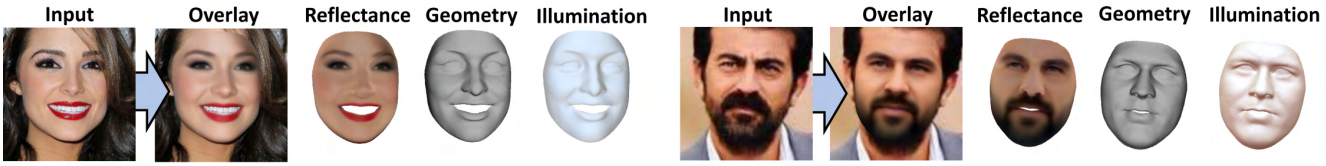
Self-supervised Multi-level Face Model Learning for Monocular Reconstruction at over 250 Hz — Supplemental Material —

Ayush Tewari[1,2] Michael Zollhöfer[1,2,3] Pablo Garrido[1,2] Florian Bernard[1,2]
Hyeongwoo Kim[1,2] Patrick Pérez[4] Christian Theobalt[1,2]
[1]MPI Informatics [2]Saarland Informatics Campus [3]Stanford University [4]Technicolor



Our novel monocular reconstruction approach estimates high-quality facial geometry, skin reflectance (including facial hair) and incident illumination at over 250 Hz. A trainable multi-level face representation is learned jointly with the feed forward inverse rendering network. End-to-end training is based on a self-supervised loss that does not require dense ground truth.

In this supplemental document we provide more details on our approach. More specifically, we discuss robust training (Sec. 1) of our architecture, we provide the runtime (Sec. 2) for training and testing, we discuss different corrective spaces that we tested (Sec. 3), we perform a study (Sec. 4) on the number of required corrective parameters, and show more results (Sec. 5).

Please note that all shown colored reconstructions of our approach are not textured with the input image, but show the color due to the reconstructed reflectance and/or illumination, for e.g. see Fig. 1. The underlying model is a multi-level face model with a base level employing a 3DMM (Base), and a final level that adds optimal learned per-vertex shape and reflectance correctives (Final). Skin reflectance is represented with a low-dimensional coefficient vector of size $580 = 500 + 80$ (500 coefficients for the correctives and 80 coefficients for the 3DMM). Thus, skin reflectance is stored using only 2.3KB (one float per coefficient). The shape is represented based on a low-dimensional vector of size $644 = 500 + 80 + 64$ (500 coefficients for the correctives, 80 coefficients for the shape identity in the 3DMM, and 64 blendshape coefficients). Thus, the geometry is also stored using only 2.6KB (one float per coefficient). In total, the complete reconstruction is efficiently represented with less than 5KB. This can be exploited for compression, i.e., if the reconstruction has to be transmitted over the internet.

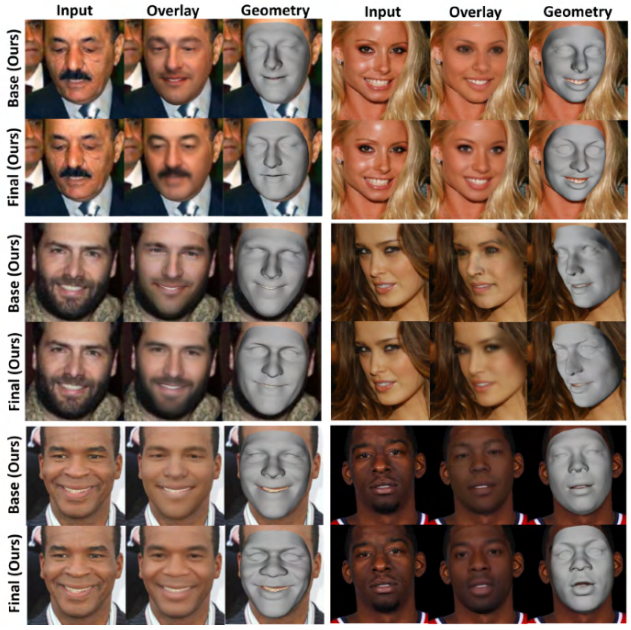


Figure 1. Jointly learning a multi-level model improves the estimated geometry and reflectance compared to the underlying 3DMM on the coarse base level. Note the better aligning nose, lips and the reconstructed facial hair.

1. Training

In the following, we describe how we train our novel architecture end-to-end based on a two stage training strategy. Training the face regressor and the corrective space jointly, in one go, turned out to be challenging. For robust training, we first pretrain our network up to the base level for 200k iterations with a learning rate of 0.01. We implemented our approach using *Caffe* [3]. Training is

based on *AdaDelta* with a batch size of 5. We use fixed weights $w_\bullet$ in all our experiments. For training the base level we use the following weights $w_{\text{photo}} = 1.9, w_{\text{reg}} = 0.00003, w_{\text{rstd}} = 0.002, w_{\text{smo}} = 0.0, w_{\text{ref}} = 0.0, w_{\text{glo}} = 0.0$ and $w_{\text{sta}} = 0.0$. In addition, we only use the photometric alignment term on the base level. Afterwards, we finetune our complete network for 190k iterations end-to-end with a learning rate of 0.001 for the base level network, 0.005 for the geometry correctives network and 0.01 for the reflectance correctives network. For finetuning, in all our experiments we instantiate our loss based on the following weights $w_{\text{photo}} = 0.2, w_{\text{reg}} = 0.003, w_{\text{rstd}} = 0.002, w_{\text{smo}} = 3.2 \cdot 10^4, w_{\text{ref}} = 13, w_{\text{glo}} = 80, w_{\text{sta}} = 0.08$, and use 500 correctives for both geometry and reflectance. Please note, the illumination estimate for rendering the base and final model is not shared between the two levels, but independently regressed. This is due to the fact that a different illumination estimate might be optimal for the coarse and final reconstruction due to the shape and skin reflectance correctives. During finetuning, all weights associated with the correctives receive a higher learning rate ($\times 100$) than the pretrained layers. We found that this two stage strategy enables robust and efficient training of our architecture.

2. Runtime Performance

We evaluate the runtime performance of our approach on an Nvidia GTX TITAN Xp graphics card. Training our novel monocular face regressor takes 16 hours. A forward pass of our trained convolutional face parameter regressor takes less than 4 ms. Thus, our approach performs monocular face reconstruction at more than 250 Hz.

3. Evaluation of the Corrective Space

Our corrective space is based on (potentially non-linear) mappings $\mathcal{F}_\bullet : \mathbb{R}^C \rightarrow \mathbb{R}^{3N}$ that map the C-dimensional corrective parameter space onto per-vertex corrections in shape or reflectance. The mapping $\mathcal{F}_\bullet(\delta_\bullet|\Theta_\bullet)$ is a function of $\delta_\bullet \in \mathbb{R}^C$ that is parameterized by $\Theta_\bullet$. In the linear case, one can interpret $\Theta_\bullet$ as a matrix that spans a subspace of the variations, and $\delta_\bullet$ is the coefficient vector that reconstructs a given sample using the basis. Let $\mathcal{L}_i(\delta) = \mathbf{M}_i\delta + \mathbf{b}_i$ be an affine/linear mapping and $\Theta_\bullet^{[i]}$ stack all trainable parameters, i.e., the trainable matrix $\mathbf{M}_i$ and the trainable offset vector $\mathbf{b}_i$. We tested different linear and non-linear corrective spaces, see Fig. 2. The affine/linear model (Linear) is given as:

$$\mathcal{F}_\bullet(\delta_\bullet|\Theta_\bullet^{[0]}) = \mathcal{L}_0(\delta_\bullet) \ . \tag{1}$$

However, in general we do not assume $\mathcal{F}_\bullet$ to be affine/linear. For this evaluation, given the non-linear function Ψ, which in our case is a ReLU non-linearity, we define a corrective model with two affine/linear layers and one non-linearity (One Non-Linearity):

$$\mathcal{F}_\bullet(\delta_\bullet|\Theta_\bullet^{[0]}, \Theta_\bullet^{[1]}) = \mathcal{L}_1(\Psi(\mathcal{L}_0(\delta_\bullet))) \ . \tag{2}$$

In addition, we also define a corrective model with three affine/linear layers and two non-linearities (Two Non-Linearities):

$$\mathcal{F}_\bullet(\delta_\bullet|\Theta_\bullet^{[0]}, \Theta_\bullet^{[1]}, \Theta_\bullet^{[2]}) = \mathcal{L}_2(\Psi(\mathcal{L}_1(\Psi(\mathcal{L}_0(\delta_\bullet))))) \ . \tag{3}$$

As can be seen in Fig. 2, the results obtained by the affine/linear and non-linear models are visually quite similar. The affine/linear model obtains a lower photometric re-rendering error, see Fig. 3. Thus, we decided for the simpler affine/linear model and use it in all our experiments.

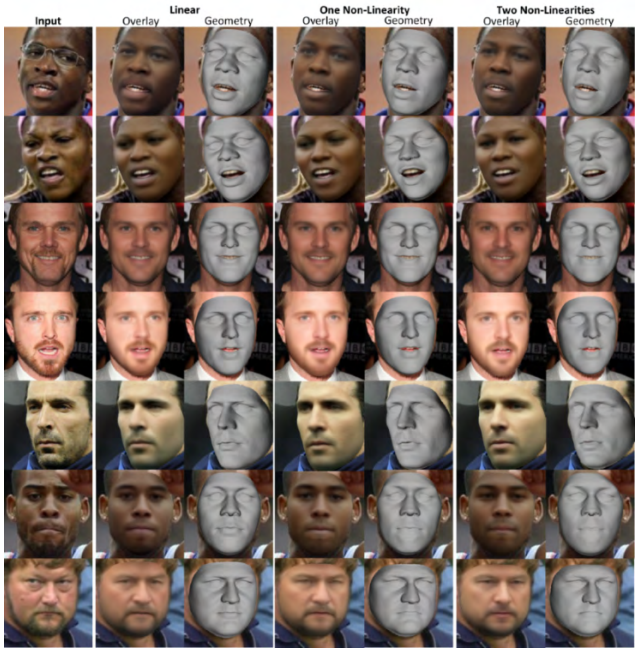


Figure 2. Comparison of linear and non-linear corrective spaces.

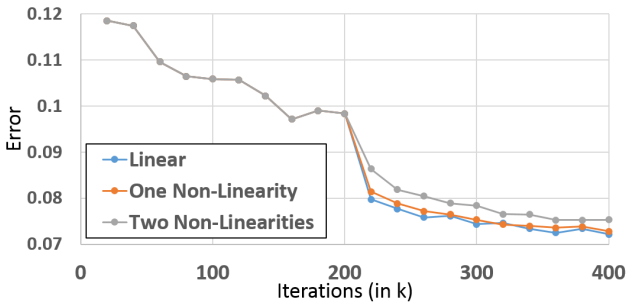


Figure 3. The affine/linear model obtains a lower photometric re-rendering error than the two tested non-linear corrective spaces. Thus, we use the affine/linear model in all our experiments.

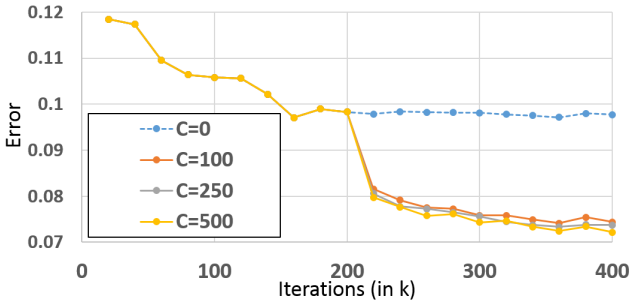


Figure 4. We trained our network with $C = 0, 100, 250$ and 500 corrective parameters for shape and reflectance. Our networks with correctives significantly improve upon the baseline network that only uses the 3DMM ($C = 0$) in terms of the photometric re-rendering error. The corrective basis with $C = 500$ parameters achieves the lowest photometric re-rendering error. Thus, we use this network for all further experiments.

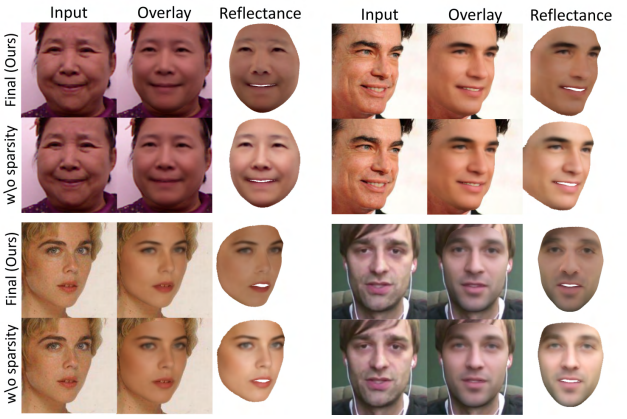


Figure 5. Removing reflectance sparsity leads to shading information being wrongly explained by reflectance variation.

Figure 6. Both local sparsity and global constancy terms help in obtaining plausible reflectance estimates.

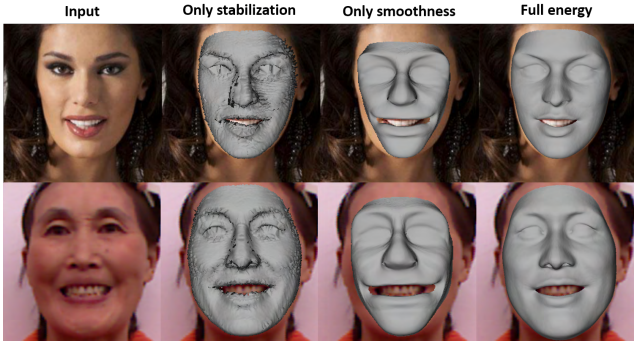


Figure 7. Both corrective smoothness and stabilization terms help in obtaining nice geometry estimates.

4. Additional Evaluation

We also report the photometric re-rendering error on a test set (2k images) for a varying number of corrective parameters. To this end, we follow our two-stage training schedule and trained our network with $C = 0, 100, 250$ and 500 corrective parameters for shape and reflectance, see Fig. 4. Our networks with correctives significantly improve upon the baseline network that only uses the 3DMM ($C = 0$) in terms of the photometric re-rendering error. The best results in terms of the lowest photometric re-rendering error are obtained by our network with $C = 500$ additional corrective parameters for shape and reflectance. Thus, we use this network for all further experiments.

We also performed an ablation study to evaluate the contribution on reconstruction quality of the different objectives of our self-supervised loss function. More specifically, we evaluated two different variations of our self-supervised loss: 1) We removed all corrective shape regularizers and 2) We removed all reflectance sparsity priors. Removing the shape regularizers (stabilization and smoothness) leads to a complete failure during training, since all corrective per-vertex displacements are independent and thus the reconstruction problem is severely underconstrained. If the reflectance sparsity priors (local and global) are removed, the network can still be trained and the overlayed reconstructions look plausible, see Fig. 5, but all shading information is wrongly explained by reflectance variation. Thus, both the used shape and reflectance priors are necessary and drastically improve reconstruction quality. An ablative analysis of the individual terms of the reflectance and shape regularizers can be found in Figs. 6 and 7. We also evaluate our estimated reflectance on synthetic images. We generate a synthetic training corpus using 80 reflectance vectors of the 3DMM. We only allow the base model of the network to regress 40 reflectance parameters, which allows the final model to learn the reflectance correctives between the base model and the ground truth, see Fig. 8. We also perform a quantitative evaluation by computing per-pixel RGB distances between the rendered reflectance image and

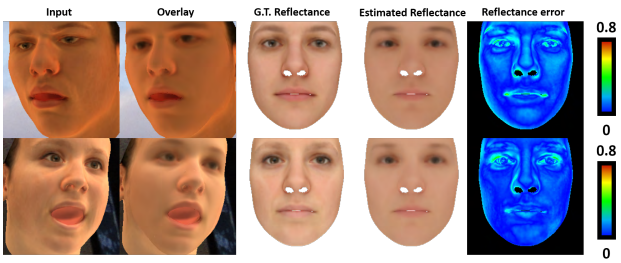


Figure 8. Our estimated reflectance is close to the ground truth reflectance for synthetic images.

the ground truth (we compensate for global shifts in reflectance). We render these images using the mean face geometry in a canonical pose. We obtain an error of 0.072 (averaged over 1k test images), which shows the accuracy of our predictions.

5. Additional Results and Comparisons

We show more results (Fig. 13) and comparisons to current optimization-based (Figs. 10 and 11) and learning-based (Figs. 14 and 15) state-of-the-art approaches. Note, in the comparison to Tewari et al. [7], we compare to their weakly supervised training, which, similar to our approach, uses a sparse set of facial landmarks for supervision. Our approach obtains high reconstruction quality and compares favorably to all of these state-of-the-art techniques. In particular, it is able to reconstruct colored surface reflectance and it robustly handles even challenging cases, such as occlusions by facial hair and make-up. For a detailed discussion of the differences to the other approaches we refer to the main document. In addition, we show more examples of limitations (Fig. 12), such as external occluders, which are baked into the recovered model. We also show more examples of the photometric re-rendering error (Fig. 9) and show that the corrective space improves the regressed shape and reflectance estimate (Fig. 1).

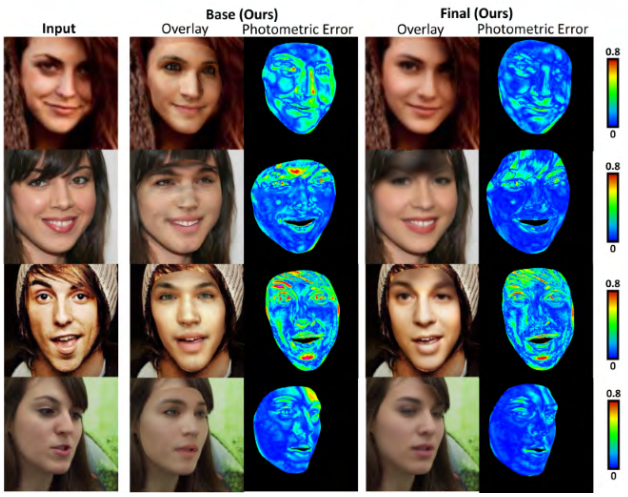


Figure 9. Euclidean photometric error in RGB space, each channel in $[0, 1]$. Our final results significantly improve fitting quality.

References

[1] J. Booth, E. Antonakos, S. Ploumpis, G. Trigeorgis, Y. Panagakis, and S. Zafeiriou. 3d face morphable models "in-the-wild". In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[2] P. Garrido, M. Zollhöfer, D. Casas, L. Valgaerts, K. Varanasi, P. Pérez, and C. Theobalt. Reconstruction of personalized 3D face rigs from monocular video. *ACM Transactions on Graphics*, 35(3):28:1–15, June 2016.

[3] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[4] E. Richardson, M. Sela, and R. Kimmel. 3D face reconstruction by learning from synthetic data. In *3DV*, 2016.

[5] E. Richardson, M. Sela, R. Or-El, and R. Kimmel. Learning detailed face reconstruction from a single image. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[6] M. Sela, E. Richardson, and R. Kimmel. Unrestricted facial geometry reconstruction using image-to-image translation. *arxiv*, 2017.

[7] A. Tewari, M. Zollöfer, H. Kim, P. Garrido, F. Bernard, P. Perez, and T. Christian. MoFA: Model-based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction. In *ICCV*, 2017.

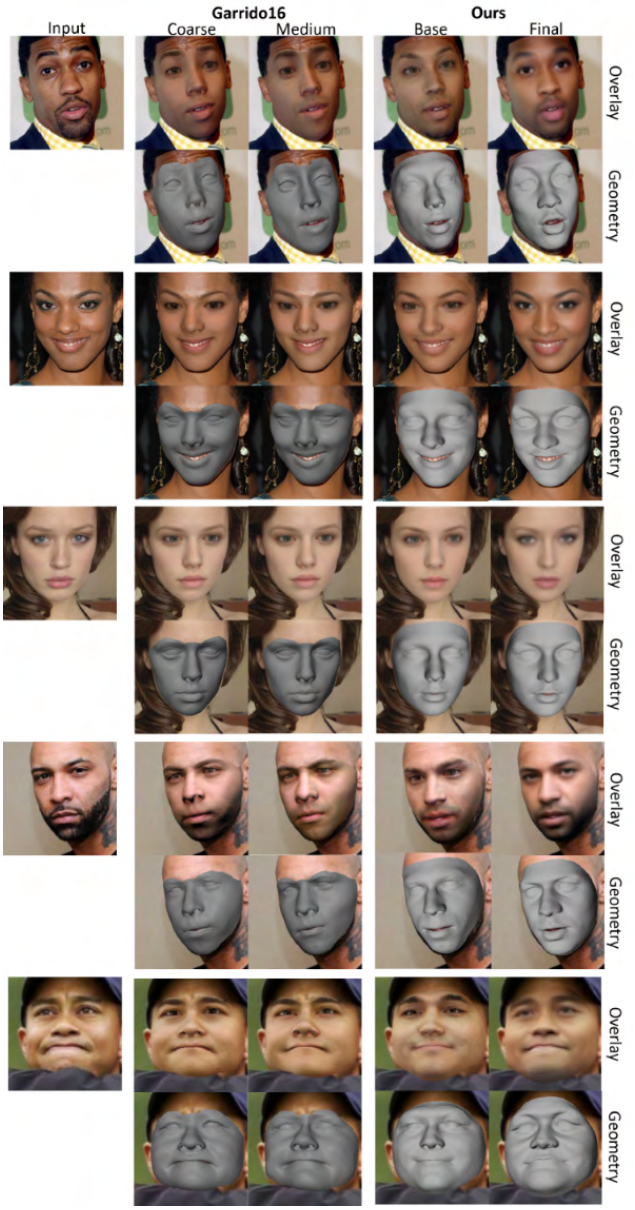


Figure 10. Comparison to Garrido et al. [2]. We achieve higher quality reconstructions, since our jointly learned model allows leaving the restricted 3DMM subspace and generalizes better than a corrective space based on manifold harmonics.

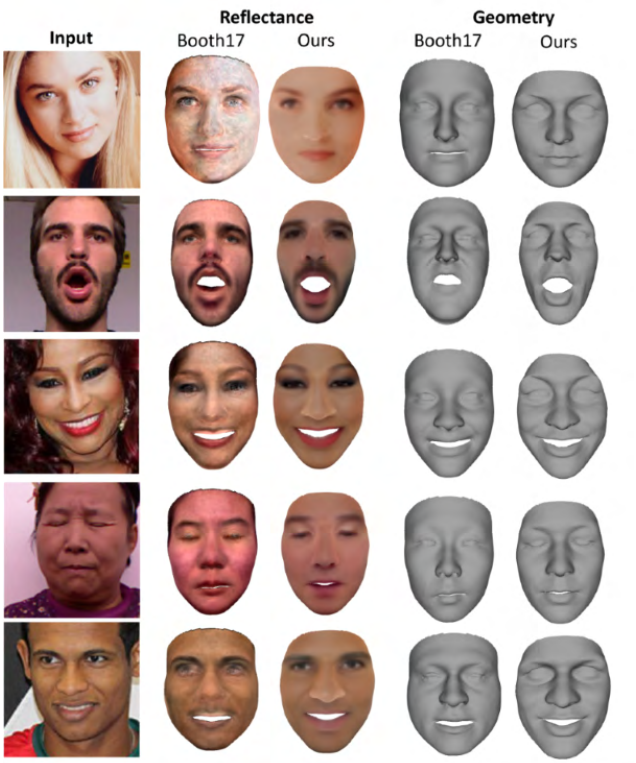


Figure 11. In contrast to the 'in-the-wild' texture model of Booth et al. [1] that contains shading, our approach yields a reflectance model. In addition, our learned optimal corrective space goes far beyond the restricted low-dimensional geometry subspace that is commonly employed.

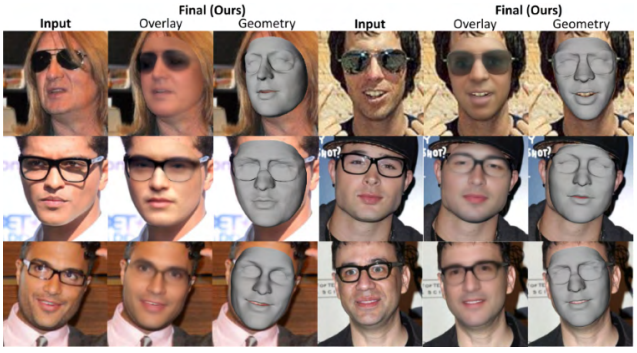


Figure 12. External occluders might be baked into the correctives.

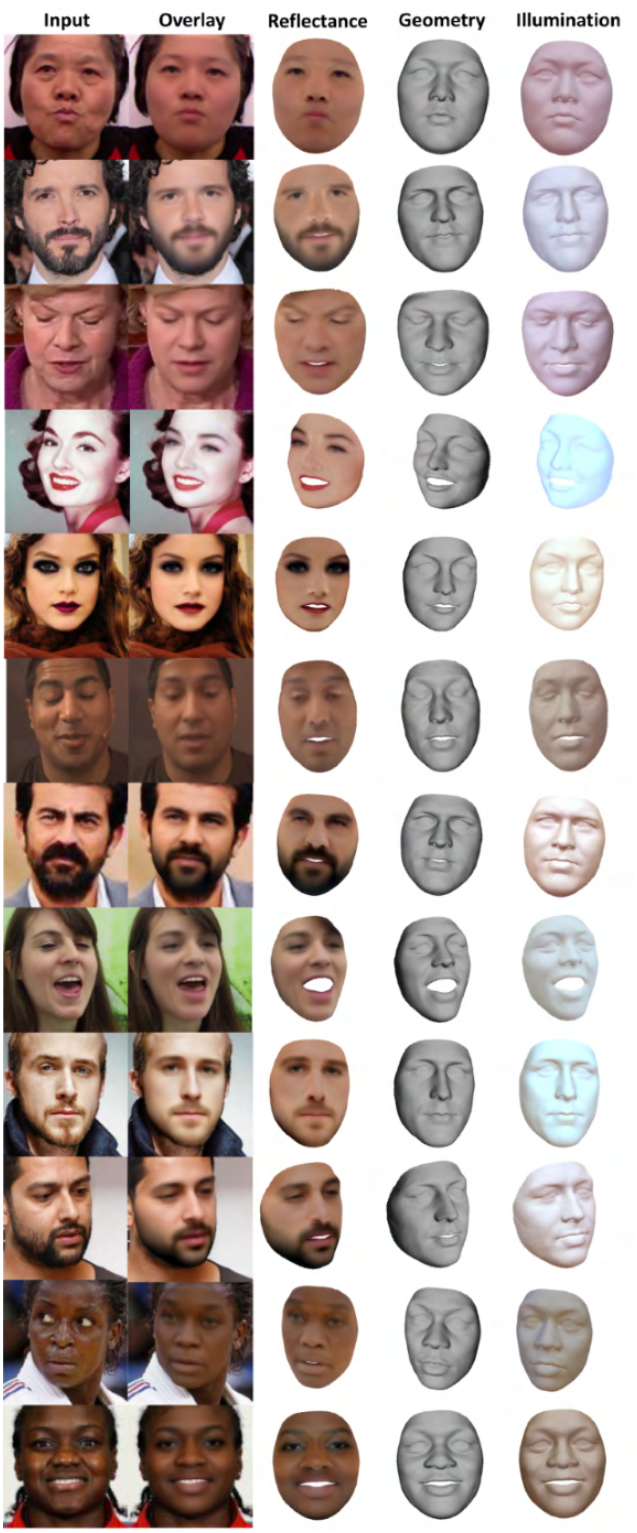


Figure 13. Our approach allows for high-quality reconstruction of facial geometry, reflectance and incident illumination from just a single monocular color image. Note the reconstructed facial hair, e.g., the beard, reconstructed make-up, and the eye lid closure, which are outside the restricted 3DMM subspace.

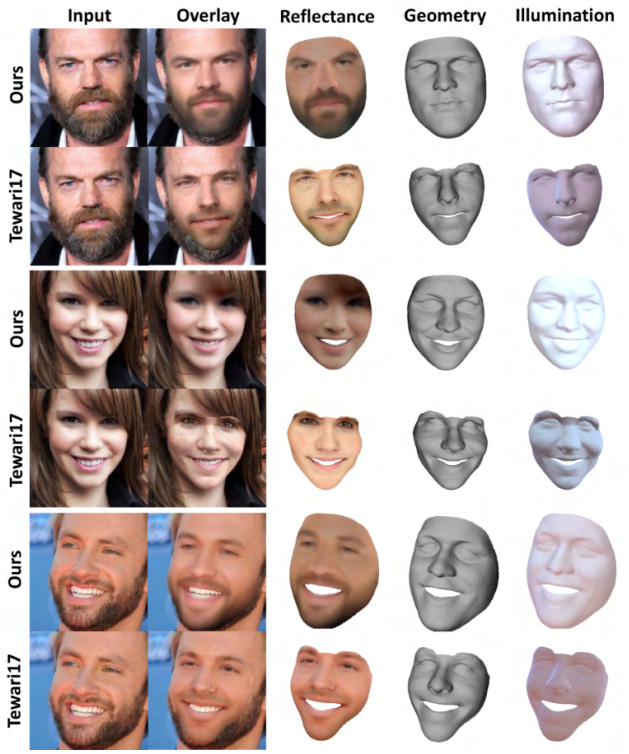


Figure 14. Comparison to Tewari et al. [7]. We achieve higher quality in terms of geometry and reflectance, since our jointly trained model allows leaving the restricted 3DMM subspace. This prevents surface shrinkage due to unexplained facial hair.

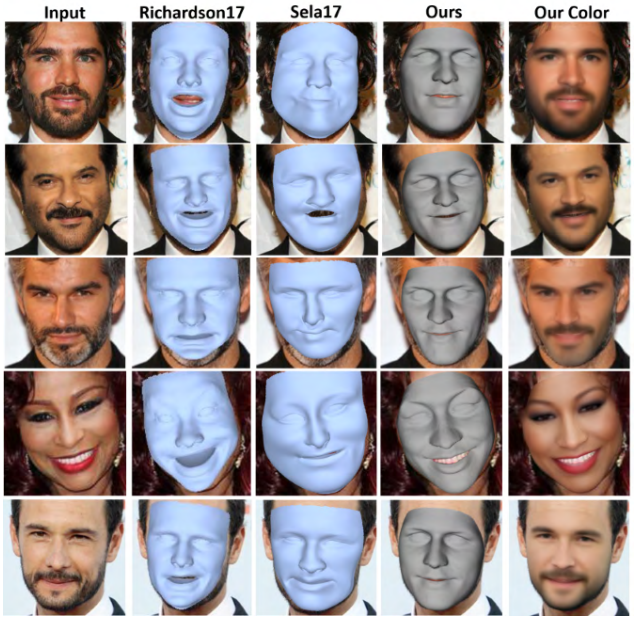


Figure 15. Comparison to Richardson et al. [4, 5] and Sela et al. [6]. They obtain impressive results for faces within the span of the synthetic training corpus, but suffer for out-of-subspace shape and reflectance variations, e.g., people with beards. Our approach is not only robust to facial hair and make-up, but learns to reconstruct such variations based on the jointly trained model.