [3], whereas we use a state-of-the-art face recognition network [2] for comparing the identities between the generated and the real images via comparing their facial feature vectors.

1.2. Fitting

During the fitting step, for a given image $\hat{\mathbf{I}}$ containing a face, we firstly crop the image so that the face is in the centre of the image. Even though, our fitting approach works well without the aforementioned step, we empirically noticed that our network performs better and diverges faster when it gets as input the cropped input image. Secondly, from the image $\hat{\mathbf{I}}$, we extract important facial information such as the identity vector $\hat{\mathbf{z}}_{ID}$, using a face recognition network [2] and the 2D facial landmarks using a facial landmark detection network [1]. We initialize the identity input vector $\mathbf{z}_{ID}$ equal with the $\hat{\mathbf{z}}_{ID}$, the expression input vector $\mathbf{z}_{exp}$ equal with the zero vector, the camera pose vector $\mathbf{z}_{cam}$ equal with the frontal view vector, whereas the illumination vector $\mathbf{z}_{ill}$ is initialized as the average illumination vector. Given the parameters $\mathbf{z} = \{\mathbf{z}_{ID}, \mathbf{z}_{exp}, \mathbf{z}_{cam}\mathbf{z}_{ill}\}$, our network $\mathcal{S}$ outputs a rendered image $\mathbf{I}$, in which we apply a face mask based on the depth produced by the depth channels, and we get a masked image $\mathbf{I}_{masked}$. We also initialize a trainable similarity matrix $\mathcal{M}$ as a vector for overlaying the reconstructed masked image I_{masked} with the input image $\hat{\mathbf{I}}$ and the overlaid image $\mathbf{I}_{over}$ is compared with the original input image $\hat{\mathbf{I}}$ via the aforementioned loss functions.

2. Additional Qualitative Results

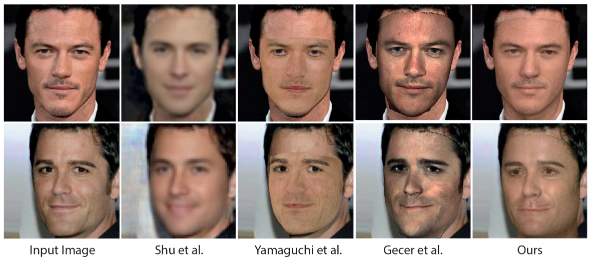


Figure 1: A qualitative comparison between our method (3DMM-RF) and [13, 11, 4]. Our approach can render high quality images whereas retrieve accurate identity, camera pose and illumination through fitting.

3DMM-RF is capable of rendering high quality images including a great variety of output subjects. In addition to the experiments in the main manuscript, in Figure 1, we compare our approach with the fitting results obtained from other state-of-the-art methods [13, 11, 4]. 3DMM-RF recreate authentic details, which are more accurate and sharper when compared with [13, 11] and more photorealistic than [4].

Moreover, in Figures 7 and 8, we showcase additional fittings, which demonstrate that our network has the ability

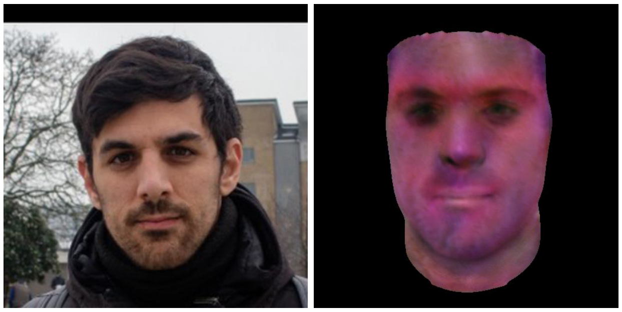
---

[1] https://github.com/lucidrains/stylegan2-pytorch

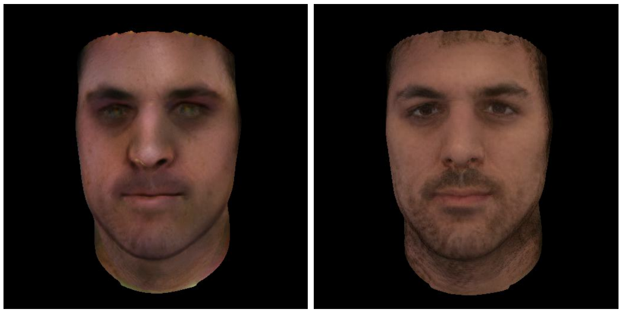
to be applied to a wide range of demographics, including a range of ethnicities and genders.

Finally, we attach a video to the supplementary materials (`764.mp4`), which shows interpolations of latent identity, pose, expression and illumination values. It displays our method's ability to disentangle the identity vector $\mathbf{z}_{ID}$, the expression vector $\mathbf{z}_{exp}$, the scene illumination vector $\mathbf{z}_{ill}$ and the camera pose vector $\mathbf{z}_{cam}$.

3. Volumetric Rendering Ablation



(a) The input image (b) Random Slice

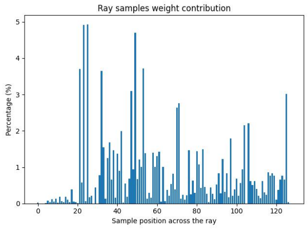


(c) Only max opacity samples (d) Ours

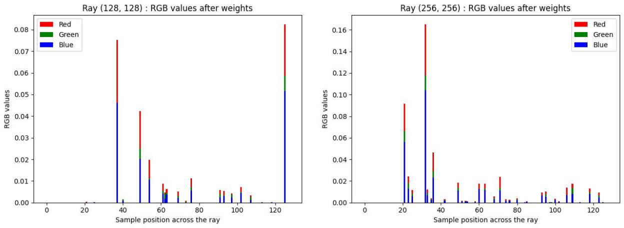
Figure 2: A qualitative comparison between different rendering approaches for reproducing the input image (Fig 2a). Fig 2b includes the results when we take a random slice of the output radiance field provided that all weights are 1 in this slice only, Fig 2c shows the output when taking only the samples with the biggest opacity across each ray, whereas Fig 2d shows our method's output.

We want to confirm that the rendered images are produced using several meaningful volumetric samples across a ray, and not just one slice of the output radiance field. To do so, we perform the following ablation studies.

Figure 2 contains the results of our first ablation study. We plot several slices of the radiance field, after assigning as 1 to their weight values, to make sure that no slice contains the whole image (Fig. 2b contains an example of these slices). Secondly, for each ray, we only consider the RGB values contained in the sample having the maximum opacity value. The rendered image based on this method is shown Fig. 2c, whereas Fig. 2d includes our method's output. Comparing those figures, it is clearly shown that the final rendered image requires more than one sample per ray to efficiently reconstruct the input image.



(a) Mean weight contribution across ray samples' position



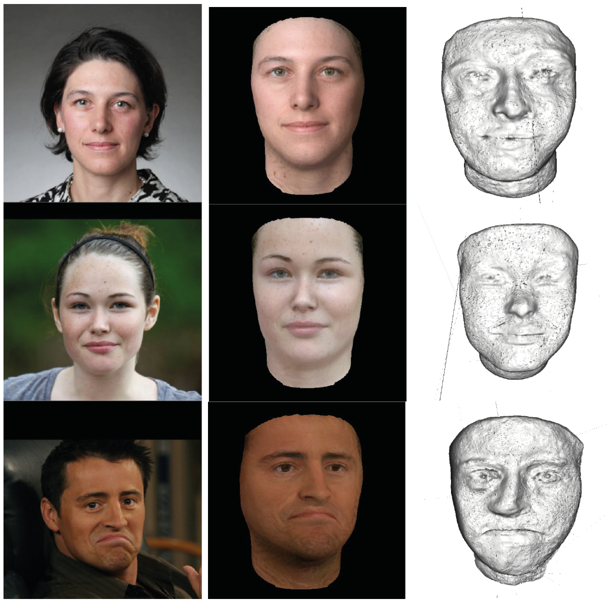
(b) Pixel (128,128) (c) Pixel (256,256)

Figure 3: Fig. 3a : The weight contribution across the rays considering only the foreground pixels. Fig. 3b & Fig.3c : For pixels (128,128), (256,256) the figure depicts the RGB values for each sample across the ray after having been multiplied by their corresponding opacity weights

On the other hand, we perform another study for confirming that various samples of the ray are contributing to the rendered image and the results are portrayed in Fig. 3. Fig. 3a portrays the mean weight contribution of the samples according to their position across the ray. It is clear depicted that their is no position across the rays that our model learns to render the final RGB value only based on that. On the other hand, Fig. 3b&3c contain the RGB values across two random rays multiplied by the opacity weights. As these figures show, the final rendered RGB value depends on the contribution of several different parts of the ray.

4. Depth estimation

Given a ray r, our synthesis network predicts an estimated ray-facial surface intersection depth D_r. Based on this depth prediction, we can extract a facial mesh by using the marching cubes algorithm [7]. Fig. 4 depicts the results of this algorithm after our network was fitted to the images in Fig. 4a. The reconstructed images are presented in Fig. 4b. Fig.4c clearly shows that we can extract a facial



(a) Input image (b) Reconstructed face (c) Mesh reconstruction

Figure 4: This figure portrays the 3D face reconstruction acquired after applying the marching cubes [7] to the predicted facial depth.

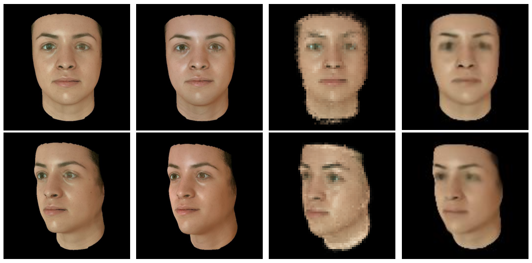
mesh after applying the marching cubes algorithms.

5. Comparison against NeRF-based Methods

We compare our method with NeRF-based methods, in order to provide insight into the comparative quality of our model's rendering, as well as with our advantages with regards to the training and rendering time. We acquire a random unseen subject, similar to the ones used in our training dataset, and generate 40 renderings of various poses, plus a frontal and side one used for testing. We compare our method against NeRF [9] and NeuS [12], using their official implementations and default settings. For NeuS, we also provide masks. Then, we fit our model on 1, instead of 40 images and compare our results in Tab. 1 and Fig. 5. As can be seen, our method outperforms both, in terms of quality, training and testing speed.

6. Failure Cases

Fig. 6 shows some common failure cases occurred during our experiments. Such cases include severe facial occlusions (e.g. glasses and hair), strong shadows and illumination, and extreme poses, which inhibit both the facial landmark annotation and the loss functions we use.



(a) Real (b) Ours (c) NeRF[9] (d) NeuS[12]

Figure 5: Qualitative comparison with NeRF-based methods: From left to right: a) frontal test image of the training dataset, b) our frontal rendering based on a single-image fitting, c) frontal rendering of NeRF[9] trained on 40 images and d) frontal rendering of NeuS[12] trained on 40 images. Please note, NeRF and NeuS are trained in low-resolution given the rebuttal's short period, and will be replaced with higher resolution training.

Method	Images	Fitting t (min)	Render t (sec)
Ours	$\mathbf{1}$	$\mathbf{5.5}$	$\mathbf{0.085}$
NeRF [9]	40	858	9.5
NeuS [12]	40	660	7

Table 1: Quantitative comparison between our method, NeRF[9] and NeuS[12], based on the image required to fit the network, fitting time (minutes) and rendering time (seconds) for a single frame.

References

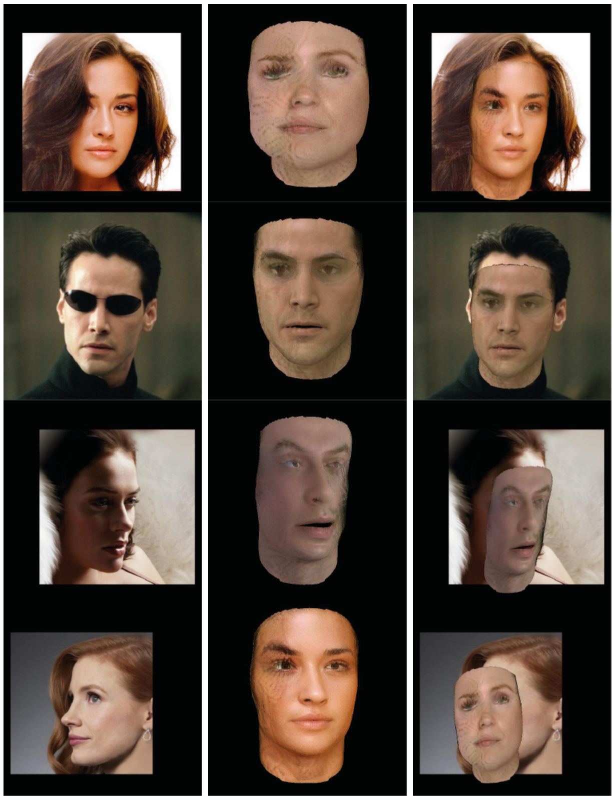
[1] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *International Conference on Computer Vision*, 2017.

[2] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June:4685–4694, 2019.

[3] Baris Gecer, Jiankang Deng, and Stefanos Zafeiriou. Ostec: One-shot texture completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7628–7638, June 2021.

[4] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. GANFIT: Generative Adversarial Network Fitting for High Fidelity 3D Face Reconstruction. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1155–1164, Long Beach, CA, USA, June 2019. IEEE.

[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Wein-



(a) Input image (b) Our prediction (c) Ours overlaid

Figure 6: Examples of common failure cases.

berger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[6] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proc. CVPR*, 2020.

[7] William E. Lorensen and Harvey E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. In Maureen C. Stone, editor, *SIGGRAPH*, pages 163–169. ACM, 1987.

[8] Lars M. Mescheder. On the convergence properties of GAN training. *CoRR*, abs/1801.04406, 2018.

[9] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, pages 405–421. Springer, 2020.

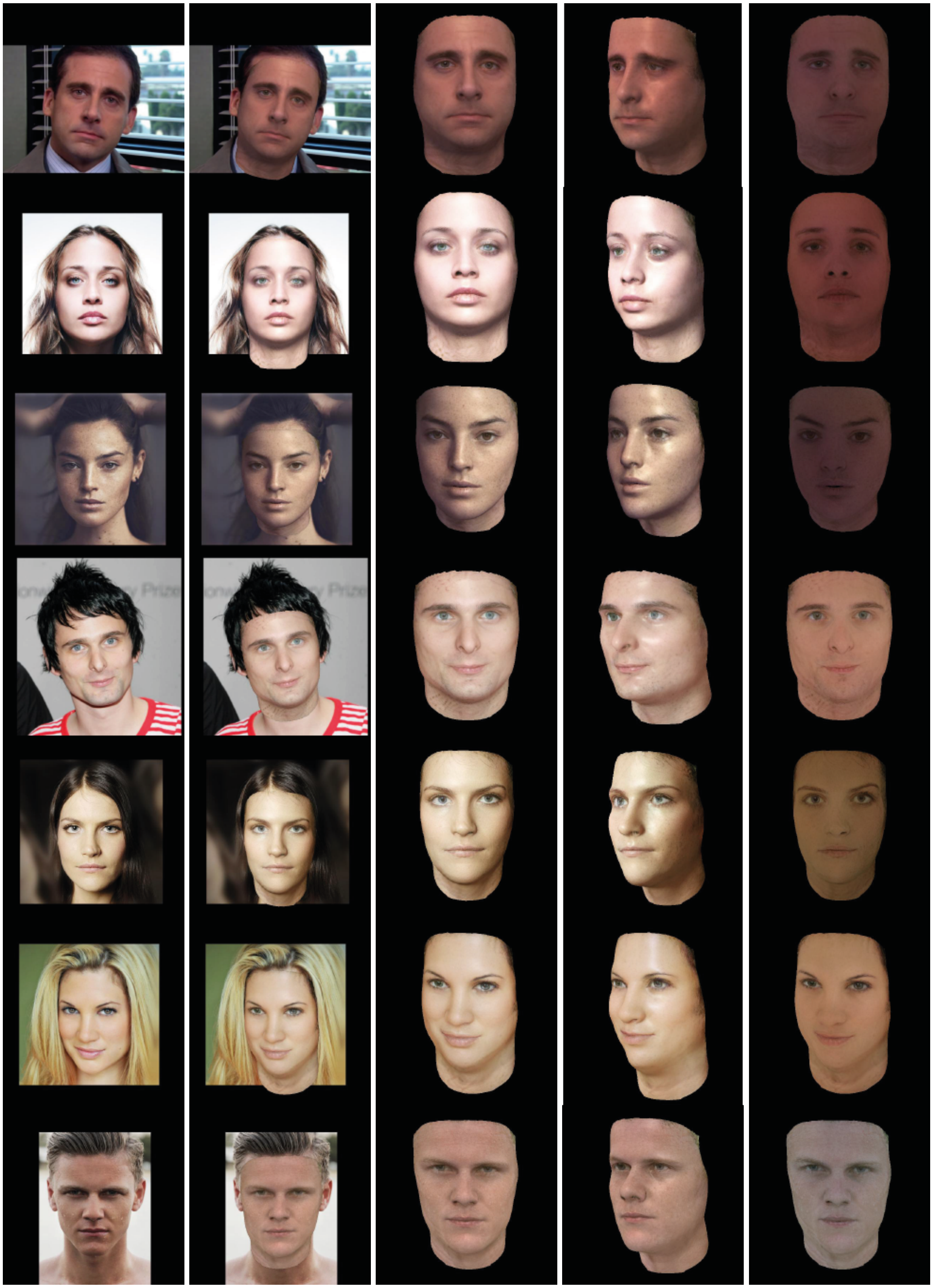
[10] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020.

[11] Zhixin Shu, Ersin Yumer, Sunil Hadap, Kalyan Sunkavalli, Eli Shechtman, and Dimitris Samaras. Neural face editing with intrinsic image disentangling. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5444–5453, 2017.

[12] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *Advances in Neural Information Processing Systems*, 34:27171–27183, 2021.

[13] Shugo Yamaguchi, Shunsuke Saito, Koki Nagano, Yajie Zhao, Weikai Chen, Kyle Olszewski, Shigeo Morishima, and Hao Li. High-fidelity facial reflectance and geometry inference from an unconstrained image. *ACM Trans. Graph.*, 37(4), jul 2018.

[14] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.



(a) Source Image

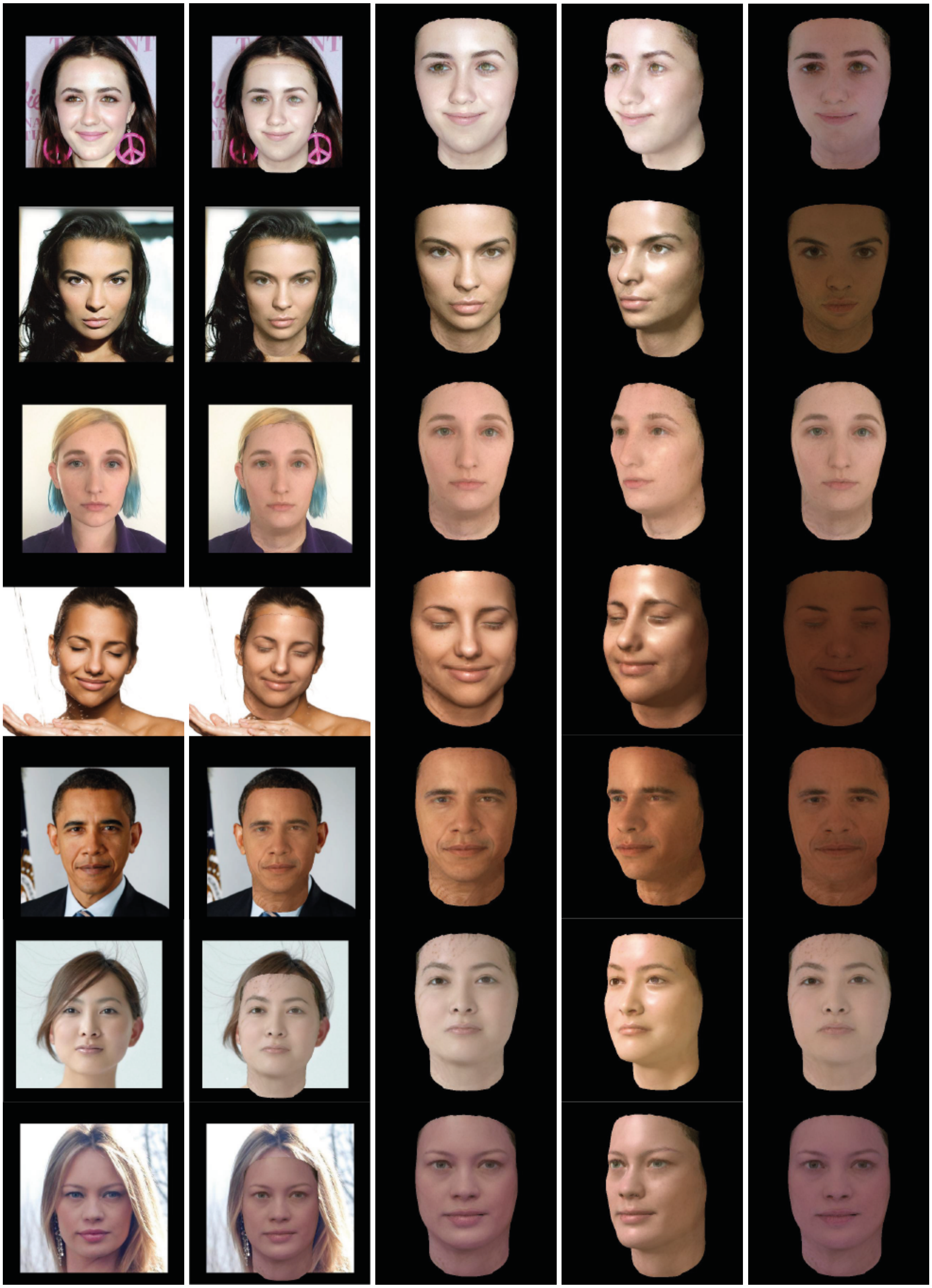
(b) Ours overlaid

(c) Ours prediction

(d) Ours Novel Angle

(e) Ours diffuse albedo

Figure 7: Our network is capable of being fitted in a great variety of ”in-the-wild” images including diverse subjects and produce high fidelity images.



(a) Source Image

(b) Ours overlaid

(c) Ours prediction

(d) Ours Novel Angle

(e) Ours diffuse albedo

Figure 8: Another example of 3DMM-RF 's ability to perform well and render high quality images.