Deformable Model-Driven Neural Rendering for High-Fidelity 3D Reconstruction of Human Heads Under Low-View Settings (Supplementary Material)

Baixin Xu[1*] Jiarui Zhang[2] Kwan-Yee Lin [3,4] Chen Qian [5] Ying He[1†]
[1] S-Lab, Nanyang Technological University [2] Peking University
[3] The Chinese University of Hong Kong [4] Shanghai Artificial Intelligence Laboratory
[5] SenseTime Research

In the appendix, we present 1) additional results of unseen identities and low-view inputs in Section A, which demonstrate that the pre-trained template serves as a good initialization and enables our method to adapt to new identities not available in the training dataset; 2) detailed results of our method and compare them with NeuS, HF-NeuS, and VolSDF on the PR-Senior and PR-Young datasets under a 10-view setting in Section B; and 3) an application on color transfer in Figure A4 to demonstrate the flexibility and potential of our geometry decomposition.

A. Experimental Results

In this section, we present additional results of unseen identities and results of sparse views.

A.1. Unseen Identities

Our method has the ability to adapt to new individuals as the pre-trained template serves as a good initialization. To verify this, we consider 3 **new** identities (Models 552, 555 and 598) and each identity is associated with only 5 views.

We adopted the pre-trained template, i.e., the one trained on 30 identities the PR-Senior and PR-Young datasets with 10 views for each identity, to learn the fine details for each identity in Stage 2. We observed that our method also produced plausible results for the unseen identities as shown in Table A2. This demonstrates that the pre-trained template can adapt to new identities.

A.2. Small Dataset

We also conducted another experiment using the 3 unseen identities as a small dataset. In the second experiment, we trained a **new** template using only the 15 images of the 3 new identities in Stage 1 and then used the template to learn the fine details for each identity in Stage 2. Without a surprise, the geometry of the newly trained template is worse than that of the pre-trained template due to significantly fewer views involved in Stage 1 training. Still, our method produced a fairly good result and none of the other methods, VolSDF, NeuS and HF-NeuS, were able to reconstruct satisfactory geometry with only 5 views as input as illustrated in Figure A2 and Figure A3.

A.3. Sparse Views

This section provides further results on sparse views as shown in Figure A1. Moreover, Table A2 also demonstrates that our approach surpasses other methods in terms of reconstructing geometry under sparse view conditions.

B. Analysis

We provide a thorough evaluation of our method and the state-of-the-art methods NeuS, VolSDF, and HF-NeuS on the PR-Senior and PR-Young datasets in this supplementary material. Our analysis includes a discussion of the strengths and weaknesses of each method and a comparison of their performance under various settings. It is worth noting that when VolSDF or HF-NeuS fails to reconstruct the geometry for certain models, we exclude them from the calculation of Chamfer distances for their methods. However, we use all models when calculating the Chamfer distances for our method and NeuS, both of which can reconstruct geometry for all 30 identities.

B.1. Comparison to VolSDF

In the 10-view setting, VolSDF generates erroneous geometry for Models 377, 383, and 401 due to the insufficient number of views. As illustrated in Figure A5, VolSDF only learns a partial geometry for the training views, resulting in poor novel view synthesis results.

Although the reconstruction quality improves with 15 views, VolSDF still fails to reconstruct Models 558 and 608.

*Project page: https://github.com/xubaixinxbx/3dheads.
†Corresponding author: Y. He (yhe@ntu.edu.sg).

Symbol	Meaning
I_i	the input images with camera parameters
f_{geo}	the Geometry Network
f_{tem}	the Template Network
f_{def}	the Deformation Network
f_{ren}	the Rendering Network
f_{dis}	the Displacement Network
$\mathbf{z}_s, \mathbf{z}_c \in \mathbb{R}^{128}$	identity-dependent latent codes for shape and color
$\mathbf{F}_{\text{def}} \in \mathbb{R}^{192}$	identity-dependent feature associated with non-rigid deformation
$\mathbf{F}_{\text{tem}} \in \mathbb{R}^{64}$	identity-independent feature associated with the template head
$\mathbf{F}_{\text{dis}} \in \mathbb{R}^{64}$	ID-dep. geometry feature associated with displacement
$\mathbf{F}_{\text{all}} \in \mathbb{R}^{320}$	the overall feature fed into the Rendering Network in Stage 2, which is the concatenation of $\mathbf{F}_{\text{def}}$, $\mathbf{F}_{\text{tem}}$, and $\mathbf{F}_{\text{dis}}$
$\mathbf{x} \in \mathbb{R}^3$	a query point in the observation space
$\mathbf{d} \in \mathbb{R}^3$	an offset vector indicating the deformation from an individual to the template
$\mathbf{x} + \mathbf{d} \in \mathbb{R}^3$	a query point in the template space
$s \in \mathbb{R}$	signed distance
$\mathbf{n}_b, \mathbf{n}_f \in \mathbb{R}^3$	normal vectors of the base and final surfaces
$\delta \in \mathbb{R}$	an implicit displacement
$c \in \mathbb{R}^3$	radiance
$C \in \mathbb{R}^3$	rgb color

Table A1. Notation Table.

Model	NeuS			HF-NeuS			VolSDF			Ours(Template on the 3 ids)			Ours(Template on 30 ids)		
	CD (10^{-4})	PSNR_t	PSNR_n	CD (10^{-4})	PSNR_t	PSNR_n	CD (10^{-4})	PSNR_t	PSNR_n	CD (10^{-4})	PSNR_t	PSNR_n	CD (10^{-4})	PSNR_t	PSNR_n
552	3.769	35.22	21.64	N.A.	35.40	13.36	8.192	33.62	23.13	1.815	33.58	26.51	1.197	34.92	25.88
555	3.614	35.37	18.97	N.A.	35.65	12.19	243.4	33.45	11.80	1.254	33.30	24.14	1.071	35.04	23.39
598	16.25	36.39	21.89	N.A.	36.30	12.81	23.76	35.63	20.27	1.056	35.70	27.02	1.020	35.50	27.39

Table A2. Performance on three unseen identities under 5 views. N.A. indicates no results successfully reconstructed.

It is possible that VolSDF was successful on these two models with 10 views, but failed on them with 15 views because we chose the input views **randomly** from the original PR dataset in order to test the robustness of various approaches. The redundant information in the given views may not be helpful for improving the reconstruction quality of VolSDF.

In the 20-view setting, VolSDF still failed on Model 571. We noticed that this model is affected by the failure reconstruction of the neck shown in Figure A9, which leads to inaccurate cropping of the face.

B.2. Comparison to NeuS & HF-NeuS

We found that NeuS was successful in reconstructing all 3D human heads in our experiments. However, due to the lack of modeling high-frequency signals, it cannot recover fine details, resulting in Chamfer distances in their results that are 1 times larger than ours. In contrast, our method can reconstruct fine details such as wrinkles, scarves, and hair, thanks to the additional degree of freedom provided by the displacement field.

HF-NeuS extends NeuS by learning a displacement field for representing high-frequency details. It typically achieves the best performance with a sufficient number of views. However, as the number of views decreases, the 3D reconstruction quality often degrades significantly. The reason is that HF-NeuS learns both the base surface and the high-frequency details at the same time. Such a learning process is unstable under a low-view setting. With only 10 views, HF-NeuS failed to reconstruct geometry for 19 out of 30 subjects, while with 15 views, it failed on Models 376, 377, 383, 396, 435, 469, 487, 491, 548, and 566. Even with 20 views, HF-NeuS still failed to reconstruct six models, which are 487, 548, 608, 399, 413, and 397. These results confirm that learning high-frequency details from low-view inputs is a challenging task.

In contrast, our method tackles this challenge by adopting a geometry-decomposition and a two-stage training framework. The template is trained on multiple persons with randomly chosen views. Although the number of views for each person is still low, the randomly selected views complement each other and provide a complete head. This template provides a good initialization for training the displacement in Stage 2.

Comparing to NeuS and HF-NeuS, our method performs consistently well in terms of geometry measure under the same low-view settings, thanks to the use of a pre-trained template and the displacement field.

While both NeuS and HF-NeuS produce high-quality RGB images for training views, which are 0.5-1.5 dB higher than ours, their novel view synthesis results are consistently

Figure A1. Results for Model 383 with only 5 views as input. The template human head was trained using 5 randomly selected views for all 30 identities of the PR-Senior and PR-Young datasets. The images of Model 383 for Stage 1 training and Stage 2 training are the same, therefore no additional views were provided.

worse than ours with a 2-4 dB lower score. This is attributed to the fact that these methods have less accurate geometry reconstruction and do not incorporate multiple views from various identities.

B.3. Summary

Our approach is specifically designed to enhance the performance of 3D reconstruction in low-view settings and complements the existing methods, such as VolSDF, NeuS and HF-NeuS, by utilizing a pre-trained template and a two-stage training framework. We do not intend to replace these methods, but rather to **improve their performance in such scenarios**.

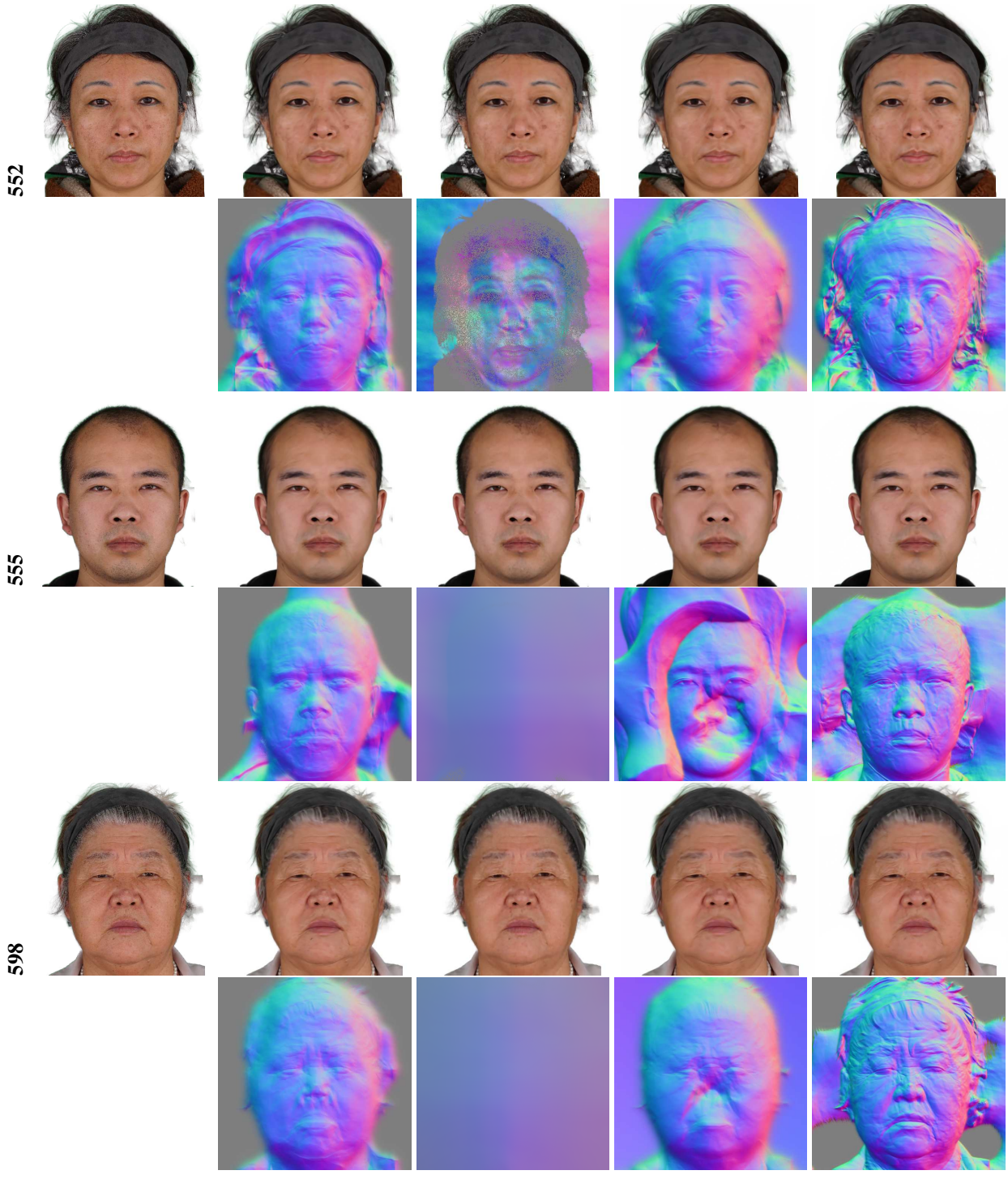

Figure A2. Training view results for 3 unseen identities (552, 555, 598) with only 5 views.

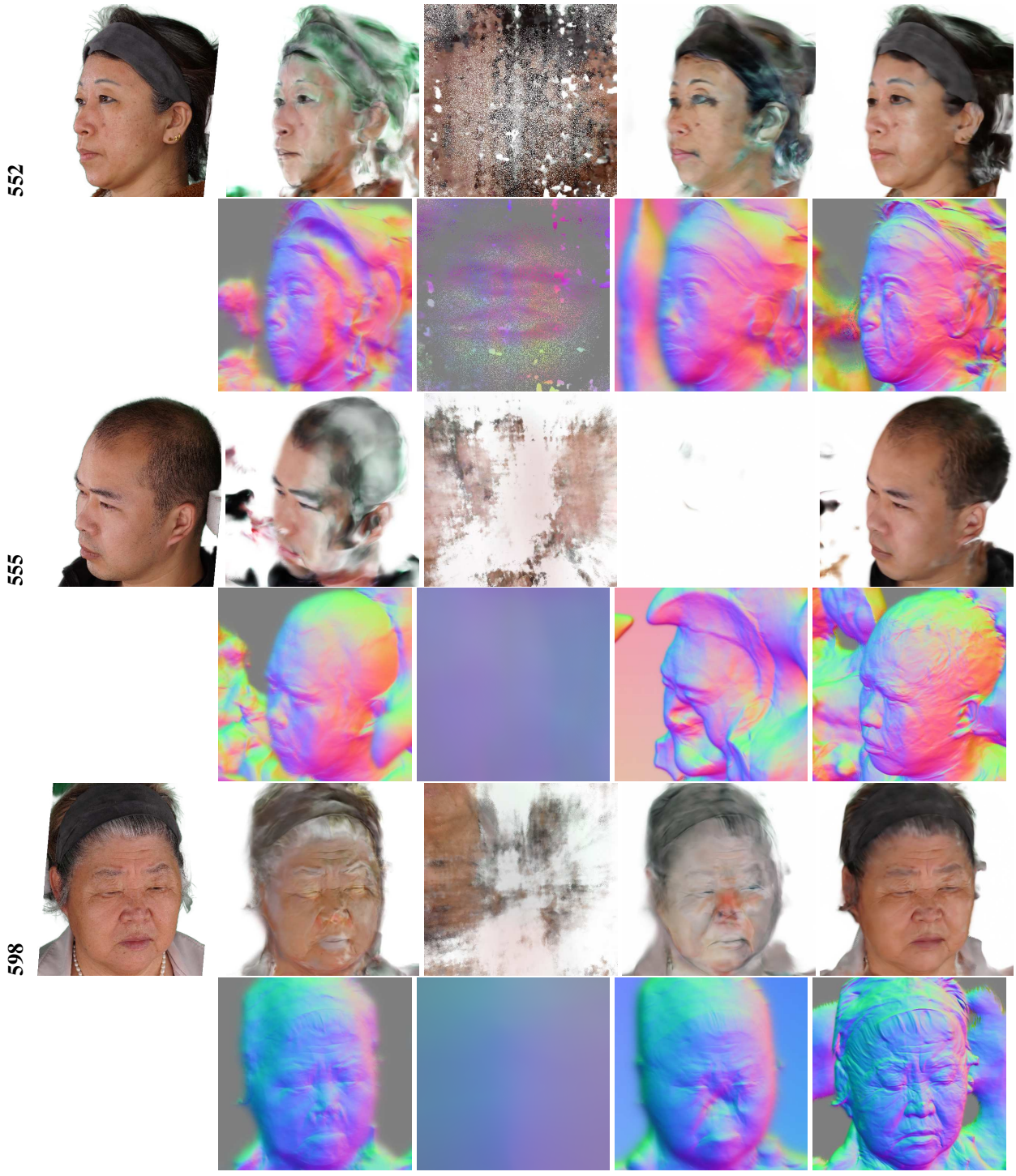

Figure A3. Novel view results for 3 unseen identities (552, 555, 598) with only 5 views.

Reference identity | Source identity | Source identity normal map | Color transfer result

Figure A4. Our approach of decomposing geometry and appearance enables us to transfer the color appearance from one model to another while keeping the geometry unchanged. In this figure, we provide two examples of transferring skin colors from a reference identity to a source identity. All models are trained under 20 views. Notably, our method can preserve small geometric features, such as speckles, which are encoded in the SDF. This is in contrast to existing image-based color transfer algorithms, which cannot differentiate between geometric features and skin colors, often transferring them together.

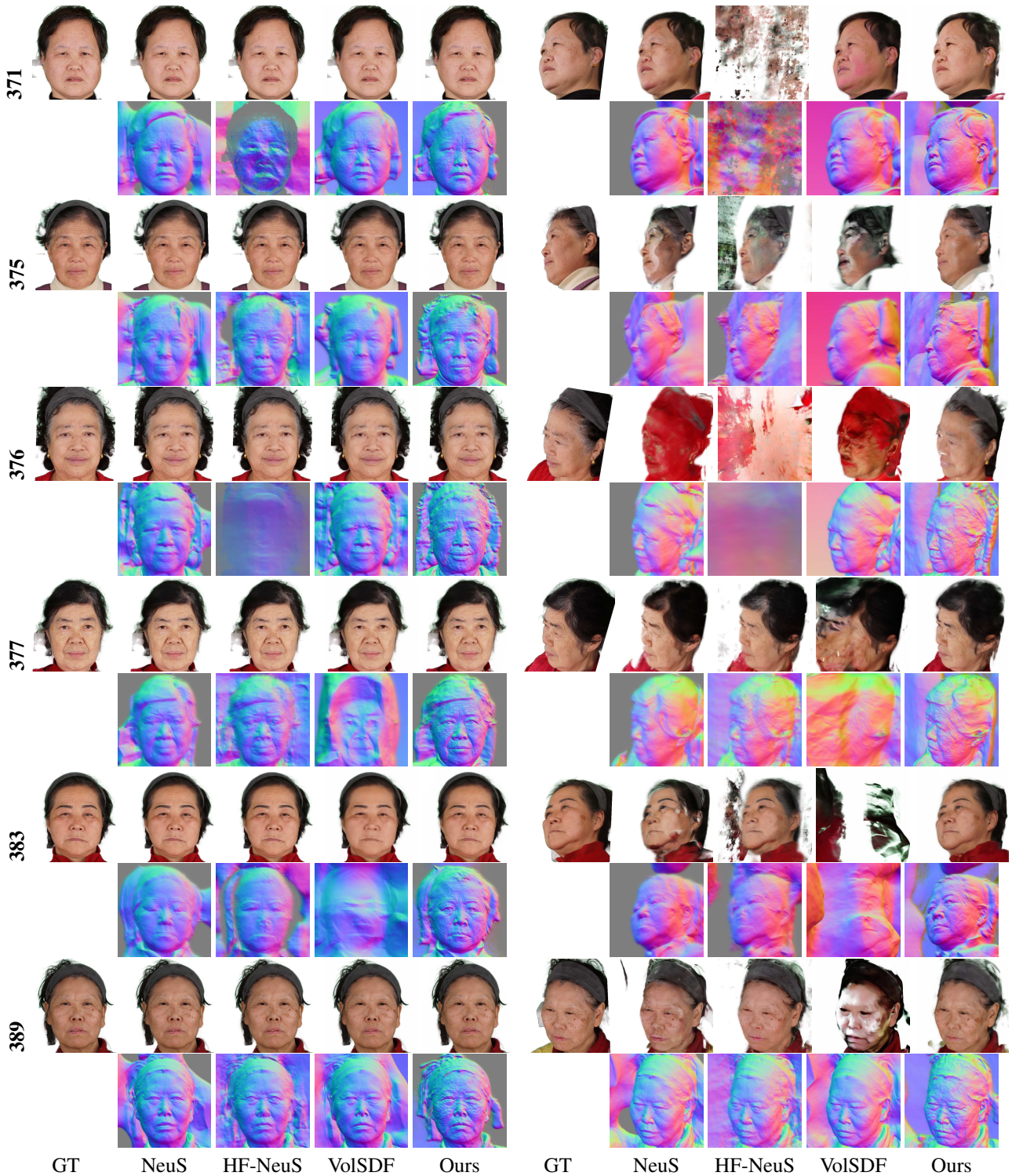

Figure A5. Comparison of various approaches under a 10-view setting (from Model 371 to Model 389). For each model, we show the results on one training view (left) and one novel view (right).

Figure A6. Comparison of various approaches under a 10-view setting (from Model 395 to Model 413). For each model, we show the results on one training view (left) and one novel view (right).

Figure A7. Comparison of various approaches under a 10-view setting (from Model 416 to Model 451). For each model, we show the results on one training view (left) and one novel view (right).

Figure A8. Comparison of various approaches under a 10-view setting (from Model 454 to Model 548). For each model, we show the results on one training view (left) and one novel view (right).

Figure A9. Comparison of various approaches under a 10-view setting (from Model 558 to Model 635). For each model, we show the results on one training view (left) and one novel view (right).