Supplemental: Monocular Reconstruction of Neural Face Reflectance Fields

Mallikarjun B R[1] Ayush Tewari[1] Tae-Hyun Oh[2] Tim Weyrich[3]
Bernd Bickel[4] Hans-Peter Seidel[1] Hanspeter Pfister[5]
Wojciech Matusik[6] Mohamed Elgharib[1] Christian Theobalt[1]
[1]Max Planck Institute for Informatics, Saarland Informatics Campus [2]POSTECH
[3]University College London [4]IST Austria [5]Harvard University [6]MIT CSAIL

1. Dataset visualization

Our relfectance field is learned from a light-stage dataset containing 350 identities, captured from 8 camera viewpoints and illuminated by 150 point light sources one at a time (see Fig. 1). The dataset was originally proposed in Weynrich *et al*. [5], with just 149 identities. The dataset we use, however, contains 201 additional identities. Throughout our work we use 300 identities for training, 10 for validation and the rest for test.

2. Network details

Please refer to Tables.1 and 2 for detailed architectures of the geometry and reflectance networks.

3. Video results

We provide a compiled video in the project page[1] consisting of following results.

3.1. Relighting with densely sampled OLATs

Even though our reflectance model is trained using only 150 point light sources, we can sample any arbitrary number of OLATs using the light direction input. In the video, we sample results for 1024 point light sources. Here, we move the light source from top to bottom in a spiral manner. We show result for 18 identities shot in-the-wild.

3.2. View dependent effects

Please watch the video, which demonstrates view dependent effects. Here, we keep the light source fixed and only change the camera pose. Note how the specular component around the nose region changes as the camera moves. Results also show changes in sub surface component as the camera pose changes (e.g. change in soft shadows).

3.3. Relighting with environment maps

Please watch the video, where we relight several identities with environment map. Here we also change the viewing angle, which shows the view dependent capablities of our method. We also relight dynamic video of the faces, which shows the generalization capablity of our method for different expressions.

4. Ablative Study: VGG-based imageNet and light feature Loss

We assess the impact of the feature losses used in our reflectance learning. One of the feature losses (L_{I}) is based on a VGG network trained on ImageNet [1]. The other feature loss $\mathcal{L}_{\mathrm{L}}$ is based on a VGG network trained to predict lighting direction of OLATs [2]. We evaluate the OLAT reconstructions error using Si-MSE on the same test dataset used in Table. 1 in the main paper. We report results for renderings with same and different input head pose. Tab. 2 shows that the best performance is obtained when both losses are used. This is especially case when rendering with a head pose different from the input.

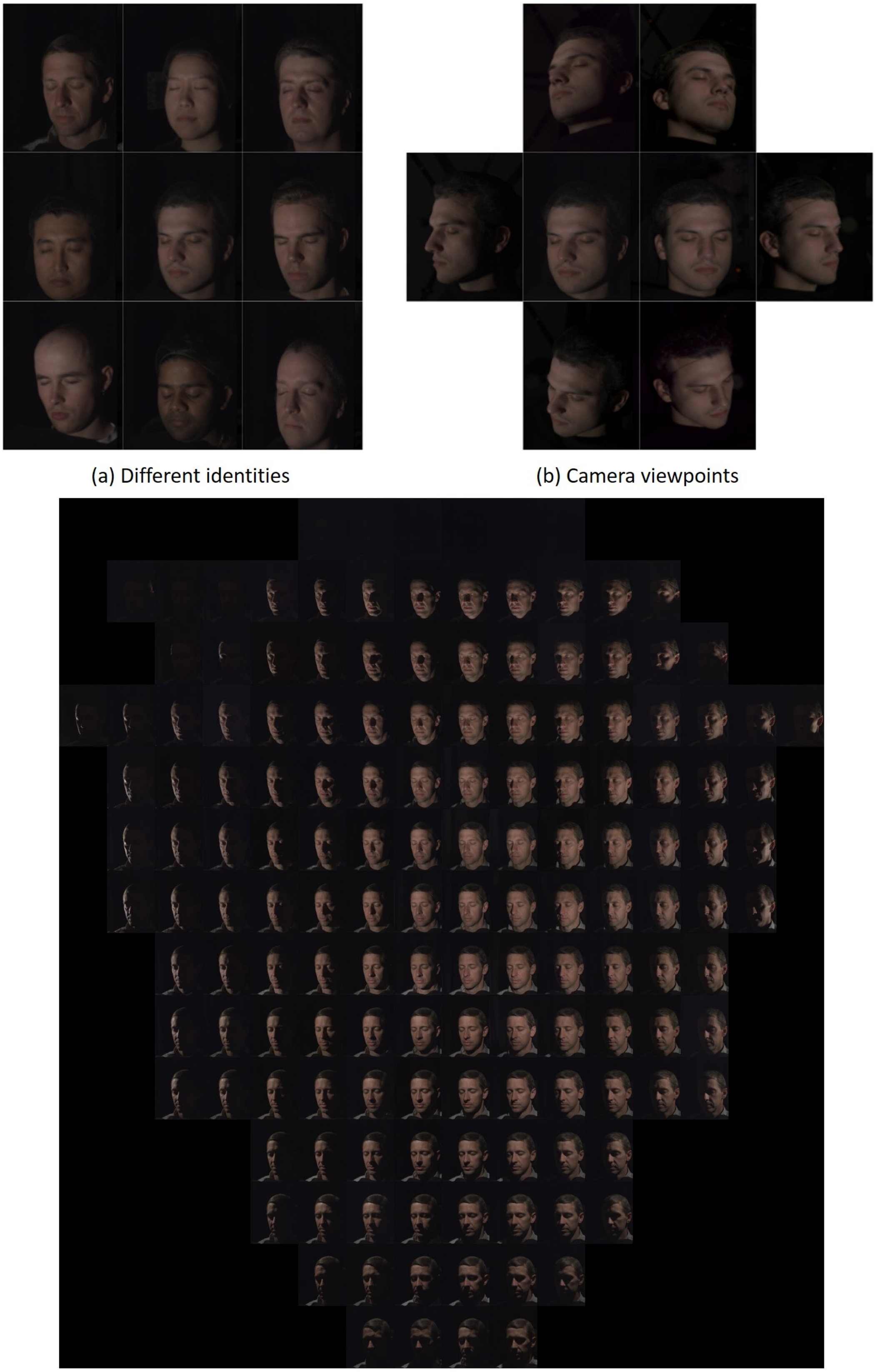
5. OLAT comparison:

We project the OLAT lights to the spherical harmonic space and perform a comparison of relighting results. Fig. 2 shows that our method can capture physically correct self-shadows and other global effects to synthesize photo realistic image.

6. High-Fidelity Facial Reflectance and Geometry Inference From an Unconstrained Image

The approach of Yamaguchi *et al*. [6] regresses diffuse and specular albedo maps of a face from an image. The ground truth during training is obtained using a light stage

[1] **project page:** http://gvv.mpi-inf.mpg.de/projects/FaceReflectanceFields/



(a) Different identities

(b) Camera viewpoints

(c) 150 point light sources

Figure 1. Our reflectance field is learned from a light stage dataset containing 350 identities (a) recorded from 8 different camera viewpoints (b) and illuminated by 150 point light sources (c).

	Layers	Output
Image (512,512,3)	MaxPool (kernel 3x3, stride 2)	*unnamed*
↑	Conv2D (kernel 11x11x96, stride 4) + ReLU	*unnamed*
↑	MaxPool (kernel 3x3, stride 2)	*unnamed*
↑	Conv2D (kernel 5x5x256, stride 1) + ReLU	*unnamed*
↑	MaxPool (kernel 3x3, stride 2)	*unnamed*
↑	Conv2D (kernel 3x3x384, stride 1) + ReLU	lowFeatures
↑	Conv2D (kernel 3x3x384, stride 2) + ReLU	*unnamed*
↑	Conv2D (kernel 3x3x256, stride 2) + ReLU	mediumFeatures
mediumFeatures	Conv2D (kernel 3x3x384, stride 1) + ReLU	*unnamed*
↑	Conv2D (kernel 3x3x256, stride 1) + ReLU	*unnamed*
↑	Fully Connected (kernel x1000) + ReLU	*unnamed*
↑	Fully Connected (kernel x1000) + ReLU	*unnamed*
↑	Fully Connected (kernel x(64+64))	shapeParam
lowFeatures	Conv2D (kernel 3x3x384, stride 1) + ReLU	intermediate
lowFeatures, intermediate	Concat	*unnamed*
↑	Conv2D (kernel 3x3x768, stride 1) + ReLU	*unnamed*
↑	Conv2D (kernel 3x3x384, stride 1) + ReLU	*unnamed*
↑	Conv2D (kernel 3x3x256, stride 1) + ReLU	*unnamed*
↑	MaxPool(kernel 3x3, stride 2)	*unnamed*
↑	Fully Connected (kernel x2048) + ReLU	*unnamed*
↑	Fully Connected (kernel x(6+64))	pose, expressionParam

Table 1. Geometry network details.

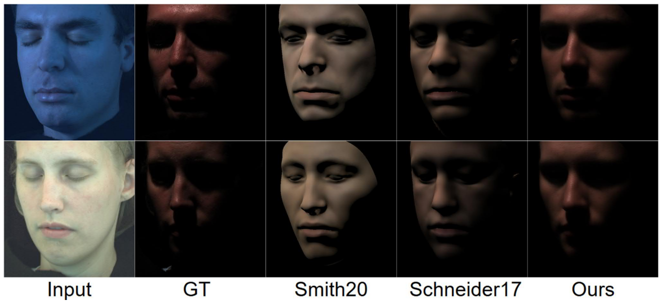


Figure 2. We use spherical harmonic approximation of point light source to synthesize comparitive method. Observe that our method captures physically correct shadows and other global illumination effects photo realistically.

dataset. Since only the diffuse and specular components are modeled, other higher view-dependent effects is ignored. Since this approach cannot predict the environment light from the image, it is difficult to compare directly using a reference image for lighting. Each approach uses a different coordinate system for representing meshes and the light, making it difficult to render the results of both methods under the same lighting. For a qualitative comparison, we manually change the lights in the scene in order to get visually similar results. As can be see in Fig. 3, our results are more natural with better sub-surface scattering and soft shadows (nose, eyes, chin). The visualization requires an expensive rendering step which computes the interactions between the lighting and albedo. Our results, on the other hand, already take the lighting into account. This offers several advantages. First, our rendering is much faster than Yamaguchi *et al.* Second, our final rendering is differentiable, which can be used as a component in any other learning task. The rendering step is not differentiable for Yamaguchi *et al.*

7. Additional comparisons

We compare to the single image relighting approach of Zhou *et al.* [8]. Since this approach uses a spherical approximation to represent the lights, our relighting results are more natural and reflect the target lighting better, see Fig 4. In addition, image relighting approaches do not model view-dependent effects and thus, cannot change the head pose.

We also compare the quality of our geometry reconstructions with MoFA [4] in Tab. 3. MoFA was trained on our training dataset. The errors are computed on the BU-3DFE dataset [7]. We use the same metric as used in Tewari *et al.* [3], where a dense correspondence map between the re-

	Layers	Output
Src Texture(512x512x3), Src Normal (512x512x3)	Concat	*unnamed*
↑	Conv2D (kernel 3x3x32, stride 1) + ReLU	skip1
↑	MaxPool (kernel 3x3, stride 2)	*unnamed*
↑	Conv2D (kernel 3x3x64, stride 1) + ReLU	skip2
↑	MaxPool (kernel 3x3, stride 2)	*unnamed*
↑	Conv2D (kernel 3x3x64, stride 1) + ReLU	skip3
↑	MaxPool (kernel 3x3, stride 2)	*unnamed*
↑	Conv2D (kernel 3x3x128, stride 1) + ReLU	skip4
↑	MaxPool (kernel 3x3, stride 2)	*unnamed*
↑	Conv2D (kernel 3x3x128, stride 1) + ReLU	skip5
↑	MaxPool (kernel 3x3, stride 2)	*unnamed*
↑	Conv2D (kernel 3x3x256, stride 1) + ReLU	skip6
↑	MaxPool (kernel 3x3, stride 2)	*unnamed*
↑	Conv2D (kernel 3x3x256, stride 1) + ReLU	skip7
↑	MaxPool (kernel 3x3, stride 2)	*unnamed*
↑	Conv2D (kernel 3x3x512, stride 1) + ReLU	skip8
↑	MaxPool (kernel 3x3, stride 2)	*unnamed*
↑, target light (1x3)	Concat	*unnamed*
↑	Conv2D (kernel 3x3x512, stride 1) + ReLU	*unnamed*
↑	UpSampling (kernel 2x2)	*unnamed*
↑, skip8	Concat	*unnamed*
↑	Conv2D (kernel 3x3x256, stride 1) + ReLU	*unnamed*
↑	UpSampling (kernel 2x2)	*unnamed*
↑, skip7	Concat	*unnamed*
↑	Conv2D (kernel 3x3x256, stride 1) + ReLU	*unnamed*
↑	UpSampling (kernel 2x2)	*unnamed*
↑, skip6	Concat	unnamed
↑	Conv2D (kernel 3x3x128, stride 1) + ReLU	*unnamed*
↑	UpSampling (kernel 2x2)	*unnamed*
↑, skip5	Concat	unnamed
↑	Conv2D (kernel 3x3x128, stride 1) + ReLU	*unnamed*
↑	UpSampling (kernel 2x2)	*unnamed*
↑, skip4	Concat	*unnamed*
↑	Conv2D (kernel 3x3x64, stride 1) + ReLU	*unnamed*
↑	UpSampling (kernel 2x2)	*unnamed*
↑, skip3	Concat	*unnamed*
↑	Conv2D (kernel 3x3x64, stride 1) + ReLU	*unnamed*
↑	UpSampling (kernel 2x2)	*unnamed*
↑, skip2	Concat	*unnamed*
↑	Conv2D (kernel 3x3x32, stride 1) + ReLU	*unnamed*
↑	UpSampling (kernel 2x2)	*unnamed*
↑, skip1	Concat	*unnamed*
↑, Target Normal (512x512x3)	Concat	*unnamed*
↑	Conv2D (kernel 3x3x32, stride 1) + ReLU	*unnamed*
↑	Conv2D (kernel 3x3x32, stride 1) + ReLU	*unnamed*
↑	Conv2D (kernel 3x3x32, stride 1) + ReLU	*unnamed*
↑	Conv2D (kernel 3x3x32, stride 1) + ReLU	*unnamed*
↑	Conv2D (kernel 3x3x32, stride 1) + ReLU	*unnamed*
↑	Conv2D (kernel 3x3x3, stride 1) + ReLU	Target Texture

Table 2. Reflectance network details.

	W/o $\mathcal{L}_\text{I}$, W/o $\mathcal{L}_\text{L}$	W/o $\mathcal{L}_\text{I}$, W $\mathcal{L}_\text{L}$	W $\mathcal{L}_\text{I}$, W/o $\mathcal{L}_\text{L}$	W $\mathcal{L}_\text{I}$, W $\mathcal{L}_\text{L}$
Same pose (Si-MSE)	0.0012 (σ=0.0009)	0.0008 (σ=0.0006)	0.0007 (σ=0.0006)	**0.0007** (σ=0.0006)
Different pose (Si-MSE)	0.0013 (σ=0.0011)	0.0009 (σ=0.0010)	0.0009 (σ=0.0009)	**0.0008** (σ=0.0009)

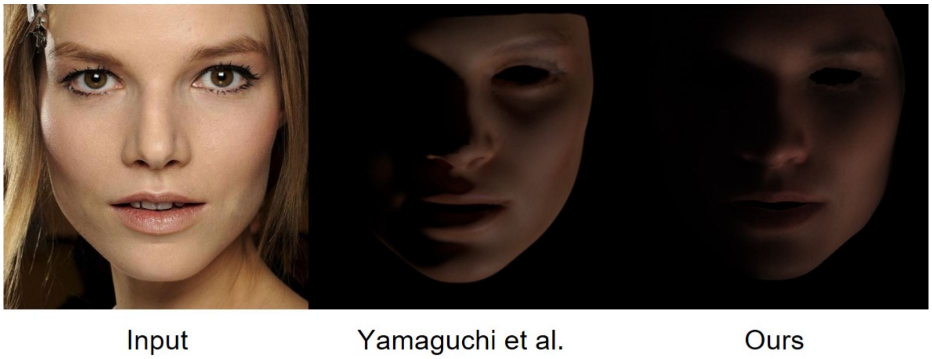


Figure 3. Comparison of our approach with the method of Yamaguchi *et al*. [6]. We obtain more natural results, since our reflectance is not limited to just the diffuse and specular components. Note that the results have been rendered under similar lights in order to be visually similar.

	MoFA	Ours
3D error	1.93(σ=0.39)	2.01(σ=0.38)

Table 3. Geometric reconstruction error (in mm) on the BU-3DFE dataset [7].

constructed and ground-truth mesh templates is precomputed for evaluation. The translation, scale, and orientation of the reconstructed and ground-truth meshes are aligned before computing the errors. While MoFA achieves slightly better numbers, its reflectance quality is limited as demonstrated in the main paper.

References

[1] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proc. ECCV*, 2016. 1

[2] Abhimitra Meka, Christian Haene, Rohit Pandey, Michael Zollhoefer, Sean Fanello, Graham Fyffe, Adarsh Kowdle, Xueming Yu, Jay Busch, Jason Dourgarian, Peter Denny, Sofien Bouaziz, Peter Lincoln, Matt Whalen, Geoff Harvey, Jonathan Taylor, Shahram Izadi, Andrea Tagliasacchi, Paul Debevec, Christian Theobalt, Julien Valentin, and Christoph Rhemann. Deep reflectance fields - high-quality facial reflectance field inference from color gradient illumination. In *ACM Transactions on Graphics (Proceedings SIGGRAPH)*, 2019. 1

[3] 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 Conference on Computer Vision and Pattern Recognition*, pages 10812–10822, 2019. 3

[4] Ayush Tewari, Michael Zollhöfer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Theobalt Christian. MoFA: Model-based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction. In *ICCV*, 2017. 3

[5] Tim Weyrich, Wojciech Matusik, Hanspeter Pfister, Bernd Bickel, Craig Donner, Chien Tu, Janet McAndless, Jinho Lee, Addy Ngan, Henrik Wann Jensen, and Markus Gross. Analysis of human faces using a measurement-based skin reflectance model. *ACM Trans. on Graphics (Proc. SIGGRAPH 2006)*, 2006. 1

[6] Shuco 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 Transactions on Graphics*, 2018. 1, 5

[7] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and M. J. Rosato. A 3d facial expression database for facial behavior research. In *International Conference on Automatic Face and Gesture Recognition (FGR06)*, pages 211–216, 2006. 3, 5

[8] Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David W. Jacobs. Deep single-image portrait relighting. In *Proc. ICCV*, 2019. 3, 6

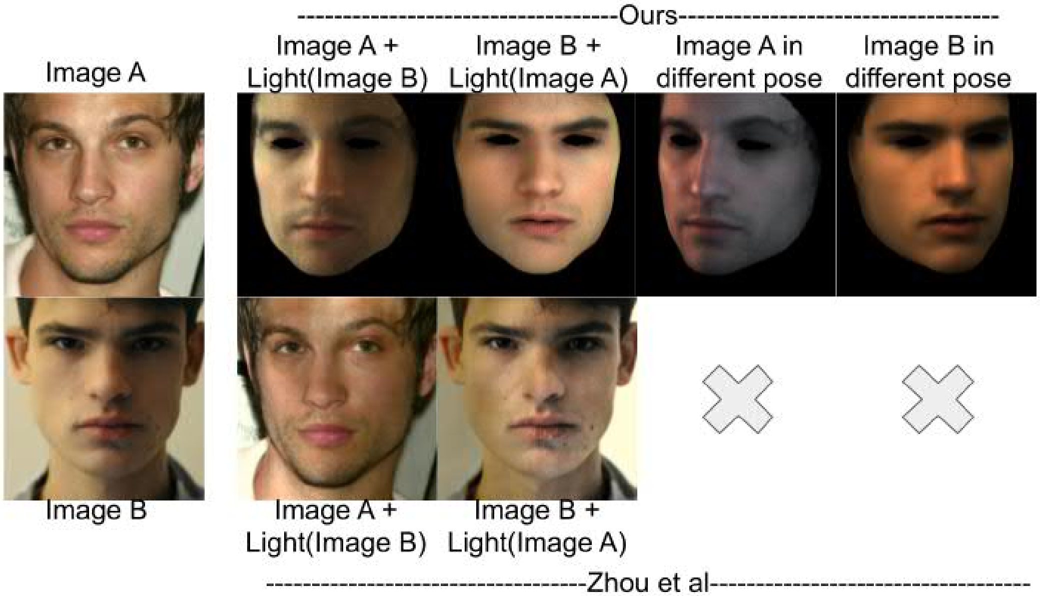

-----------------------------------Ours-----------------------------------
Image A
Image A + Light(Image B)
Image B + Light(Image A)
Image A in different pose
Image B in different pose
Image B
Image A + Light(Image B)
Image B + Light(Image A)
-----------------------------------Zhou et al-----------------------------------

Figure 4. Comparison of our approach with the single image relighting method of Zhou *et al*. [8]. Our relighting results better capture the light in the reference image. Zhou *et al*. do not capture view-dependent effects and can thus not change the head pose.