MoRF: Morphable Radiance Fields for Multiview Neural Head Modeling (Supplementary Material)

DAOYE WANG, DisneyResearch|Studios, Switzerland and ETH Zurich, Switzerland
PRASHANTH CHANDRAN, DisneyResearch|Studios, Switzerland and ETH Zurich, Switzerland
GASPARD ZOSS, DisneyResearch|Studios, Switzerland
DEREK BRADLEY, DisneyResearch|Studios, Switzerland
PAULO GOTARDO, DisneyResearch|Studios, Switzerland

CCS Concepts: • **Computing methodologies** → **Rendering**; **Neural networks**; **Volumetric models**.

ACM Reference Format:
Daoye Wang, Prashanth Chandran, Gaspard Zoss, Derek Bradley, and Paulo Gotardo. 2022. MoRF: Morphable Radiance Fields for Multiview Neural Head Modeling (Supplementary Material). In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings (SIGGRAPH '22 Conference Proceedings), August 7–11, 2022, Vancouver, BC, Canada.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3528233.3530753

1 METHOD DETAILS

1.1 Architecture Details

The detailed architectures of the three main MLPs in MoRF are illustrated in Fig. 1. The design of our deformation MLPs is similar to the one in [Park et al. 2021], while the canonical NeRF is designed after [Mildenhall et al. 2020], except for the addition of the input canonical identity code and another output branch for the separate specular distribution. To model this distribution smoothly, we use a K-th order spherical harmonics representation, with $K = 3$ (9 coefficients). For the identity network, we trained with D-dimensional **id** codes where we set $D = 4$, and the output $\mathbf{id}_\mathrm{w}$ and $\mathbf{id}_\mathrm{c}$ have 128 dimensions. Position encoding is done with 8 frequency bands.

1.2 Deformation Details

Following Park et al. [2021], we use the well-known axis-angle rotation representation and Rodrigues' formula to deform a point $\mathbf{x}$ to $\mathbf{x}'$ given translation $\mathbf{t}$, rotation axis $\mathbf{r}$ and rotation angle θ. Let $\mathbf{x}' = \mathbf{x}_\mathrm{rot} + p$, where

$$\mathbf{x}_\mathrm{rot} = \mathbf{x}\cos(\theta) + (\mathbf{r}\times\mathbf{x})\sin(\theta) + \mathbf{r}(\mathbf{r}\cdot\mathbf{x})(1-\cos(\theta)) \quad (1)$$

$$p = \mathbf{t}\sin(\theta) + (\mathbf{r}\times\mathbf{t})(1-\cos(\theta)) + \mathbf{r}(\mathbf{r}\cdot\mathbf{t})(\theta-\sin(\theta)) \quad (2)$$

1.3 Fitting MoRF to a New Subject

As mentioned in the main text, during fitting, we first quickly fit an **id** to initialize a pair of $\mathbf{id}_\mathrm{w}$ and $\mathbf{id}_\mathrm{c}$ codes, which are then optimized further in their own latent subspaces. We fit **id** for 100 iterations, and we fit $\mathbf{id}_\mathrm{w}$ and $\mathbf{id}_\mathrm{c}$ for 1000 iterations. In this fitting stage, we freeze the network weights and only update the latent codes. Optimization on a single NVidia GTX3090 takes less than a minute for **id**, and around 10 minutes for $\mathbf{id}_\mathrm{w}$ and $\mathbf{id}_\mathrm{c}$.

As we don't expect a pre-trained MoRF to faithfully represent image detail such as particular identity features that are unique to an arbitrary person not seen during training, this fitting stage is applied to input images that are downscaled by a factor of 6, and only looks at coarse rendering (we resize the camera frame accordingly). To obtain more robust fits to images of novel subjects, whose pixels may correspond to "outliers" that the pre-trained model cannot represent, we also adapt our rendering loss $\mathcal{L}_{RGB}$ (main text) and replace its L2-norm with the L1-norm. Additionally, following established practice when fitting StyleGAN models, we also apply a VGG-16 perceptual loss $\mathcal{L}_{LPIPS}$ as done in [Abdal et al. 2019],

$$\mathcal{L}_{LPIPS} = \lambda_P \sum_c \left\| \Phi(\mathcal{I}'_c) - \Phi\Big(\sum_i \omega_{pi}\big(\mathbf{c}_{dpi} + \mathbf{c}_{spi}^T \mathrm{SH}(\mathbf{r})\mathbf{1}_{[c\notin C_{\mathcal{X}^\mathcal{P}}]}\big)\Big)\right\|_F, \quad (3)$$

which is similar to our $\mathcal{L}_{RGB}$ loss in the main text. Here, $\lambda_P = 0.05$ and we first render the complete image before applying the perceptual loss: $\Phi(\cdot)$ denotes the set of feature activations from layers *conv1-1*, *conv1-2*, *conv3-3* of a pre-trained VGG-16 network; and $||\cdot||_F$ is the Frobenious norm. If the input includes a registered 3D mesh, we also apply our deformation loss $\mathcal{L}_{DF}$ (main text), unmodified.

We also enforce $\mathcal{L}_{ID}$ in the initial stage. Then, when fitting $\mathbf{id}_\mathrm{w}$ and $\mathbf{id}_\mathrm{c}$, we modify this loss slightly to

$$\mathcal{L}'_{ID} = \lambda_{ID}\Big(\big\|\mathbf{id}_\mathrm{w} - \mathbf{id}_\mathrm{w}^0\big\|_2^2 + \big\|\mathbf{id}_\mathrm{c} - \mathbf{id}_\mathrm{c}^0\big\|_2^2 \Big). \quad (4)$$

Above, $[\mathbf{id}_\mathrm{w}^0, \mathbf{id}_\mathrm{c}^0] = ID(\mathbf{id}^0)$ are the codes initially predicted from the optimized $\mathbf{id}^0$ code fed into MoRF's ID network.

1.4 Tuning MoRF to a New Subject

Our network tuning is a simplification of the method in [Roich et al. 2021]. After fitting $\mathbf{id}_\mathrm{w}$ and $\mathbf{id}_\mathrm{c}$ to the unseen subject, we can freeze these codes and further optimize the network weights to obtain more faithful fits. We fine-tune the pre-trained MoRF MLPs for another 2000 iterations (1% of the total iterations in training a standard NeRF for a given subject). On a single NVidia GTX3090, this fitting stage takes about 30 minutes (less than 5% of the total time to train a standard NeRF for this subject from scratch).

For faithfully fitting, this stage operates on input images in their original resolution and uses our unmodified $\mathcal{L}_{RGB}$, $\mathcal{L}_{DF}$, and matting loss (on density), as in the main text. Here, we do not use the depth-map component of the density loss. Empirically, we obtained

SIGGRAPH '22 Conference Proceedings, August 7–11, 2022, Vancouver, BC, Canada

ACM ISBN 978-1-4503-9337-9/22/08…$15.00
https://doi.org/10.1145/3528233.3530753

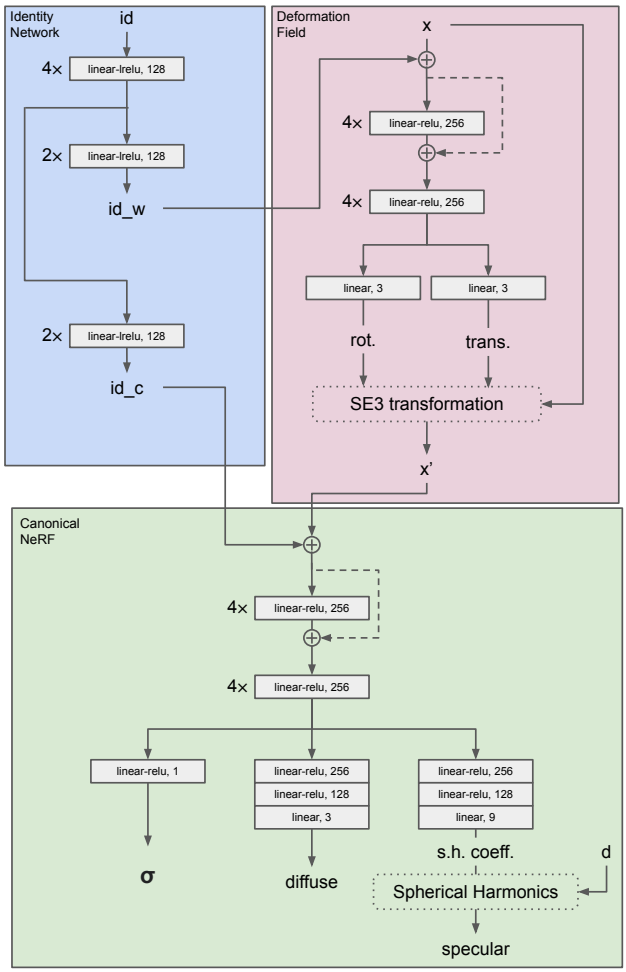

Fig. 1. Architecture detail of the three main MLPs in MoRF.

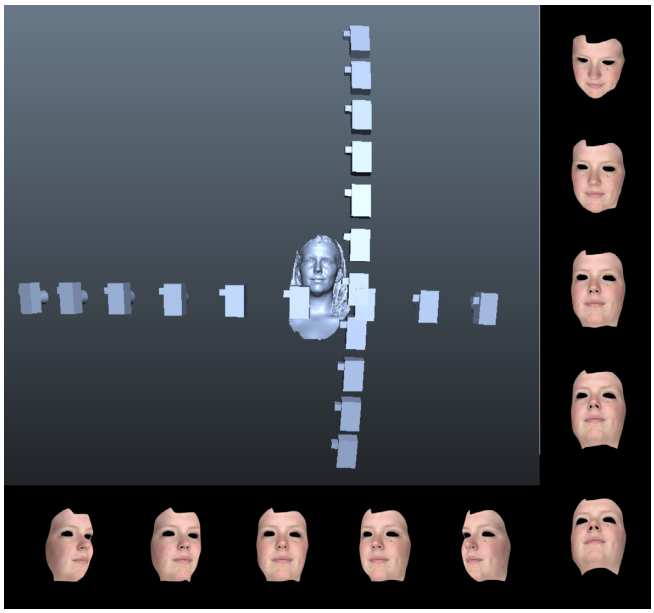

Fig. 2. Additional synthetic views generated for validation, using traditional techniques: 3D mesh from multiview stereo, high-quality appearance maps and traditional skin rendering [Riviere et al. 2020]. A set of 22 views were left out of training, used for validation.

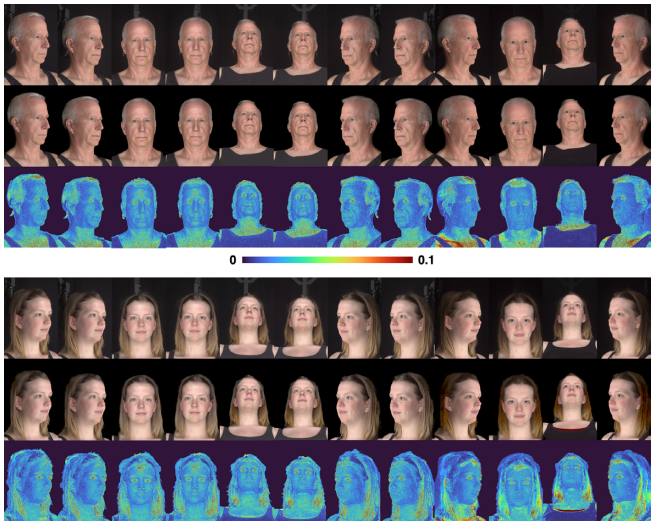

Fig. 3. Rendering error for all 12 real images (top) for two training subjects, used to supervise MoRF. Errors are overall small and predominantly on hair areas, due to complex hair (dis)occlusion and the small number of real training images depicting hair.

better convergence and results when enabling the perceptual loss $\mathcal{L}_{LPIPS}$ only in the first 10% of the fine-tuning iterations.

2 ADDITIONAL RESULTS

In the following, we present additional evaluations and results that supplement our main paper. Experiments include an evaluation on novel view renders, an ablation of different architectures, and an evaluation of the depth estimation in MoRF. We further highlight the reconstructions of all 15 subjects in our dataset, and illustrate the effect of pairwise mixing of the deformation and canonical identity codes. Finally, we show results of fitting and tuning a pre-trained MoRF to 5 novel subjects not seen during training.

2.1 Evaluation of Novel View Renders

To train MoRF, we augment the training data using synthetic images that are obtained from a combination of traditional techniques, such as multiview stereo (MVS) reconstructions, high-quality capture of appearance maps (*e.g.*, diffuse and specular albedo), and skin rendering via ray tracing [Riviere et al. 2020]. Besides synthesizing 35 novel views for training, as shown in Fig. 5(b) of the main text, we also synthesized 22 additional views for each subject, which were left out of training and reserved for validation. This validation set includes sequences of views that go left to right, top to bottom, as illustrated in Fig. 2. Just as during training, these synthetic images

Fig. 4. Comparison of MoRF and two variants of its architecture and training losses: rendering quality when interpolating between two training subjects deteriorates when removing the deformation loss (V2) and deformation network (V1) altogether; note artifacts on the sides of the face, center columns.

Fig. 5. When the canonical appearance code $\mathbf{id}_c$ is used to condition only the output color branches, but not density, the canonical NeRF variant cannot generate densities to model different hair styles (bottom) and fails to properly represent the real training subjects (top).

depict only facial skin areas, as traditional techniques cannot capture the other head components with high quality (thus motivating the development of neural rendering alternatives, such as ours). After training, we use our single MoRF model to render all 15 subjects under the same 22 validation views. We then compare our neural renderings against the well-established, high-quality ray tracing method, as measured by a PSNR of 36.13 on the validation views (skin areas only). In comparison, PSNR for the full head on the training images (including hair areas) was slightly lower, 35.98. Color-coded error maps for real training images are shown in Fig. 3, showing that errors are overall small and predominantly located on hair areas, which are supervised by only a small number of real views during training. Note that the skin areas of the face are assigned more ray samples during training (due to augmentation with synthetic training images) and thus present lower error.

2.2 Ablation of Different Architectures and Losses

Note that the architecture of MoRF is similar to that of the Nerfies method in [Park et al. 2021], with two main differences: (1) MoRF includes an additional Identity Network, and (2) MoRF's $\mathbf{id}_c$ appearance code is also used to condition the density output. We now compare MoRF against two variants of its design: variant V1 does not include the deformation network and is thus closer in spirit to [Schwarz et al. 2020]; the second variant, V2, includes the deformation component but is not explicitly supervised using our deformation loss (i.e., using cross-subject semantic correspondences from the registered, 3D template mesh). For this comparison, MoRF and its V1, V2 variants were trained on 4 subjects, over 200K iterations. While we observed similar PSNR values on the rendered validation views (respectively, 35.75, 35.86, and 35.9), we found that MoRF behaves better than V1 and V2 when interpolating **id** codes in between training subjects, Fig. 4. Finally, for the canonical NeRF, we also found that the Nerfies variant of this MLP (also similar to that in "NeRF in the Wild" [Martin-Brualla et al. 2021]) could not model the variable hair styles presented by our training subjects (all subjects were modeled with short hair). This is because this variant uses the input appearance code $\mathbf{id}_c$ to condition only the output color values, but not the output density. Therefore, the network is unable to generate new density for long hair in areas that had previously been occupied by empty space, Fig. 5 (i.e., the warp MLP alone could not prevent this issue).

2.3 Density (Depth) Supervision

MoRF is trained with losses derived from traditional multiview stereo (MVS) reconstructions, which are used to constrain the density field output by the canonical NeRF for each subject. Note that the density field is used to derive depth maps from this neural model [Mildenhall et al. 2020]. Fig. 6 shows examples of the resulting differences between the depth maps rendered using MoRF and depth maps from traditional MVS: both solutions largely agree on facial skin areas (RMS difference of 5mm), while MoRF is free to improve depth estimates on the areas where MVS confidence is low (*e.g.*, hair). Also of note, the depth estimates by MoRF also diverge from the MVS reconstruction on the neck areas, which we have found to be less accurate in the 3D meshes used to provide supervision during training.

2.4 Spherical Harmonics Constraints

We have empirically verified that our use of a spherical harmonics (SH) subspace better constrains the specular radiance output branch when training from few camera views (i.e., by regularizing the output and enforcing *smoothness across viewpoints*). In particular, for the specular radiance on scalp hair, we only have supervision from 8 camera views (4 narrow-baseline stereo pairs). Here we show an ablation in which we replace the SH output branch with an MLP with positional-encoded view direction as input, as in the original NeRF. We compare the specular signal estimated with and without SH in Fig. 7. For synthesized novel views, we can see that the SH output of MoRF is better constrained, while the MLP output without SH shows spurious fluctuation in underconstrained head areas and

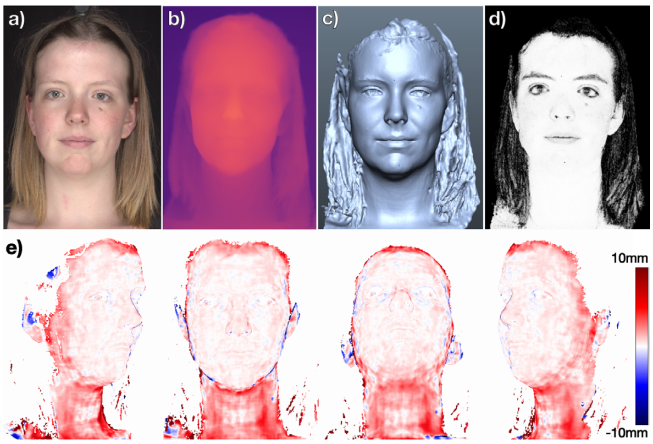

Fig. 6. Results of depth supervision based on traditional multiview stereo (MVS): (a) example training subject; (b) final depth rendered by MoRF; (c) 3D mesh reconstructed via MVS, used to constrain the density values output by MoRF; (d) MVS confidence (high on skin, low on hair areas); and (e) four views showing depth differences (close agreement) between the results from MoRF and from MVS on areas where MVS has high confidence. MoRF is free to improve depth on the other areas.

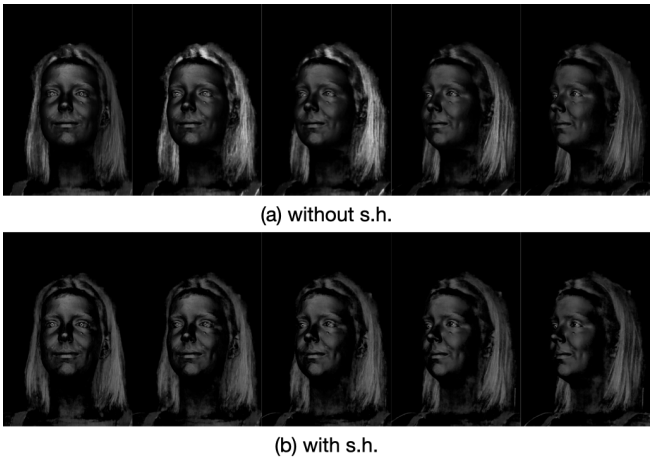

Fig. 7. Comparison of MoRF with (a) positional-encoded view direction as network input and (b) with spherical harmonics. We can see that in (a), the specular signals on hair is inconsistent, while spherical harmonics does not suffer from the artifact.

viewpoints. Another advantage of our SH-based architecture is that the Canonical NeRF becomes omnidirectional and can be densely sampled only once (and cached), then allowing for fast rendering under arbitrary views [Yu et al. 2021]. This is a functionally that we intend to explore in future work.

2.5 Modeling Facial Expressions

The seminal work by Blanz and Vetter [1999] presented a 3D face model that also had two codes, for shape and texture (appearance),

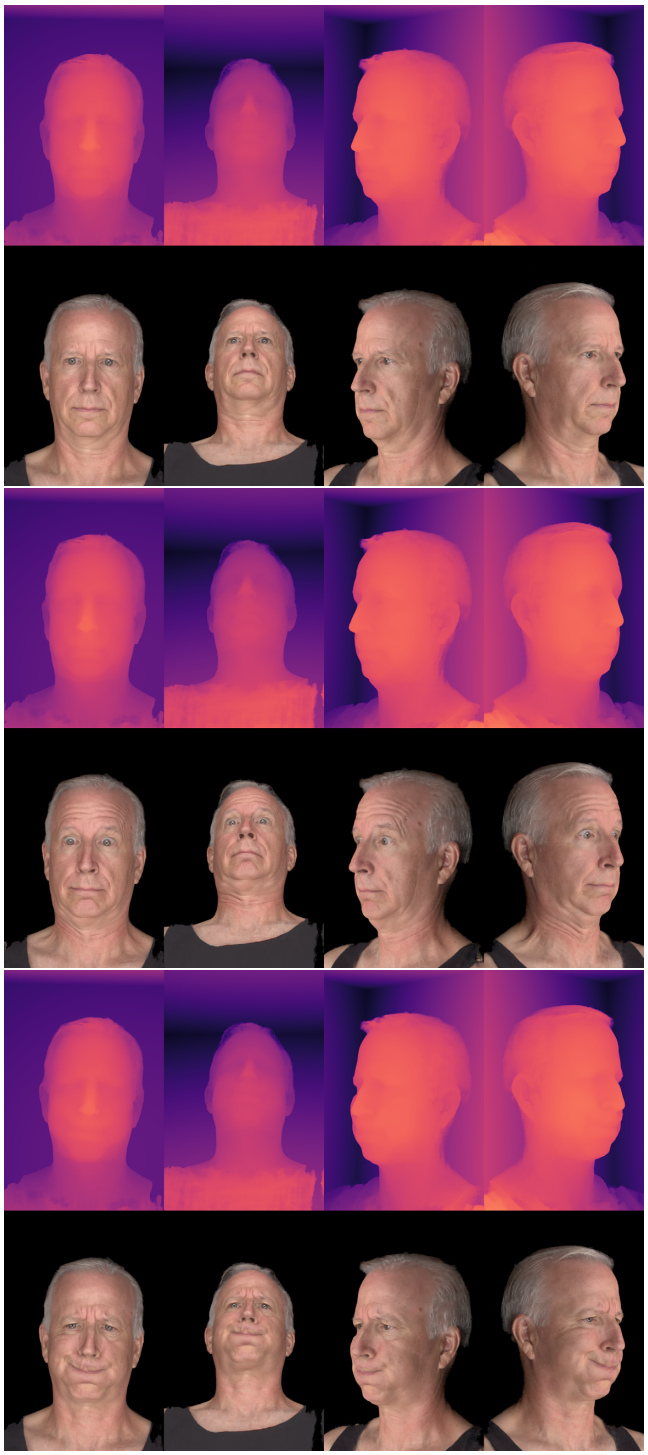

Fig. 8. As a proof of concept that MoRF can also naturally model different facial expressions for each subject, we train MoRF normally, but using training data from a single subject under three different facial expressions (neutral, surprise, frown). Here we show these expressions as rendered by the MoRF model in 4 views.

which can be used to encode variations across identity or expressions. Here, MoRF is focused on disentangling geometry and appearance for modeling vastly different identities and appearances. Nevertheless, modeling facial expressions (with smaller variations in appearance) is a natural next step, which we demonstrate next (see also the relation between MoRF and Nerfies, as noted above). We now show a preliminary result demonstrating that MoRF can be easily extended to model changes due to facial expression (a simpler case when compared to the abrupt changes in appearance across different subject identities). This is done simply by considering additional training data for non-neutral expressions, without modifying MoRF's architecture. As a proof of concept, we train our MoRF architecture normally but on a training dataset that included three different expressions of a single subject. Fig. 8 shows these expressions as rendered by the single MoRF model in four different views.

2.6 Modeling Quality of MoRF versus Single-Subject NeRF

We also compare MoRF's reconstruction quality to that of a regular NeRF trained on a single subject. Given enough data, the simple NeRF can overfit this single person, while MoRF will focus on doing an overall good job for several subjects, simultaneously, but no subject in particular. We compare MoRF's performance against single-subject NeRF using the first subject as shown in Fig. 9 (top). MoRF's PSNR on validation views is 35.84, and on the real training image is 33.91. For the single-subject NeRF, PSNR on validating views is 37.98 and on the real training image it is 35.79. Although PSNR for MoRF is slightly lower than for the single-subject NeRF, visual quality is still very similar. Note that MoRF and this single-subject NeRF also have similar network capacity.

2.7 All 15 Subjects

For all the 15 subjects used to train our single MoRF model, we show examples of novel "turntable" views in in Fig. 9, Fig. 10, and Fig. 11 (with other views shown in the supplementary video). For each subject, we see a pair of real images used for training, where one image is cross-polarized (captures only diffuse reflections) and the other real image below it is parallel-polarized (captures both diffuse and specular components). Next to them, we see five synthetic novel views rendered in three different layers, which separately show only diffuse colors, only specular colors, and full reflected color (diffuse plus specular). Note that not only are these renderings consistent across views but also realistically capture the diffuse and specular reflection of the different head components such as hair, eyes, and skin areas. Finally, the estimated depth map is also rendered for each view.

2.8 Pairwise Code Mixing

As described in the main text, a simple way to generate novel, synthetic subjects using MoRF is to simply mix and match the deformation and canonical identity codes, $\mathbf{id}_{\mathrm{w}}$ and $\mathbf{id}_{\mathrm{c}}$, taken from two real training subjects. The results of this simple operation carried out on the full set of 15×15 pairs of training subjects in our dataset is shown in Fig. 12 (over 200 new synthetic subjects). The figure should be interpreted as mixing the appearance ($\mathbf{id}_{\mathrm{c}}$) of the person at the top with the shape ($\mathbf{id}_{\mathrm{w}}$) of the person on the left. Despite the small number of training subjects, note the variety of novel synthetic subjects that can be generated with MoRF via such simple mixing. All these subjects can be rendered realistically under novel views, with full separation of diffuse and specular layers (see additional views in supplementary video).

2.9 MoRF Fitting and Tuning Results

We also compute results of pre-training MoRF on 15 subjects and then fitting and tuning the model to images of 5 novel subjects not seen at training, effectively producing a NeRF from as few as one image and an optional 3D mesh. Fig. 13 shows the synthesized "turntable" views of these 5 out-of-sample subjects, after fitting and tuning the network to *11 input views* and the registered 3D mesh. For each subject, the second image on the left column shows the real frontal view *left out of the fit*, with rendering PSNR of 32.8, 35.9, 34.2, 31.9, 32.0, respectively (average 33.3) after 2K tuning iterations. Finally, Fig. 14 shows the results of tuning MoRF to *a single (frontal) input view* and the registered 3D mesh of each new subject; the second image on the left column now shows the real view *included in the fit* for each subject. Rendering PSNR, computed over the other 11 images held out, is lower in this case, 24.3, 27.6, 28.0, 24.5, 27.6, respectively. After *removing the 3D mesh from the input*, Fig. 15, these PSNR values drop slightly from an average of 26.4 (with) to 25.8 (without). Still, noticeable errors are seen on face silhouettes (*e.g.*, shape of nose). In practice, our mesh topology could be adapted to work with any method that estimates a 3D mesh from a single face image [Feng et al. 2021, 2018; Guo et al. 2020]. Finally, results of only the initial **id**-fit, without the tuning step, are shown in Fig. 16, Fig. 17, and Fig. 18.

REFERENCES

Rameen Abdal, Yipeng Qin, and Peter Wonka. 2019. Image2StyleGAN: How to embed images into the StyleGAN latent space?. In *Proc. ICCV*. IEEE, 4432–4441.

Volker Blanz and Thomas Vetter. 1999. A morphable model for the synthesis of 3D faces. In *Siggraph*, Vol. 99. 187–194.

Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. 2021. Learning an Animatable Detailed 3D Face Model from In-the-Wild Images. *ACM Trans. Graphics (Proc. SIGGRAPH)* 40, 4 (2021).

Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. 2018. Joint 3D Face Reconstruction and Dense Alignment with Position Map Regression Network. In *Proc. ECCV*.

Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. 2020. Towards Fast, Accurate and Stable 3D Dense Face Alignment. In *Proc. ECCV*.

Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. 2021. NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. In *Proc. CVPR*.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *ECCV*.

Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. 2021. Nerfies: Deformable Neural Radiance Fields. *Proc. ICCV* (2021).

Jérémy Riviere, Paulo Gotardo, Derek Bradley, Abhijeet Ghosh, and Thabo Beeler. 2020. Single-Shot High-Quality Facial Geometry and Skin Appearance Capture. *ACM Trans. Graphics (Proc. SIGGRAPH)* 39, 4, Article 81 (2020), 12 pages.

Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. 2021. Pivotal Tuning for Latent-based Editing of Real Images. *arXiv preprint arXiv:2106.05744* (2021).

Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. 2020. GRAF: Generative Radiance Fields for 3D-Aware Image Synthesis. In *Proc. NeurIPS*.

Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. 2021. PlenOctrees for Real-time Rendering of Neural Radiance Fields. In *Proc. ICCV*.

(a) real (b) rendered diffuse-only, full color novel views (c) rendered depth, specular components

Fig. 9. Synthesized "turntable" views using a single MoRF model to reproduce training subjects 1-5 (out of 15). Each pair of rows shows: (a) cross- and parallel-polarized real training images; (b) novel views rendered in diffuse color only (odd rows) and full color (even rows); and (c) rendered depth and specular components. This figure is best seen on a computer screen.

(a) real (b) rendered diffuse-only, full color novel views (c) rendered depth, specular components

Fig. 10. Synthesized "turntable" views using a single MoRF model to reproduce training subjects 6-10 (out of 15). Each pair of rows shows: (a) cross- and parallel-polarized real training images; (b) novel views rendered in diffuse color only (odd rows) and full color (even rows); and (c) rendered depth and specular components. This figure is best seen on a computer screen.

(a) real (b) rendered diffuse-only, full color novel views (c) rendered depth, specular components

Fig. 11. Synthesized "turntable" views using a single MoRF model to reproduce training subjects 11-15 (out of 15). Each pair of rows shows: (a) cross- and parallel-polarized real training images; (b) novel views rendered in diffuse color only (odd rows) and full color (even rows); and (c) rendered depth and specular components. This figure is best seen on a computer screen.

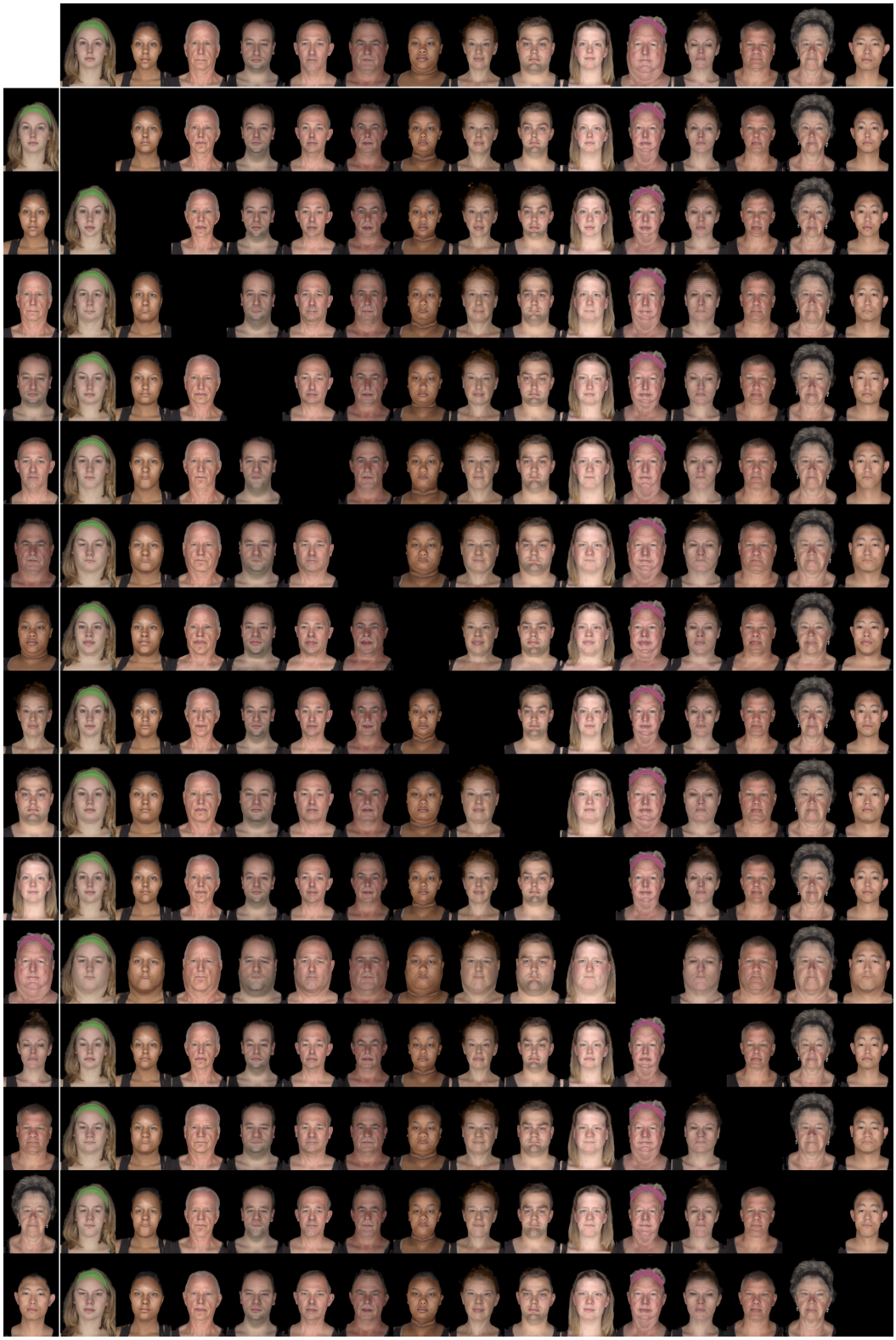

Fig. 12. Novel synthetic subjects generated by mixing codes between all pairs of training subjects: the canonical identity code $\mathbf{id}_c$ is taken from the subject at the top, while the deformation code $\mathbf{id}_w$ is taken from the subject on the left (best seen on computer screen).

(a) real (b) rendered diffuse-only, full color novel views (c) rendered depth, specular components

Fig. 13. Synthesized "turntable" views of the 5 out-of-sample new subjects, after fitting and tuning the pre-trained MoRF network to **11 input views** and the registered 3D mesh. For each subject, the second image in (a) shows the real frontal view left out of the fit. Rendering PSNR is 32.8, 35.9, 34.2, 31.9, 32.0, respectively.

(a) real (b) rendered diffuse-only, full color novel views (c) rendered depth, specular components

Fig. 14. Synthesized "turntable" views of the 5 out-of-sample new subjects, after fitting and tuning the pre-trained MoRF network to **a single frontal view (second image in (a))** and the registered 3D mesh. Rendering PSNR, computed over the other 11 images held out, is 24.3, 27.6, 28.0, 24.5, 27.6, respectively.

(a) real (b) rendered diffuse-only, full color novel views (c) rendered depth, specular components

Fig. 15. Synthesized "turntable" views of the 5 out-of-sample new subjects, after fitting and tuning the pre-trained MoRF network to **a single frontal view (second image in (a))**, **without** the registered 3D mesh. Rendering PSNR, computed over the other 11 images held out, is 23.7, 27.0, 27.4, 24.1, 26.6, respectively.

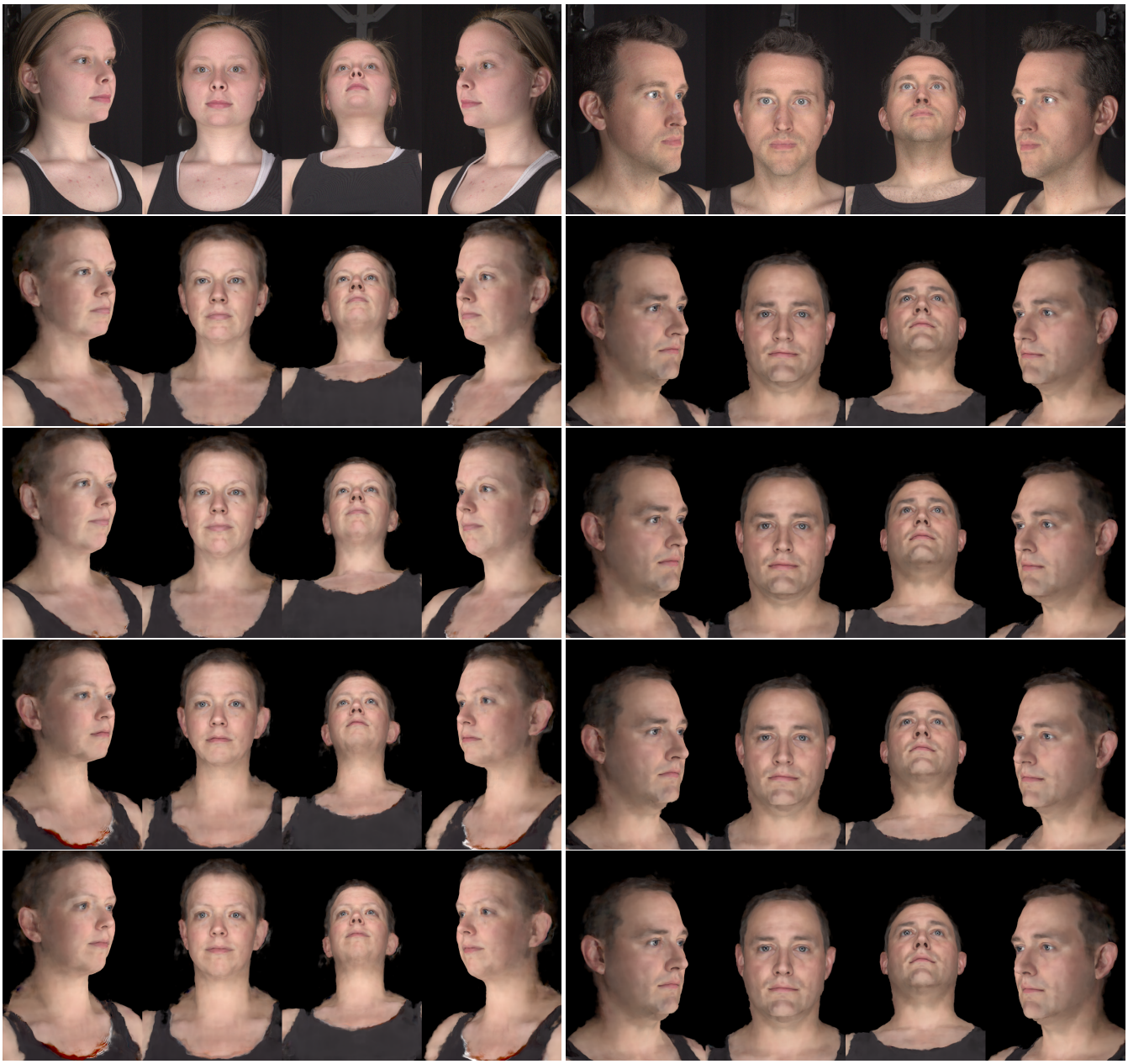

Fig. 16. Simple **id**-fitting results for out-of-sample subjects 1 and 2. From top to bottom: 4 of the 12 real views, fit to 12 views (with and without the input 3D mesh), and fit to 1 frontal view (with and without 3D mesh).

Fig. 17. Simple **id**-fitting results for out-of-sample subjects 3 and 4. From top to bottom: 4 of the 12 real views, fit to 12 views (with and without the input 3D mesh), and fit to 1 frontal view (with and without 3D mesh).

Fig. 18. Simple **id**-fitting results for out-of-sample subject 5. From top to bottom: 4 of the 12 real views, and fit to 12 views (with and without the input 3D mesh), fit to 1 frontal view (with and without 3D mesh).