Supplemental Material for
Photorealistic Facial Texture Inference Using Deep Neural Networks

Shunsuke Saito[*†§] Lingyu Wei[*†§] Liwen Hu[*†] Koki Nagano[‡] Hao Li[*†‡]

[*]Pinscreen [†]University of Southern California [‡]USC Institute for Creative Technologies

Appendix I. Additional Results

Our main results in the paper demonstrate successful inference of high-fidelity texture maps from unconstrained images. The input images have mostly low resolutions, non-frontal faces, and the subjects are often captured in challenging lighting conditions. We provide additional results with pictures from the annotated faces-in-the-wild (AFW) dataset [10] to further demonstrate how photorealistic pore-level details can be synthesized using our deep learning approach. We visualize in Figure 9 the input, the intermediate low-frequency albedo map obtained using a linear PCA model, and the synthesized high-frequency albedo texture map. We also show several views of the final renderings using the Arnold renderer [13]. We refer to the accompanying video for additional rotating views of the resulting textured 3D face models.

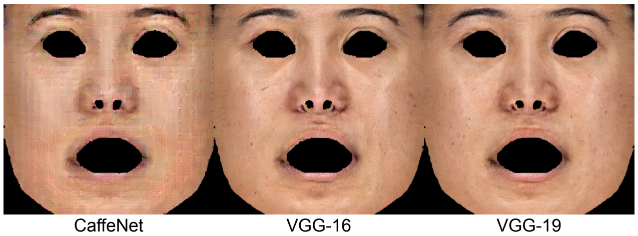

Figure 1: Comparison between different convolutional neural network architectures.

Evaluation. As Figure 1 indicates, other deep convolutional neural networks can be used to extract mid-layer feature correlations to characterize multi-scale details, but it seems that deeper architectures produce fewer artifacts and higher quality textures. All three convolutional neural networks are pre-trained for classification tasks using images from the ImageNet object recognition dataset [4]. The results of the 8 layer CaffeNet [2] show noticeable blocky artifacts in the synthesized textures and the ones from the 16 layer VGG [12] are slightly noisy around boundaries, while the 19 layer VGG network performs the best.

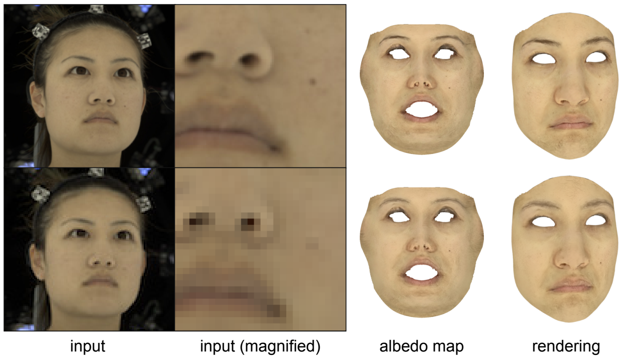

Figure 2: Even for largely downsized image resolutions, our algorithm can produce fine-scale details while preserving the person's similarity.

We also evaluate the robustness of our inference framework for downsized image resolutions in Figure 2. We crop a diffuse lit face from a Light Stage capture [5]. The resulting image has 435×652 pixels and we decrease its resolution to 108×162 pixels. In addition to complex skin pigmentations, even the tiny mole on the lower left cheek is properly reconstructed from the reduced input image using our synthesis approach.

Comparison. We provide in Figure 3 additional visualizations of our method when using the closest feature correlation, unconstrained linear combinations, and convex combinations. We also compare against a PCA-based model fitting [3] approach and the state-of-the-art visio-lization framework [9]. We notice that only our proposed technique using convex combinations is effective in generating mesoscopic-scale texture details. Both visio-lization and the PCA-based model result in lower frequency textures and less similar faces than the ground truth. Since our inference also fills holes, we compare our synthesis technique with a general inpainting solution for predicting unseen face regions. We test with the widely used PatchMatch [1] technique as illustrated in Figure 4. Unsurprisingly, we observe unwanted repeating structures and semantically wrong fillings since this method is based on low-level vision cues.

§- indicates equal contribution

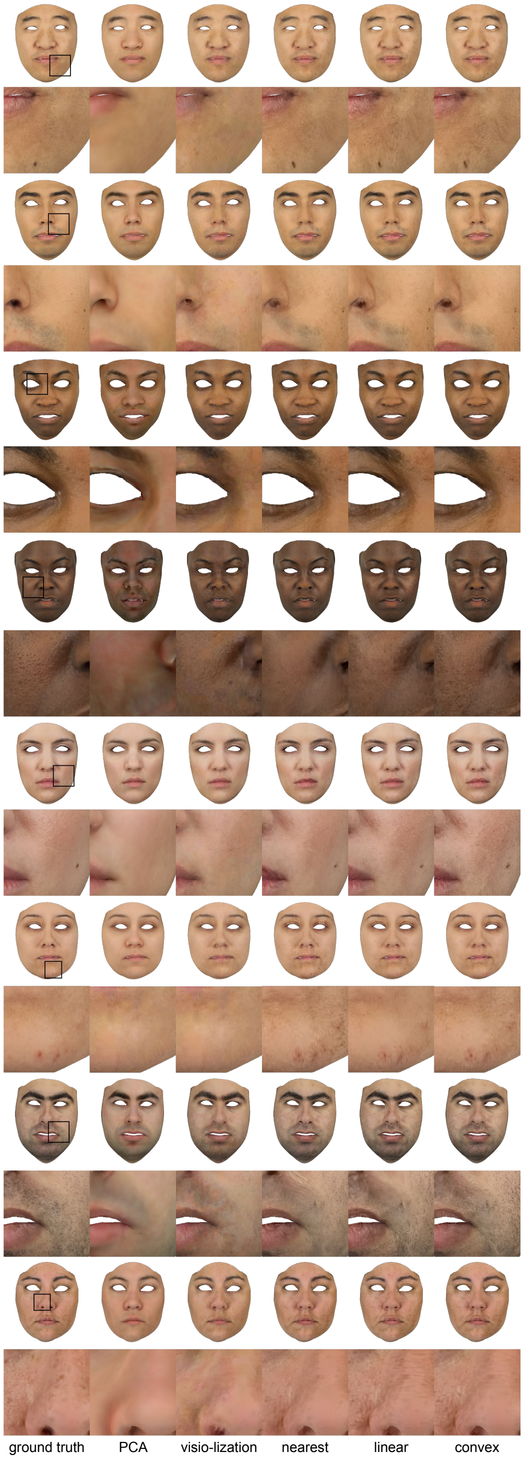

Figure 3: Comparison between PCA-based model fitting [3], visio-lization [9], our method using the closest feature correlation, our method using unconstrained linear combinations, and our method using convex combinations.

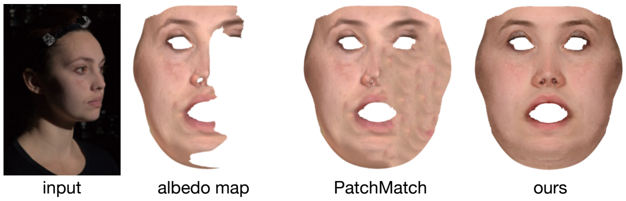

Figure 4: Comparison with PatchMatch [1] on a partial input data.

Appendix II. User Study Details

This section gives further details and discussions about the two user studies presented in the paper. Figures 5 and 7 also show the user interfaces that we deployed on Amazon Mechanical Turk (AMT).

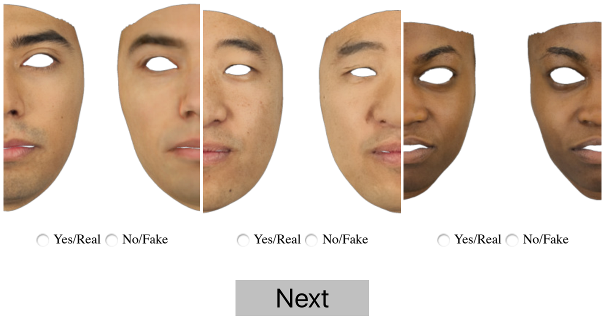

Figure 5: AMT user interface for user study A.

User Study A: Photorealism and Alikeness. We recall that method (1) is obtained using PCA model fitting, (2) is visio-lization, (3) is our method using the closest feature correlation, (4) our method using unconstrained linear combinations, and (5) our method using convex combinations. We randomly select 11 photographs from the Chicago Face Database [8] for this evaluation, and downsize/crop their resolution from 2444 × 1718 to 512 × 512 pixels. At the end we apply one iteration of Gaussian filtering of kernel size 5 to remove all the facial details. We show the turkers a left and right side of a face and inform them that the left side is always the ground truth. The right side has a 50% chance of being computer generated. The task consists of deciding whether the right side is "real" and identical to the ground truth, or "fake". We summarize our analysis with the box plot in Figure 6 using 150 turkers. Only 65.6% of the real images on the right have been correctly marked as "real". This is likely due to the fact that the turkers know that only 50% are real, which affects their confidence in distinguishing real ones from digital reconstructions. Results based on PCA model fittings have few occurrences of false positives, which indicates that turkers can reliably identify

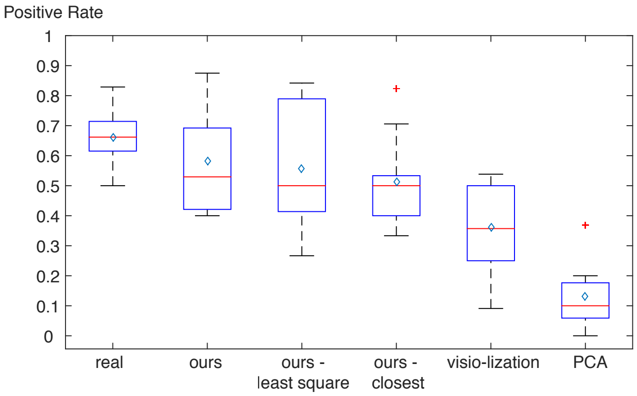

Figure 6: Box plots of 150 turkers rating whether the image looks realistic and identical to the ground truth. Each plot contains the positive rates for 11 subjects in the Chicago Face Database.

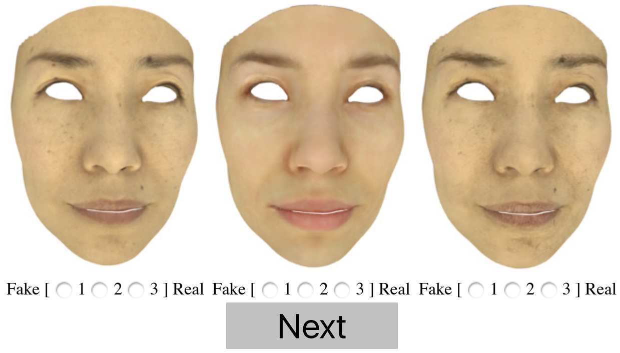
Fake [○ 1 ○ 2 ○ 3] Real
Fake [○ 1 ○ 2 ○ 3] Real
Fake [○ 1 ○ 2 ○ 3] Real
Next

Figure 7: AMT user interface for user study B.

them. The generated faces using visio-lization also appear to be less realistic and similar than those obtained using variations of our method. For the variants of our method, (3), (4), and (5), we measure similar means and medians, which indicates that non-technical turkers have a hard time distinguishing between them. However, method (4) has a higher chance than variant (3) to be marked as "real", and the convex combination method (5) achieves the best results as they occasionally notice artifacts in (4). Notice how the left and right sides of the face are swapped in the AMT interface to prevent users from comparing texture transitions.

User Study B: Our method vs. Light Stage Capture. We used three subjects (due to limited availability) and randomly perturbed their head rotations to produce more rendering samples. To obtain a consistent geometry for the Light Stage data, we warped our mesh to fit their raw scans using non-rigid registration [6]. All examples are rendered using full-on diffuse lighting and our input image to the inference framework has a resolution of 435×652 pixels. We asked 100 turkers to sort 3 sets of renderings, one for each of the three subjects. Surprisingly, we found that 56% think that ours are superior in terms of realism than those obtained from the Light Stage, 74% of the turkers found the results of (2) to be more realistic than (3), and 72% think that ours is superior to (3). We believe that over 20% of the turkers who believe that (3) is better than the two other methods are outliers. After removing these outliers, we still have 57% who believe that our results are more photoreal than those from the Light Stage. We believe that our synthetically generated fine-scale details confuse the turkers for subjects that have smoother skins in reality.

Appendix III. Frequently Asked Questions

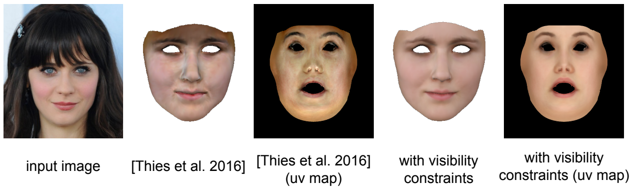
input image
[Thies et al. 2016]
[Thies et al. 2016] (uv map)
with visibility constraints
with visibility constraints (uv map)

Figure 8: The effect of using visibility constraints when estimating the PCA-based albedo map.

Q: What is the role of the visibility term in Section 3?
A: Our PCA-based model fitting does share similarities with [14], but we introduce an important visibility term E_c based on the visibility of pixels $p \in \mathcal{M}$, which is obtained using the two-stream segmentation network introduced in [11]. Without this term, it is not possible to recover facial appearances reliably in the presence of occlusions (due to hairstyles, hands, etc.), as shown in Fig. 8.

Q: Why does it make sense to represent local structures as Gramian matrices rather than as filter responses directly?
A: Multi-scale features including local structures are intrinsically represented by distributions of activations in a CNN. A theoretical proof that shows that synthesizing images using Gramian matrices is equivalent to minimizing the difference of feature distributions (Maximum Mean Discrepancy with second order polynomial kernels) between two images has been recently presented by Li et al [7]. Reconstructing filter responses F directly is not possible since they contain spatial information and linearly combining them would yield non-facial features.

Q: Why doesn't it make sense to train the network exclusively on a face dataset?
A: Negative (i.e., non-facial) samples are necessary for the network to discern between facial and non-facial features. In fact, a significant amount of non-facial images is needed in order to ensure that Gramian matrices of faces are embedded in a sufficiently low-rank manifold in feature space. If we only train the classification network with faces, the blending of multiple subjects will yield non-facial features. We have conducted extensive experiments and the use of face-only classification networks significantly underperform the use of a general one, such as VGG-19.

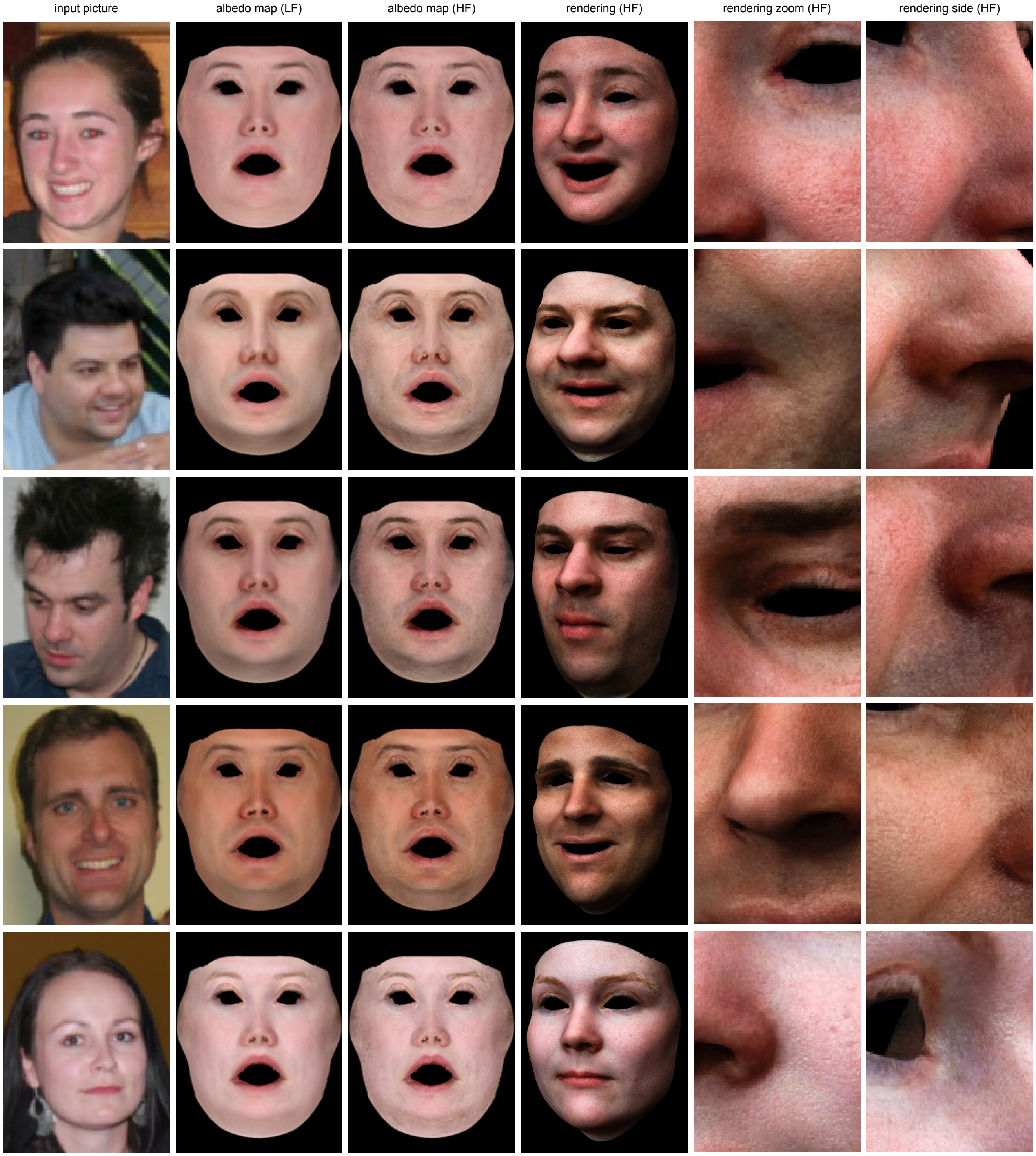

Figure 9: Additional results with images from the annotated faces-in-the-wild (AFW) dataset [10].

References

[1] C. Barnes, E. Shechtman, A. Finkelstein, and D. Goldman. Patchmatch: a randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics-TOG*, 28(3):24, 2009.

[2] Berkeley Vision and Learning Center. Caffenet, 2014. `https://github.com/BVLC/caffe/tree/master/models/bvlc_reference_caffenet`.

[3] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[5] A. Ghosh, G. Fyffe, B. Tunwattanapong, J. Busch, X. Yu, and P. Debevec. Multiview face capture using polarized spherical gradient illumination. *ACM Trans. Graph.*, 30(6):129:1–129:10, 2011.

[6] H. Li, B. Adams, L. J. Guibas, and M. Pauly. Robust single-view geometry and motion reconstruction. *ACM Transactions on Graphics (Proceedings SIGGRAPH Asia 2009)*, 28(5), 2009.

[7] Y. Li, N. Wang, J. Liu, and X. Hou. Demystifying neural style transfer. *arXiv preprint arXiv:1701.01036*, 2017.

[8] D. S. Ma, J. Correll, and B. Wittenbrink. The chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods*, 47(4):1122–1135, 2015.

[9] U. Mohammed, S. J. D. Prince, and J. Kautz. Visio-lization: Generating novel facial images. In *ACM SIGGRAPH 2009 Papers*, pages 57:1–57:8. ACM, 2009.

[10] D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *IEEE CVPR*, pages 2879–2886, 2012.

[11] S. Saito, T. Li, and H. Li. Real-time facial segmentation and performance capture from rgb input. In *ECCV*, 2016.

[12] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[13] Solid Angle, 2016. `http://www.solidangle.com/arnold/`.

[14] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *IEEE CVPR*, 2016.