Supplementary Material for Learning Detailed Face Reconstruction from a Single Image

1. Supplementary Qualitative Results

In Figure 1 we present additional qualitative comparisons. First, note how our network correctly infers the face alignment without any external information, producing similar alignment to the state-of-the-art alignment from [3]. The proposed method is able to produce fine facial details, as opposed to [6, 8], while being more robust to different expressions compared to the template-based method of [4]. Figure 2 shows additional reconstructions.

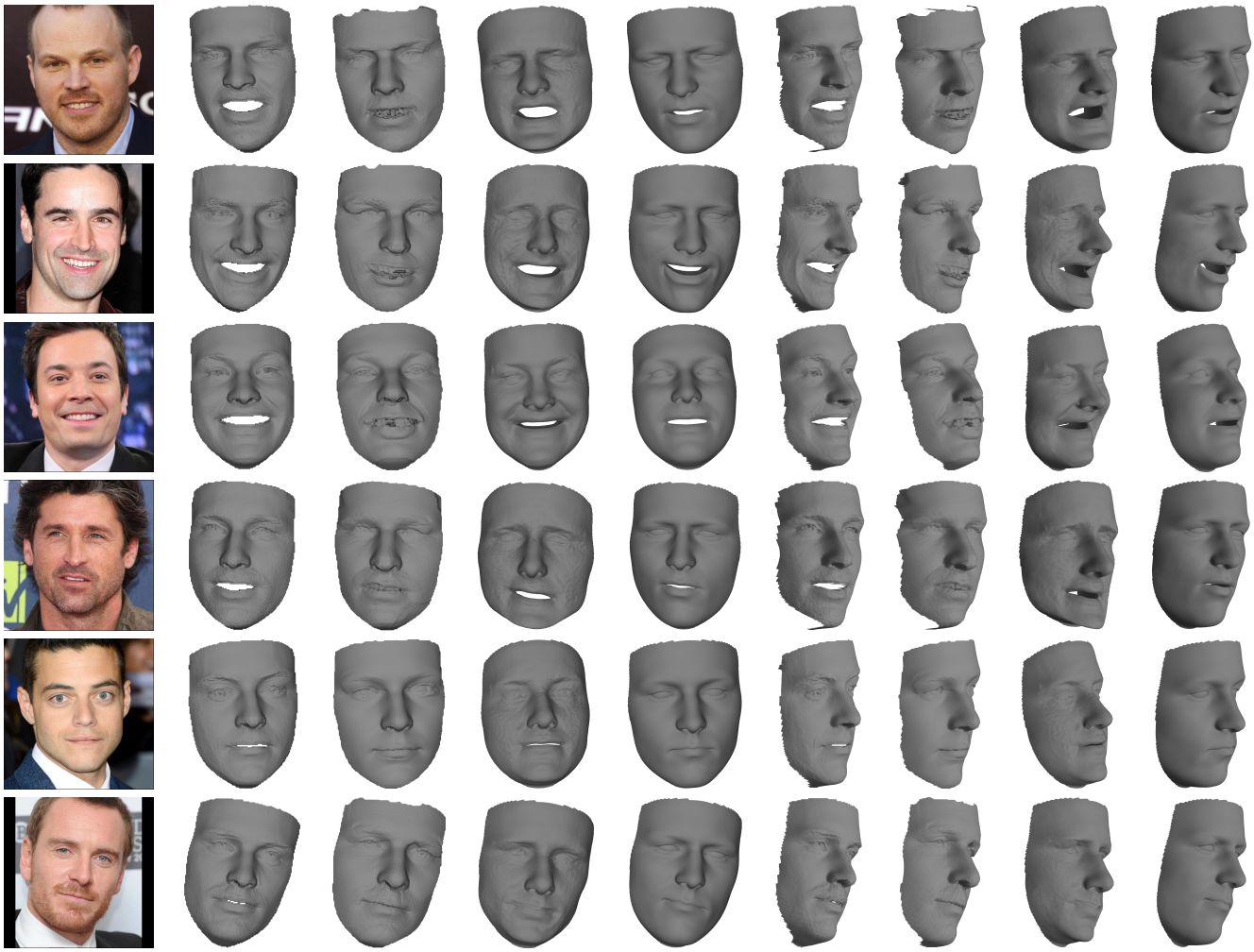

Figure 1: Additional qualitative results.

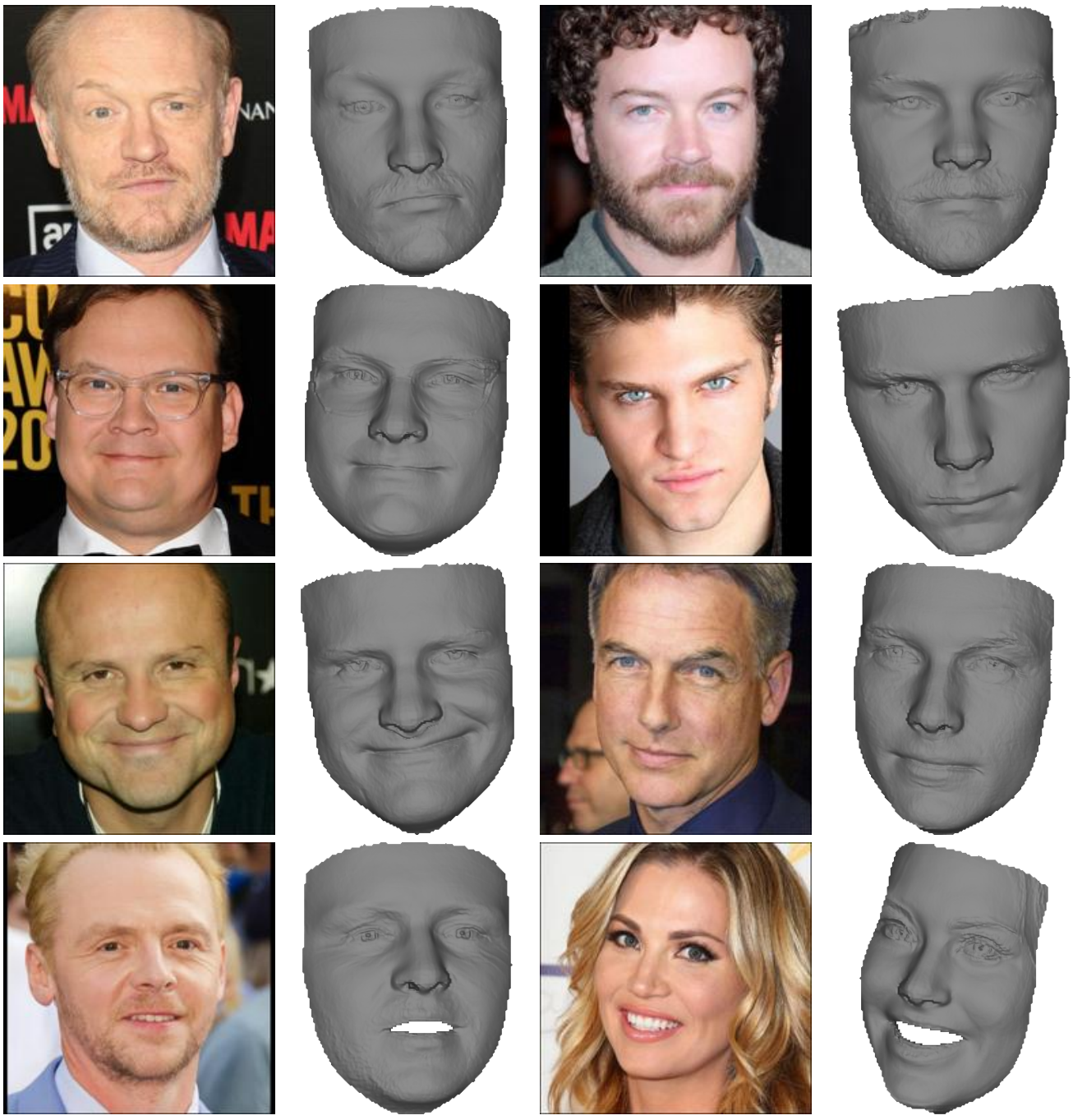

Figure 2: Additional qualitative results

2. Synthetic Data

In Figure 3 sampled synthetic examples are visualized, where random backgrounds are used. The rendered faces differ extensively in their geometry, texture, illumination, and reflectance properties.

Figure 3: Synthetic data samples.

2.1. Generalizing from Synthetic Data

The usage of synthetic data is becoming a prominent approach for learning problems on 3D data [7, 2, 1]. While data generation grants us flexibility and allows generation of large-scale datasets, there are still some limitations for synthetic data. As noted in [6], while training on synthetic faces generally produces plausible results on *in-the-wild* images, the network might fail when the input contains details that are not seen in the synthetic dataset, such as glasses or facial hair. In Figure 4 we show how our method handles such examples compared to [6]. From the results we can see that both methods show some robustness to eyeglasses, even when the eyes themselves are occluded. Regarding facial hair, one can see that a dominant beard might confuse both methods and make them misalign the chin or the mouth. Still, our method is able to produce more viable results than those of [6].

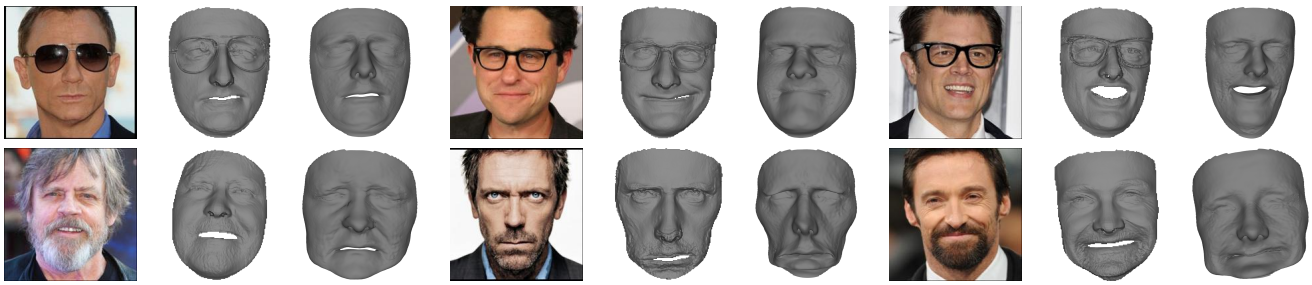

Figure 4: Generalization from synthetic data

3. Further Analysis

Next, we present a few additional experiments conducted on the different elements of the proposed network.

3.1. CoarseNet

A key property of iterative networks is convergence. To validate that CoarseNet meets this requirement, we calculated the average change in the output of CoarseNet between different iterations. As can be seen in Figure 5, the network indeed converges after a few iterations.

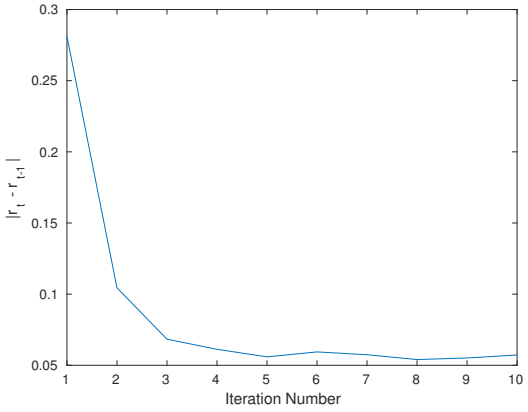

Figure 5: Convergence graph. The average MSE as a function of the iteration number.

3.2. FineNet

As detailed in the paper, FineNet starts with a set of convolutional blocks from the VGG Face Net [5], each followed by a pooling layer. The output of these blocks is then connected together to form a set of dense feature maps. While using more VGG blocks could possibly provide more data for the final prediction, it would also result in a larger network, increasing the overall training and runtime complexity. As Shape-from-Shading mainly relies on local features we choose to truncate the network after the third pooling operation. As shown in Figure 6, while using only a single block results in discontinuances and artifacts, using two or more blocks produces reasonable results.

Another interesting property of FineNet presented in the paper is its robustness to different input sizes, allowing it to extract more details when a high-resolution input is given. Note that the same does not hold for CoarseNet which uses a fixed averaging operator. However, as CoarseNet recovers only the coarse geometry it does not require a high-resolution input and would not benefit from it. In practice, we always scale the input given to CoarseNet to 200×200, while feeding FineNet with inputs in the desired scale.

Figure 6: Network depth. From left to right, the input image, and FineNet results using 1 to 4 blocks from VGG Face Net.

4. Supplementary Quantitative Analysis

Here, we demonstrate a quantitative analysis of the performance of the proposed method. The absolute error heat maps in Figure 7 present the typical error distribution of the proposed method versus those of other techniques [4, 6, 8].

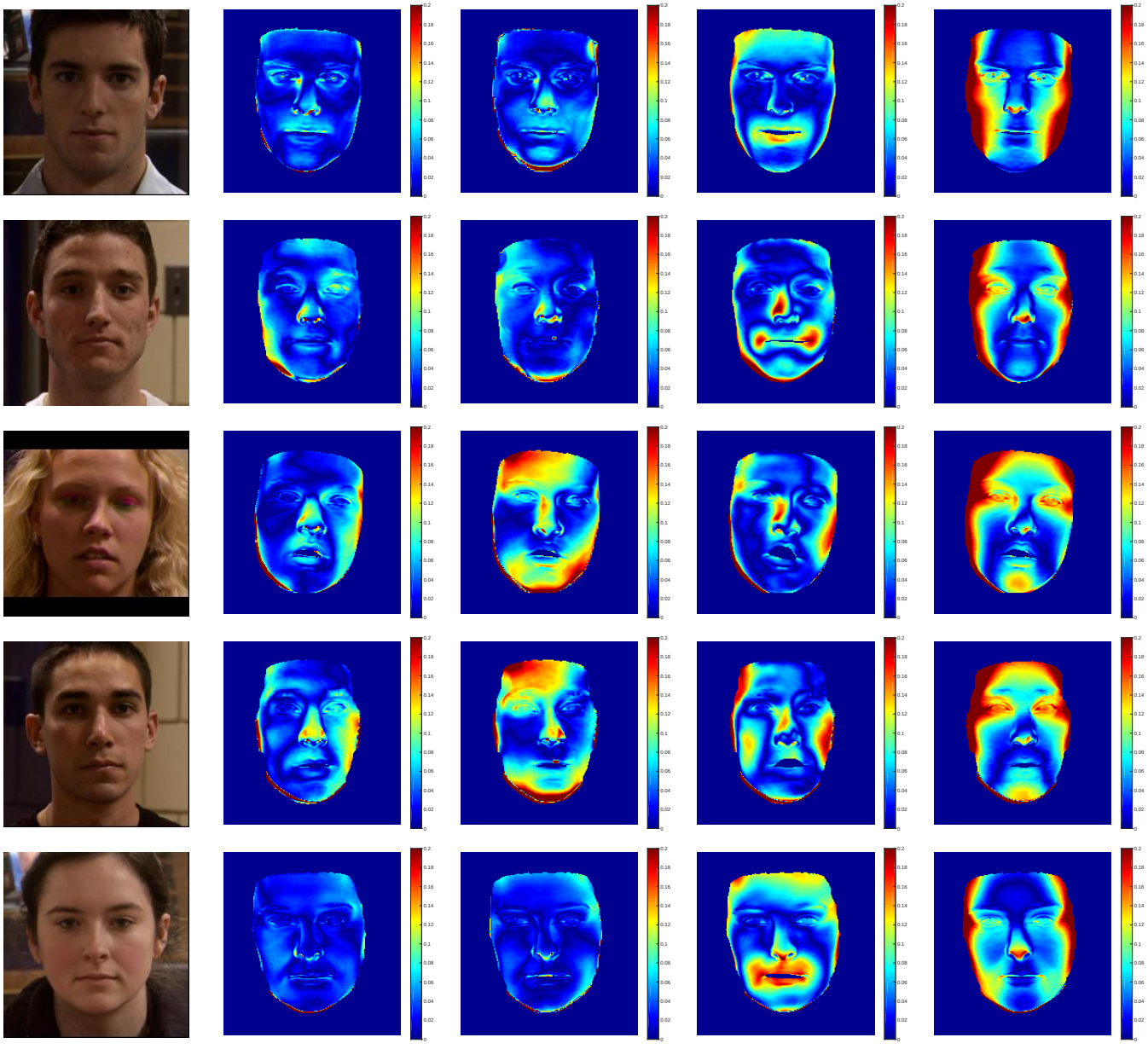

Figure 7: Absolute depth error heat maps of different methods.

References

[1] W. Chen, H. Wang, Y. Li, H. Su, D. Lischinsk, D. Cohen-Or, B. Chen, et al. Synthesizing training images for boosting human 3D pose estimation. *arXiv preprint arXiv:1604.02703*, 2016. 3

[2] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese. 3D-R2N2: A unified approach for single and multi-view 3D object reconstruction. *arXiv preprint arXiv:1604.00449*, 2016. 3

[3] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 1

[4] I. Kemelmacher-Shlizerman and R. Basri. 3D face reconstruction from a single image using a single reference face shape. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(2):394–405, 2011. 1, 5

[5] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, pages 41.1–41.12, 2015. 4

[6] E. Richardson, M. Sela, and R. Kimmel. 3D face reconstruction by learning from synthetic data. In *3D Vision (3DV), 2016 International Conference on*, pages 460–469. IEEE, 2016. 1, 3, 5

[7] H. Su, C. R. Qi, Y. Li, and L. J. Guibas. Render for cnn: Viewpoint estimation in images using cnns trained with rendered 3D model views. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2686–2694, 2015. 3

[8] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li. High-fidelity pose and expression normalization for face recognition in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 787–796, 2015. 1, 5