Supplementary Material for "Structure-aware Editable Morphable Model for 3D Facial Detail Animation and Manipulation"

Jingwang Ling[1], Zhibo Wang[1], Ming Lu[2], Quan Wang[3], Chen Qian[3], and Feng Xu[1]

[1] BNRist and school of software, Tsinghua University
[2] Intel Labs China
[3] Sensetime Research, China

A Overview

In this supplementary material we present:

- Effectiveness of LPIPS[7] as a displacement map metric
- Quantitative comparison using LPIPS
- User study
- Additional discussion on wrinkle line editing
- More qualitative results
- Additional qualitative comparison on in-the-wild images
- Examples of extracted detail line maps and distance fields

Please also refer to our video for animation results.

B Effectiveness of LPIPS[7] as a displacement map metric

We use LPIPS[7] to evaluate the similarity between displacement maps, because it is more consistent with the visual similarity perceived by humans, compared to traditional metrics such as L1 Loss. To investigate this, we select a wrinkle on a displacement map, delete it or move it, and generate two modified displacement maps, which we refer to as "absent wrinkles" and "misaligned wrinkles", respectively. We evaluate the L1 Loss and LPIPS between the modified displacement maps and the ground truth. When evaluating LPIPS, we normalize the displacement value to range $[-1, 1]$ and convert grayscale to RGB. We also invite 23 participants and ask them which is more similar to the ground truth. The results are shown in Fig. 1. We find that LPIPS considers misaligned wrinkles to be closer to the ground truth, which agrees with human judgements, while L1 Loss disagrees with humans. Therefore, we believe LPIPS is more suitable to measure displacement map similarity.

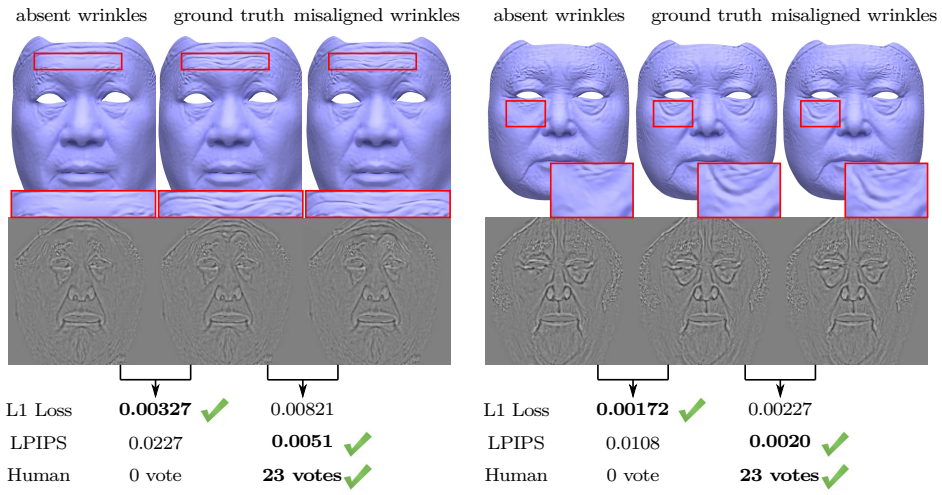


Fig. 1: LPIPS agrees with humans on the similarity of facial details, while L1 Loss disagrees with humans.

C Quantitative comparison using LPIPS

We use LPIPS to quantitatively compare the displacement maps generated by our method, FaceScape[6] and DECA[1]. The comparison is performed on randomly selected 618 test samples from the dataset from [6], which has samples of the same person with different expressions. For each sample, all the methods (ours, FaceScape and DECA) first obtain the detail representation from an input image. Then, original details are generated from the representation and used to evaluate the reconstruction error in LPIPS. The detail representation is also combined with target expression parameters to generate displacement maps with other target expressions. We evaluate the editing error between the generated and reference displacement maps in LPIPS. Because DECA's mesh topology is different from the dataset from [6], we perform non-rigid registration and then extract the displacement maps in the way described in our paper. FaceScape is known to work better with neutral expression inputs, so we separately report the errors using neutral expression inputs and using inputs with non-neutral expressions. The results are shown in Table 1.

Our model achieves the lowest error both in reconstruction and editing. The generated details of DECA are visually plausible, but their quantitative errors are higher than ours, possibly because their method is only trained on 2D image data. In DECA's paper, a similar phenomenon is reported that after adding details, their mesh reconstruction error increases. The results indicate that our model is able to both more accurately represent input details and generate dynamic details better matching the input person's identity, with either neutral or non-neutral input expressions.

Table 1: Quantitative comparison with FaceScape and DECA.

	Ours	FaceScape	DECA
neutral recon	**0.1447**	0.1784	0.4023
neutral edit	**0.1719**	0.1991	0.4020
non-neutral recon	**0.1511**	0.1991	0.4039
non-neutral edit	**0.1759**	0.1980	0.4021

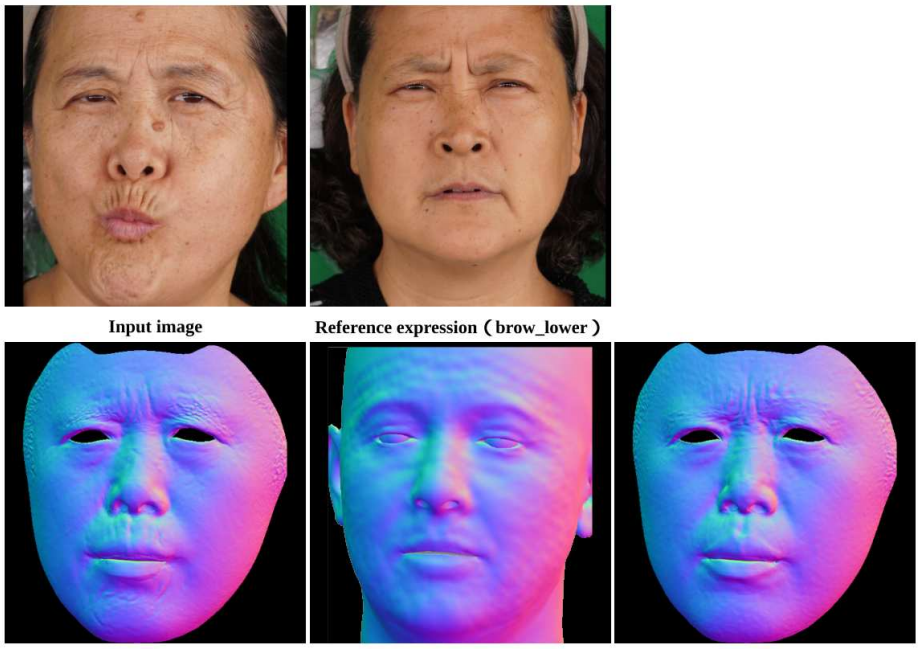


Fig. 2: A user study example. We use normal map rendering for each method.

D User study

To compare with FaceScape and DECA, we conduct a user study to measure: (1) how well each method preserves the input identity, (2) how well it conveys the target expression the user wants to change to, and (3) the overall generation quality. First, we generated 297 samples from the dataset from [5], each containing an input image, a reference image with a different expression, and the editing results generated by different methods. In generating these samples, while we use DECA's original renderer to better visualize their results in the qualitative study, we used a normal map rendering shown in Fig. 2 for all the methods to render details without bias. The results are also randomly shuffled. Then, 20 randomly selected samples were provided to each participant, and they rated in the three aspects mentioned above from 1 to 5 (higher is better). In total, we

Table 2: User study results vs. FaceScape[6] and DECA[1]. FaceScape is expectedly better at preserving identity, at the cost of not animating details to convey target expression. Our method is considered the best in overall quality.

	Ours	FaceScape	DECA
Same identity	3.33	**3.37**	1.37
Convey target expression	**3.40**	3.01	1.87
Overall better	**3.44**	3.18	1.36

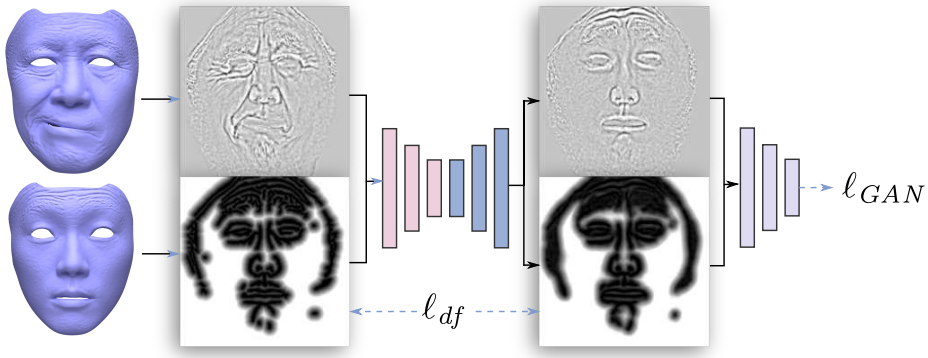


Fig. 3: Training pipeline for wrinkle line editing.

collected 282 valid responses from 15 participants. The average ratings are shown in Table 2. Notice that FaceScape treats the dynamic details as static ones and wrongly keeps them for other expressions. Since all the details are kept, it may lead to "better" identity preservation as a side effect. Our method is better than FaceScape in conveying target expressions and is considered the best in overall quality.

E Additional discussion on wrinkle line editing

The key to achieving wrinkle line editing is using mismatched distance field and displacement map in training, as shown in Fig. 3. Specifically, the input displacement map represents the original details. As the distance field is from another random face, it is used to mimic the user editing. ℓ_{df} supervises the output distance field to be consistent with the input, keeping the output wrinkle structure consistent with the edits. ℓ_{GAN} keeps the output displacement map consistent with the output distance field, thus translating the distance field to the final details represented by the displacement map.

F More qualitative results

Here we present more extreme results from Feng et al.[4], NoW[2], and CelebA-HQ[3] datasets in Fig. 4, where more varieties in skin tones and head poses are well handled. Note that we cannot handle extreme profile poses as the used large-scale 3DMM fitting fails, which is beyond the scope of this paper on detail modeling.

G Additional qualitative comparison on in-the-wild images

Our method can reconstruct and manipulate details from an in-the-wild image. In Fig. 5, 6 and 7, we show more comparison results on the images from CelebA-HQ[3] dataset. We generate more diverse dynamic details corresponding to the reference expression, and can properly animate the dynamic details in the input image. As a morphable model-based method, we are also more robust than FaceScape in handling occlusions like facial hair.

H Examples of extracted detail line maps and distance fields

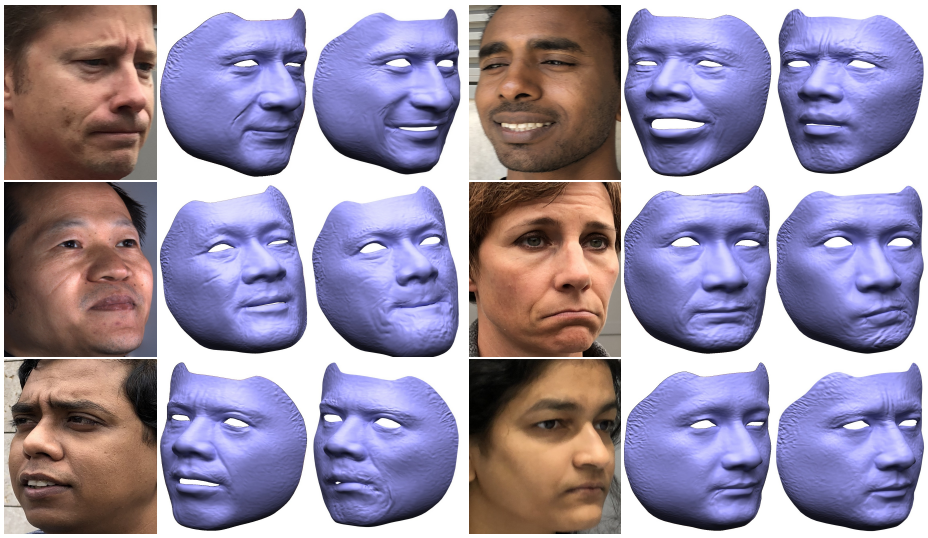
We use a distance field as the detail structure representation in our model. We obtain the distance field by extracting lines from the displacement map, and converting the lines to a distance field. More examples of the input scans, displacement maps, extracted line maps and distance fields are shown in Fig. 8.

References

1. Feng, Y., Feng, H., Black, M.J., Bolkart, T.: Learning an animatable detailed 3d face model from in-the-wild images. ACM Transactions on Graphics (TOG) **40**(4), 1–13 (2021)
2. Feng, Z.H., Huber, P., Kittler, J., Hancock, P., Wu, X.J., Zhao, Q., Koppen, P., Rätsch, M.: Evaluation of dense 3d reconstruction from 2d face images in the wild. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). pp. 780–786. IEEE (2018)
3. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. In: International Conference on Learning Representations (2018)
4. Sanyal, S., Bolkart, T., Feng, H., Black, M.J.: Learning to regress 3d face shape and expression from an image without 3d supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7763–7772 (2019)
5. Wang, Z., Yu, X., Lu, M., Wang, Q., Qian, C., Xu, F.: Single image portrait relighting via explicit multiple reflectance channel modeling. ACM Transactions on Graphics (TOG) **39**(6), 1–13 (2020)
6. Yang, H., Zhu, H., Wang, Y., Huang, M., Shen, Q., Yang, R., Cao, X.: Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 601–610 (2020)

7. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)

(a) More diverse skin tones.



(b) More diverse head poses.

Fig. 4: Input images, our reconstruction and editing results.

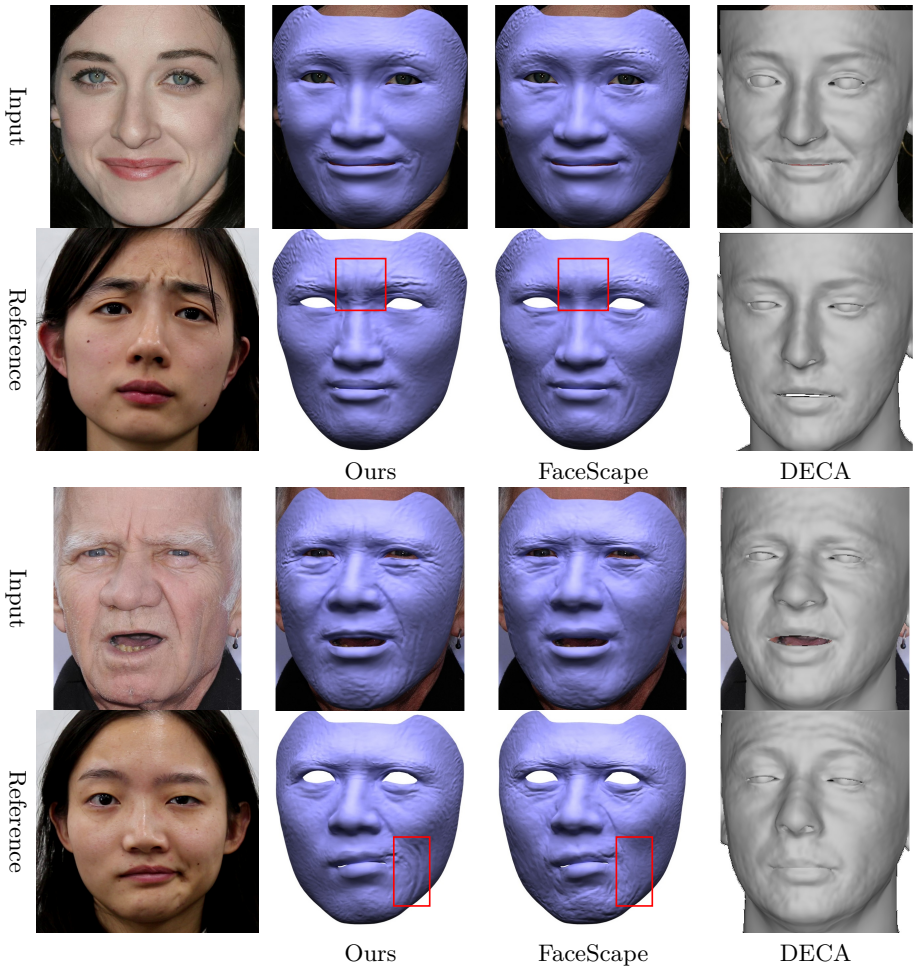


Fig. 5: Qualitative results on CelebA-HQ.

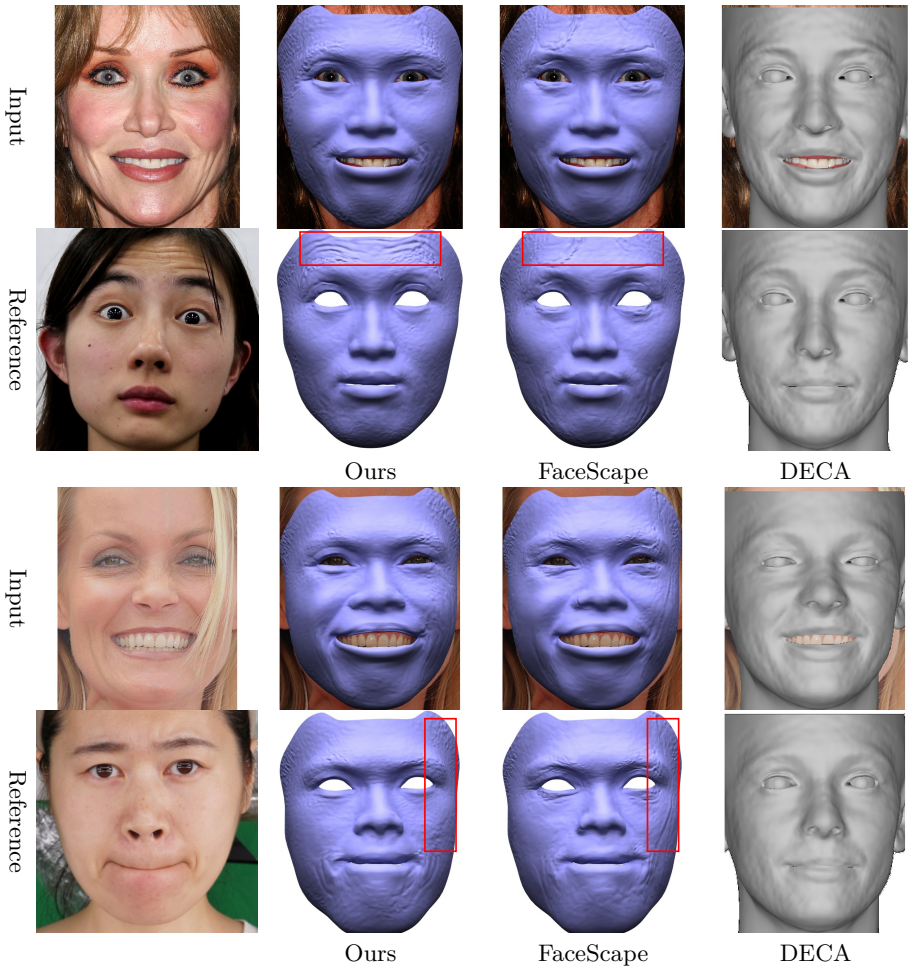


Fig. 6: Qualitative results on CelebA-HQ.

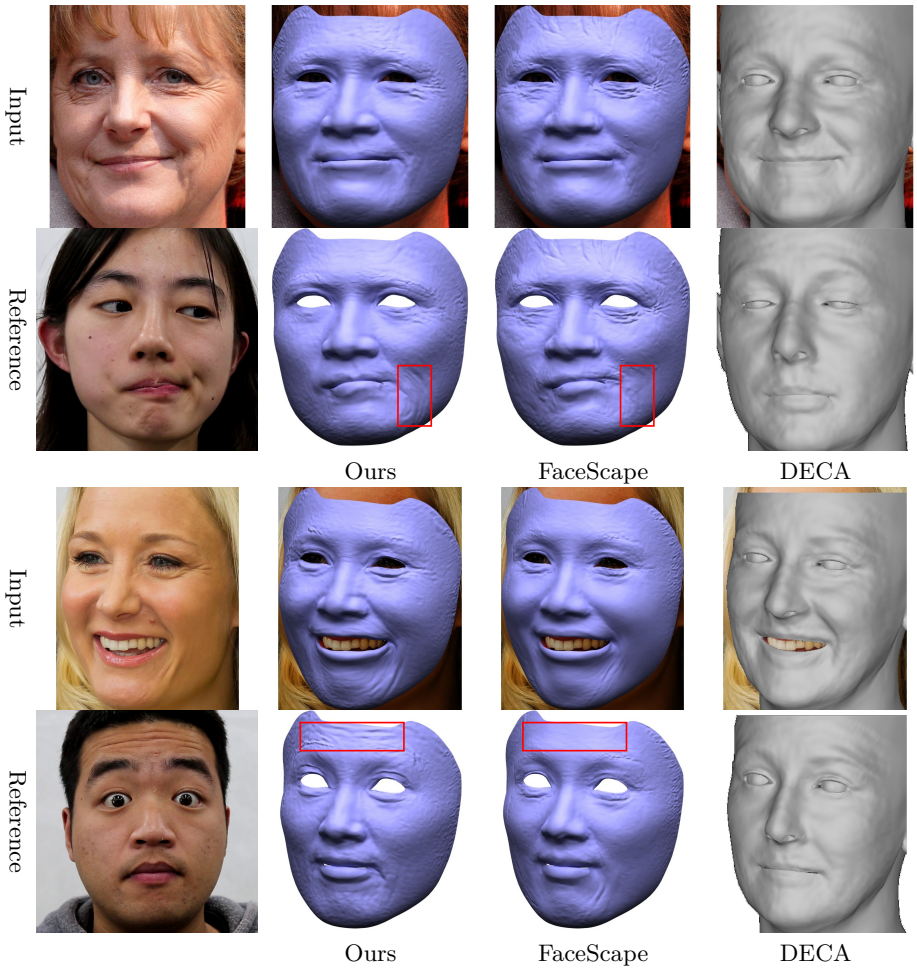


Fig. 7: Qualitative results on CelebA-HQ.

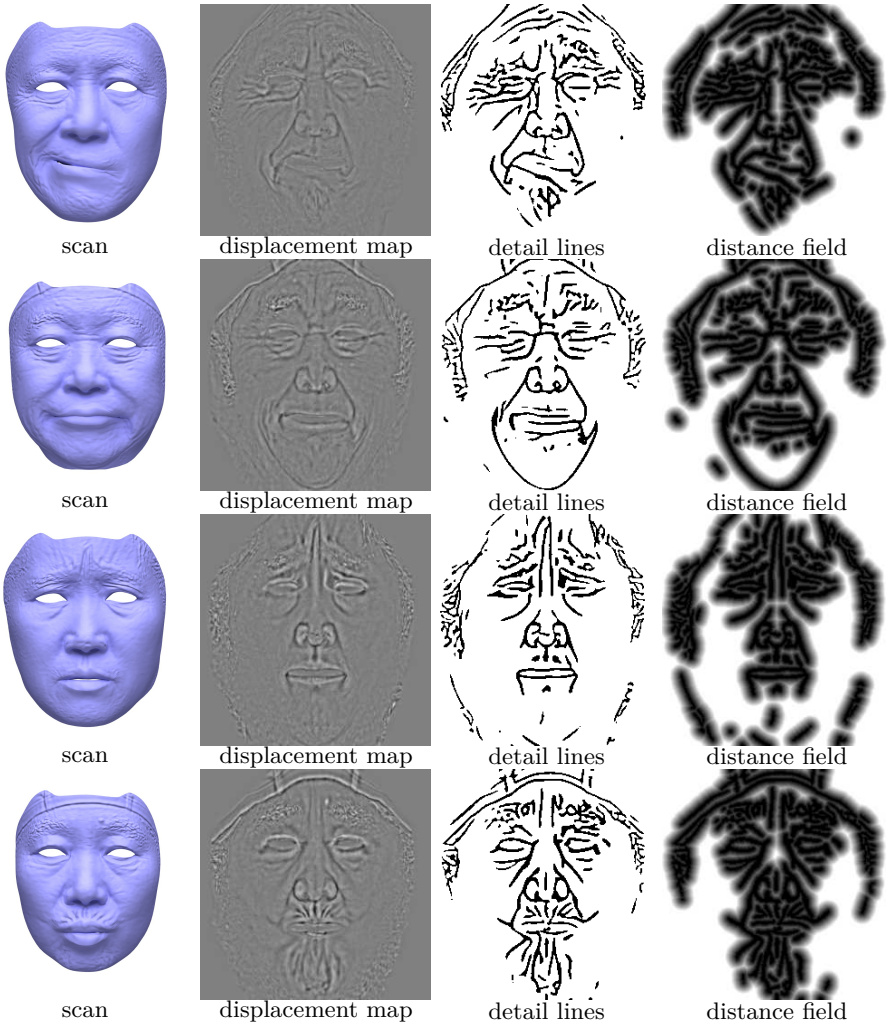


Fig. 8: Scans, displacement maps, extracted detail lines and distance fields.