Supplementary Materials for

FitMe: Deep Photorealistic 3D Morphable Model Avatars

Alexandros Lattas Stylianos Moschoglou Stylianos Ploumpis
Baris Gecer Jiankang Deng Stefanos Zafeiriou

Imperial College London, UK

{a.lattas,s.moschoglou,s.ploumpis,b.gecer,j.deng16,s.zafeiriou}@imperial.ac.uk

1. Model Manipulation

Given enough samples from the trained BRDF-GAN generator (10000 in our case), the latent space $\mathbf{W}$ of a style-based generator, can be analyzed with PCA [11]. We expect the first principal components to expose interpretable controls over features of the reconstruction. In Fig. 1 we show how the first three components correspond roughly to the skin tone (given our augmentation), gender and age variations. These show that our model learns a meaningful latent space and enable us to perform semantic manipulations directly on the reflectance UV maps.

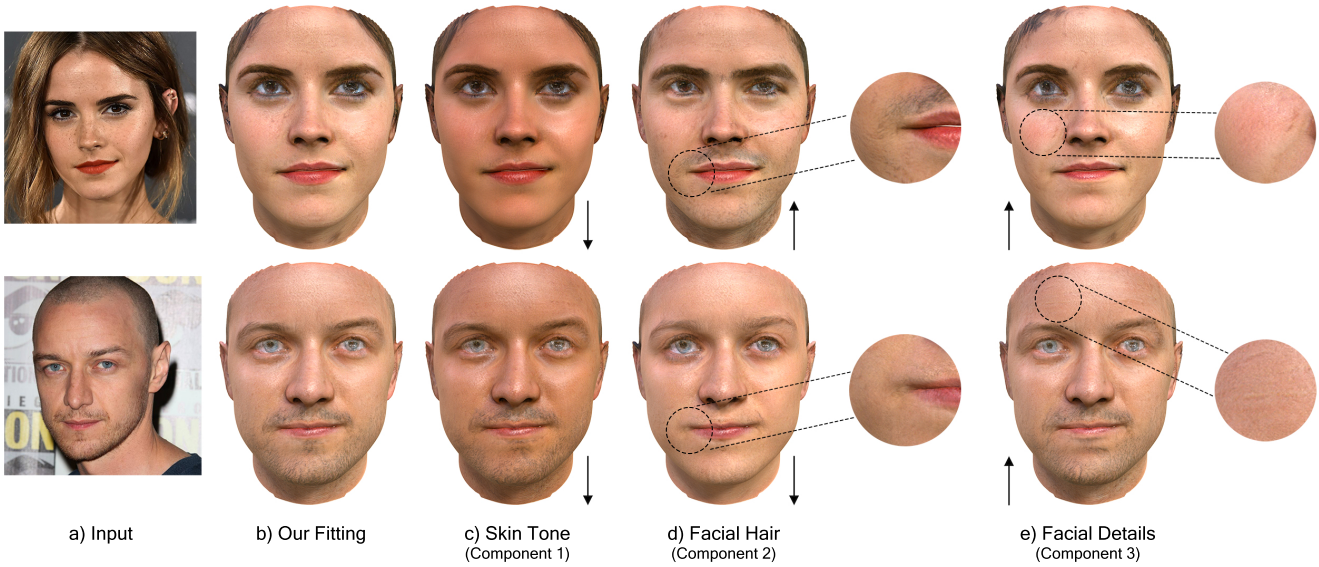
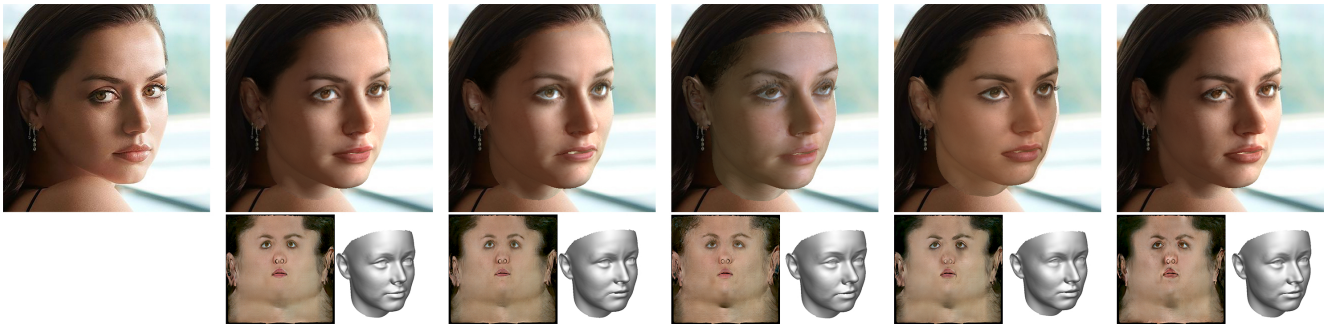


Figure 1. Manipulation of a BRDF-GAN projection. By performing PCA on the latent space of BRDF-GAN, we can manipulate the most important components, which roughly correspond to c) skin tone, d) gender and e) details (wrinkles, freckles). The arrows indicate the direction of the manipulation of the PCA component.

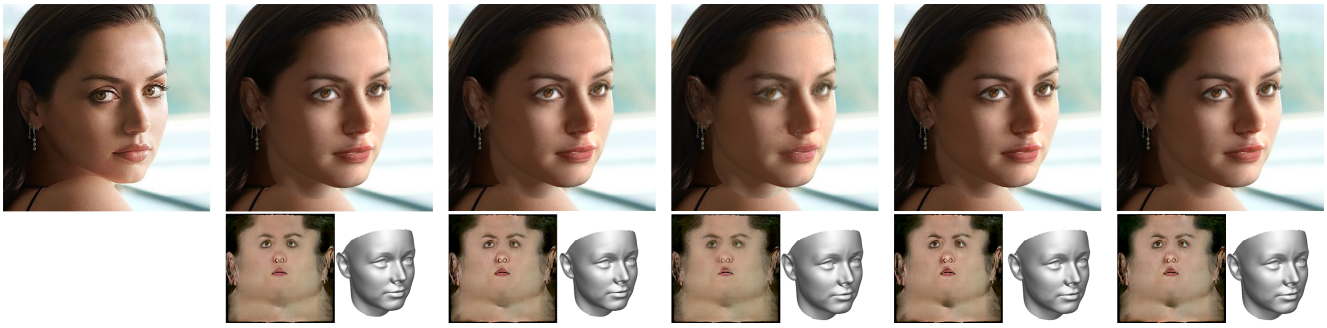
2. Ablation Study

We perform three ablation studies to validate our architectural and optimization choices. The first ablation, shown in Fig. 2, shows the effect of the losses used in the GAN inversion part of our method. For each example, we remove one of the following losses: the identity loss $\mathcal{L}_{ID}$, the perceptual loss $\mathcal{L}_{per}$, the photometric loss $\mathcal{L}_{ph}$ and the $\mathbf{W}$ regularization loss $\mathcal{L}_{\mathbf{W}}$. Removing the landmark loss $\mathcal{L}_{lan}$ or the shape and expression regularization losses $\mathcal{L}_s, \mathcal{L}_e$, fails the optimization, as the shape fitting is misplaced.



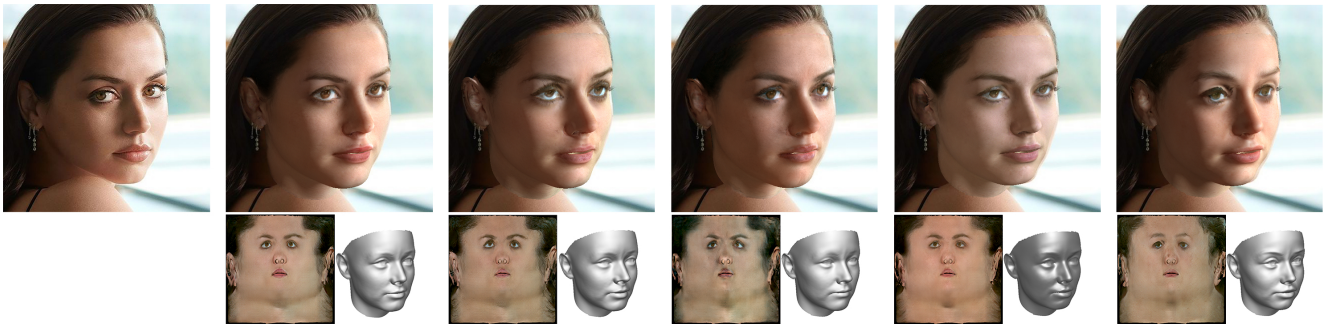
a) Input b) Ours c) no $\mathcal{L}_{ID}$ d) no $\mathcal{L}_{per}$ e) no $\mathcal{L}_{ph}$ f) no $\mathcal{L}_{\mathbf{W}}$

Figure 2. Ablation study on the losses of the **GAN inversion** optimization. For each example, we remove the shown loss from the optimization, and show the rendered result $\mathbf{I}_R$ (top), the diffuse albedo $\mathbf{A}_D$ (bottom-left) and the shape $\mathbf{S}$ (bottom-right). It is apparent how all the proposed losses are required, in order to obtain facial shape and reflectance with high identity and visual similarity, while also maintaining an albedo without scene illumination.



a) Input b) Ours c) no $\mathcal{L}_{ph}$ d) no $\mathcal{L}_{LP}$ e) no $\mathcal{L}_{flip}$ f) no $\mathcal{L}_{\kappa}$

Figure 3. Ablation study on the losses of the **GAN tuning** optimization. For each example, we remove the shown loss from the optimization, and show the rendered result $\mathbf{I}_R$ (top), the diffuse albedo $\mathbf{A}_D$ (bottom-left) and the shape $\mathbf{S}$ (bottom-right). Despite most cases showing a highly optimized rendered result, without the proposed losses combination, the albedo absorbs residual scene illumination, not captured by our rendering, in cases c), e), f), Also, since we do not optimize the shape $\mathbf{S}$ during tuning, the shape remains the same in these examples.



a) Input b) Ours c) no GAN tuning d) $\mathbf{W}+$ instead of $\mathbf{W}$ [14] e) Pytorch3D [15] shader f) $\mathbf{z}$ fitting instead of $\mathbf{W}$

Figure 4. Ablation study on **method choices**. For each example, make the shown change during the optimization, and show the rendered result $\mathbf{I}_R$ (top), the diffuse albedo $\mathbf{A}_D$ (bottom-left) and the shape $\mathbf{S}$ (bottom-right). c), e)and f) show the importance of $\mathbf{W}$ fitting, our photorealistic facial shading and the GAN tuning. d) shows that $\mathbf{W}+$ optimization, as performed by [14], achieves great likeness in the albedo, but does not maintain the semantics of the facial reflectance.

Moreover, in Fig. 3, we perform a similar ablation for the losses used during the GAN tuning part of our method. The removal of the photometric $\mathcal{L}_{ph}$ and LPIPS $\mathcal{L}_{LP}$ losses, shows their importance in achieving higher likeness to the input

image, The removal of the flip $\mathcal{L}_{flip}$ and chromaticity $\mathcal{L}_{\kappa}$ losses shows their need in maintaining an albedo without scene illumination, when compared to vanilla pivotal tuning [16].

Finally, in Fig. 4, we show an ablation study on choices regarding the optimization method. In Fig. 4 c), we show the optimized result, without our GAN tuning, Fig. 4 d) shows an extended latent space $\mathbf{W}+$ optimization, rather than $\mathbf{W}$, which means that each slice of $\mathbf{W}$ is separately optimized for each layer of the generator network and is suggested by [14]. Such an approach also achieves great likeness in the optimized rendering, but disturbs the statistics of the generator, so that scene illumination is absorbed by the albedo, and the network cannot be accurately manipulated (Fig. 1). Similar observations are also reported by [16]. In Fig. 4 e) we show the full optimization (inversion and tuning), while using the Pytorch3D [15] implementation of Phong shader, which also only supports the diffuse albedo $\mathbf{A}_D$, no spatial variation in specular roughness and no subsurface scattering optimization. Fig. 4 f) shows the optimization of the latent variable $\mathbf{z}$ which is passed through a mapping network before given to the generator. Such a fitting is more constrained and cannot reach a satisfying identity similarity.

3. Dataset Augmentation

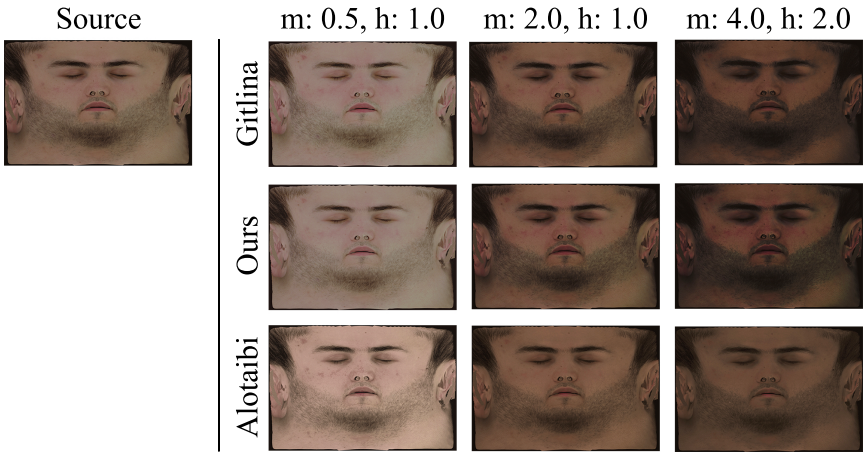


Figure 5. Skin tone augmentation comparisons, with adjusted melanin m and hemoglobin h, between (top) LightStage-measured look-up table melanin-hemoglobin of Gitlina et al. [10], (middle) our proposed masked histogram matching and (bottom) inverse-rendering based manipulation of Alotaibi and Smith [17].

We perform a comparison of our proposed histogram matching albedo augmentation, against a LightStage captured melanin-hemoglobin manipulation method [10], and a melanin-hemoglobin manipulation based on inverse rendering [2]. Our method achieves comparable performance to physics-based skin models [1, 10], without requiring the calculation of such complex interactions, as shown in Fig. 5.

4. Additional Results

We present additional results of multi-view capturing in Figs. 6,7,8. In each case we ask the subject to take three unconstrained mobile phone images, from the front and the side. We show that these are enough to create an accurate photorealistic avatar of the subject, using our method. In total, capturing takes less than 10 seconds and processing less than a minute.

Additional to the quantitative comparison on identity similarity presented in the main manuscript, we also perform and present here a quantitative comparison on shape reconstruction. Following the benchmark proposed by GANFit [6], we reconstruct the 53 subjects from the MICC Florence 3D Faces dataset (MICC) [3] with our facial shape and measure their distance (in nm) from the dataset shapes, using the open-source benchmark code of GANFit [6]. MICC includes three subcategories, "cooperative", "indoor" and "outdoor", on which we report of findings separately. Instead of 5 images [6], we only use 3 random images from each video for the reconstruction. We then perform dense alignment and measure the point-to-plane distance and present our findings in Tab. 1.

Our method performs similarly in shape reconstruction to GANFit [6] and GANFit++ [7], and significantly better when compared to previous methods [5, 8, 18] and Fast-GANFit [7]. GANFit reconstructions are slightly closer to the ground truth shapes. This can be explained by the fact that we optimize both the shape mesh and the texture normals, and thus a part of

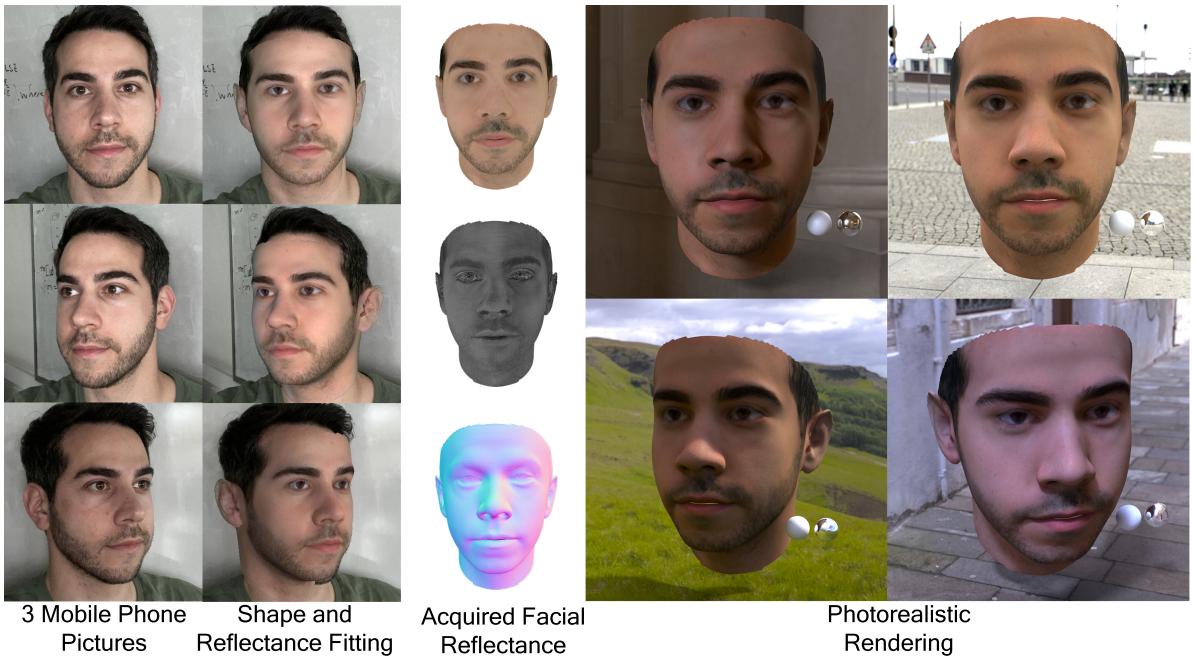


Figure 6. Additional multi-view capture results, using our method. From left to right: a) three input images, b) rendered fitting using our method, c) diffuse albedo $\mathbf{A}_D$, specular albedo $\mathbf{A}_S$ and normals $\mathbf{N}_S$, d) rendered shape $\mathbf{S}$ and e) rendered results on various environments.

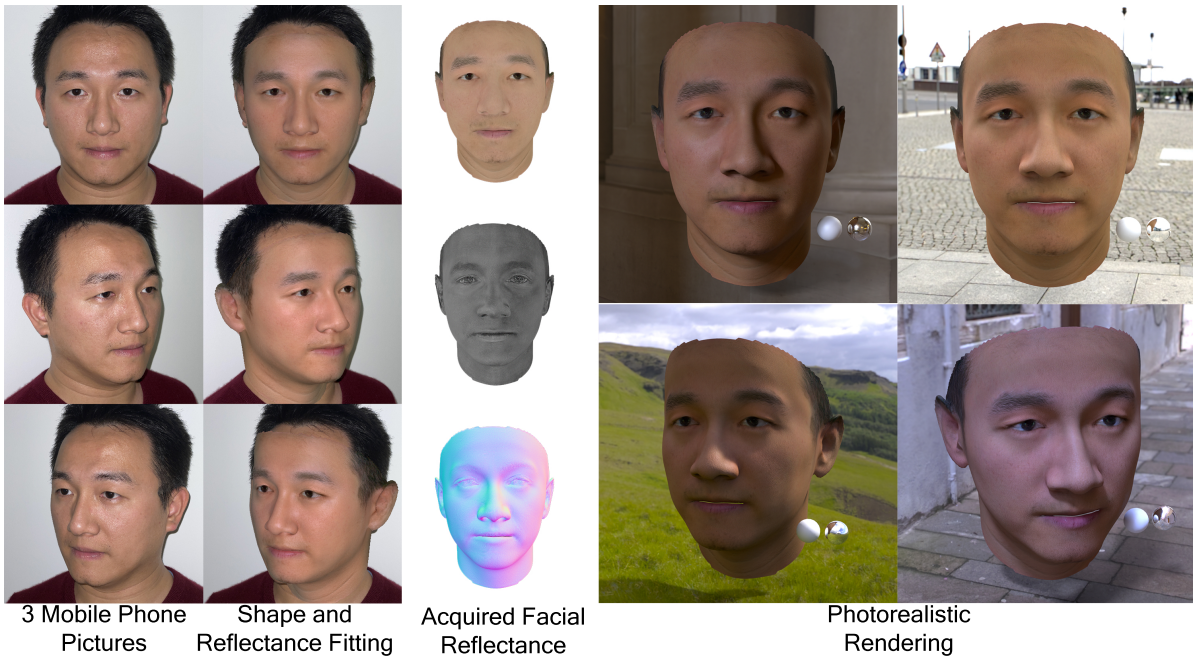


Figure 7. Additional multi-view capture results, using our method. From left to right: a) three input images, b) rendered fitting using our method, c) diffuse albedo $\mathbf{A}_D$, specular albedo $\mathbf{A}_S$ and normals $\mathbf{N}_S$, d) rendered shape $\mathbf{S}$ and e) rendered results on various environments.

the shape information is explained in the normals domain rather than in the actual shape space. For a fair comparison, we only compared the meshes, potentially missing details. Finally, as shown in the main manuscript, our method scores first in identity similarity, while also acquiring relightable reflectance textures.

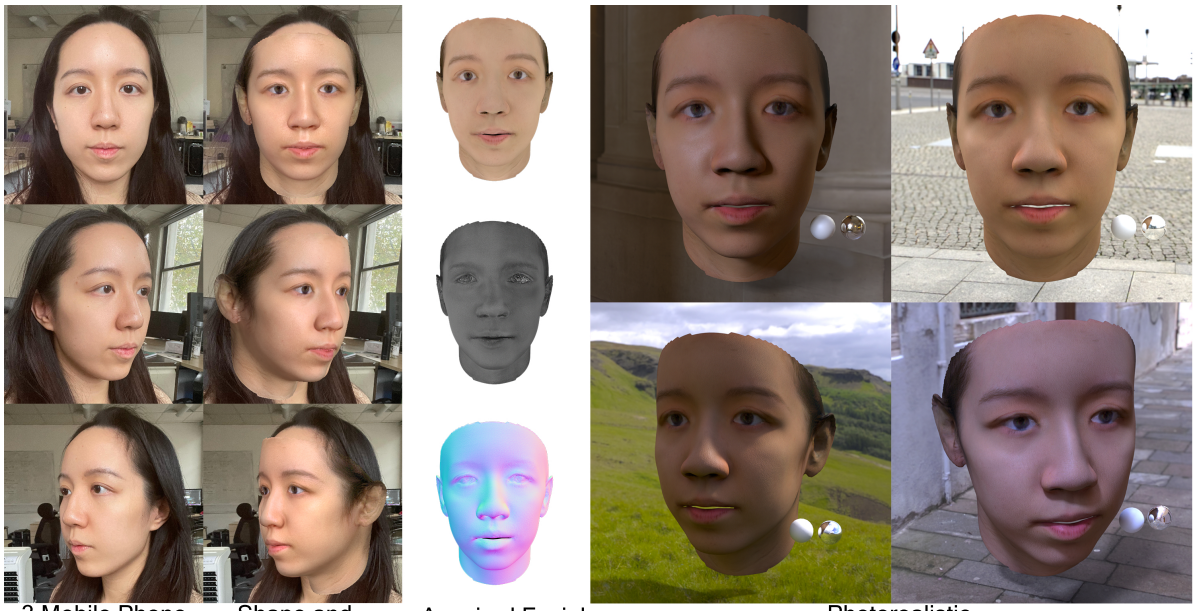


Figure 8. Additional multi-view capture results, using our method. From left to right: a) three input images, b) rendered fitting using our method, c) diffuse albedo $\mathbf{A}_D$, specular albedo $\mathbf{A}_S$ and normals $\mathbf{N}_S$, d) rendered shape $\mathbf{S}$ and e) rendered results on various environments.

Method	**Cooperative** Mean $\pm$ Std.	**Indoor** Mean $\pm$ Std.	**Outdoor** Mean $\pm$ Std.
Tran et al. [19]	1.93 ± 0.27	2.02 ± 0.25	1.86 ± 0.24
Booth et al. [5]	1.82 ± 0.29	1.85 ± 0.22	1.63 ± 0.16
Genova et al. [9]	1.50 ± 0.14	1.50 ± 0.11	1.48 ± 0.11
GANFit [6]	0.95 ± 0.10	0.94 ± 0.10	0.94 ± 0.10
GANFit++ [7]	0.94 ± 0.17	0.92 ± 0.14	0.94 ± 0.19
Fast-GANFit [7]	1.11 ± 0.25	0.98 ± 0.15	1.16 ± 0.18
Ours	0.95 ± 0.18	0.97 ± 0.20	0.98 ± 0.21

Table 1. Quantitative benchmark comparison of shape reconstruction, on the MICC Florence 3D Faces dataset, using point-to-plane distance, as used in the open-source benchmark of [7]. we compare our results with the reported results of Tran et al. [19], Booth et al. [5], Genova et al. [9], GANFit [6], GANFit++ [7] and Fast-GANFit [7].

5. Implementation Details

As briefly described in the main manuscript, FitMe implementation builds on the public repository of StyleGAN2-ADA [12], in pytorch, both for the generator and the discriminator. However, we make the following changes. a) The generator $\mathcal{G}$ is branched on the last convolutional blocks, which is achieved feeding the output of the last single-branch layer, to 3 different copies of the last branched module. The generator follows the skip-connections architecture of [12]. b) The discriminator is also branched, and follows the resnet architecture of [12]. The output of each branch is concatenated, and fed to the last convolutional block and the fully connected layers.

For the differentiable photorealistic rendering, we create a shader based on the Blinn-Phong model [4], following AvatarMe++ [13]. The implementation of the model is done by extending the shader and shading classes in Pytorch3D [15].

Our optimization is based on a number of hyperparameters λ_i, each corresponding to a loss described in the paper. We find these empirically, and present them below. For the inversion, we use a learning rate of $l_{inv} = 1 \times 10^{-2}$ and for the tuning we use a learning rate of $l_{pti} = 8 \times 10^{-4}$. In Table 2 and Table 3 we present the values which we find most optimal, when optimizing for crop and rendering at 512×512 pixels, 250 iterations for inversion and 30 iterations for tuning.

hyper-param	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7
corresponding loss	$\mathcal{L}_{lan}$	$\mathcal{L}_{ph}$	$\mathcal{L}_{ID}$	$\mathcal{L}_{per}$	$\mathcal{L}_{W}$	$\mathcal{L}_{s}$	$\mathcal{L}_{e}$
value	100	0.5	1.0	25.0	5×10^{-2}	5×10^{-4}	5×10^{-4}

Table 2. Hyper-parameter values and their corresponding loss terms, used for the GAN-inversion part of the proposed method.

hyper-param	λ_8	λ_9	λ_{10}	λ_{11}
corresponding loss	$\mathcal{L}_{LP}$	$\mathcal{L}_{ph}$	$\mathcal{L}_{flip}$	$\mathcal{L}_{\kappa}$
value	2.0	0.5	0.8	0.35

Table 3. Hyper-parameter values and their corresponding loss terms, used for the GAN-tuning part of the proposed method.

References

[1] Carlos Aliaga, Christophe Hery, and Mengqi Xia. Estimation of spectral biophysical skin properties from captured rgb albedo. *arXiv preprint arXiv:2201.10695*, 2022. 3

[2] Sarah Alotaibi and William AP Smith. A biophysical 3d morphable model of face appearance. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 824–832, 2017. 3

[3] Andrew D. Bagdanov, Alberto Del Bimbo, and Iacopo Masi. The florence 2d/3d hybrid face dataset. In *Proceedings of the 2011 Joint ACM Workshop on Human Gesture and Behavior Understanding*, J-HGBU '11, page 79–80, New York, NY, USA, 2011. ACM. 3

[4] James F Blinn. Models of light reflection for computer synthesized pictures. In *Proceedings of the 4th annual conference on Computer graphics and interactive techniques*, pages 192–198, 1977. 5

[5] James Booth, Anastasios Roussos, Stefanos Zafeiriou, Allan Ponniah, and David Dunaway. A 3d morphable model learnt from 10,000 faces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5543–5552, 2016. 3, 5

[6] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1155–1164, 2019. 3, 5

[7] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. Fast-ganfit: Generative adversarial network for high fidelity 3d face reconstruction. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):4879–4893, 2021. 3, 5

[8] Kyle Genova, Forrester Cole, Aaron Maschinot, Aaron Sarna, Daniel Vlasic, and William T Freeman. Unsupervised training for 3d morphable model regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8377–8386, 2018. 3

[9] Kyle Genova, Forrester Cole, Aaron Maschinot, Aaron Sarna, Daniel Vlasic, and William T. Freeman. Unsupervised training for 3d morphable model regression. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 5

[10] Yuliya Gitlina, Giuseppe Claudio Guarnera, Daljit Singh Dhillon, Jan Hansen, Alexander Lattas, Dinesh Pai, and Abhijeet Ghosh. Practical measurement and reconstruction of spectral skin reflectance. In *Computer graphics forum*, volume 39, pages 75–89. Wiley Online Library, 2020. 3

[11] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *Advances in Neural Information Processing Systems*, 33:9841–9850, 2020. 1

[12] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in Neural Information Processing Systems*, 33:12104–12114, 2020. 5

[13] Alexandros Lattas, Stylianos Moschoglou, Stylianos Ploumpis, Baris Gecer, Abhijeet Ghosh, and Stefanos Zafeiriou. Avatarme++: Facial shape and brdf inference with photorealistic rendering-aware gans. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9269–9284, 2021. 5

[14] Huiwen Luo, Koki Nagano, Han-Wei Kung, Qingguo Xu, Zejian Wang, Lingyu Wei, Liwen Hu, and Hao Li. Normalized avatar synthesis using stylegan and perceptual refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11662–11672, 2021. 2, 3

[15] 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. 2, 3, 5

[16] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Transactions on Graphics (TOG)*, 42(1):1–13, 2022. 3

[17] 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. 3

[18] Luan Tran, Feng Liu, and Xiaoming Liu. Towards high-fidelity nonlinear 3d face morphable model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1126–1135, 2019. 3

[19] Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gérard 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. 5