Pix2NPHM: Learning to Regress NPHM Reconstructions From a Single Image

Simon Giebenhain[1] Tobias Kirschstein[1] Liam Schoneveld[2]
Davide Davoli[3*] Zhe Chen[2] Matthias Nießner[1]

[1]Technical University of Munich [2]Woven by Toyota [3]Toyota Motor Europe

This supplementary material contains additional information about our inference-time optimization routine (Sec. 1), and, finally, we present addition qualitative result comparisons and ablations in Sec. 2.

Moreover, we highly encourage the reviewers to watch our supplementary video for improved result visualizations using rendered camera trajectories, and video tracking results.

1. Inference-Time Optimization

At inference-time we assume initial camera parameters to be provided, *e.g.* estimated using Pixel3DMM [3]. Additionally, we require facial segmentation masks, *e.g.* provided by FaRL [8], in order to mask out the background and occluders such as glasses, hats and hands.

Once these things have been pre-computed, we initialize our optimization procedure using our feed-forward predictions, and minimize equation 10 of the main paper. We optimize for 100 steps using $\lambda_{\mathrm{n}} = 5.0$, $\lambda_{\mathrm{p}} = 1.0$, $\lambda_{\text{reg}} = 1.0$, $\lambda_{\text{id}}^{\mathcal{R}} = 0.5$ and $\lambda_{\text{ex}}^{\mathcal{R}} = 2.0$.

2. Additional Results and Ablations

In the remainder of this supplementary document we provide additional qualitative result comparisons on the SVFR-NeRSemble benchmark [3] and NoW benchmark [5]. Furthermore, we include more results of our ablation experiments in Sec. 2.2.

We refer to our supplemental video for result comparisons rendered from a camera trajectory for better 3D perception.

2.0.1. NeRSemble [3]

For additional *posed* reconstruction comparisons against recent SotA baselines we refer to Fig. 1. Similarly, we show additional *neutral* reconstruction results in Fig. 2

2.1. NoW [5]

Next to the controlled reconstruction scenarios on NeRSemble, we present additional qualitative results on the in-the-wild captures from the validation set of the NoW benchmark [5] in Fig. 3.

2.2. Ablations

Additional qualitative ablations results are shown in Figs. 4 and 5 for *posed* and *neutral* reconstructions on the SVFR NeRSemble benchmark, respectively.

Fig. 6 shows ablation results on the NoW validation set.

Furthermore, we show different ablation experiments to the main paper in Fig. 7 (posed, NeRSemble) and Fig. 8 (neutral, NoW).

Additionally, Fig. 9 shows ablative comparisons on in-the-wild image from the FFHQ dataset [4].

2.3. In-the-Wild Results

Finally, Fig. 10 shows more in-the-wild results on FFHQ [4]. For in-the-wild tracking results we refer to our supplementary video.

References

[1] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (ToG)*, 40(4): 1–13, 2021. 3

[2] Simon Giebenhain, Tobias Kirschstein, Markos Georgopoulos, Martin Rünz, Lourdes Agapito, and Matthias Nießner. Mononphm: Dynamic head reconstruction from monocular videos. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3, 4

[3] Simon Giebenhain, Tobias Kirschstein, Martin Rünz, Lourdes Agapito, and Matthias Nießner. Pixel3dmm: Versatile screen-space priors for single-image 3d face reconstruction, 2025. 1, 3, 4, 5

[4] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 1

[5] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael Black. Learning to regress 3D face shape and expression from an image without 3D supervision. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 7763–7772, 2019. 1, 10

[*]Providing contracted services for Toyota

[6] Felix Taubner, Prashant Raina, Mathieu Tuli, Eu Wern Teh, Chul Lee, and Jinmiao Huang. 3D face tracking from 2D video through iterative dense UV to image flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1227–1237, 2024. 3

[7] Tianke Zhang, Xuangeng Chu, Yunfei Liu, Lijian Lin, Zhendong Yang, Zhengzhuo Xu, Chengkun Cao, Fei Yu, Changyin Zhou, Chun Yuan, et al. Accurate 3d face reconstruction with facial component tokens. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9033–9042, 2023. 3

[8] Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18697–18709, 2022. 1

[9] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. In *European conference on computer vision*, pages 250–269. Springer, 2022. 4, 5

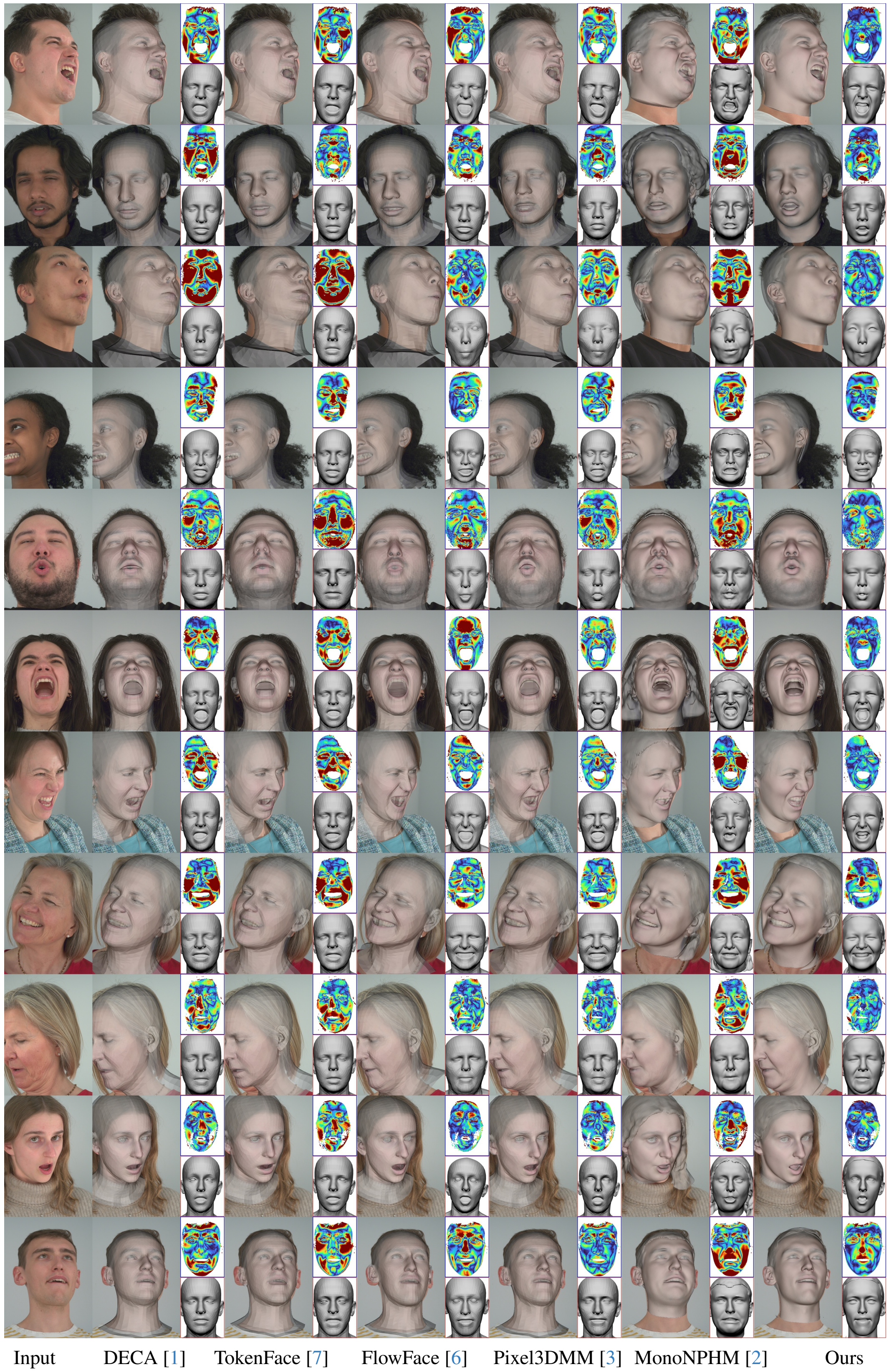

Figure 1. **Posed Reconstruction:** We show overlays of the reconstructed meshes to judge the reconstruction alignment. Insets with a blue border depict L_2-Chamfer distance as an error map, rendered from a frontal camera. Red insets show the reconstructed mesh from the same camera. All our figures are best viewed digitally and zoomed-in.

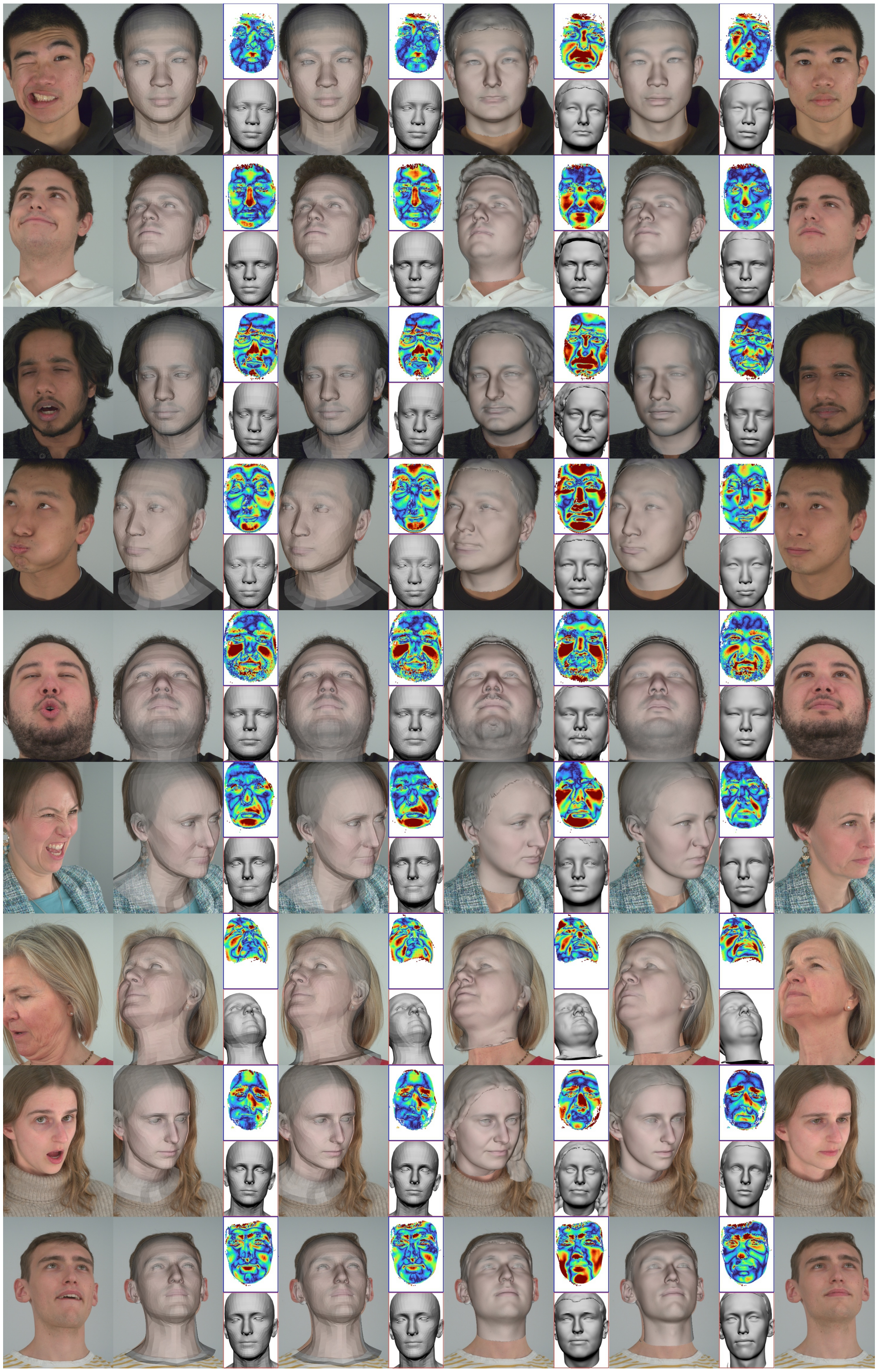

Figure 2. **Neutral Reconstruction, NeRSemble:** Comparison against available SotA methods on top of neutral reference image.

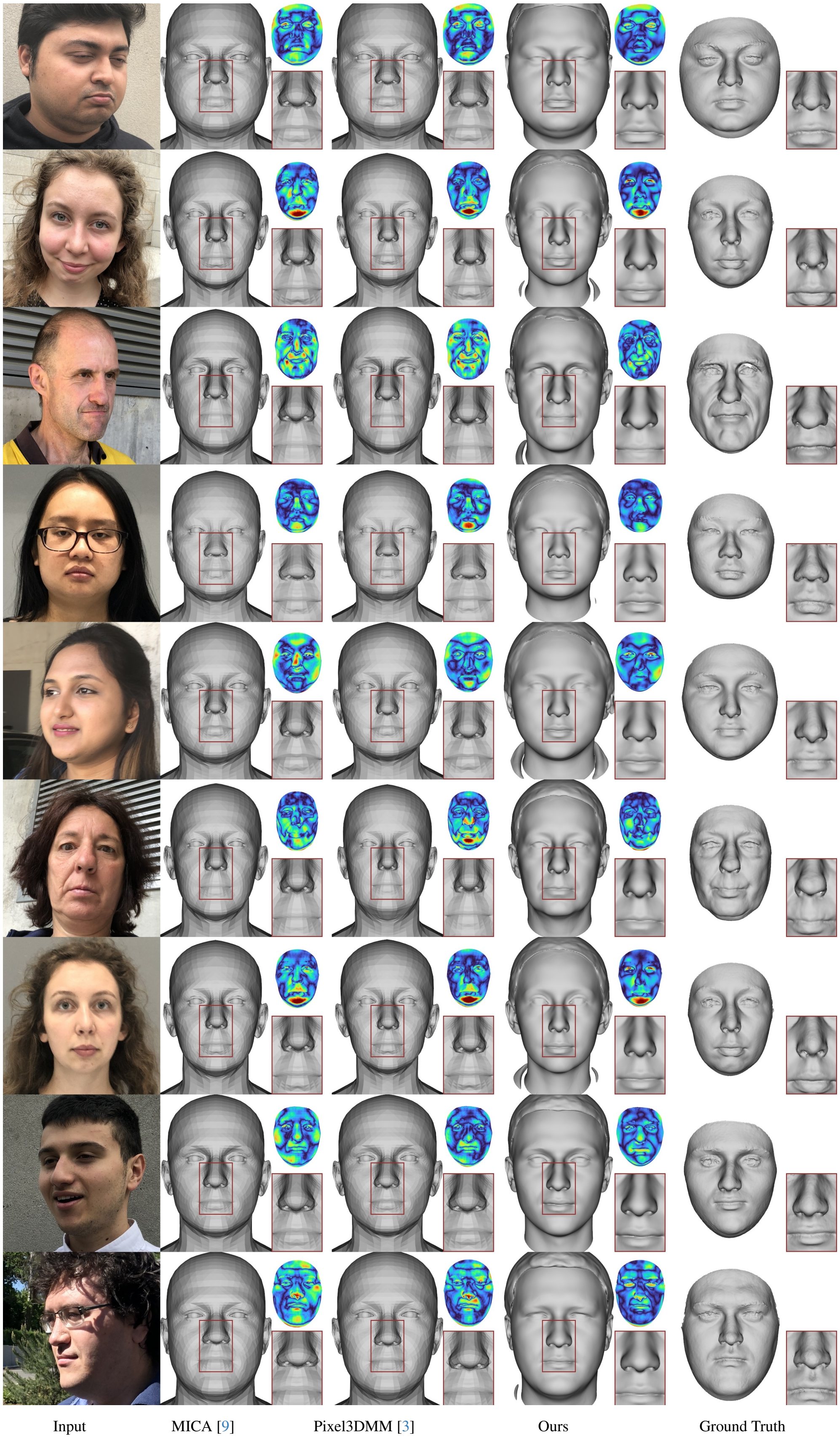

Figure 3. **Neutral Reconstruction, NoW:** We show frontal mesh renderings in comparison to the ground truth mesh, as well as, error maps and zoom ins.

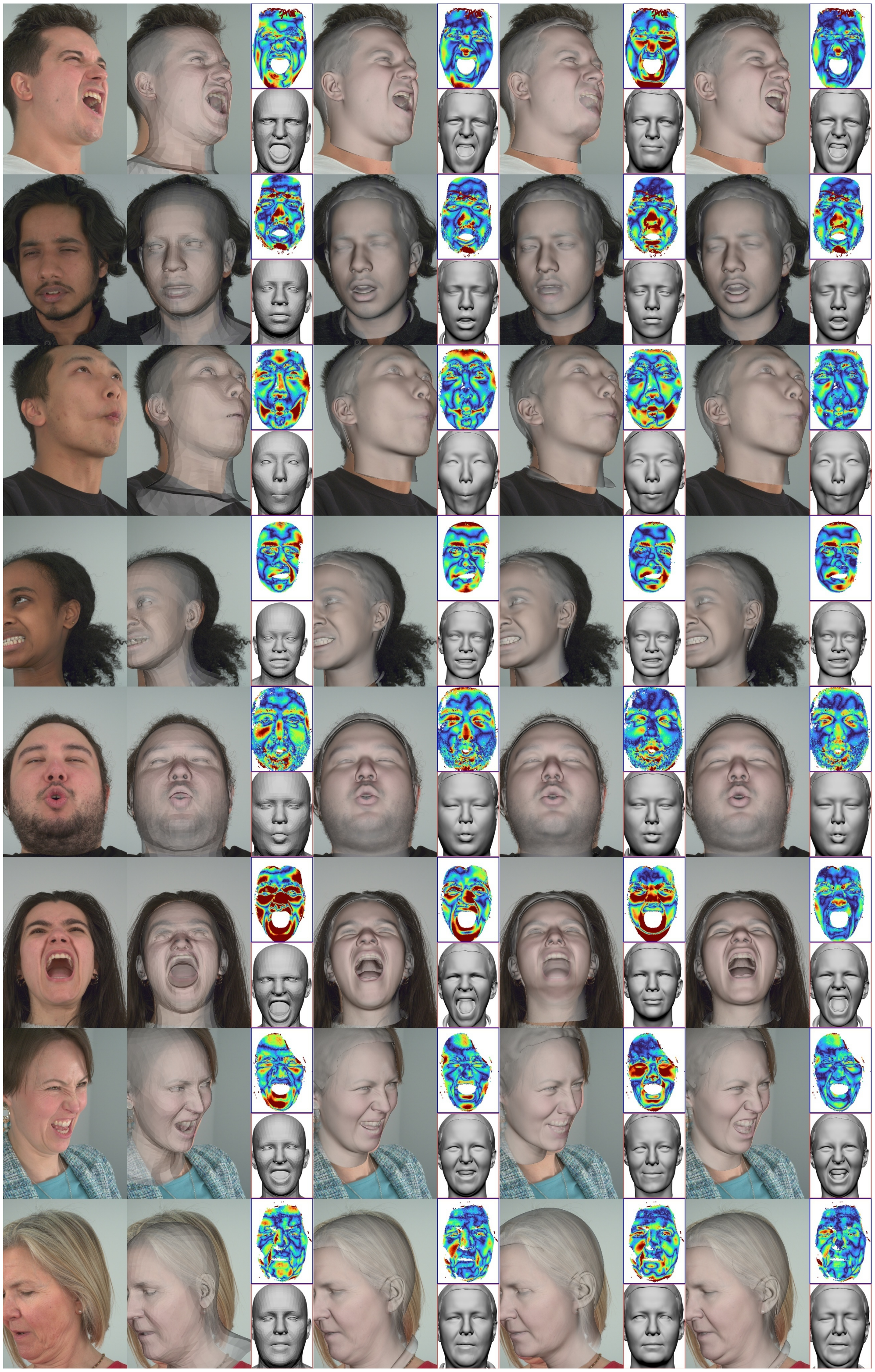

Figure 4. **Ablations, Posed:** NPHM feed-forward predictions exhibt more details compared to FLAME. Wihtout the feed-forward initialization our optimization sometimes fails to reconstruct extreme expressions (*e.g*. see rows 1 and 3).

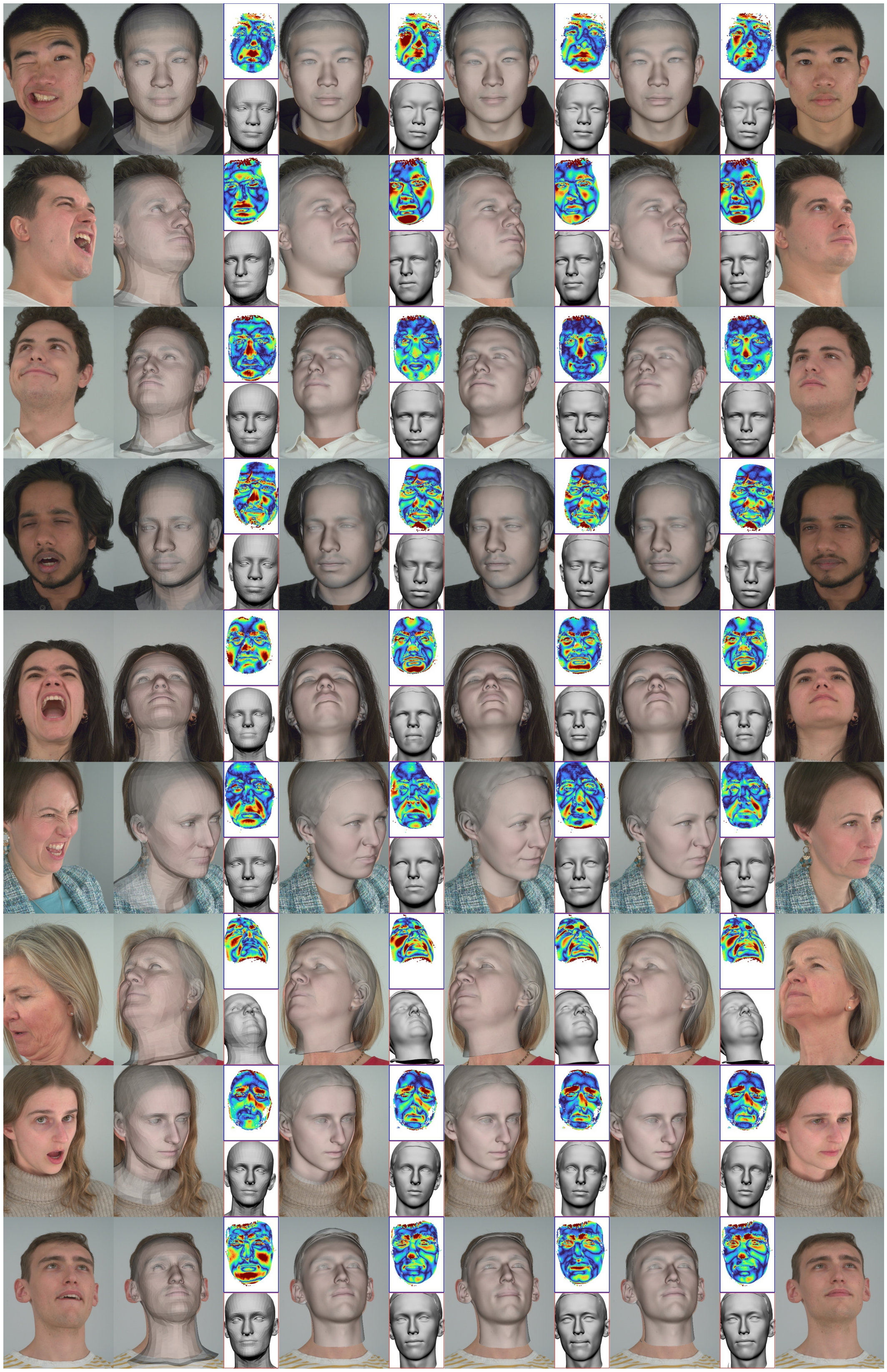

Figure 5. **Ablations, Neutral:** Without the feed-forward prior, our optimization cannot properly disentangle identity and expression.

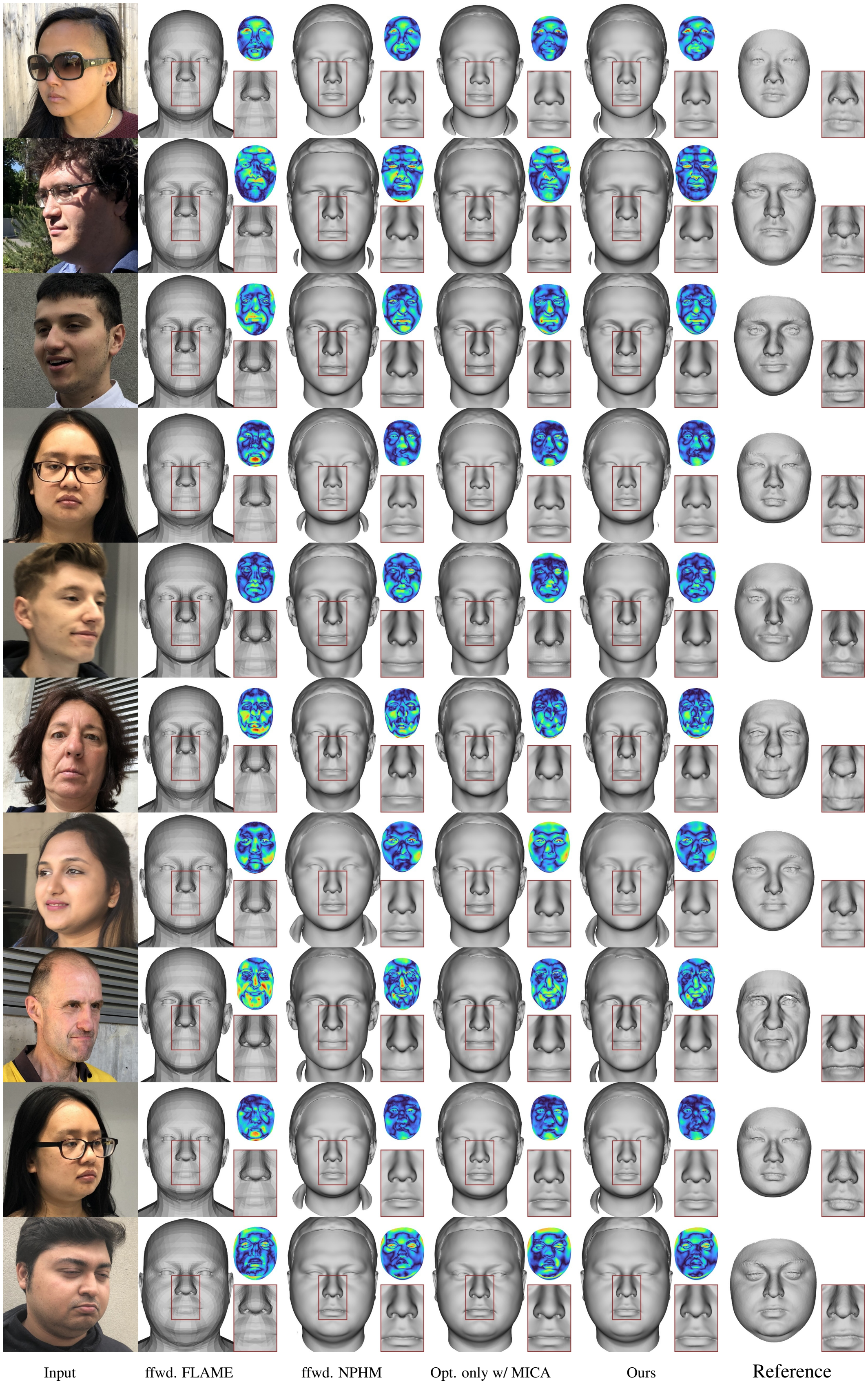

Figure 6. **Ablations, NoW.**

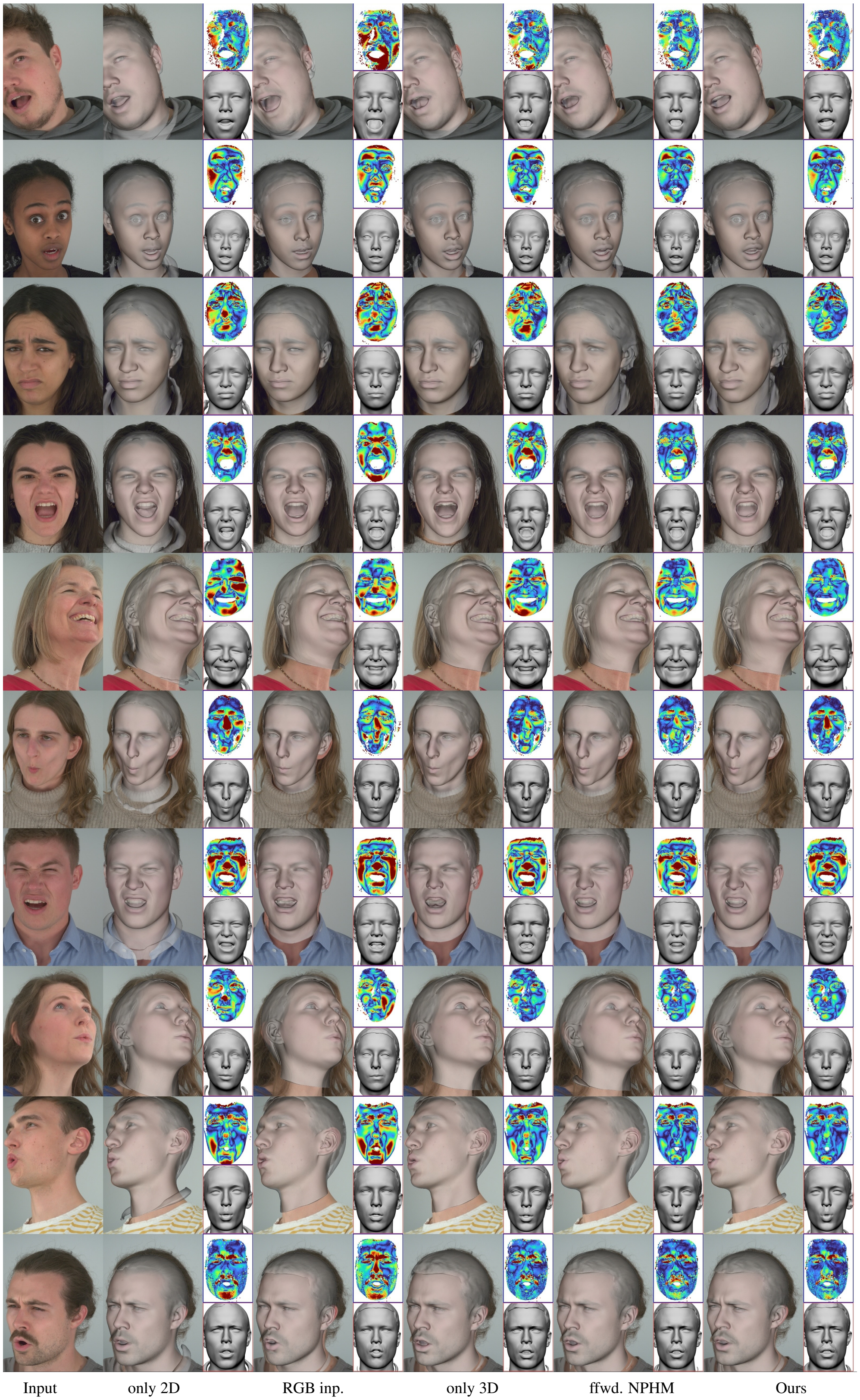

Figure 7. **Ablations, Posed:** Here we ablate different training dataset types, and different input types against our proposed feed-forward predictor and full method.

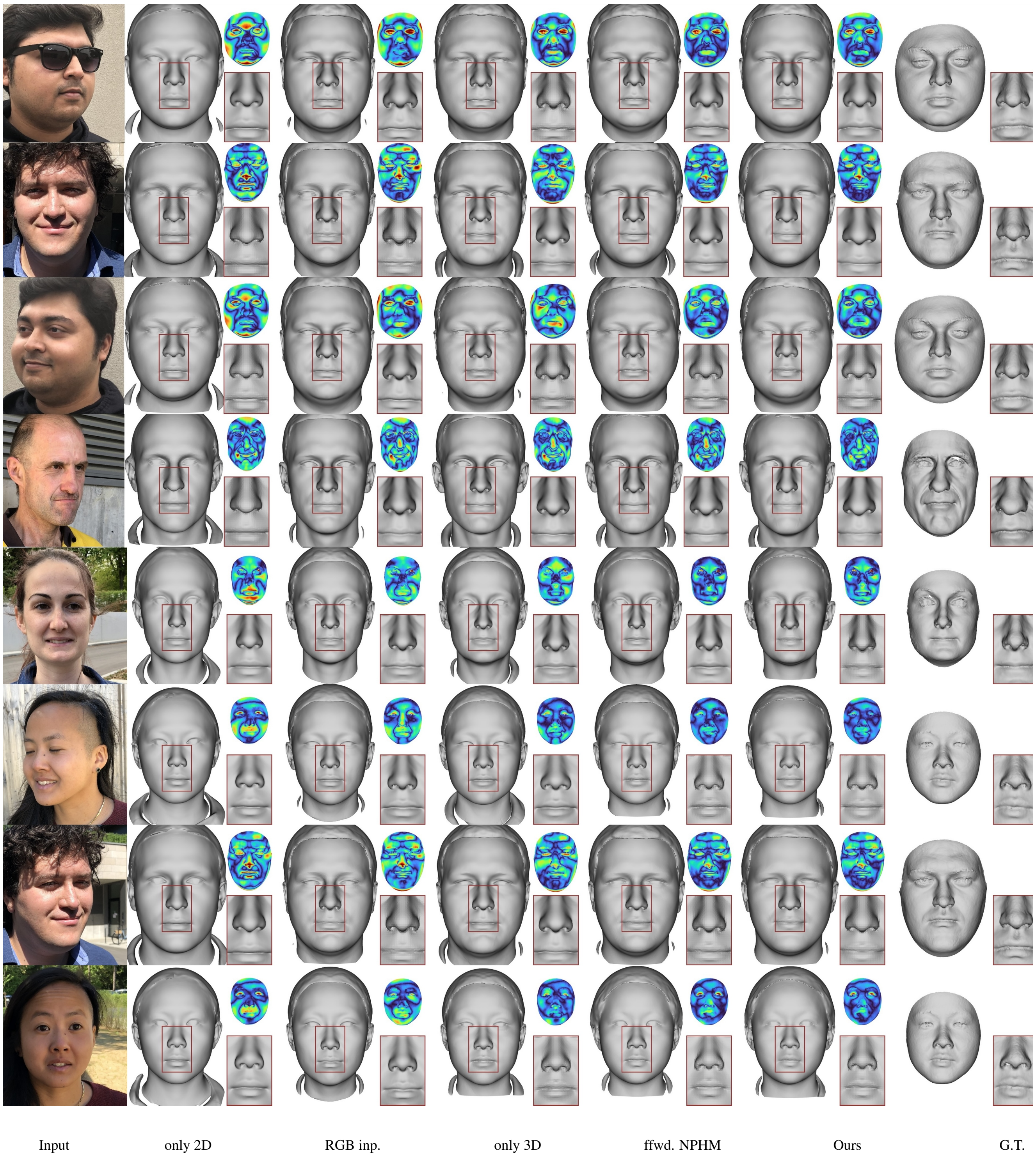

Figure 8. **Ablations, Neutral on NoW [5]:** Here we ablate different training dataset types, and different input types against our proposed feed-forward predictor and full method.

Figure 9. **In-the-Wild Ablations**

Figure 10. **In-the-Wild Reconstruction**