InverseFaceNet: Deep Monocular Inverse Face Rendering — Supplemental Material —

Hyeongwoo Kim [1, 2] Michael Zollhöfer [1, 2, 3] Ayush Tewari [1, 2]
Justus Thies [4] Christian Richardt [5] Christian Theobalt [1, 2]

[1] Max-Planck-Institute for Informatics [2] Saarland Informatics Campus
[3] Stanford University [4] Technical University of Munich [5] University of Bath

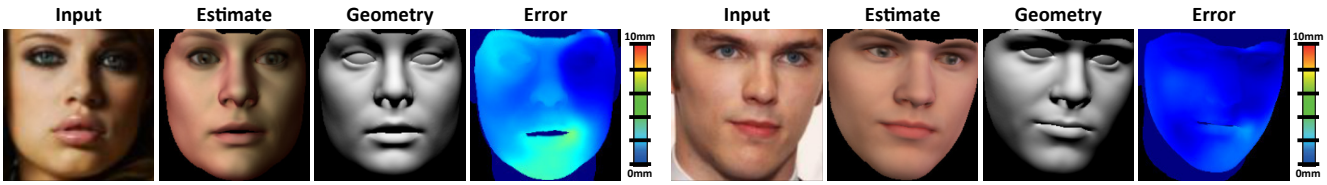

Figure 1. Our single-shot deep inverse face renderer *InverseFaceNet* obtains a high-quality geometry, reflectance and illumination estimate from just a single input image. We jointly recover the facial pose, shape, expression, reflectance and incident scene illumination. *From left to right:* the input photo, our estimated face model, its geometry, and the pointwise Euclidean geometry error compared to Garrido et al. [3].

This document provides additional discussion, comparisons and results for our approach. We discuss technical details on how we evaluate the self-supervised bootstrapping approach using a synthetic test set in Section 1. Also, we provide quantitative and qualitative comparisons in Sections 2 and 3, respectively, to further demonstrate the accuracy and effectiveness of our approach. Finally, we demonstrate the robustness of our approach on a wide range of challenging face images in Section 4.

1. Self-Supervised Bootstrapping

To evaluate the strength of our self-supervised bootstrapping step in the training loop, we use synthetic validation images, as it is difficult to acquire the ground-truth parameters for real-world images. This section explains in more detail the evaluation shown in Section 7.2 and Figure 4 of the main document, in particular the image sets used for training, bootstrapping and validation.

We first generate a set of 50,000 training images with a parameter distribution that has little variation; the mouth, for instance, is not opening much. We then modify the distribution with a bias and more variation in face expression and color to simulate real-world images, and generate two sets of 5,000 images each for bootstrapping and validation. The difference between the image sets is clearly visible in Figure 2.

In this evaluation, InverseFaceNet uses a set of 5,000 images without the corresponding parameters for self-supervised bootstrapping. The initialization, used weights and number of training iterations are explained in the main document. For evaluation, we visualize the face parameters estimated from premature to fully domain-adapted networks, i.e., along the bootstrapping iterations, in the testing phase as shown in Figure 3. In addition, we compute the model-

Figure 2. Images used for training (top) and testing (bottom) in the bootstrapping evaluation. The synthetic examples for testing and bootstrapping are sampled from a wider distribution than the training images. Thus, there is more variation in face shape, expression and color, such as mouth opening and colored illumination.

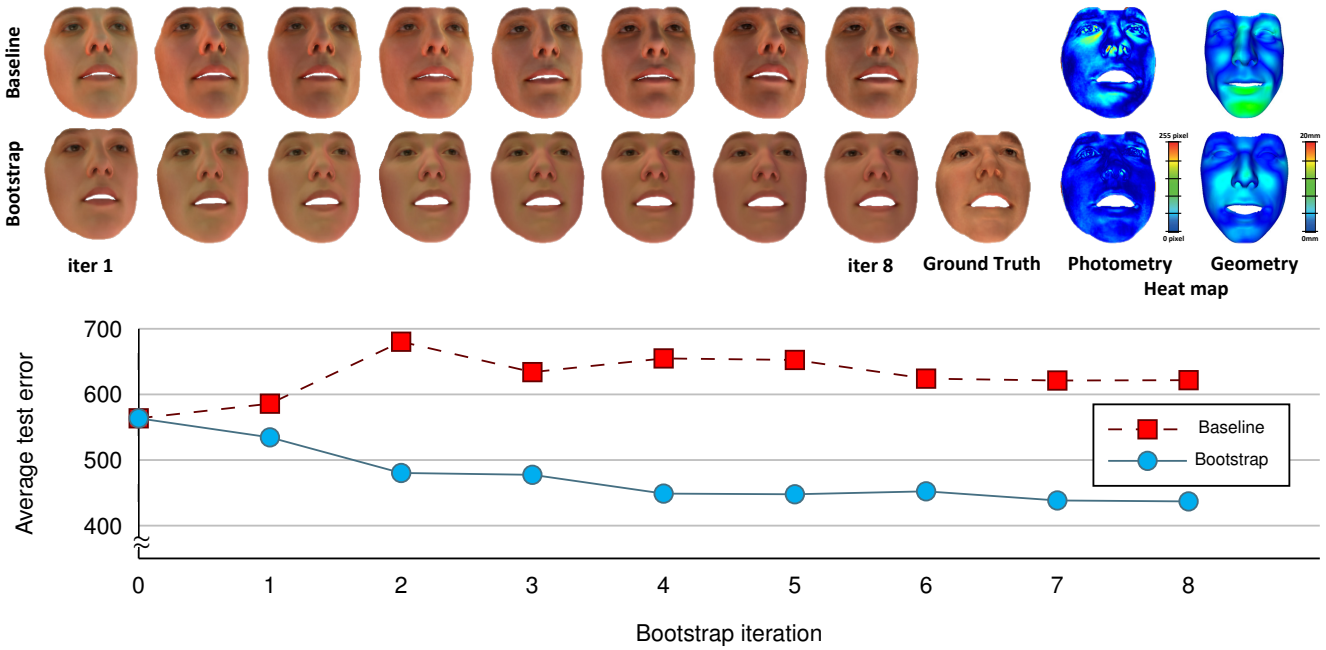

Figure 3. Comparison of baseline and bootstrapping approaches on a synthetic test corpus with higher parameter variation than in the used training corpus (also synthetic). **Top:** Reconstructions of an unseen input image after different numbers of bootstrapping iterations. Notice how the reconstructions with bootstrapping gradually converge towards the ground-truth face model (right), e.g., opening the mouth, while the baseline approach does not improve visibly over time. The last two columns visualize the photometric ($2\times$ scaled) and geometric errors at the final bootstrapping step. The mean photometric error is 16.74 pixels in the L_1-norm distance for the baseline method, and 12.11 pixels after bootstrapping. The Hausdorff distance is 2.56 mm and 2.01 mm for the baseline method and after bootstrapping, respectively. **Bottom:** Model-space parameter loss for the baseline and bootstrapping approaches. While our domain-adaptive bootstrapping approach continuously decreases the error by adapting the parameter distribution based on a higher variation corpus without available ground truth, the baseline network overfits to the training data and fails to generalize to the unseen data.

space parameter loss of the validation image set. With the visual and numeric metrics, the performance of bootstrapped InverseFaceNet is compared against a vanilla AlexNet without bootstrapping. The decrease of the model-space loss via bootstrapping substantiates that the parameter distribution of the training set is automatically adapted to better match the image set used for bootstrapping, i.e., more mouth opening is added to the initial training set. This is in contrast to the regressed face with a closed mouth, and non-decreasing model-space parameter error by the baseline method, which estimates the best possible parameters only within the initial training set. On the basis of this evaluation, we conclude that our self-supervised bootstrapping approach results in better generalization to unseen input images in the real-world scenario. For an evaluation on real-world face images, we refer to the main document.

2. Additional Quantitative Evaluation

In addition to the quantitative evaluation on *FaceWarehouse* [1] in the main document, we here evaluate and compare our approach on a challenging video sequence (300 frames of *Volker* [13]). As ground-truth geometry, we use the high-quality binocular reconstructions of Valgaerts et al. [13]. Our approach outperforms Tewari et al. [11] on this sequence, and comes close to the optimization-based results of Garrido et al. [3], which is orders of magnitude slower than our approach (2 minutes vs our 9.4 ms).

Table 1. Quantitative evaluation of the geometric accuracy on 300 frames of the *Volker* dataset [13].

	Ours	Garrido et al. [3]	Tewari et al. [11]
Error	2.10 mm	1.96 mm	2.94 mm
SD	0.42 mm	0.35 mm	0.28 mm

3. Qualitative Evaluation

In the following, we show additional results and comparisons that unfortunately did not fit into the limited space of the main document. Specifically, we compare to the approaches of Richardson et al. [8], Sela et al. [9], Jackson et al. [5], Tran et al. [12], Tewari et al. [11], Garrido et al. [2] and Garrido et al. [3] on a variety of challenging face datasets, including *LFW* (Labeled Faces in the Wild) [4], *300-VW* (300 Videos in the Wild) [10], *CelebA* [6], *FaceWarehouse* [1], *Volker* [13] and *Thomas* [2]. The results are shown in Figure 4 to Figure 9.

We compare to the results of Richardson et al.'s 'CoarseNet' [8] and Sela et al.'s aligned template mesh [9]

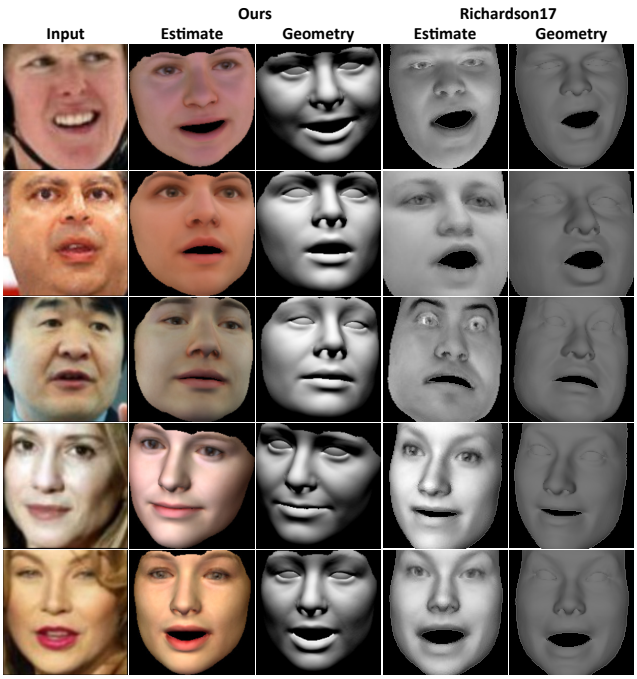

Figure 4. Qualitative comparison to Richardson et al. [8] on *LFW* [4]. Note that our reconstruction results are colored, and better fit the face shape and mouth expressions of the input images.

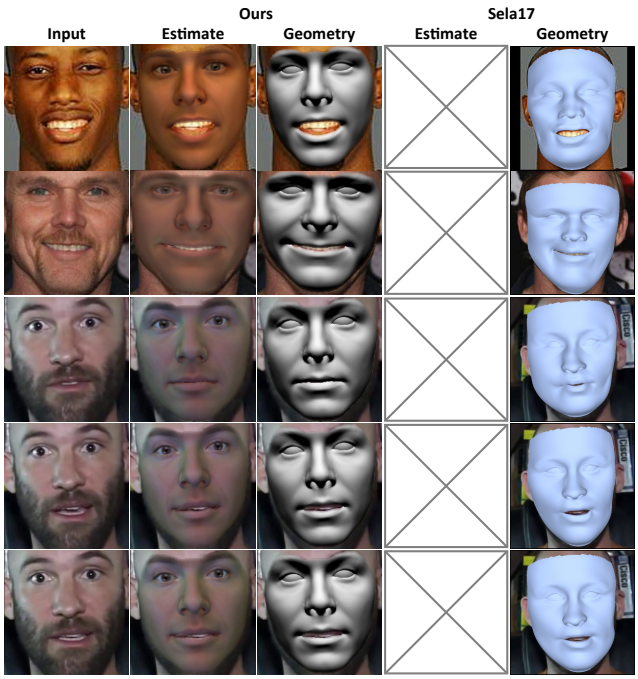

Figure 5. Qualitative comparison to Sela et al. [9] on *CelebA* [6] (top 2 rows) and *300-VW* [10] datasets. From top to bottom: our approach reconstructs facial reflectance and reliable shape, while theirs does not recover reflectance or illumination, and suffers from global shape distortion.

as we are interested in comparing the reconstructed parametric face models. As can be seen in Figures 4 and 5, we obtain similar or even higher quality results than these two state-of-the-art approaches. Note that their approaches do not require landmarks for initial cropping, but they are significantly slower due to their iterative regression strategy [8] or the involved non-rigid registration [9], and do not recover color reflectance. In contrast, our approach provides a one-shot estimate of all face model parameters.

Jackson et al. [5] recover coarse volumetric reconstructions, and do not reconstruct facial appearance or illumination (Figure 6). In contrast to Richardson et al. [7, 8] and Jackson et al. [5], our approach obtains an estimate of the colored skin reflectance and illumination.

The approach of Tran et al. [12] is targeted at face recognition, and thus does not recover the facial expression and illumination (Figure 7). Our results are comparable to Tewari et al. [11], but we avoid the geometric shrinking seen in Figure 8. Notice that their estimated geometry is visibly thinner than the input faces. Our approach also obtains similar quality results (Figure 9) as the optimization-based approaches by Garrido et al. [2, 3], while being several orders of magnitude faster. For a detailed discussion, we refer the reader to the main document.

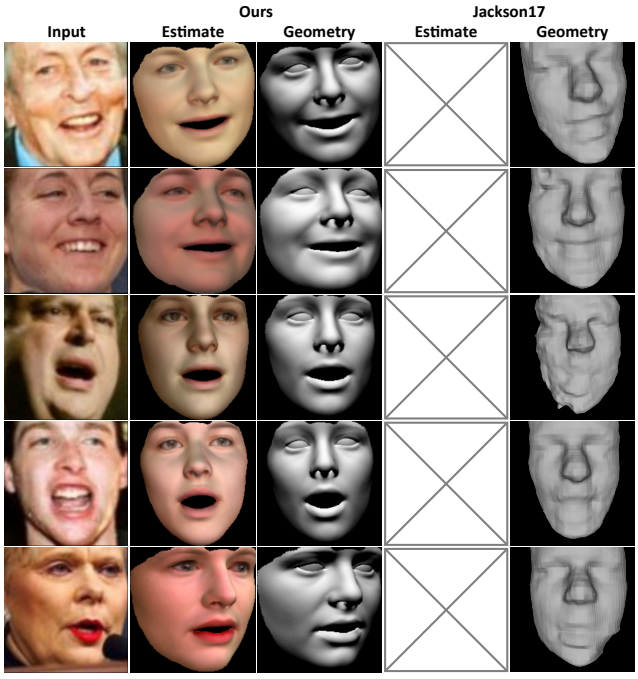

Figure 6. Qualitative comparison to Jackson et al. [5] on *LFW* [4]. Our reconstruction results include reflectance and illumination, and better fit the face shape.

4. Additional Results

Our approach works well even for the challenging images shown in Figure 10 with different head orientations (rows one and two), challenging expressions (rows three to five), and variation in skin reflectance (rows four to six). Our approach provides perceptually more plausible reconstructions

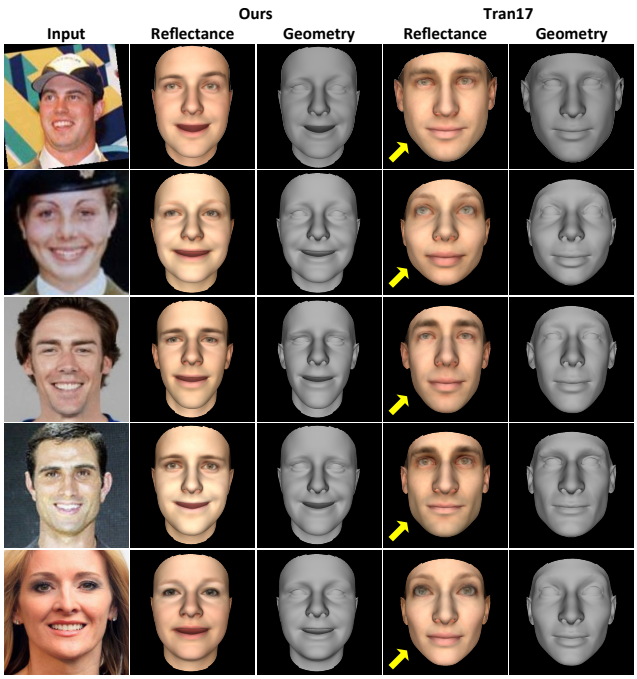

Figure 7. Qualitative comparison to Tran et al. [12] on images of the *CelebA* [6] (top 3 rows) and *LFW* [4] (rest) datasets: our approach reconstructs expressions, while theirs cannot recover this dimension (arrows).

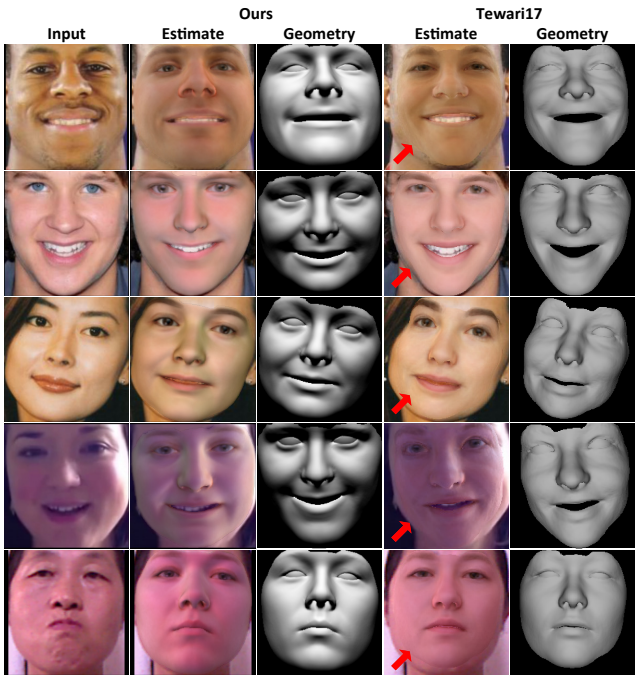

Figure 8. Qualitative comparison to Tewari et al. [11]: our approach reconstructs the facial outline accurately, while theirs suffers from shrinking artifacts (arrows).

due to our novel model-space loss and the self-supervised bootstrapping that automatically adapts the parameter distribution to match the real world. For more results and a detailed discussion, we refer the reader to the main document.

References

[1] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou. FaceWarehouse: A 3D facial expression database for visual computing. *IEEE TVCG*, 20(3):413–425, 2014.

[2] P. Garrido, L. Valgaerts, C. Wu, and C. Theobalt. Reconstructing detailed dynamic face geometry from monocular video. *ACM ToG*, 32(6):158:1–10, 2013.

[3] 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 ToG*, 35(3):28:1–15, 2016.

[4] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[5] A. S. Jackson, A. Bulat, V. Argyriou, and G. Tzimiropoulos. Large pose 3D face reconstruction from a single image via direct volumetric CNN regression. In *ICCV*, 2017.

[6] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *ICCV*, pages 3730–3738, 2015.

[7] E. Richardson, M. Sela, and R. Kimmel. 3D face reconstruction by learning from synthetic data. In *3DV*, pages 460–469, 2016.

[8] E. Richardson, M. Sela, R. Or-El, and R. Kimmel. Learning detailed face reconstruction from a single image. In *CVPR*, pages 5553–5562, 2017.

[9] M. Sela, E. Richardson, and R. Kimmel. Unrestricted facial geometry reconstruction using image-to-image translation. In *ICCV*, pages 1585–1594, 2017.

[10] J. Shen, S. Zafeiriou, G. G. Chrysos, J. Kossaifi, G. Tzimiropoulos, and M. Pantic. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *ICCV Workshops*, pages 1003–1011, 2015.

[11] A. Tewari, M. Zollhöfer, H. Kim, P. Garrido, F. Bernard, P. Pérez, and C. Theobalt. MoFA: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *ICCV*, pages 3735–3744, 2017.

[12] A. T. Tran, T. Hassner, I. Masi, and G. Medioni. Regressing robust and discriminative 3D morphable models with a very deep neural network. In *CVPR*, pages 1493–1502, 2017.

[13] L. Valgaerts, C. Wu, A. Bruhn, H.-P. Seidel, and C. Theobalt. Lightweight binocular facial performance capture under uncontrolled lighting. *ACM ToG*, 31(6):187:1–11, 2012.

Figure 9. Qualitative comparison to optimization-based approaches [2, 3] on the *Thomas* dataset [2].

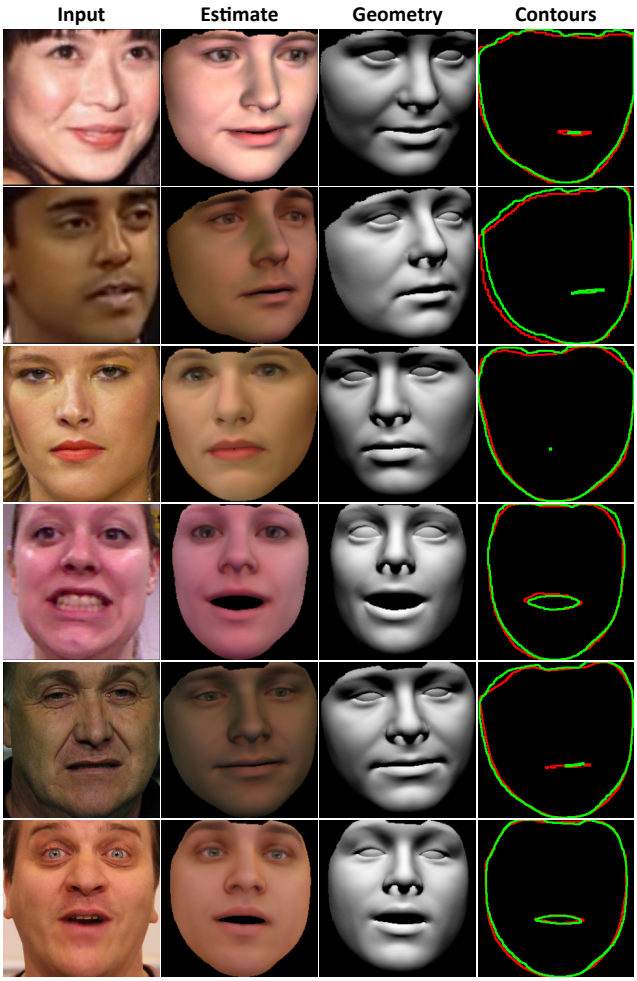

Figure 10. Qualitative results on several datasets. Left to right: input image, our estimated face model and geometry, and contours (red: input mask, green: ours). Top to bottom: *LFW* [4], *300-VW* [10], *CelebA* [6], *FaceWarehouse* [1], *Volker* [13] and *Thomas* [2]. Our approach achieves high-quality reconstructions of geometry as well as skin reflectance from just a single input image.