Supplementary material for: Faster, Better and More Detailed: 3D Face Reconstruction with Graph Convolutional Networks

Shiyang Cheng[1], Georgios Tzimiropoulos[1,3], Jie Shen[2,*], and Maja Pantic[2]

[1] Samsung AI Center, Cambridge. {shiyang.c,georgios.t}@samsung.com
[2] Imperial College London. {js1907,m.pantic}@imperial.ac.uk
[3] Queen Mary University of London.

1 Network architecture

For the 2D image encoder in coarse mesh reconstruction network, we manage to adopt both ResNet-50 [1] and MobileNetV2 [2] for our purpose, with almost the same parameter settings and training procedure. We illustrate our system again in Fig. 1 for ease of reference.

We denote spiral convolution layer with h hops, t filters and v vertices as SPConv(h,t,v), pooling and unpooling by a factor of p as Unpool(p) and Pool(p) respectively. FC layer is written as FC(d), where d is the output feature dimension. It has be mentioned that, in Fig. 1 as well as the following paragraphs, we take ResNet-50 as the example. Our spiral mesh decoder takes a 256-d feature embedding from image encoder as input, and has the following structure:
FC(52*128)→Unpool(4)→SPConv(1,128,208)→Unpool(4)→SPConv(1,64,832)
→Unpool(4)→SPConv(2,32,3326)→Unpool(4)→SPConv(2,16,13304)
→Unpool(4)→SPConv(2,8,53215)→SPConv(2,3,53215).

For simplicity, we describe the image feature sampling, adaptation and summation as one operation, viz. AddF(r,d,v), where r represents the spatial resolution of image features. Our refinement network uses a concatenation of coarse face normals and per vertex RGB values as input, and can be described as:
SPConv(2,8,53215)→Pool(4)→SPConv(2,16,13304)→Pool(4)
→SPConv(2,32,3326)→AddF(64,32,3326)→Unpool(4)→SPConv(2,16,13304)
→AddF(128,16,13304)→Unpool(4)→SPConv(2,8,53215)→SPConv(2,3,53215).

ELU [3] activation is used after each spiral convolution and fully connected layer, except for the last layer before getting the coarse mesh or per vertex shape displacement prediction.

2 Detailed Inference Time and Model Size Comparison

A detailed comparison of inference time and model size between our approach and existing methods are given in Tab. 1. The tests were conducted on a machine

* Corresponding author.

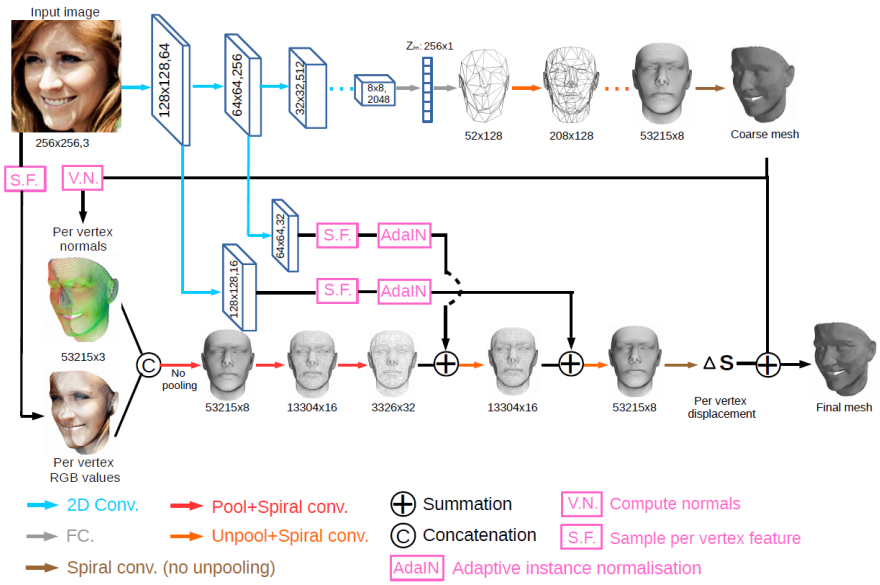

Fig. 1. Overview of our system. It consists of two connected sub-networks: (1) a coarse 3D facial mesh reconstruction network with a CNN encoder and a GCN decoder; (2) a GCN-based mesh refinement network for recovering the fine facial details. We also device a feature sampling and adaptation layer which injects fine details from the CNN encoder to the refinement network.

with an Intel Core i7-7820X CPU @3.6GHz, a GeForce GTX 1080 graphics card, and 96GB of main memory. For all methods we used the implementation provided by the original authors during this experiment (except of CMD [4], the code of which has not be released by the authors, so we used the result reported in their paper). The inference time reported in the table is averaged over all test images in Florence[5]. As shown in Tab. 1, our approach is among the fastest, taking only 10.8 ms / 6.2 ms (when using ResNet50 / MobileNet v2, respectively) to reconstruct a 3D face. Our method also has the smallest model size when using MobileNet-v2 as the image encoder. As demonstrated by this experiment, our approach is among the fastest, taking only 10.8 ms / 6.2 ms (when using ResNet50 / MobileNet v2, respectively) to reconstruct a 3D face. Our method also has the smallest model size when using MobileNet-v2 as the image encoder.

For CMD [4], we just show the values reported in their paper as the authors did not release their code and model.

Table 1. Detail comparison of inference time and model size

Method	Step	Time (ms)	Model size (MB)
DF^{2}Net [6]	End-to-end	40.4	222
Extreme3D [7]	3DMM fitting	54.8	167
	Bump map regression	176.1	194
	Recover 3D mesh	14328.9	142
	Total	14558.9	503
DFDN [8]	Proxy estimation	11788.1	309
	Detail synthesis	26974.6	1673
	Total	38762.7	1982
3DDFA [9, 10]	End-to-end	6.7	45
PRNet [11]	End-to-end	19.4	153
VRN [12]	Volumetric regression	16.4	1415
	Isosurface extraction	220.2	0
	Total	236.6	1415
Pix2vertex [13]	Depth image regression	20.5	832
	Correspondence map regression	19.5	832
	Mesh reconstruction	244055.7	0
	Detail reconstruction	3960.8	0
	Total	248056.6	1663
CMD [4]	End-to-end	3.1	93
Ours (ResNet50)	End-to-end	10.8	209
Ours (MobileNetV2)	End-to-end	6.2	37

3 Qualitative visual comparisons of image in-the-wild

In Fig. 2, we compared our method with 7 state-of-the-art methods, they included VRN [12], DF^{2}Net [6], Pix2vertex [13], extreme 3D face reconstruction [7], DFDN [8], 3DDFA [9] and PRNet [11]. From the figure, we observe that SfS-based methods (Pix2vertex, [7], DFDN and DF^{2}Net) have a better capability to recover subtle details, however, they are not quite robust, this is particularly evident when dealing with profile faces. On the contrary, the volumetric method VRN, non-parametric method PRNet and 3DMMs based method 3DDFA are much more robust under different illumination and poses, nonetheless, they can hardly produce any fine details on the mesh. Overall, our method is the best being both robust and able to capture fine facial details at the same time. Also, it has to be mentioned that, four of these methods (namely, DF^{2}Net, Pix2vertex, [7] and DFDN) collected additional high resolution data to boost their ability to model details, however, our method relied solely on 300W-LP database [9] and managed to reconstruct comparable level of facial details.

4 Reconstruction results of AFLW2000 database

In Fig. 3, we demonstrate comprehensive visualisation results on AFLW2000 database [9], due to the limited space, we plot the first 500 fitting results from the database. According to Fig. 2, we drew 3 representative methods (extreme 3D face reconstruction [7], DF^{2}Net [6] and PRNet [11]) that achieve good performance in terms of robustness and visual quality, and we visually compared our method with them. It can be clearly seen that our method are more accurate and robust than the other methods, meanwhile, our method is capable of generating high-fidelity facial details.

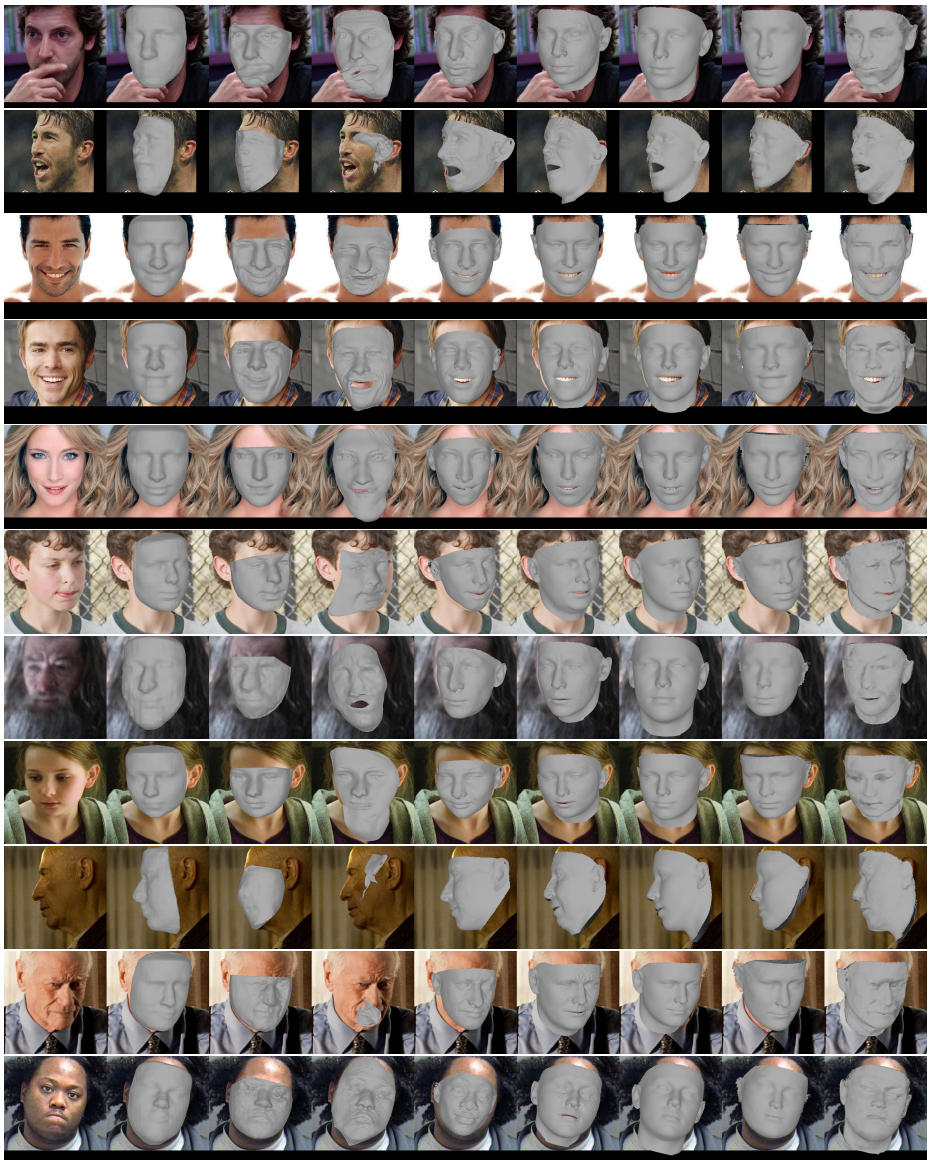

Fig. 2. Qualitative comparisons with 7 different methods on images in-the-wild. Please zoom in to check.

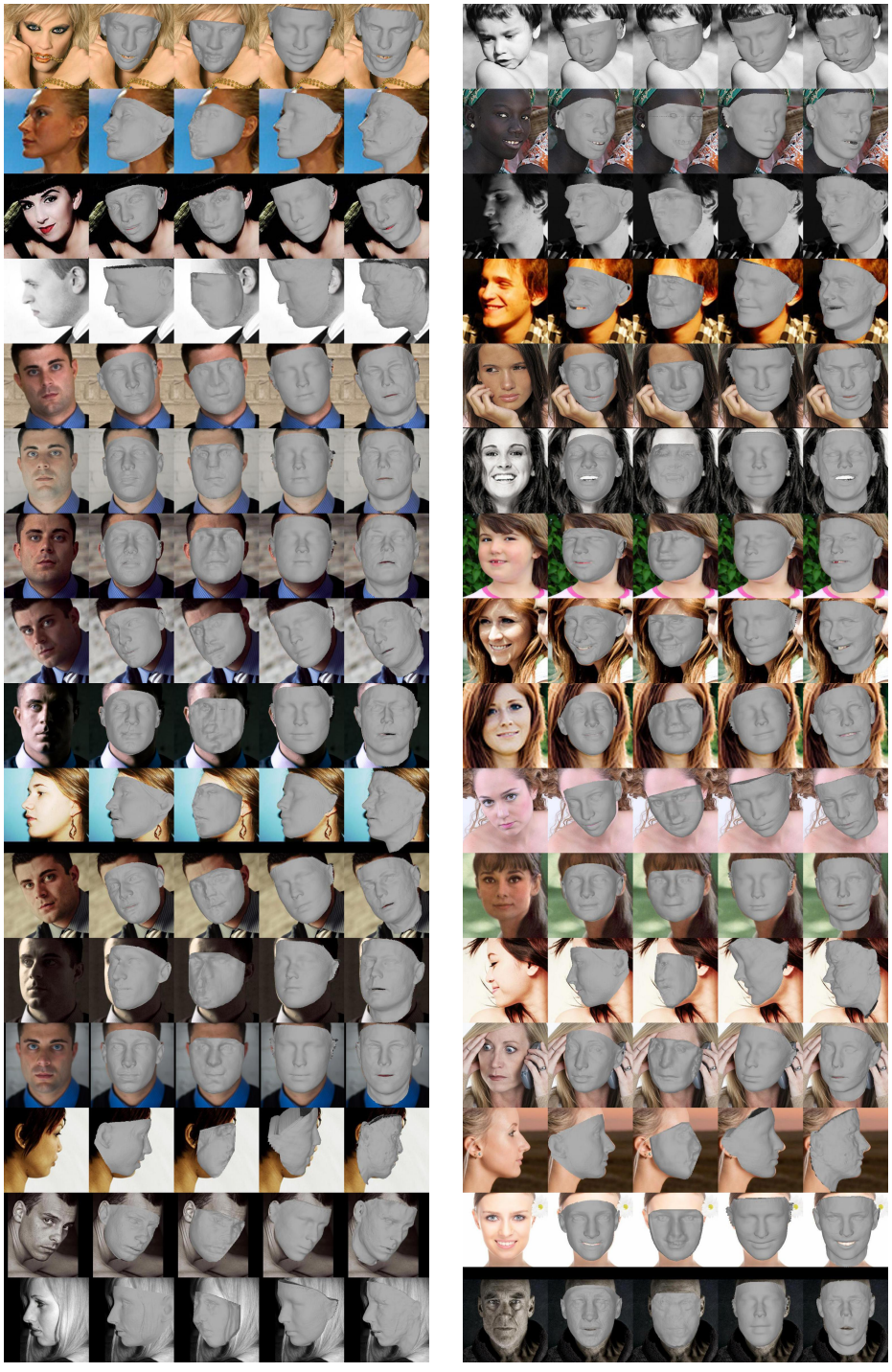

Fig. 3. Results of first 500 images from AFLW2000. Please zoom in to check.

Image [7] DF^{2}Net PRNet Ours Image [7] DF^{2}Net PRNet Ours

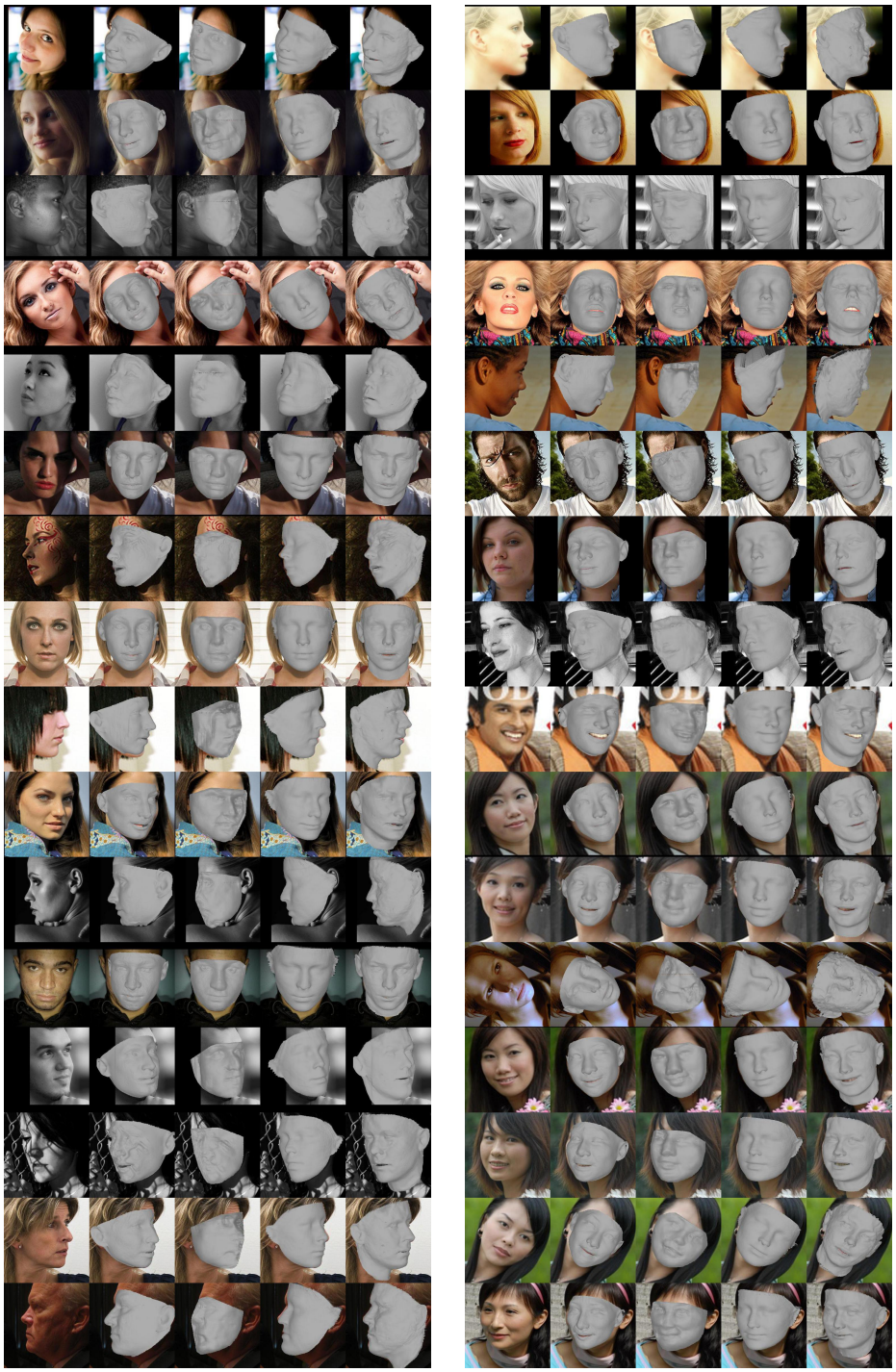

Image [7] DF^{2}Net PRNet Ours Image [7] DF^{2}Net PRNet Ours

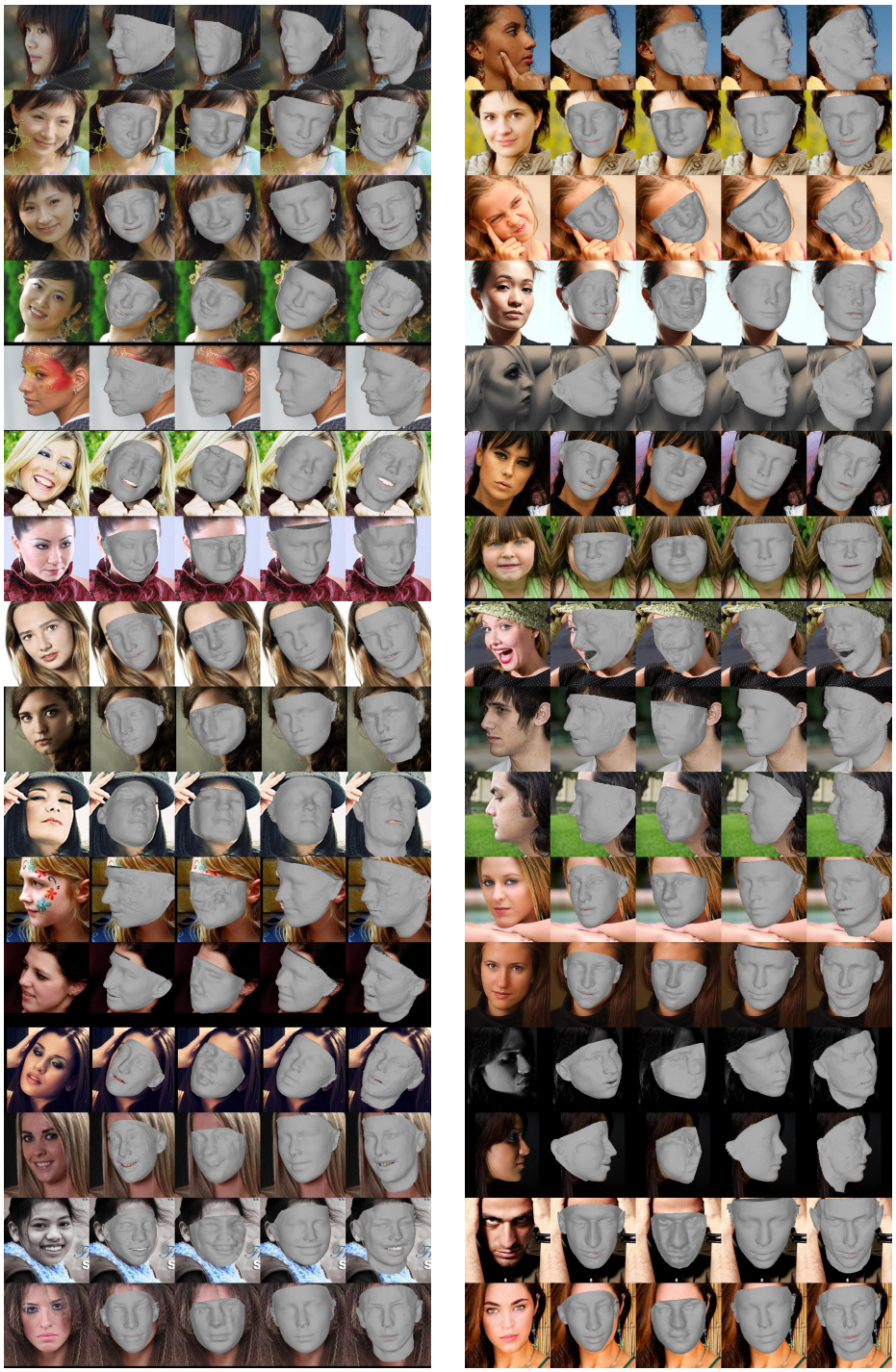

Image [7] DF^{2}Net PRNet Ours Image [7] DF^{2}Net PRNet Ours

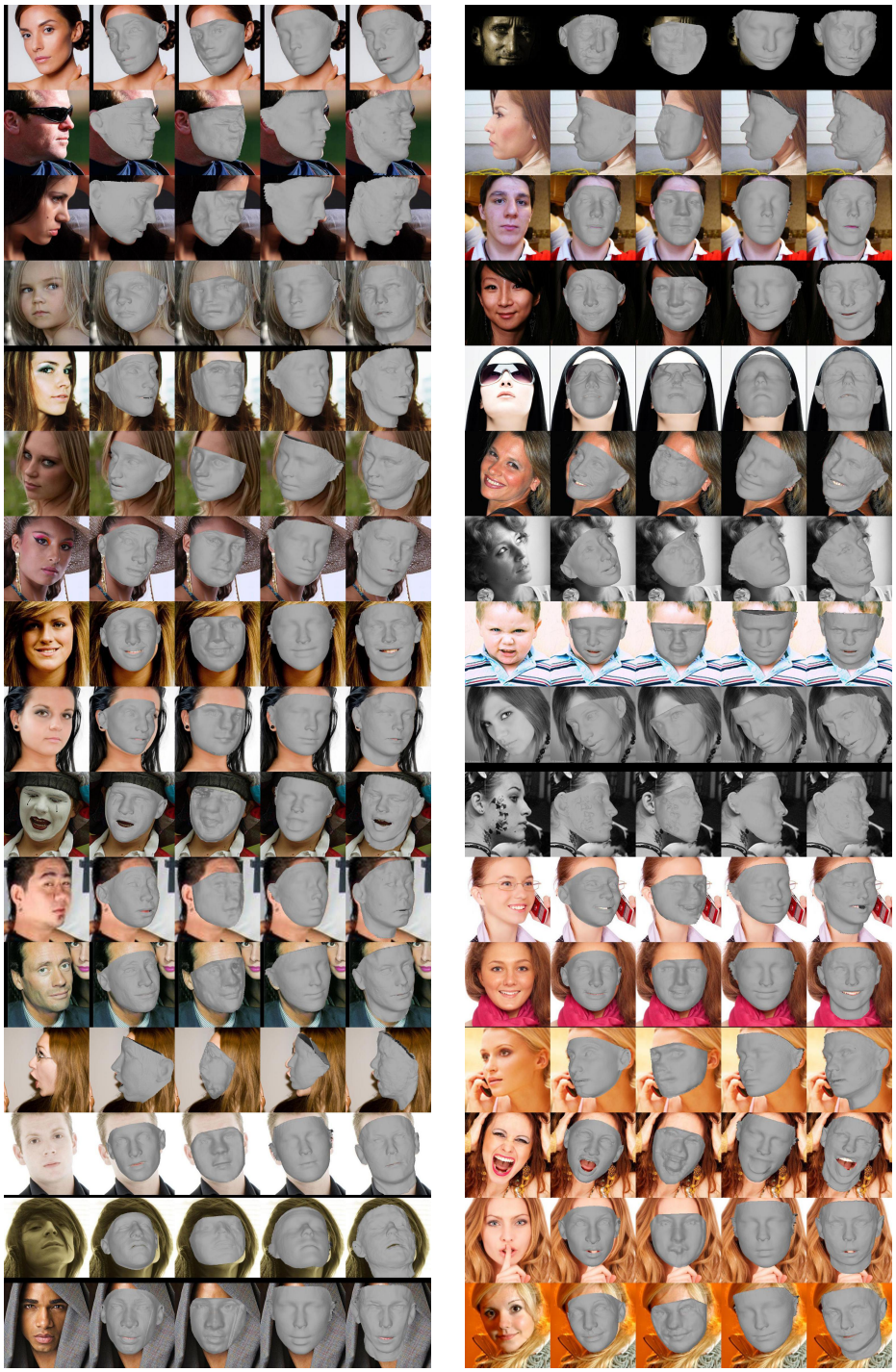
Image
[7]
DF^{2}Net
PRNet
Ours
Image
[7]
DF^{2}Net
PRNet
Ours

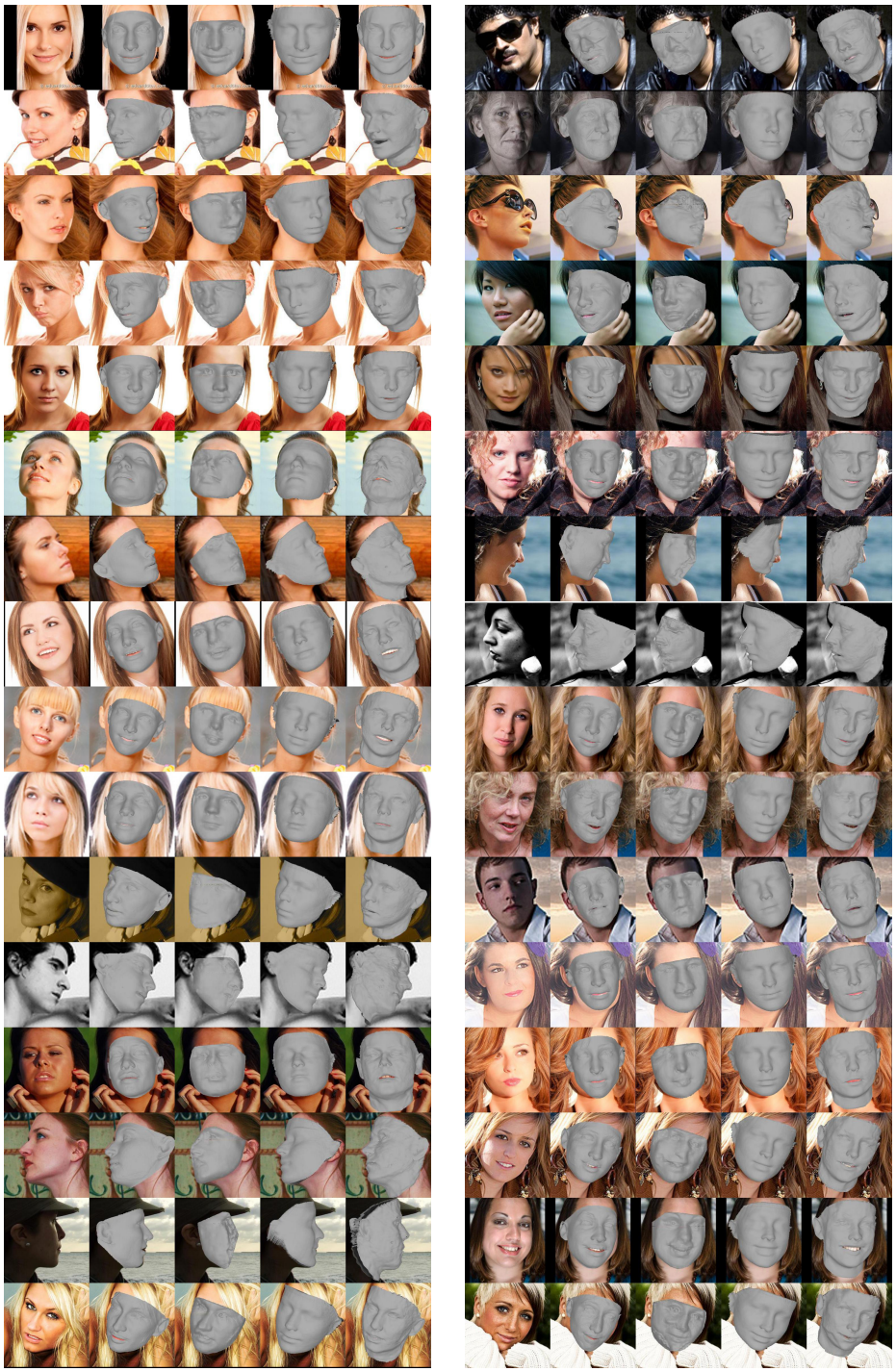

Image [7] DF^{2}Net PRNet Ours Image [7] DF^{2}Net PRNet Ours

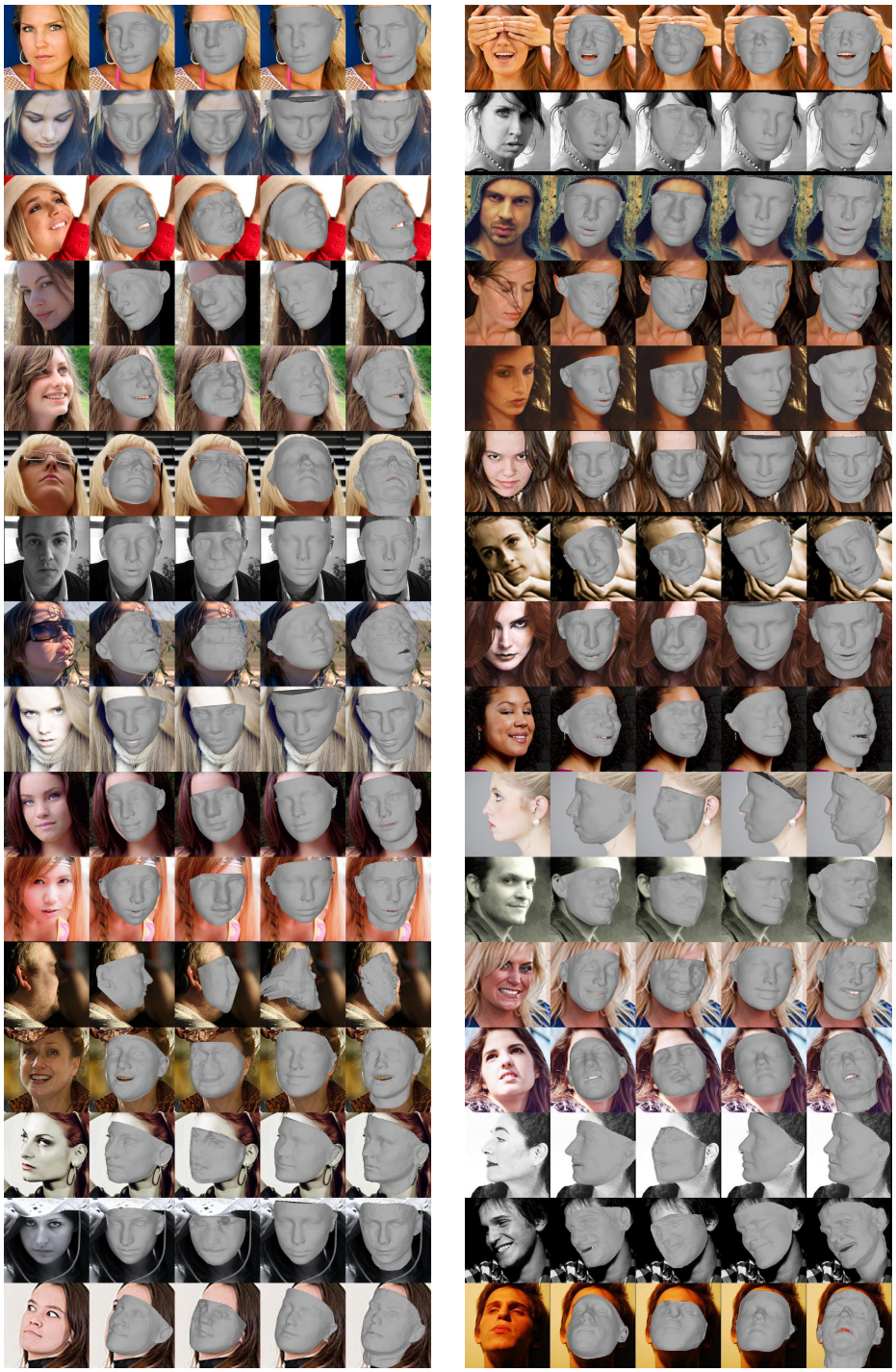
Image
[7]
DF^{2}Net
PRNet
Ours
Image
[7]
DF^{2}Net
PRNet
Ours

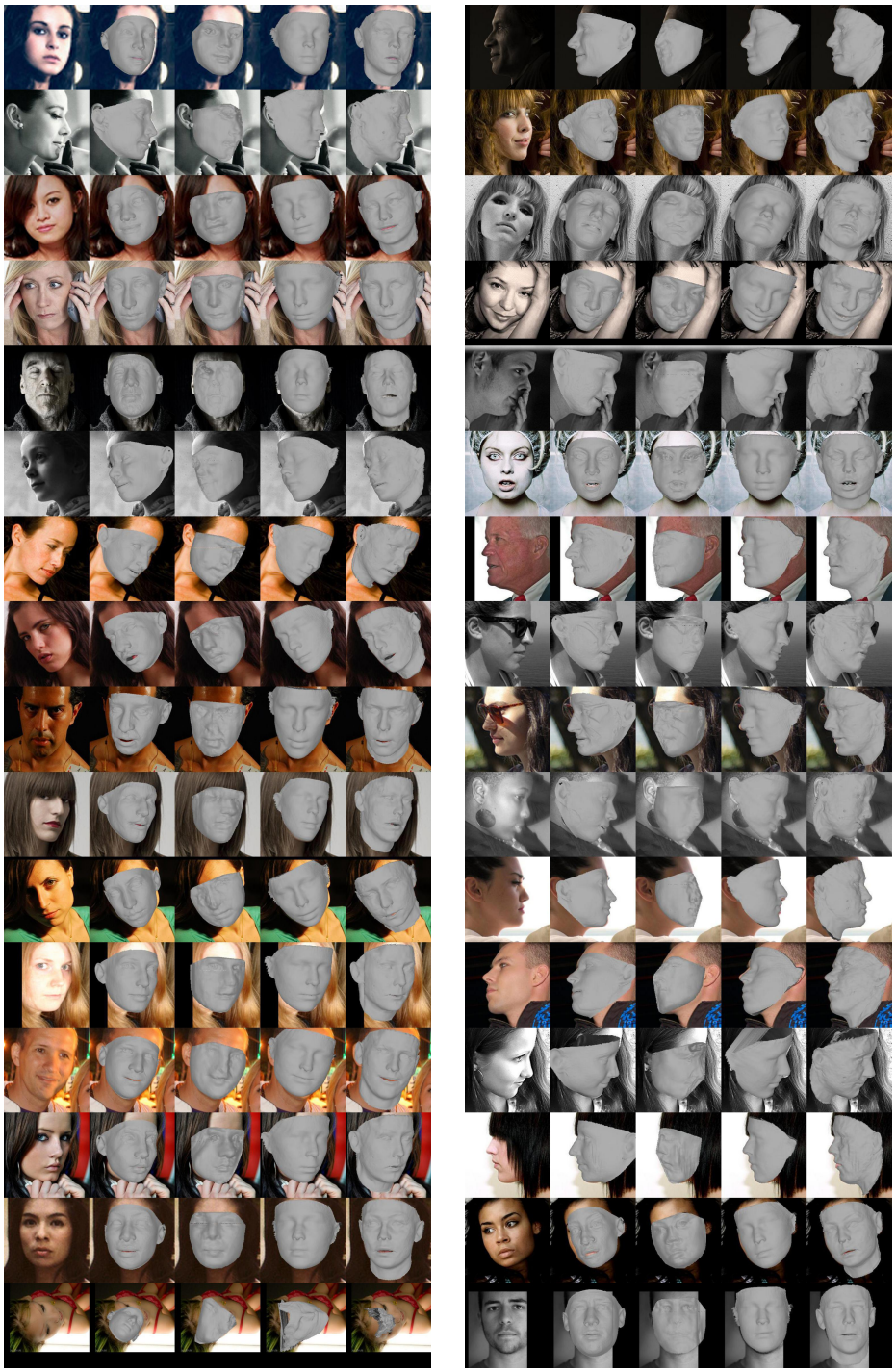

Image [7] DF^{2}Net PRNet Ours Image [7] DF^{2}Net PRNet Ours

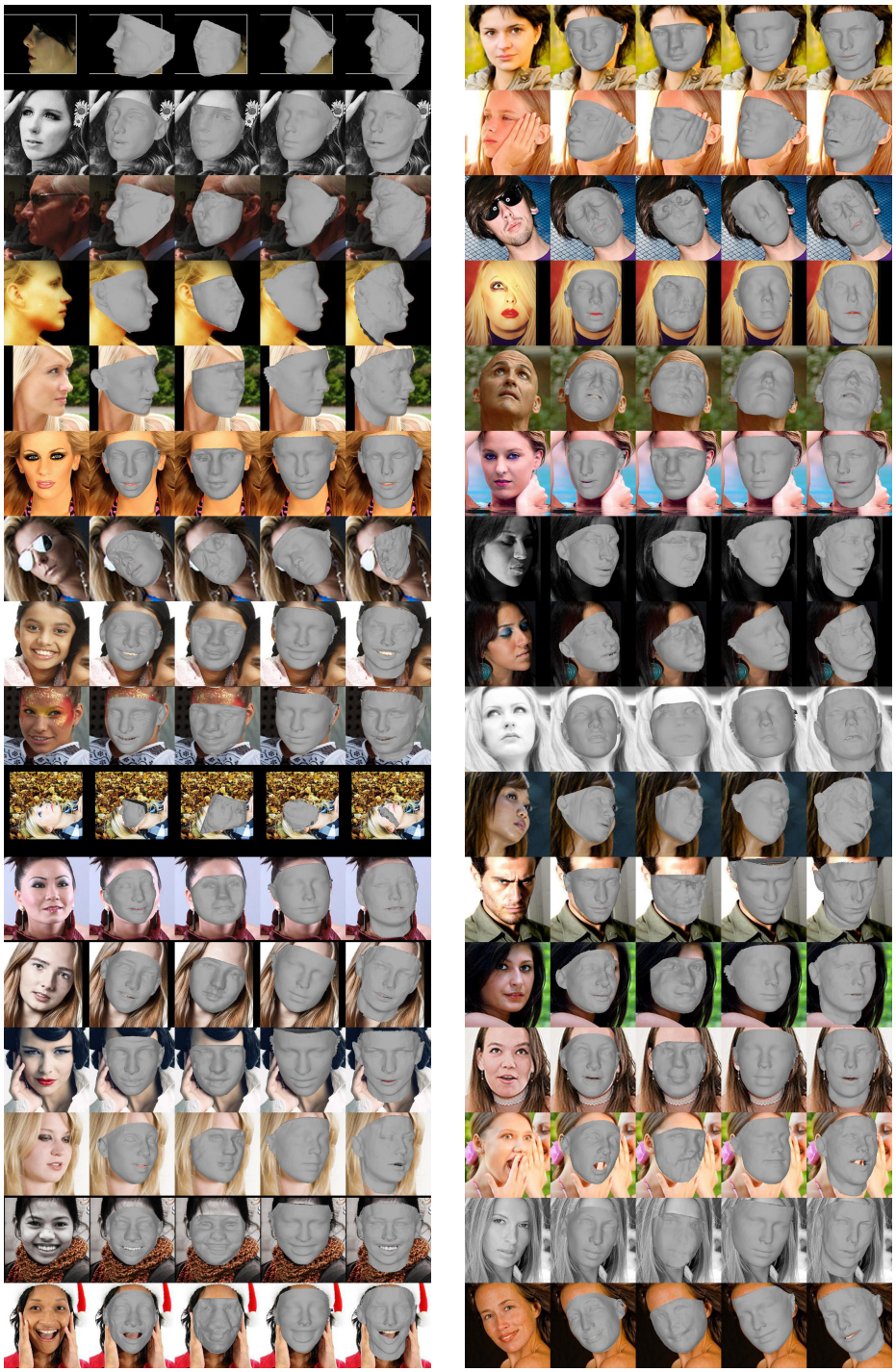
Image
[7]
DF^{2}Net
PRNet
Ours
Image
[7]
DF^{2}Net
PRNet
Ours

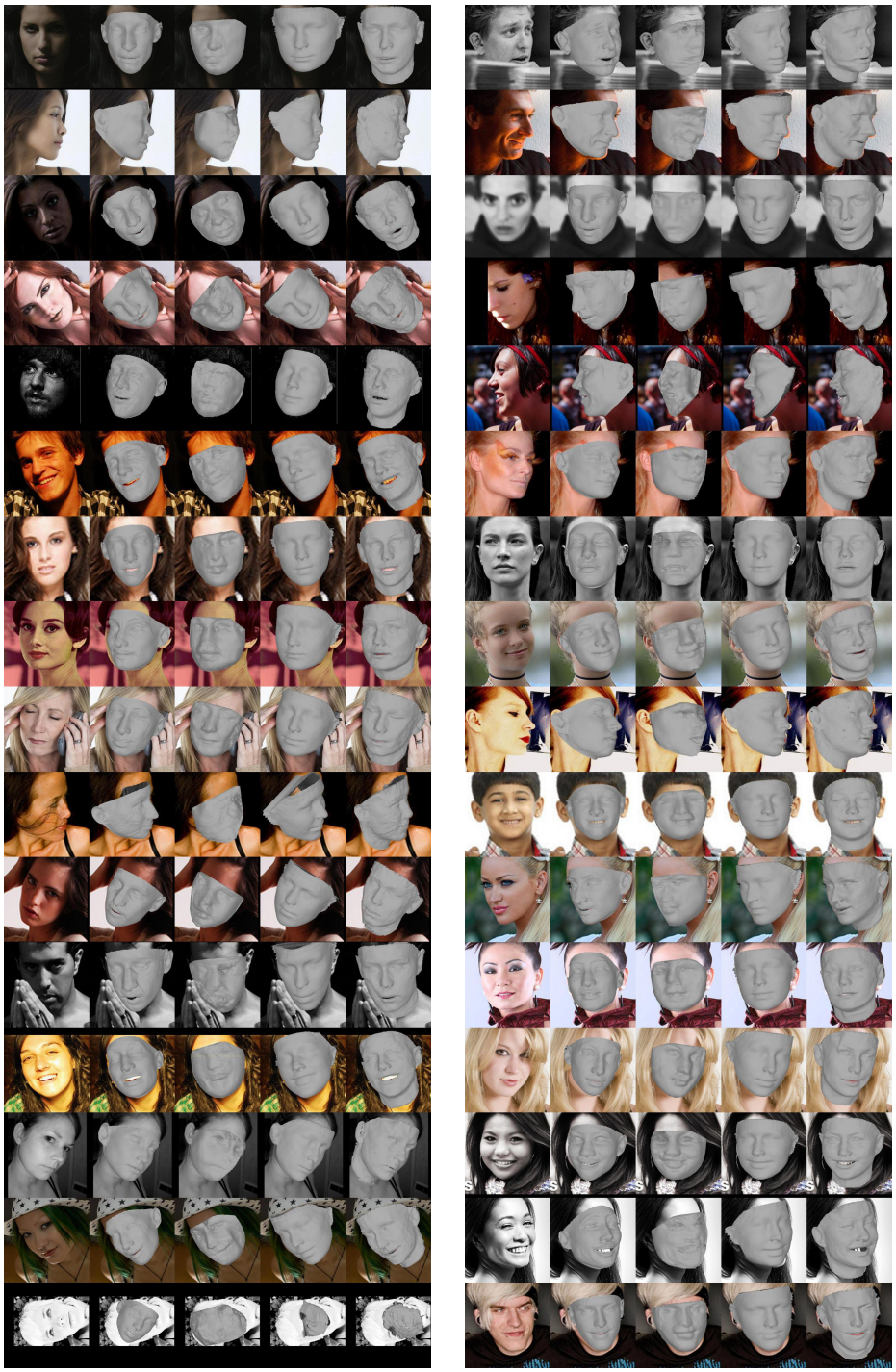

Image [7] DF^{2}Net PRNet Ours Image [7] DF^{2}Net PRNet Ours

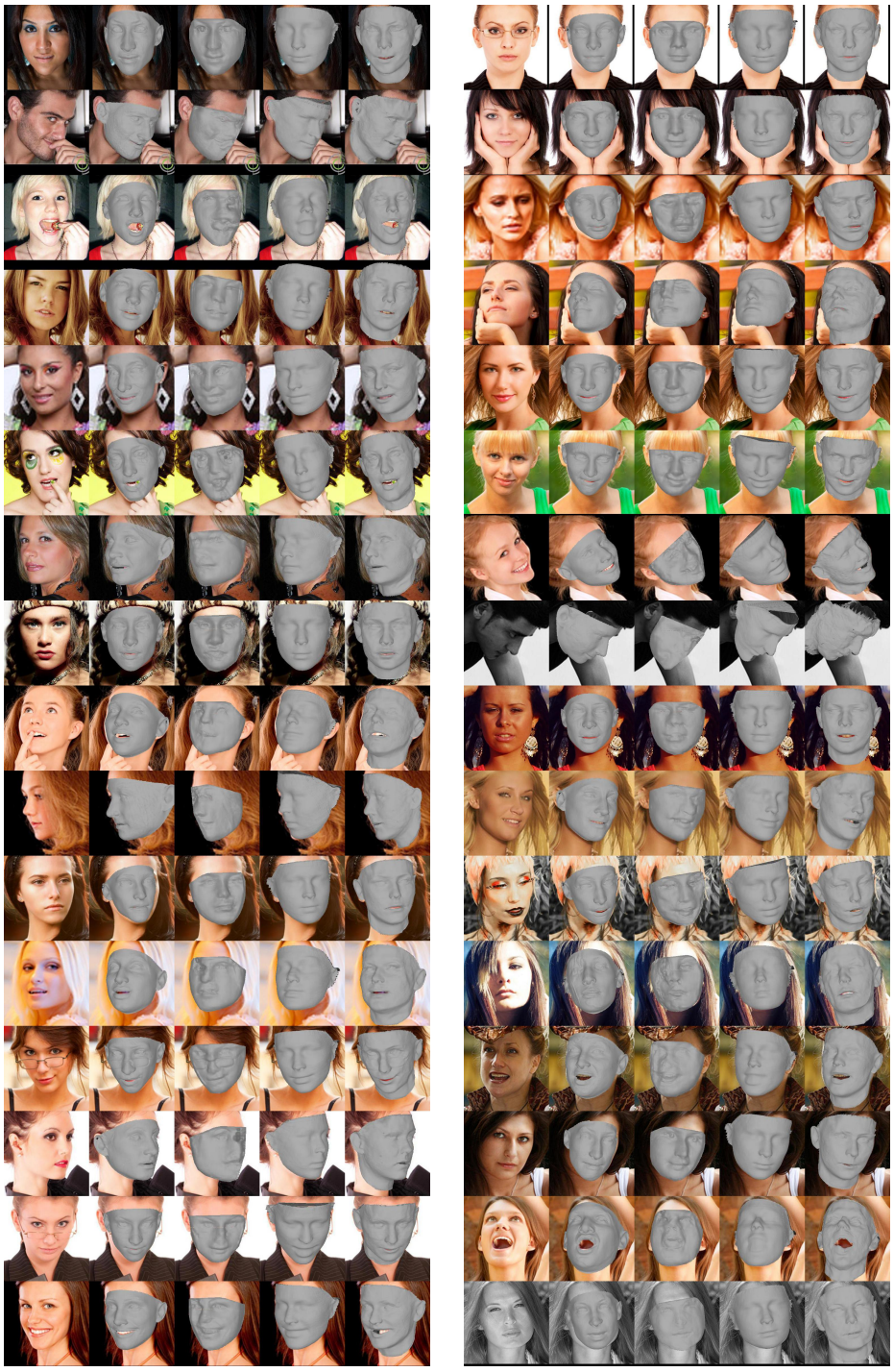
Image
[7]
DF^{2}Net
PRNet
Ours
Image
[7]
DF^{2}Net
PRNet
Ours

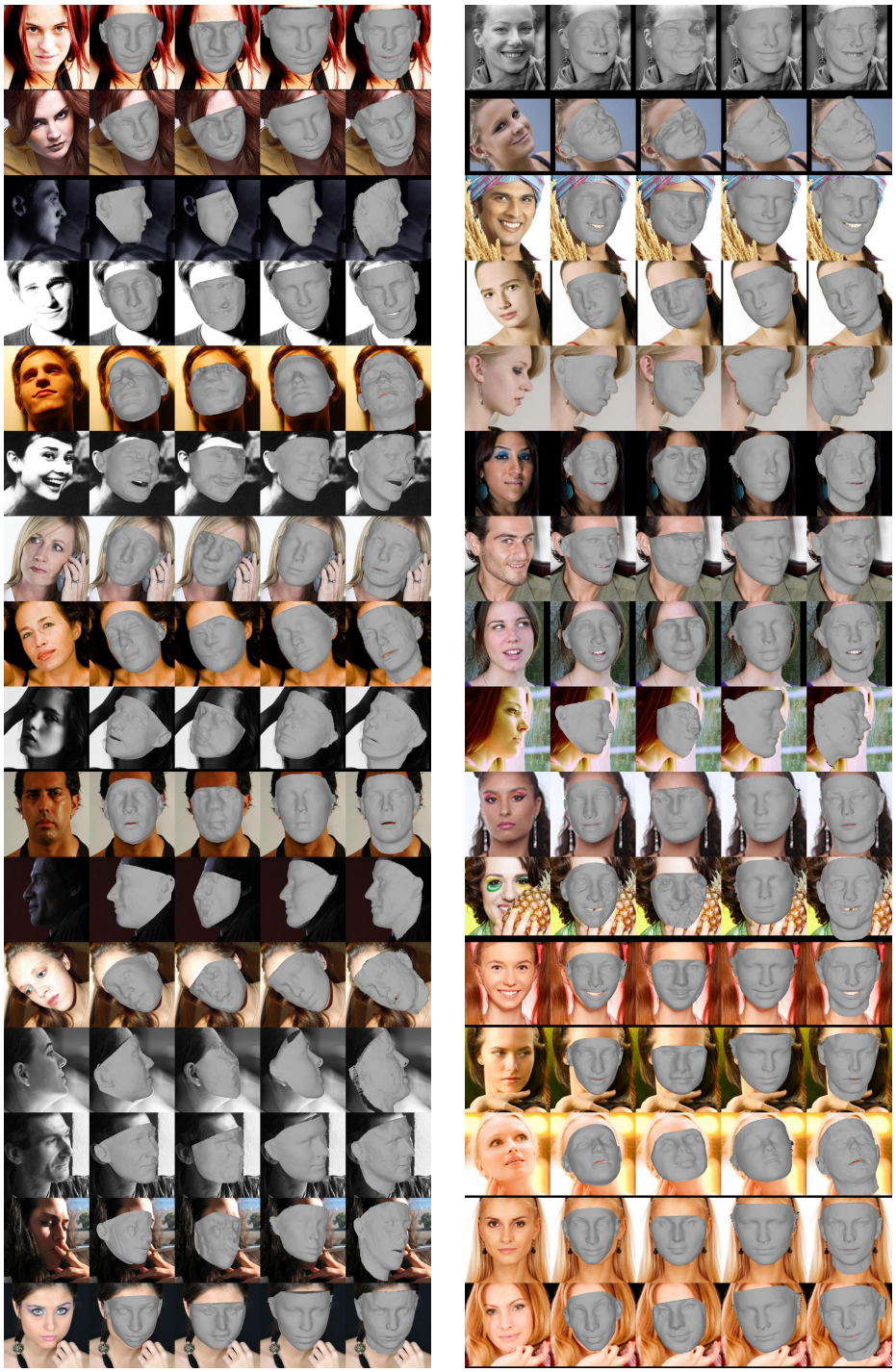

Image [7] DF^{2}Net PRNet Ours Image [7] DF^{2}Net PRNet Ours

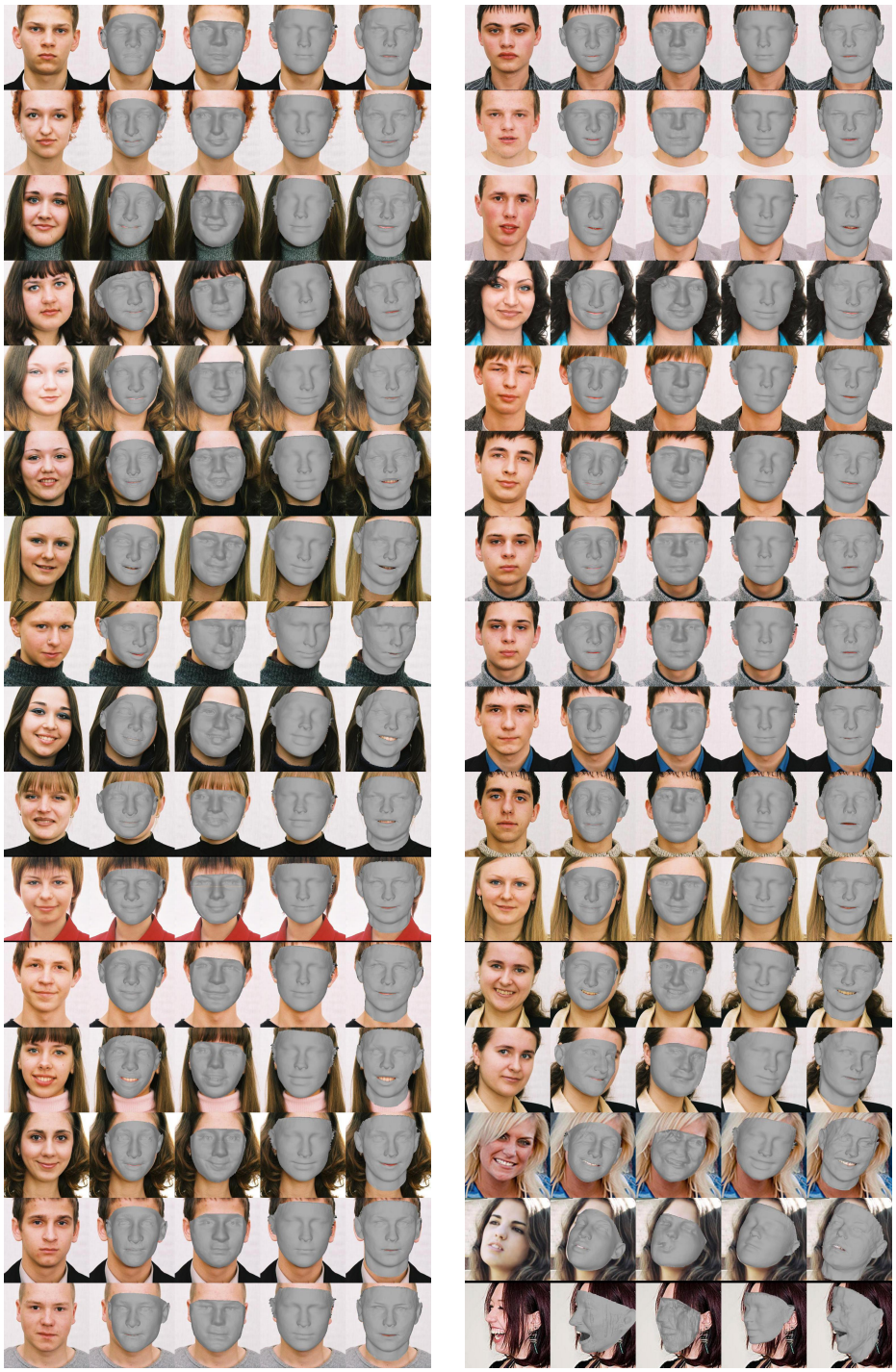
Image
[7]
DF^{2}Net
PRNet
Ours
Image
[7]
DF^{2}Net
PRNet
Ours

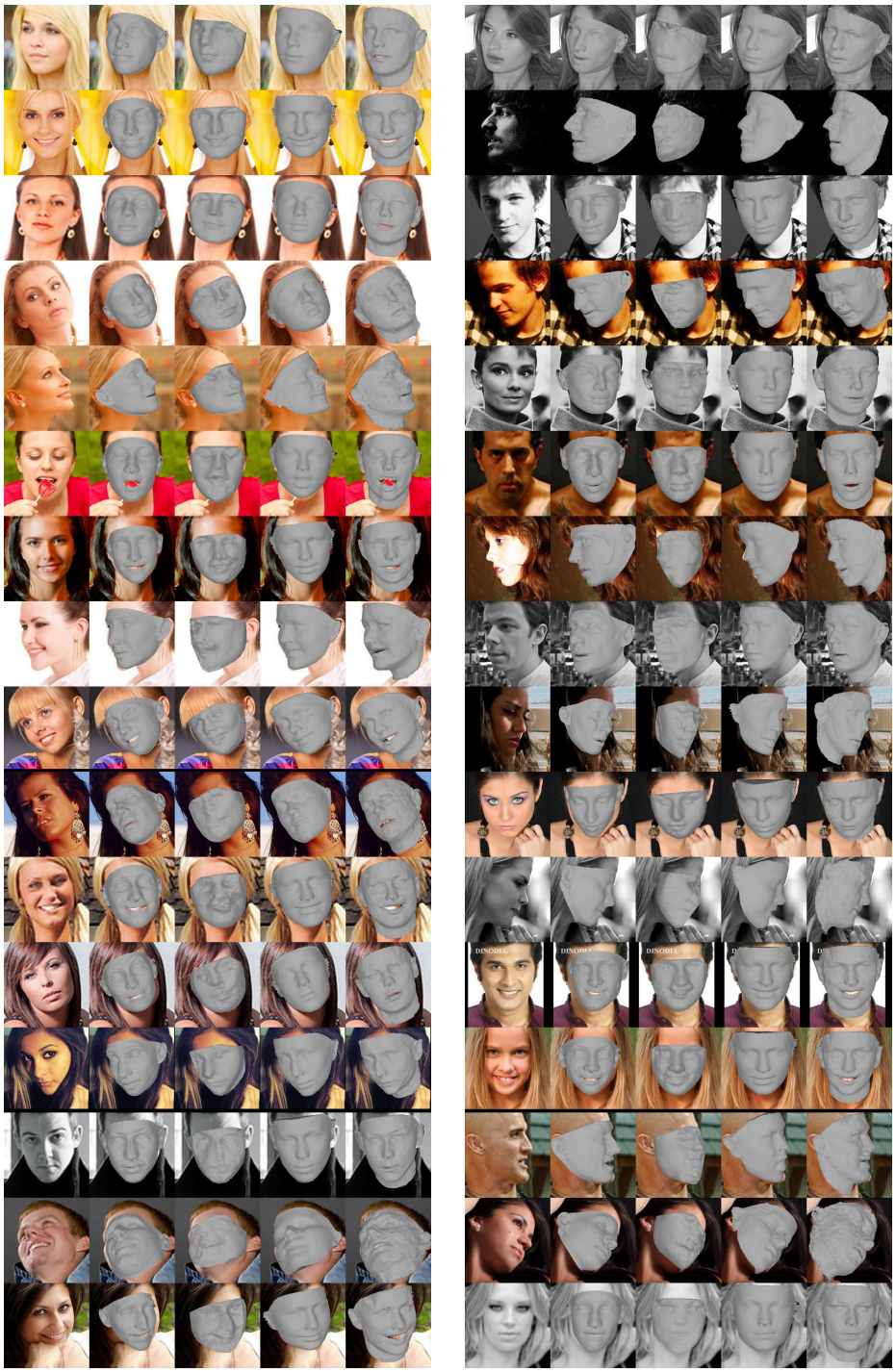
Image
[7]
DF^{2}Net
PRNet
Ours
Image
[7]
DF^{2}Net
PRNet
Ours

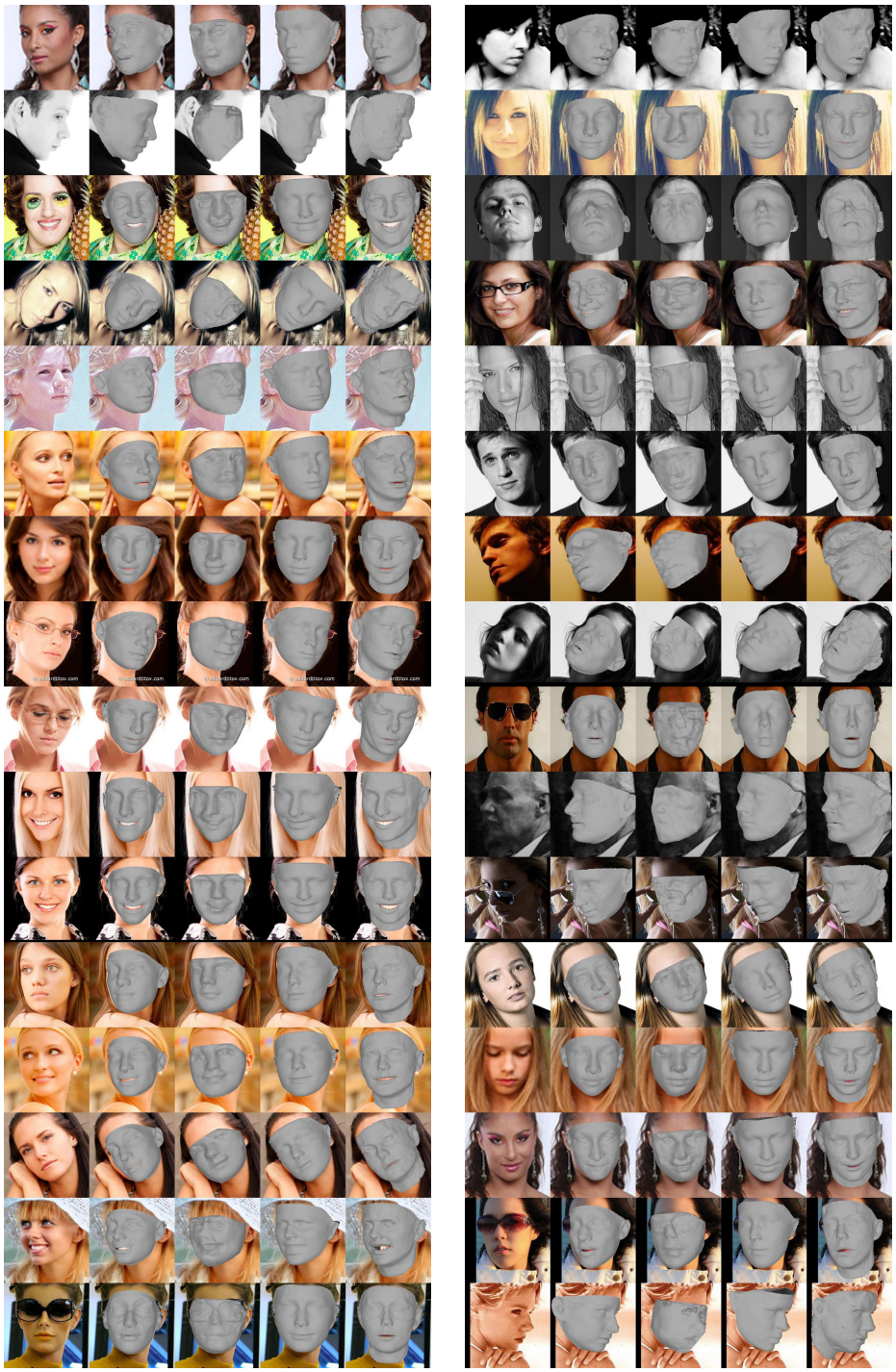

Image [7] DF^{2}Net PRNet Ours Image [7] DF^{2}Net PRNet Ours

Image [7] DF^{2}Net PRNet Ours Image [7] DF^{2}Net PRNet Ours

References

1. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 770–778
2. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation. arXiv preprint arXiv:1801.04381 (2018)
3. Djork-Arné, C., Unterthiner, T., Hochreiter, S.: Fast and accurate deep network learning by exponential linear units (elus). In: Proceedings of the International Conference on Learning Representations (ICLR). Volume 6. (2016)
4. Zhou, Y., Deng, J., Kotsia, I., Zafeiriou, S.: Dense 3d face decoding over 2500fps: Joint texture & shape convolutional mesh decoders. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 1097–1106
5. Bagdanov, A.D., Masi, I., Del Bimbo, A.: The florence 2d/3d hybrid face datset. In: Proc. of ACM Multimedia Int.'l Workshop on Multimedia access to 3D Human Objects (MA3HO'11). (2011)
6. Zeng, X., Peng, X., Qiao, Y.: Df2net: A dense-fine-finer network for detailed 3d face reconstruction. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 2315–2324
7. Tran, A.T., Hassner, T., Masi, I., Paz, E., Nirkin, Y., Medioni, G.G.: Extreme 3d face reconstruction: Seeing through occlusions. In: CVPR. (2018) 3935–3944
8. Chen, A., Chen, Z., Zhang, G., Mitchell, K., Yu, J.: Photo-realistic facial details synthesis from single image. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 9429–9439
9. Zhu, X., Lei, Z., Li, S.Z., et al.: Face alignment in full pose range: A 3d total solution. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017)
10. Jianzhu Guo, X.Z., Lei, Z.: 3ddfa. https://github.com/cleardusk/3DDFA (2018)
11. Feng, Y., Wu, F., Shao, X., Wang, Y., Zhou, X.: Joint 3d face reconstruction and dense alignment with position map regression network. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 534–551
12. Jackson, A.S., Bulat, A., Argyriou, V., Tzimiropoulos, G.: Large pose 3d face reconstruction from a single image via direct volumetric cnn regression. In: Proceedings of the IEEE International Conference on Computer Vision. (2017) 1031–1039
13. Sela, M., Richardson, E., Kimmel, R.: Unrestricted facial geometry reconstruction using image-to-image translation. In: Proceedings of the IEEE International Conference on Computer Vision. (2017) 1576–1585