Supplementary Materials of Learning Neural Proto-face Field for Disentangled 3D Face Modeling In the Wild

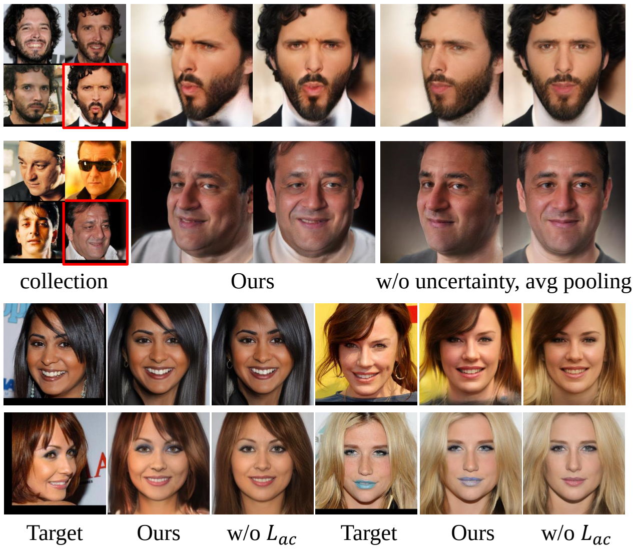

Figure 1. Analysis on the uncertainty and appearance consistency loss. Red box signs the target image of the collection.

1. More Implementation Details

Here we provide more details of the implementation. For the EG3D model, we use the one pretrained on FFHQ [3] dataset with the volume-rendering size of 128 × 128. For the deformation network Φ_{Δ}^{j} on j-th stage, we use a 3-layer MLP to extend the representation of expression coefficient β_i from 64 to 256, then concatenate it with $\bar{\mathbf{h}}^j$. The concatenated representation is then fed into another 3-layer MLP to generate the deformation code $\Delta s_i^j \in \mathbb{R}^{512}$. To evaluate the point-to-plane distance between the predicted geometry and ground truth, we first manually select 7 different facial key point on both meshes, and perform rigid align using ICP method. Then the metric can be calculated.

2. More Ablation Study

Here we provide more ablation study on the components of the proposed method. First, we compare our full method with the one using average pooling instead of uncertainty to perform style code fusion, and illustrate the results in Fig. 1. We observe that without uncertainty modeling, the average-pooling model cannot well represent the identity feature of the target image after NPF fitting, providing improper facial shapes and appearance. Besides, the expressions cannot be suitably predicted, either. In contrast, our method well models the facial geometry and corresponding expressions. Note that, the facial shape is actually a blended shape with identity and expression. As a result, leveraging uncertainty adaptively digs more suitable ID consistency to boost the deformation modeling.

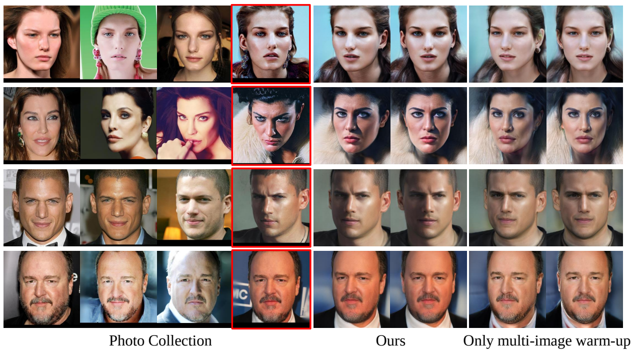

Figure 2. Analysis on the NPF fitting setting.

Then, we also perform analysis on the appearance consistency loss $\mathcal{L}_{ac}$ in Fig. 1. We observe that using $\mathcal{L}_{ac}$, the appearance details such as the hair, eye shadow and lipstick keep consistent to the target. On the contrary, discarding $\mathcal{L}_{ac}$ loses such features during rotation. Further, we also perform comparison between our full method and that only using multi-image warm-up during NPF fitting, and illustrate the results in Fig. 2. We observe that only using multi-image warm-up loses specific details, appearance and shadows. The reason behind is that these scene-specific features are not commonly shared within the photo collection, and multi-image warm-up urges the model to learn a mean condition. In contrast, our full method utilizes target-image fitting after the warm-up procedure, thus the specific lighting conditions, make-ups and other details can be recovered.

Finally, we use different approaches to get the style code $\mathbf{s}_i$, and illustrate their effects in Table 1. We leverage two widely-used methods PSP [4] and E4E [6] to predict the style code from each image. We observe that using the di-

No.	method	p2p (mm) ↓	MAD (deg.) ↓	IDE ↓
(1)	Ours	1.77	11.87	0.257
(2)	Ours-PSP [4]	1.80	11.89	0.260
(3)	Ours-E4E [6]	1.82	11.86	0.263

Table 1. Comparison with different style code optimization method. Similar performance is obtained.

No.	method	p2p (mm) ↓	MAD (deg.) ↓	IDE ↓
(1)	Ours	1.77	11.87	0.257
(2)	relaxed-6	1.83	12.21	0.296
(3)	relaxed-4	1.90	12.35	0.303
(4)	relaxed-2	2.57	12.50	0.389
(5)	single-image PTI [5]	2.35	12.10	0.482

Table 2. Analysis on the relaxed photo collection.

rect scheme [3] to optimize the style code obtains similar accuracy to that of PSP and E4E. Such a phenomenon reveals that our method is robust and insensitive.

3. Limitation

Multi-image input: Here we discuss the limitation of our method. As introduced in the Sec. 5 of main article, our method requires multiple images of a same identity as input. Although the in-the-wild image collections are easy to get, such a setting may limit the application in some extreme conditions. Here we provide a possible solution by building a 'relaxed' photo collection which is composed of images with similar facial shapes. We first predict identity coefficient for each image using the pretrained 3D face network [1], then select 5 nearest faces by calculating the coefficient difference between the target image and selected one. Combined with the original target image, we treat the 6 photos as a relaxed collection and use our proposed method to recover the target face. The results are shown in Table 2, where the flag '-n' means the relaxed collection contains n images. We observe that with 6 images, the relaxed collection gets only slightly weaker performance than our standard setting and still outperforms the single-image baseline. This reveals that even images with similar facial shapes still provide reliable priors to complement for the monocular ambiguity. The performance decreases with 4 images. When the collection has only two images, the performance suffers from the ambiguity and shape conflict between the original image and the selected one. Note that, even the nearest facial image has the lowest coefficient difference, it may not be the nearest one in physical world due to the inaccuracy of the 3D face network. In summary, with a suitable size of relaxed photo collection, our method still provides more accurate reconstruction performance than the single-image baseline.

Unusual conditions: Another limitation is the degradation on processing unusual expressions, lighting and artifacts. As illustrated in Fig. 3, we observe that our method suffers from extreme challenging conditions that hardly appear in the dataset, especially when a facial image contains several challenging factors. The reason behind is multi-aspect. On the one hand, recovering challenging expression is actually an opening problem in 3D face reconstruction, even the state-of-the-art model [1] cannot well handle it. As a result, our method is influenced by the degradation from 3D face model. On the other hand, the pretrained EG3D model also limits the performance. Although trained on FFHQ dataset that has various portrait images, the model still has difficulty recovering the artifacts, expressions and lighting appearance that have no common features within the dataset. One possible solution is to implement more targeted data to boost the EG3D model's learning. Using GAN-based image generation to synthesize images with various expressions as training samples is also a possible solution. For the harsh lighting conditions, the model could be constrained by using a lighting prior as input. In summary, our method still outperforms the EG3D + PTI [5] baseline under unusual conditions, and produces clearer and more reasonable face reconstructions.

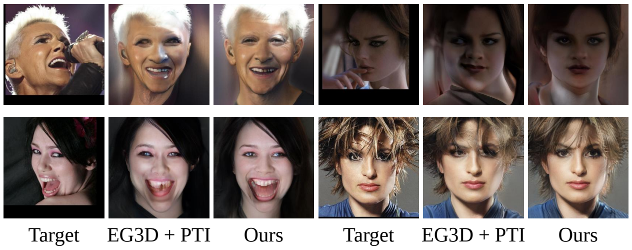

Figure 3. Limitations of our method.

4. More Results

Here we also provide more qualitative results of our method. As illustrated in Fig. 4, our method significantly outperforms EG3D + PTI baseline and HeadNeRF [2] on the extreme conditions of poses, artifacts, lighting and appearance, obtaining much more robust performance. Further, in Fig. 5 we observe that our method provides reliable reconstruction on both geometry and texture.

References

[1] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *CVPRW*, 2019. 2

[2] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. Headnerf: A real-time nerf-based parametric head model. In *CVPR*, pages 20374–20384, 2022. 2

[3] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019. 1, 2

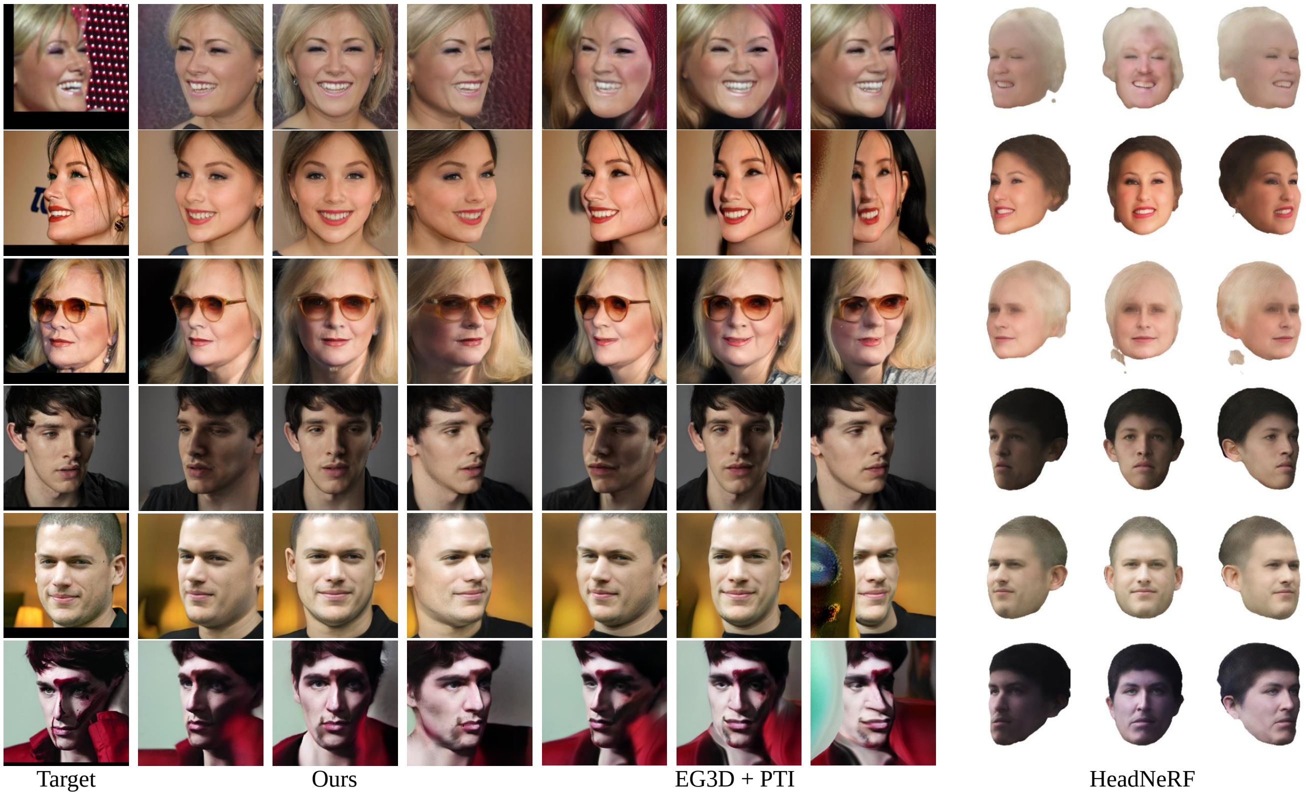

Figure 4. More comparisons between our method and the state-of-the-art.

[4] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *CVPR*, pages 2287–2296, 2021. 1, 2

[5] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Transactions on Graphics (TOG)*, 42(1):1–13, 2022. 2

[6] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021. 1, 2

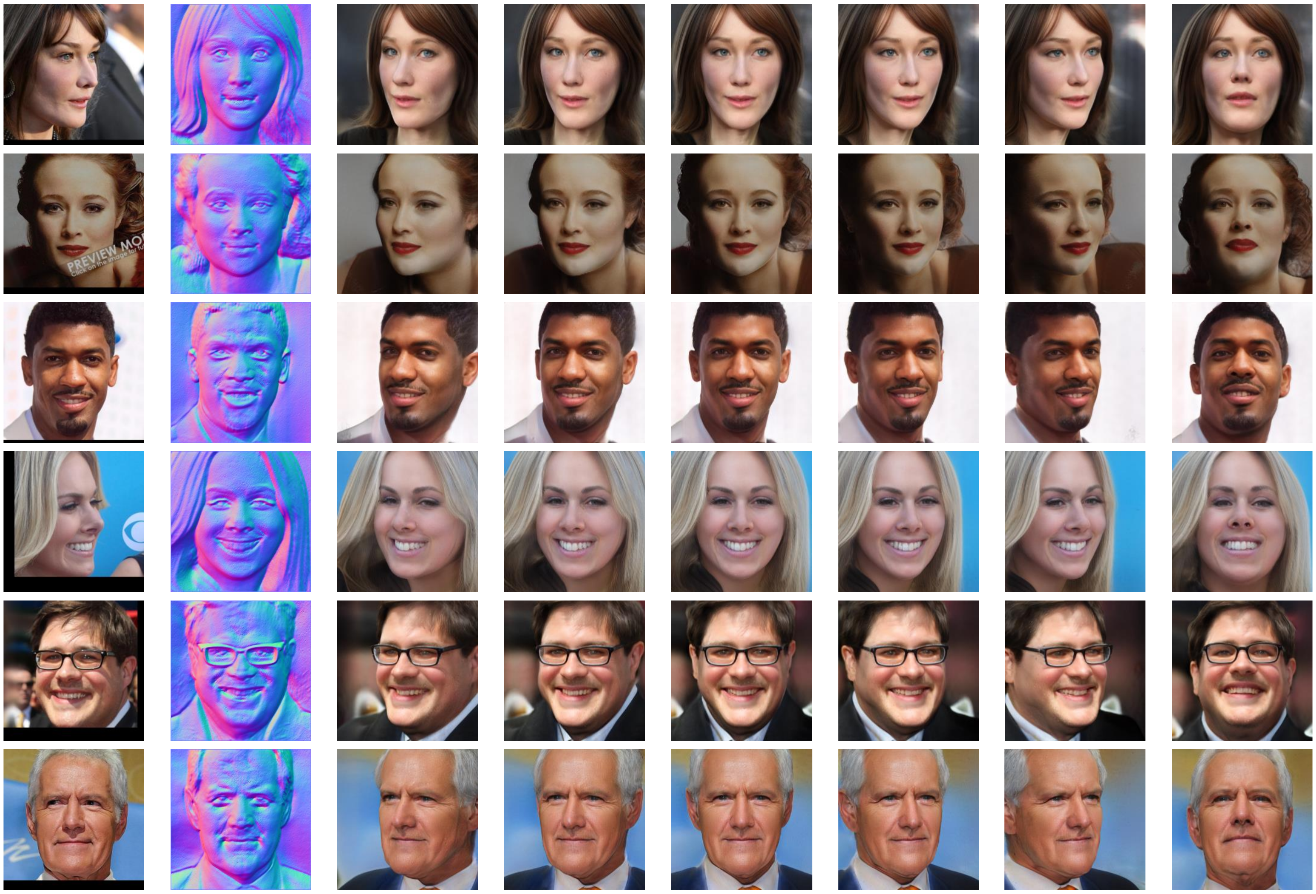

Figure 5. More reconstruction results of our method.