Supplementary of Generalizable One-shot 3D Neural Head Avatar

In this document, we first discuss the broader impacts of this work in Sec. 1. We then show additional experiments including the comparison with another concurrent baseline (*i.e*. Next3D [20]), and more ablation studies in Sec. 2. We further demonstrate more qualitative results in Sec. 3. Additional details of the proposed framework and the evaluation process are described in Sec. 4 and Sec. 5. Finally, we discuss preliminaries of the 3DMMs in Sec. 6 and the limitations of the proposed method in Sec. 7, respectively. We strongly encourage reviewers to watch the supplementary video which includes the head avatar reconstruction and animation results.

1 Broader Impact

The proposed framework has the potential to make significant contributions to various fields such as video conferencing, entertainment industries, and virtual reality. It offers a wide range of applications, including but not limited to animating portraits for film/game production, or reducing transmission costs in video conferences through self-reenactment that only requires transmitting a portrait image with compact motion vectors. However, the proposed method may present important ethical considerations. One concerning aspect is the possibility of generating "deepfakes", where manipulated video footage portrays individuals saying things they have never actually said. This misuse could lead to serious privacy infringements and the spread of misinformation. We do not advocate for such activities and instead underscore the need to build guardrails to ensure safe use of talking-head technology, such as [19; 33; 6; 2; 5].

2 Additional Experiments

2.1 Comparison to Next3D

Evaluation details. As discussed in Sec.2.3 in the main submission, controllable 3D-aware avatar generation models [22; 26; 21; 17; 27; 14; 34; 20] can be combined with GAN inversion methods [18; 8; 28; 25] to achieve head avatar animation. In this section, we substantiate this idea and compare our method against it. Specifically, we combine pivotal tuning [18] for GAN inversion with a SOTA avatar generation and animation model – Next3D [20]. Next3D is a pure generative model based on EG3D [4]. It takes a FLAME [15] avatar mesh with desired expression as input and synthesizes a photo with the target expression and random appearance. To animate a given portrait, Next3D first uses pivotal tuning [18] to map the portrait image to the latent space of its generator and then animates the portrait using expressions extracted from the target video. We use the publicly available implementation[1] of Next3D and the pivotal tuning code kindly provided by Next3D's authors. We refer to this baseline as "Next3D-PTI" in the following. Since pivotal tuning of each image takes about *5 minutes*, it is infeasible for us to carry out cross-identity reenactment on the CelebA dataset [13], which includes more than ten thousand image pairs. Thus, we compare with "Next3D-PTI" for same-identity and cross-identity reenactment on the HDTF [32] dataset, as well as cross-identity reenactment from the videos in HDTF to images in CelebA, as described in Sec.4.4 in the main submission.

[1] https://github.com/MrTornado24/Next3D

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Qualitative evaluations. The qualitative comparison of Next3D-PTI with our method, as well as other baselines can be found in Sec. 3.3, Sec. 3.4, Sec. 3.5 and the supplementary video.

Quantitative evaluations. We first show the quantitative comparisons between our method and Next3D-PTI on the HDTF dataset [32] for same-identity and cross-identity reenactment in Table 1. Our method consistently outperforms Next3D-PTI for both tasks. We further carry out motion transfer from the videos in HDTF to 60 randomly selected images in the CelebA dataset [13] and demonstrate the quantitative evaluation results in Table 2. Our method performs favorably in preserving identity and matching target pose (*i.e*. better CSIM and APD score) compared to the Next3D-PTI, while being comparable in modeling expression and capturing photo-realistic details (*i.e*. comparable AED and FID). It is also worth noting that our method takes *0.6 second* to reconstruct and animate a 3D head avatar from an unseen single-view image, while the pivotal tuning process alone takes about *5 minutes* to encode a portrait image to the latent space of the generator in Next3D.

Table 1: **Comparison with Next3D-PTI on the HDTF dataset [32].**

Methods	Same-Identity Reenactment									Cross-Identity Reenactment			
	PSNR↑	SSIM↑	CSIM↑	AED↓	APD↓	AKD↓	LPIPS↓	L1↓	FID↓	CSIM↑	AED↓	APD↓	FID↓
Next3D-PTI [20]	19.89	0.813	0.645	0.137	0.035	**1.449**	0.180	0.053	41.66	0.581	0.291	0.045	101.8
Ours	**22.15**	**0.868**	**0.789**	**0.129**	**0.010**	2.596	**0.117**	**0.037**	**21.60**	**0.643**	**0.263**	**0.018**	**47.39**

Table 2: Cross-identity reenactment between the HDTF dataset [32] and the CelebA dataset [13].

Methods	CSIM↑	AED↓	APD↓	FID↓
Next3D-PTI [20]	0.483	**0.266**	0.042	**56.01**
Ours	**0.551**	0.274	**0.017**	59.48

2.2 More Ablation Studies

The synthetic training dataset. Due to the long-tail distribution in the training data, our model fails to synthesize some rare expressions such as jaw opening realistically. This issue could be partially resolved by training our model using more balanced data. To this end, we replace EG3D [4] with Next3D [20] to produce paired data for training as discussed in Sec.4.1 in the main paper. As shown in Fig. 1, the model learned using data synthesized from Next3D has more natural jaw opening expression.

Figure 1: **Comparison of models learned with different synthetic dataset.** We demonstrate the geometry of the jaw opening expression by extracting meshes from the animated tri-plane using the Marching Cubes algorithm.

Details of the linear expression branch. We show the architecture of an alternative expression branch design in Fig. 2 (a). As described in Sec.4.5 of the main submission, this design draws inspirations from 3DMMs and learns 64 tri-plane-based expression bases denoted as $\{E_1, ..., E_{64}\}$. To produce the target expression tri-plane, it linearly combines the learnable bases by $T_c = \theta_t^1 E_1 + ... + \theta_t^{64} E_{64}$, where θ_t are the target expression coefficients extracted from the target image by the 3DMM [7]. As shown in Fig.5 and Table.4 in the main submission, this design produces unrealistic mouth regions during the animation.

Expression branch using coefficients. An intuitive design for the expression branch is to utilize an encoder that directly maps the target 3DMM expression coefficients to an expression tri-plane. We investigate this design by using two kinds of encoders: i) We simply use linear layers followed by transpose convolutional layers to map the expression coefficient vector to an expression tri-plane. We

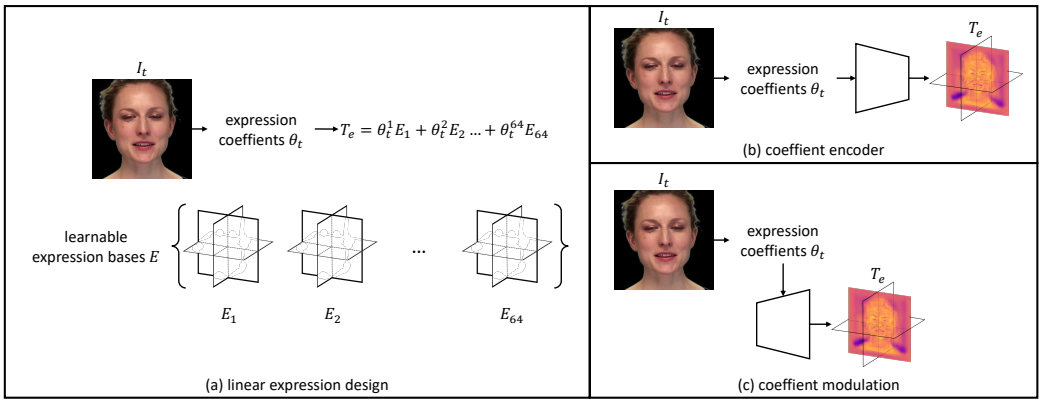

Figure 2: **Ablation variants.** See Sec. 2.2 for details.

Table 3: **Ablation studies.**Blue text highlights the inferior performance of the variants.

Methods	3D Portrait Reconstruction					Cross-Identity Reeanct			
	L1↓	LPIPS↓	PSNR↑	SSIM↑	FID↓	CSIM↑	AED↓	APD↓	FID↓
modulation	0.030	0.130	25.01	0.841	11.80	0.558	0.280	0.018	22.84
encoder	0.028	0.121	25.69	0.850	9.840	0.538	0.264	0.017	18.88
fine-tune high-res	0.028	0.114	25.87	0.870	9.177	0.597	0.278	0.017	21.51
ours	0.030	0.116	24.77	0.861	10.47	0.599	0.276	0.017	17.36

dub this design as the "encoder" design and show its architecture in Fig. 2(b). ii) We use the generator from StyleGAN2 [11] as our encoder. Specifically, we first map the target expression coefficients to a set of style vectors, the encoder takes a constant tensor as input and modulates the features at each layer using the style vectors produced from the expression coefficients. We denote this design as "modulation" and show its structure in Fig. 2(c).

As shown in Table 3, for the cross-identity reenactment evaluation on CelebA [13], both alternative expression branch designs discussed above have lower CSIM score. This is because the expression coefficients are orthogonal to the identity in the source image. By taking the expression coefficients alone as input, the model has less information of the source identity and suffers from identity preservation while animating.

Fine-tuning end-to-end in stage II. As discussed in Sec.3.5 in the submission, in order to preserve multi-view consistency, we only fine-tune the super-resolution module while freezing other parts in Stage II. We verify the effectiveness of this choice by conducting an ablation study where we instead fine-tune end-to-end in Stage II.

Table 3 demonstrates the quantitative evaluation of this variant model. Though it has better reconstruction results from the observed view, it has worse FID score for the task of cross-identity reenactment. This indicates this variant synthesizes less realistic animation results at the target pose.

3 More Qualitative Results

In this section, we show more qualitative results on the CelebA [13] and HDTF [32] datasets.

3.1 Cross-identity Reenactment on CelebA

In Fig. 4, Fig. 5, Fig. 6 and Fig. 7, we present more visualizations of cross-identity reenactment on the CelebA dataset.

3.2 3D Portrait Reconstruction on CelebA.

In Fig. 8, Fig. 9, Fig. 10 and Fig. 11, we show portrait reconstruction results by the proposed method visualized in different views .

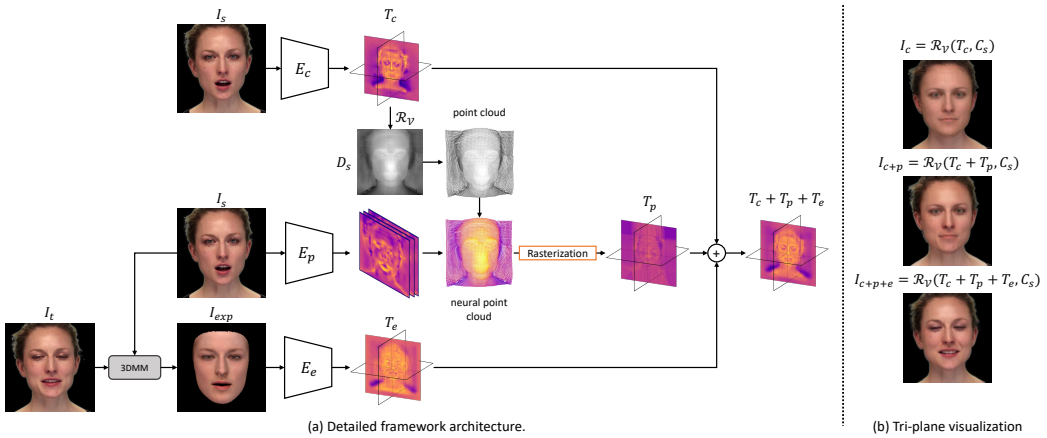

Figure 3: **Framework architecture.** (a) We present an extended version of Fig.1(a) in the submission with more framework details. The volumetric rendering and super-resolution block are left out. (b) Visualization of tri-plane combinations.

3.3 Cross-identity Reenactment from HDTF to CelebA

We visualize more cross-identity reenacment results from the videos in HDTF to the images in CelebA, including comparison with the baselines [12; 30; 20] in Fig. 12, Fig. 13, Fig. 14, Fig. 15 and the supplementary video.

3.4 Same-identity Reenactment on HDTF

Fig. 16 shows the qualitative results of same-identity reenactment on the HDTF dataset, as well as comparison with OTAvatar [16], ROME [12], StyleHeat [30] and Next3D-PTI [20].

3.5 Cross-identity Reenactment on HDTF

We present more qualitative results of cross-identity reenactment on the HDTF dataset, as well as comparison with OTAvatar [16], ROME [12], StyleHeat [30] and Next3D-PTI [20] in Fig. 17.

4 More Implementation Details

Detailed network architecture. In Fig. 3, we present the detailed architecture of the proposed method. As discussed in Sec.3 of the main submission, our model includes three branches that capture the coarse geometry, detailed appearance and expression, respectively. Specifically, the encoder E_c in the canonical branch takes a source image of size $3 \times 512 \times 512$ as input and outputs a feature map of size $256 \times 128 \times 128$. By passing the feature map through four convolution layers and one transpose convolution layer, we obtain a canonical tri-plane of size $3 \times 32 \times 256 \times 256$. The encoder E_p in the appearance branch takes the source image as input and outputs a feature map of size $256 \times 128 \times 128$. Through the "Lifting" and "Raterization" process introduced in Sec.3.2 of the main submission, we produce an appearance tri-plane of size $3 \times 32 \times 256 \times 256$. Furthermore, to prevent the expression in the source image from leaking into the final animation, we use an off-the-shelf face parsing network [31][2] to mask out the eye and mouth regions before providing the source image to the encoder E_p. Finally, the encoder E_e in the expression branch is similarly designed as E_c, except that it takes the frontal-view 3DMM rendering with the target expression as input. All three encoders (*i.e.* E_c, E_p, E_e) use a pre-trained SegFormer [24] model, up to the classifier layer. We adopt the tri-plane decoder proposed by [4] to map the interpolated tri-plane feature to color and density for each 3D point. For the super-resolution block, we fine-tune a pre-trained GFPGAN [23] model without modifying its architecture.

[2] https://github.com/zllrunning/face-parsing.PyTorch

5 More Evaluation Details

We explain details of how we evaluated the baseline methods and the proposed method in this section.

3D portrait reconstruction. For the ROME method [12], we use the publicly available code and model[3], which renders 256^2 images. For fair comparison, we resize our prediction and the ground truth images from 512^2 to 256^2. Since the synthesis from ROME is not pixel-to-pixel aligned to the input image, we rigidly transform the ground truth image and our image such that they align with ROME's predictions. To this end, we apply the Procrustes process [9; 29] to align our prediction/ground truth image to ROME's prediction using facial landmarks detected by [3]. We also replace the white background in ROME's implementation to a black one to match the ground truth images and our predictions. We compare with ROME on all 29,954 high-fidelity images in CelebA [13] for 3D portrait reconstruction.

Since the synthesis results by HeadNeRF [10][4] and our method are pixel-to-pixel aligned to the input image, we do not carry out further alignment when comparing to HeadNeRF. However, as discussed in Sec.4.2 of the main submission, applying HeadNeRF on all 29,954 images in CelebA is computationally infeasible. Thus we apply our method and HeadNeRF on a randomly sampled subset that includes 3000 images from the CelebA dataset.

Reenactment. We compare with ROME [12], StyleHeat [30][5] and OTAvatar [16] for same-identity and cross-identity reenactment. We leave HeadNeRF out on the HDTF dataset since it is impossible to test it on tens of thousands frames due to the time-consuming optimization for each frame. Moreover, OTAvatar [16] is a concurrent work and up to the date of this submission, only its partial code[6] that allows for comparison on the HDTF dataset alone has been released publicly. So we do not compare with it on motion transfer from the HDTF dataset to the CelebA dataset. For fair comparison, we align predictions from all methods to the target image using the Procrustes process discussed above. Note that synthesis from all methods have a black background and are readily comparable after the alignment.

6 Preliminaries of 3DMMs

We exploit the geometry prior from a 3DMM [1] that represents the shape and texture of a portrait by:

$$\begin{aligned} S &= \bar{S} + B_{id}\alpha + B_{exp}\beta \\ T &= \bar{T} + B_{tex}\delta \end{aligned} \tag{1}$$

where $\bar{S}, \bar{T}$ are the mean shape and texture of human faces, B_{id}, B_{exp}, B_{tex} are the shape, expression and texture bases, and α, β, δ are coefficients that linearly combine the shape, expression and texture bases, respectively. Since we mainly utilize the shape and expression components in the 3DMM in this work, we ignore its texture and illumination modules and simply denote the rendering operation from a camera view C as $I, M = \mathcal{R}_M(\alpha, \beta, C)$, where I is the rendered image, and M is the rendered mask that only includes the facial region.

7 Limitations

Teeth and pupil reconstruction. 3D head avatar reconstruction and animation is a highly challenging task. The proposed method takes the first step to produce high-fidelity results. However, to generalize to any portrait image, one dilemma is that the expression of the source portrait and target image could be arbitrary, which introduces various challenging scenarios. For instance, the source portrait image could have a closed mouth while the target expression has an open mouth (*e.g.* the second row of Fig. 4). In this case, the model should hallucinate correct inner mouth regions. Yet, in other cases, the inner mouth is visible in the source portrait (*e.g.* the sixth row in Fig. 4). To resolve

[3] https://github.com/SamsungLabs/rome
[4] https://github.com/CrisHY1995/headnerf
[5] https://github.com/FeiiYin/StyleHEAT
[6] https://github.com/theEricMa/OTAvatar

this dilemma, in this work, our model simply always hallucinates the inner mouth of the individual through our expression branch. As a result, the hallucinated mouth could deviate from the one in the source image. In other words, our model cannot accurately reconstruct teeth and lips, as shown in Fig. 18. The same analysis applies to pupils. Since we have no prior knowledge of whether the eyes in the portrait image are open or closed, our model always hallucinates the pupils through the expression branch instead of reconstructing the ones in the source image. We leave this limitation to future works.

Figure 4: **Cross-identity reenactment on CelebA.**

Figure 5: **Cross-identity reenactment on CelebA.**

Figure 6: **Cross-identity reenactment on CelebA.**

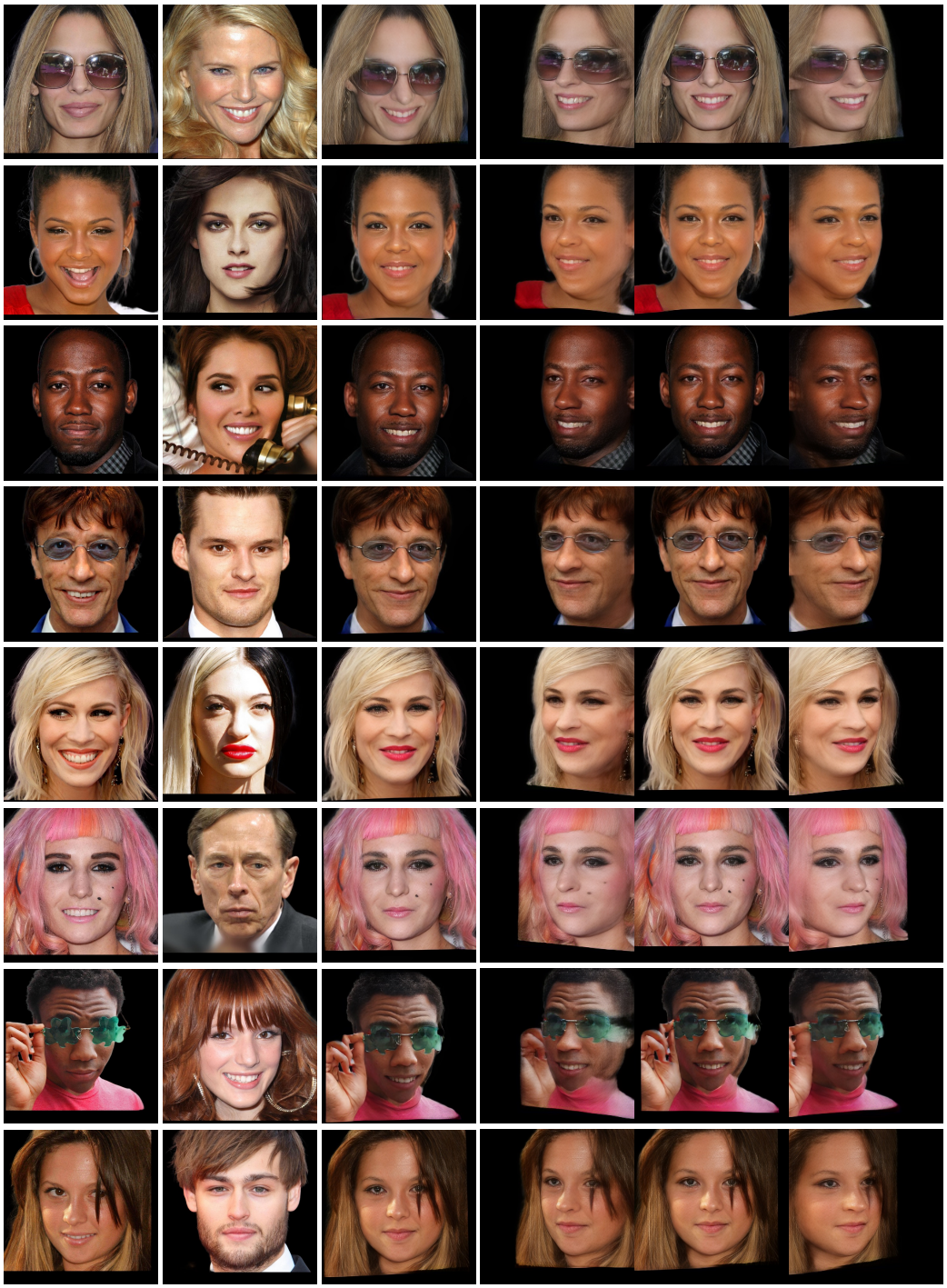

Figure 7: **Cross-identity reenactment on CelebA.**

Figure 8: **3D reconstruction on CelebA.**

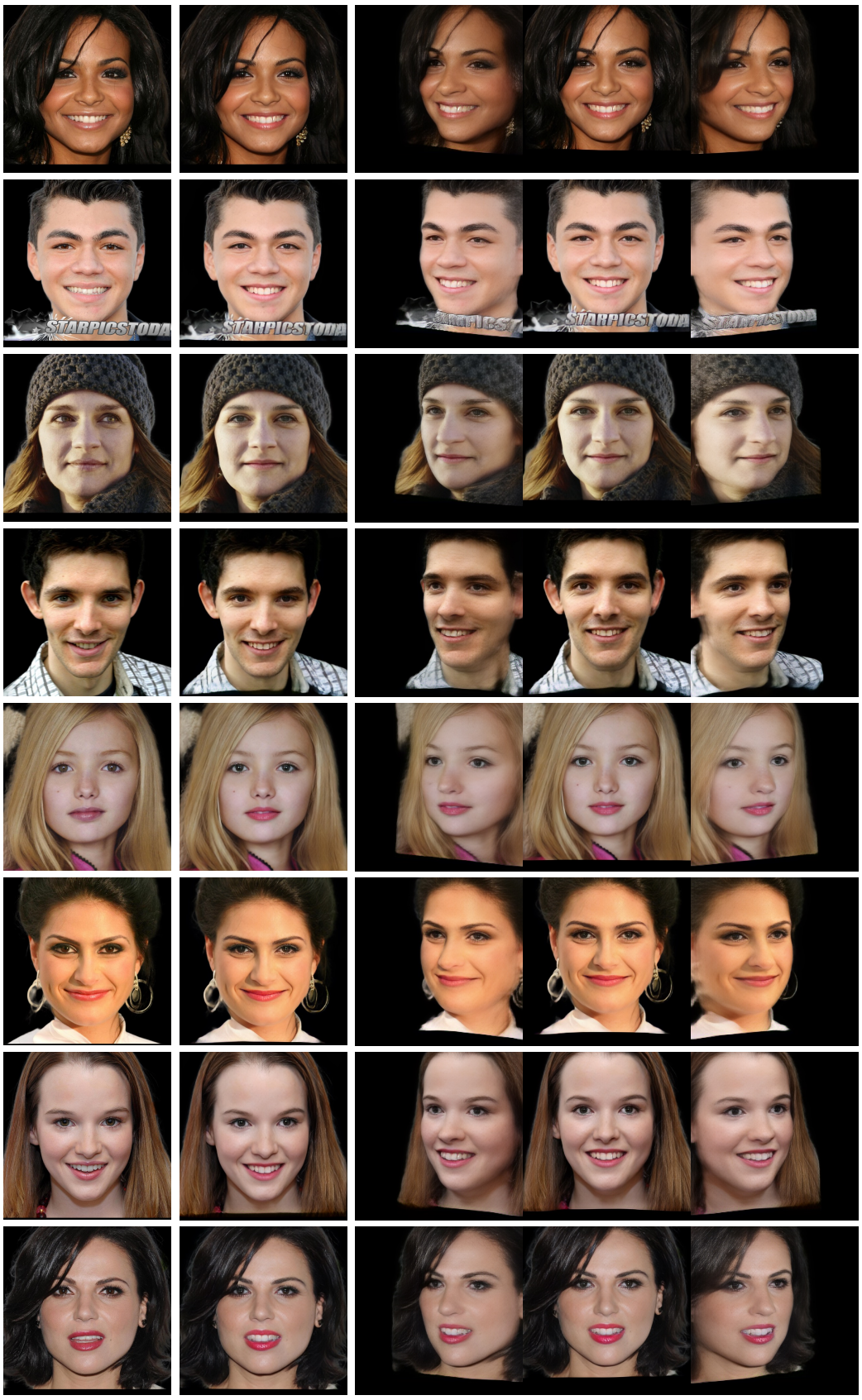

Figure 9: **3D reconstruction on CelebA.**

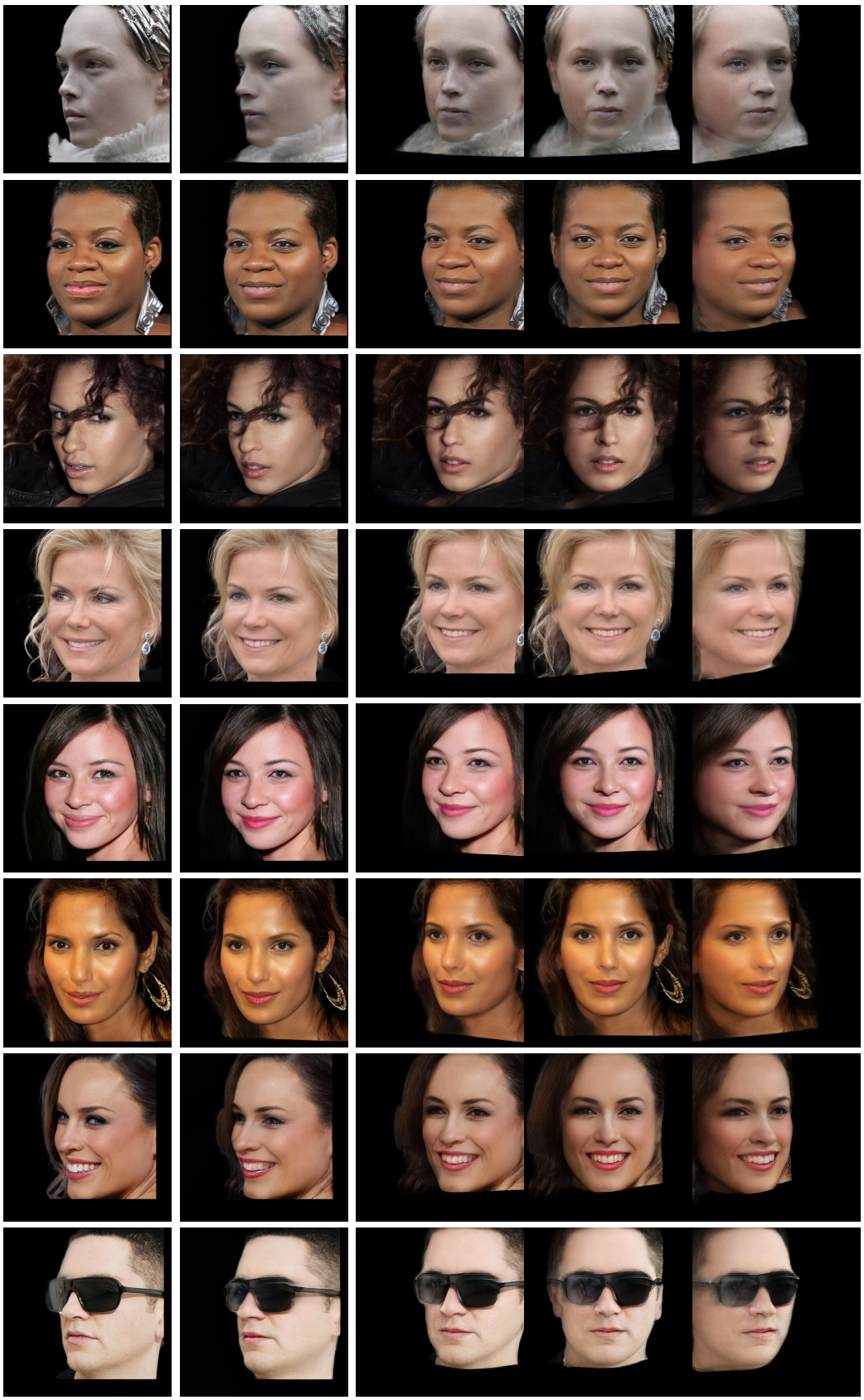

Figure 10: **3D reconstruction on CelebA.**

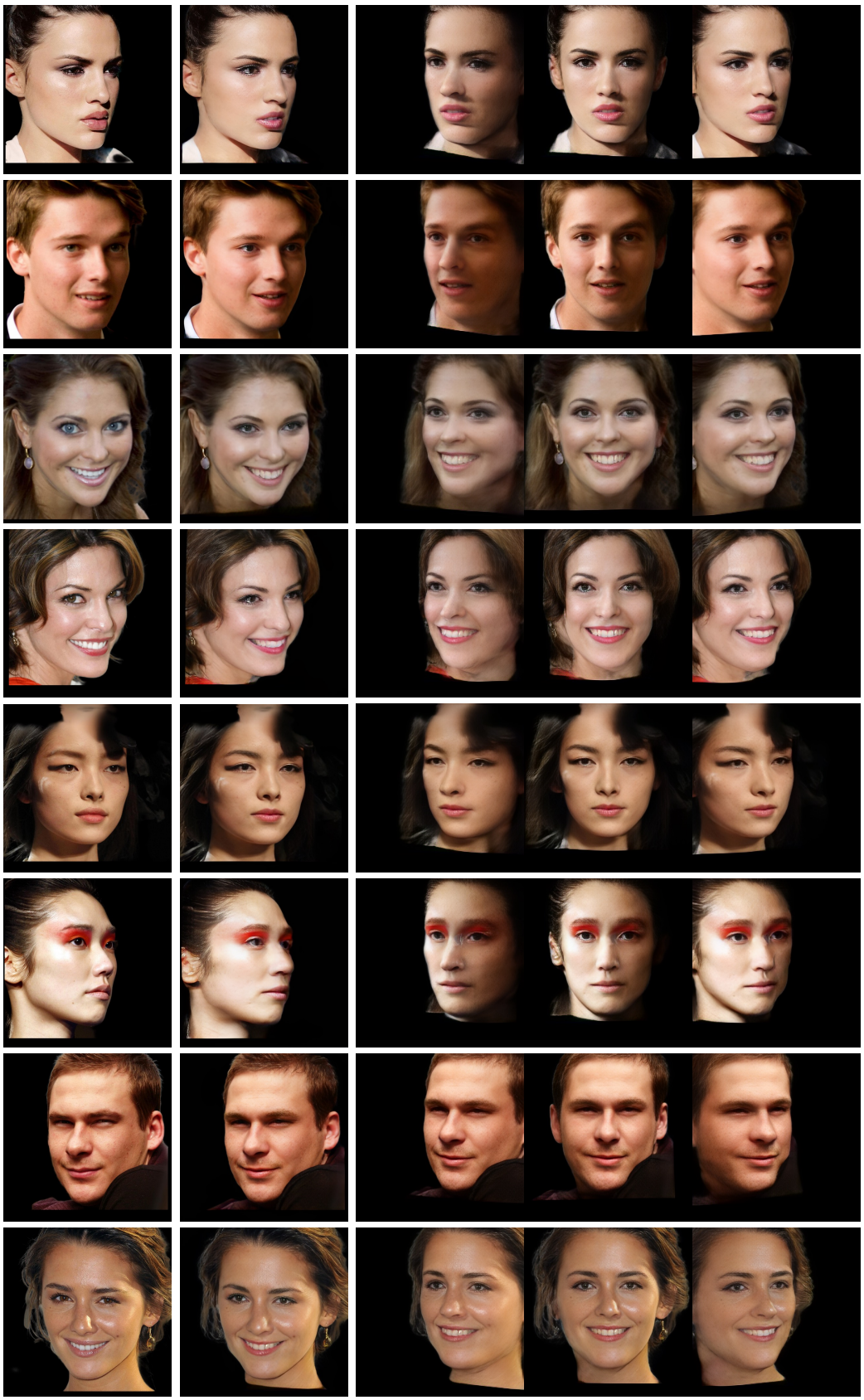

Figure 11: **3D reconstruction on CelebA.**

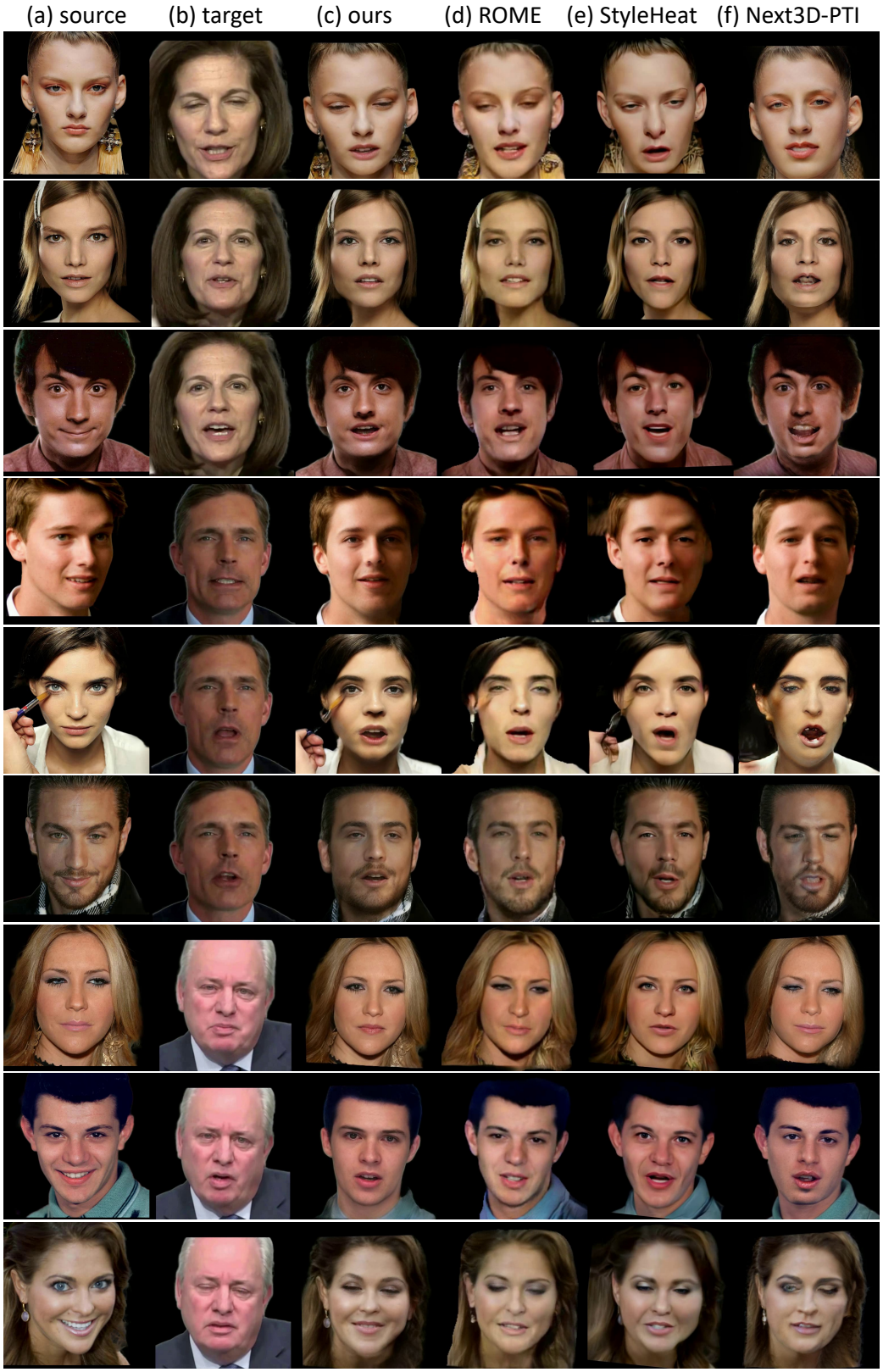

Figure 12: **Cross-identity reenactment from HDTF to CelebA.**

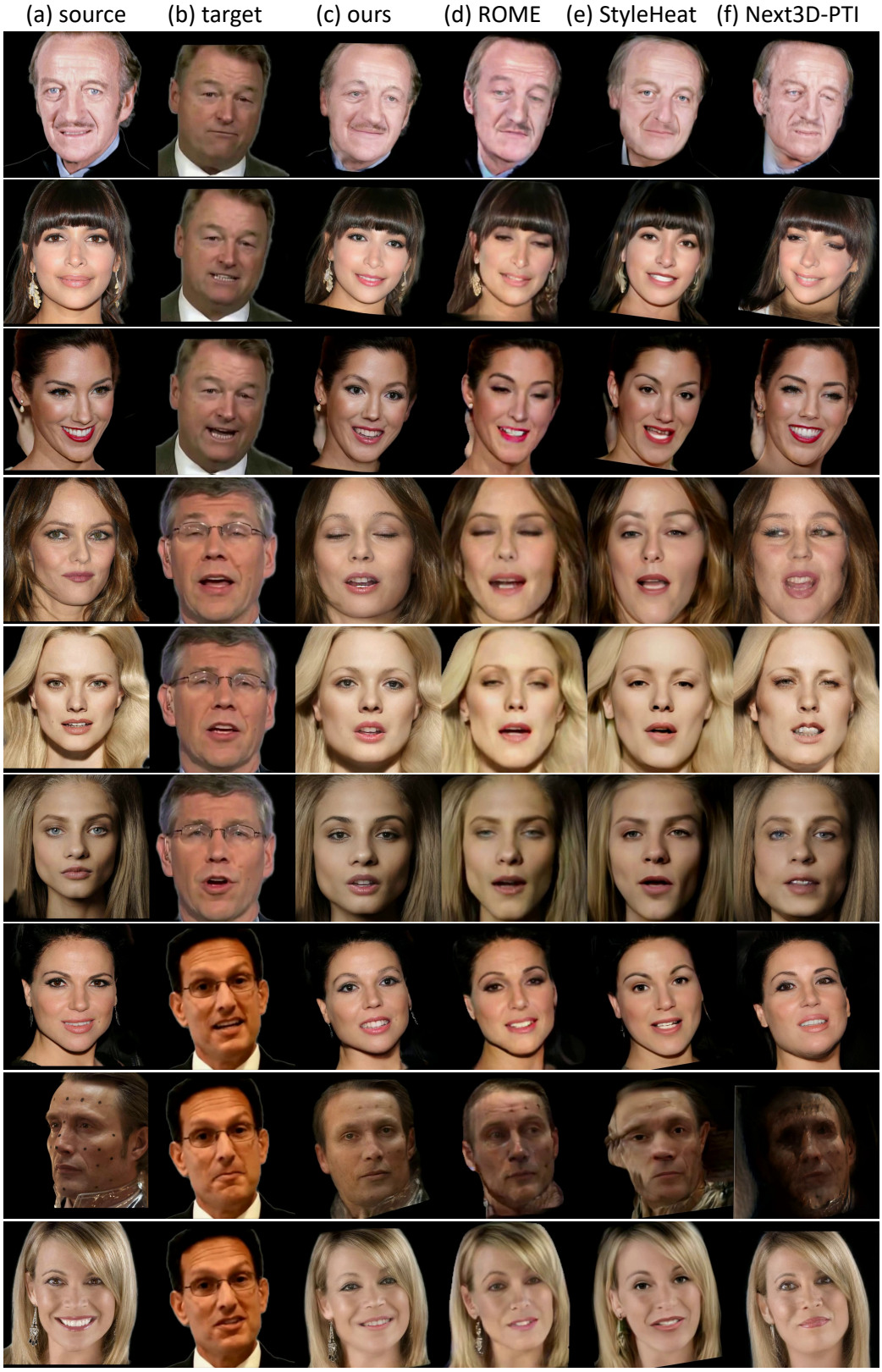

Figure 13: **Cross-identity reenactment from HDTF to CelebA.**

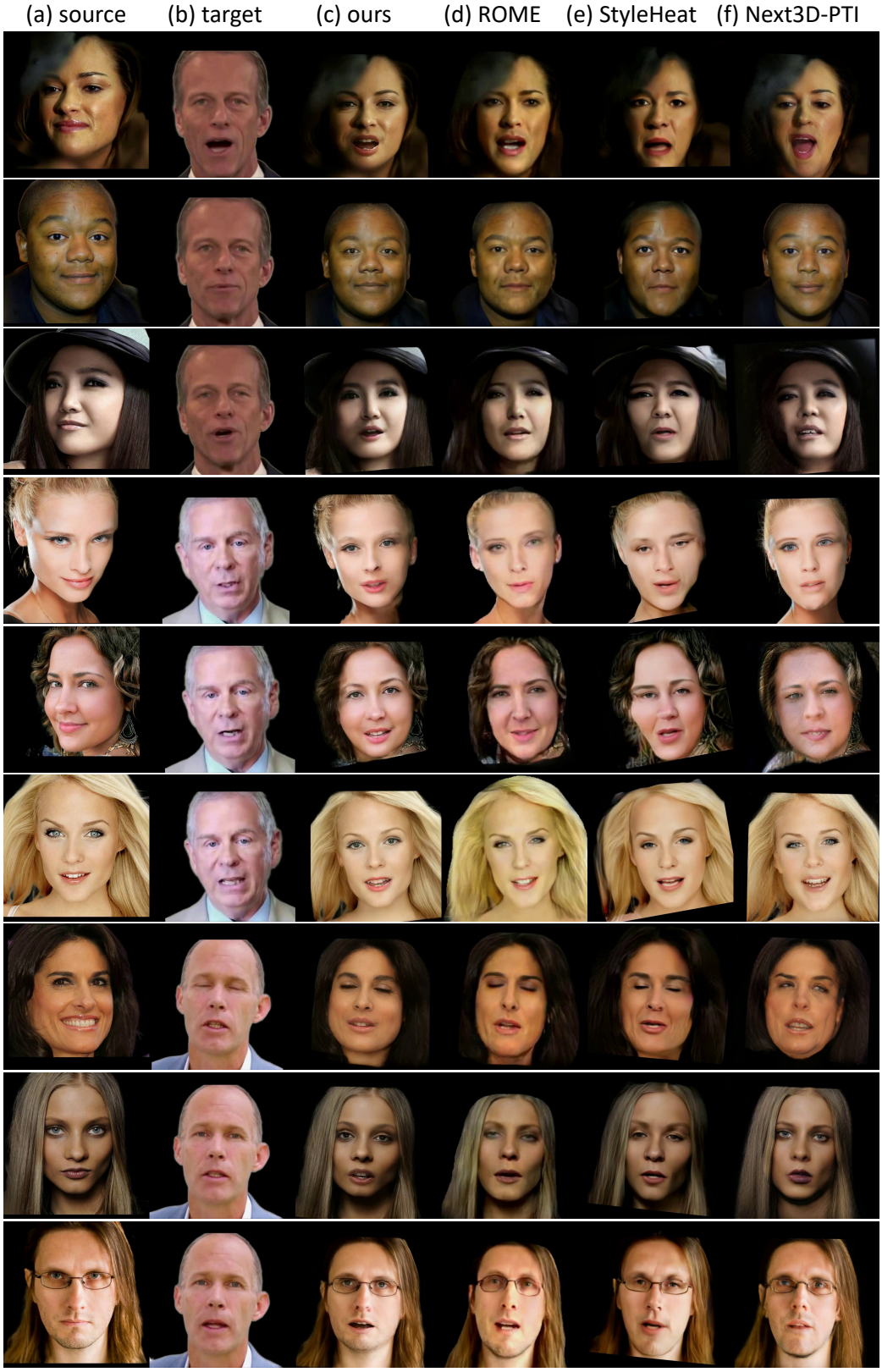

Figure 14: **Cross-identity reenactment from HDTF to CelebA.**

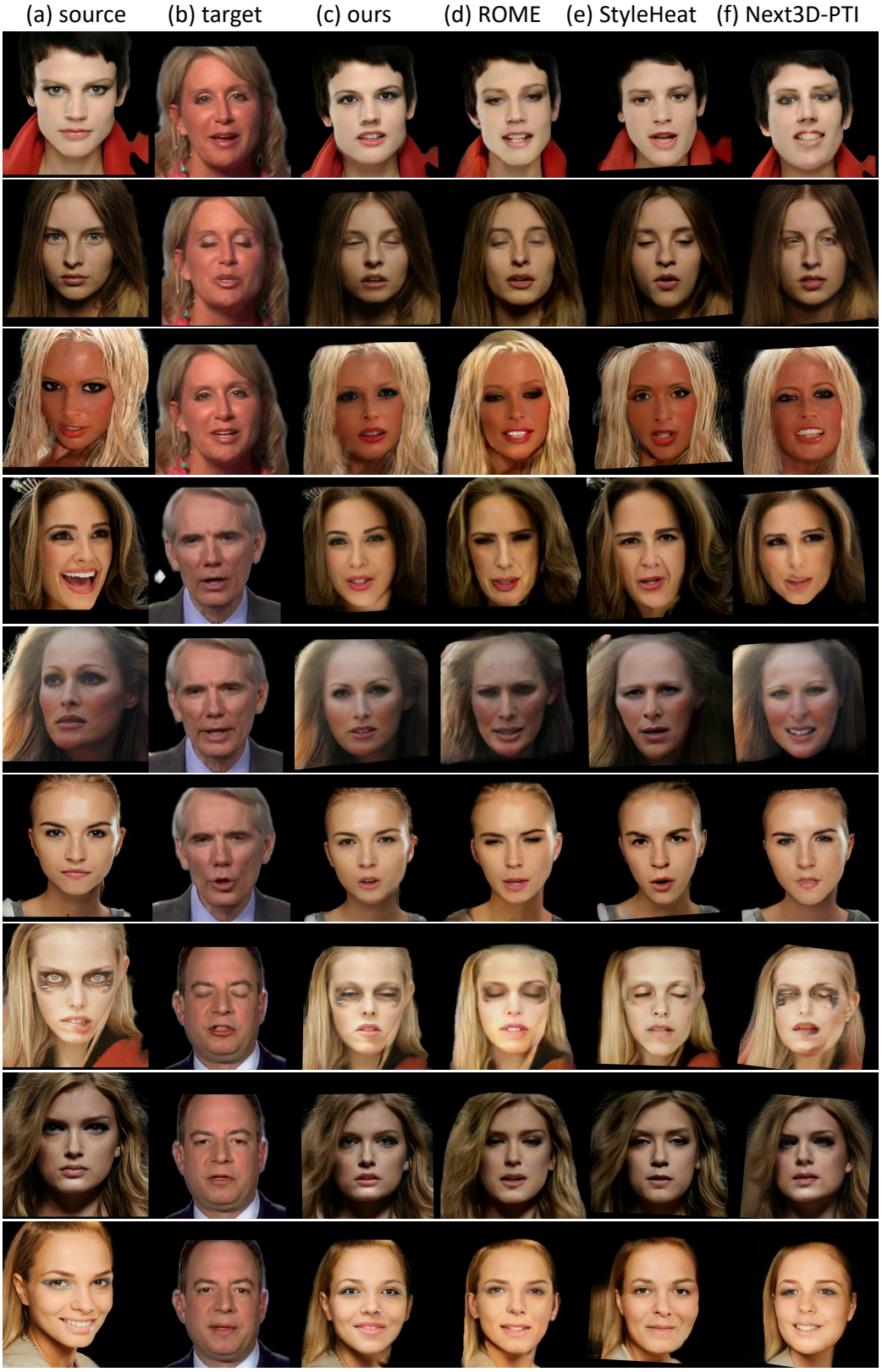

Figure 15: **Cross-identity reenactment from HDTF to CelebA.**

Figure 16: **Same-identity reenactment on HDTF.**

Figure 17: **Cross-identity reenactment on HDTF.**

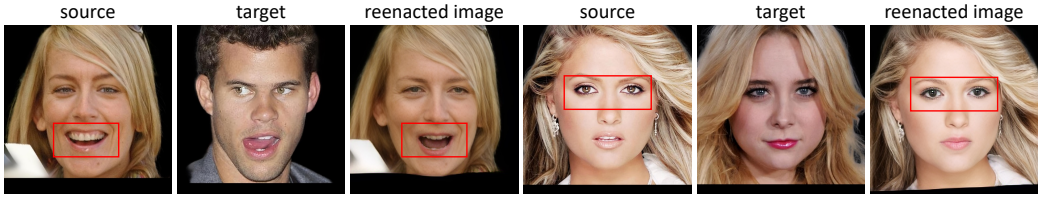

Figure 18: **Failure cases.**

References

[1] *A 3D Face Model for Pose and Illumination Invariant Face Recognition*, 2009. 5

[2] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE international workshop on information forensics and security (WIFS)*, 2018. 1

[3] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *ICCV*, 2017. 5

[4] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In *CVPR*, 2022. 1, 2, 4

[5] Davide Cozzolino, Andreas Rössler, Justus Thies, Matthias Nießner, and Luisa Verdoliva. Id-reveal: Identity-aware deepfake video detection. In *CVPR*, 2021. 1

[6] Davide Cozzolino, Justus Thies, Andreas Rössler, Christian Riess, Matthias Nießner, and Luisa Verdoliva. Forensictransfer: Weakly-supervised domain adaptation for forgery detection. *arXiv preprint arXiv:1812.02510*, 2018. 1

[7] 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

[8] Anna Frühstück, Nikolaos Sarafianos, Yuanlu Xu, Peter Wonka, and Tony Tung. VIVE3D: Viewpoint-independent video editing using 3D-aware GANs. In *CVPR*, 2023. 1

[9] John C Gower. Generalized procrustes analysis. *Psychometrika*, 1975. 5

[10] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. Headnerf: A real-time nerf-based parametric head model. In *CVPR*, 2022. 5

[11] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 3

[12] Taras Khakhulin, Vanessa Sklyarova, Victor Lempitsky, and Egor Zakharov. Realistic one-shot mesh-based head avatars. In *ECCV*, 2022. 4, 5

[13] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *CVPR*, 2020. 1, 2, 3, 5

[14] Yeonkyeong Lee, Taeho Choi, Hyunsung Go, Hyunjoon Lee, Sunghyun Cho, and Junho Kim. Exp-gan: 3d-aware facial image generation with expression control. In *ACCV*, 2022. 1

[15] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 2017. 1

[16] Zhiyuan Ma, Xiangyu Zhu, Guojun Qi, Zhen Lei, and Lei Zhang. Otavatar: One-shot talking face avatar with controllable tri-plane rendering. *arXiv preprint arXiv:2303.14662*, 2023. 4, 5

[17] Maryam Sadat Mirzaei, Kourosh Meshgi, Etienne Frigo, and Toyoaki Nishida. Animgan: A spatiotemporally-conditioned generative adversarial network for character animation. In *ICIP*, 2020. 1

[18] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Trans. Graph.*, 2021. 1

[19] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In *ICCV*, 2019. 1

[20] Jingxiang Sun, Xuan Wang, Lizhen Wang, Xiaoyu Li, Yong Zhang, Hongwen Zhang, and Yebin Liu. Next3d: Generative neural texture rasterization for 3d-aware head avatars. *arXiv preprint arXiv:2211.11208*, 2022. 1, 2, 4

[21] Keqiang Sun, Shangzhe Wu, Zhaoyang Huang, Ning Zhang, Quan Wang, and HongSheng Li. Controllable 3d face synthesis with conditional generative occupancy fields. *arXiv preprint arXiv:2206.08361*, 2022. 1

[22] Junshu Tang, Bo Zhang, Binxin Yang, Ting Zhang, Dong Chen, Lizhuang Ma, and Fang Wen. Explicitly controllable 3d-aware portrait generation. *arXiv preprint arXiv:2209.05434*, 2022. 1

[23] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *CVPR*, 2021. 4

[24] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *NeurIPS*, 2021. 4

[25] Jiaxin Xie, Hao Ouyang, Jingtan Piao, Chenyang Lei, and Qifeng Chen. High-fidelity 3d gan inversion by pseudo-multi-view optimization. *arXiv preprint arXiv:2211.15662*, 2022. 1

[26] Eric Zhongcong Xu, Jianfeng Zhang, Junhao Liew, Wenqing Zhang, Song Bai, Jiashi Feng, and Mike Zheng Shou. Pv3d: A 3d generative model for portrait video generation. In *ICLR*, 2023. 1

[27] Hongyi Xu, Guoxian Song, Zihang Jiang, Jianfeng Zhang, Yichun Shi, Jing Liu, Wanchun Ma, Jiashi Feng, and Linjie Luo. Omniavatar: Geometry-guided controllable 3d head synthesis. *arXiv preprint arXiv:2303.15539*, 2023. 1

[28] Yiran Xu, Zhixin Shu, Cameron Smith, Jia-Bin Huang, and Seoung Wug Oh. In-n-out: Face video inversion and editing with volumetric decomposition. *arXiv preprint arXiv:2302.04871*, 2023. 1

[29] Fei Yang, Qian Zhang, Chi Zheng, and Guoping Qiu. In-the-wild facial expression recognition in extreme poses. In *International Conference on Graphic and Image Processing (ICGIP 2017)*, 2018. 5

[30] Fei Yin, Yong Zhang, Xiaodong Cun, Mingdeng Cao, Yanbo Fan, Xuan Wang, Qingyan Bai, Baoyuan Wu, Jue Wang, and Yujiu Yang. Styleheat: One-shot high-resolution editable talking face generation via pre-trained stylegan. In *ECCV*, 2022. 4, 5

[31] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *ECCV*, 2018. 4

[32] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *CVPR*, 2021. 1, 2, 3

[33] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *CVPR*, 2021. 1

[34] Peiye Zhuang, Liqian Ma, Sanmi Koyejo, and Alexander Schwing. Controllable radiance fields for dynamic face synthesis. In *3DV*, 2022. 1