Portrait4D-v2: Pseudo Multi-View Data Creates Better 4D Head Synthesizer ——Supplementary Material——

Yu Deng, Duomin Wang, and Baoyuan Wang

Xiaobing.AI
https://yudeng.github.io/Portrait4D-v2/

A More Implementation Details

Learning the 3D synthesizer Ψ_{3d}**.** We use GenHead [4] trained on FFHQ [5] at 512^2 to synthesize multi-view supervisions for learning Ψ_{3d}. For high-quality data synthesis, we extract FLAME parameters from real images in FFHQ and VFHQ [10] as input conditions to GenHead. For camera pose, we follow [4] to sample it from a pre-defined uniform distribution, with pitch from $[-0.25, 0.65]$ *rad*, yaw from $[-0.78, 0.78]$ *rad*, and roll from $[-0.25, 0.25]$ *rad*. We also sample the camera radius from $[3.65, 4.45]$, and the camera look-at position from $[-0.01, 0.01] \times [-0.01, 0.01] \times [0.02, 0.04]$. The camera intrinsics are fixed with a 12° field of view (fov).

We use the balancing weights of [4] for each loss term in Eq. (2) in the main paper during training. At the first 1M images, we deactivate the adversarial loss $\mathcal{L}_{adv}$ and fix the learnable network parameters within the renderer $\mathcal{R}$, where we use an MLP for radiance decoding and a StyleGAN2-based [2,6] super-resolution module for synthesizing the final images of high resolution. After seeing 1M images, we activate $\mathcal{L}_{adv}$ and update all parameters in $\mathcal{R}$ along with Ψ_{3d}. $\mathcal{L}_{depth}$ and $\mathcal{L}_T$ are removed at this stage. Besides, we use a neural rendering resolution of 64^2 at the first 1M images, and linearly increase it to 128^2 during the next 1M-image period, while the final image resolution is consistently set to 512^2. We sample 48 coarse points and 48 fine points per ray following [2] for volume rendering. We adopt an Adam optimizer [7] with a learning rate of $1e-4$ and a batch size of 32, and train Ψ_{3d} to see 10M images in total, which takes 8 days on 8 Tesla A100 GPUs with 80GB memory.

Learning the 4D synthesizer Ψ**.** We initialize Ψ using the pre-trained weights of Ψ_{3d} except the motion-related cross-attention layers that are initialized at random. We then use a training split of 10K video clips from VFHQ [10] and sample 50 frames per clip to learn Ψ. During each iteration, we randomly choose 2 frames in a video clip as the source and driving, respectively, and send the driving image to the pre-trained Ψ_{3d} to obtain its pseudo multi-view frames, where we use a similar camera distribution for rendering as described in the above section. We adopt the balancing weights used in learning Ψ_{3d} for the four loss terms in Eq. (5) in the main paper. We use a neural rendering resolution of

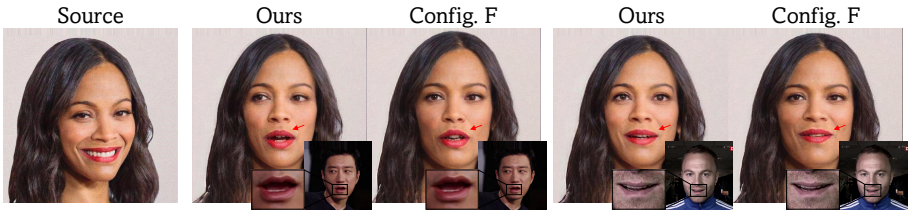

Fig. I: Comparisons between configurations using different motion representations. Ours using implicit 2D motion embedding learned from large-scale video data surpasses config. F using 3DMM at nuanced expression control (*e.g.*, see lip and teeth motions).

128^2 throughout the training process. We train Ψ to see 6M images, using Adam with a learning rate of $2.5e-4$ for the motion-related layers and $1e-4$ for the remaining, and a batch size of 32, which also takes 8 days on 8 A100 GPUs.

Neck pose rotation. We do not utilize the motion embedding [9] to model the neck pose. Instead, we introduce an explicit 3D rotation field to tackle it following [4]. More specifically, we pre-define a voxel grid of 24^3 and compute the 3D rotation of each grid point via a Surface Field (SF) [1] derived from a FLAME mesh at run time. Then, for each 3D point, we obtain its rotation via interpolating those of its nearest grid points. This way, the 3D rotation for arbitrary number of points can be efficiently calculated. Besides, the control of facial expressions and neck pose can be well disentangled.

B More Synthesis Results

Please see the *project page* for more head synthesis results of our method, where we use in-the-wild images collected from the Internet as the source and video sequences from CelebV-Text [11] as the driving.

C More Comparisons

We provide more comparisons between our method and the prior art in the *project page*. We use images from VFHQ test set as the source and those from CelebV-Text as the driving.

D More Visual Analysis

In this section, we offer a more comprehensive analysis of the advantages of our proposed framework, which combines implicit 2D motion representation with multi-view training data.

As mentioned in the main paper, we employ an implicit expression representation learned from 2D data (*i.e.*, PD-FGC [9]) to capture subtle motions difficult

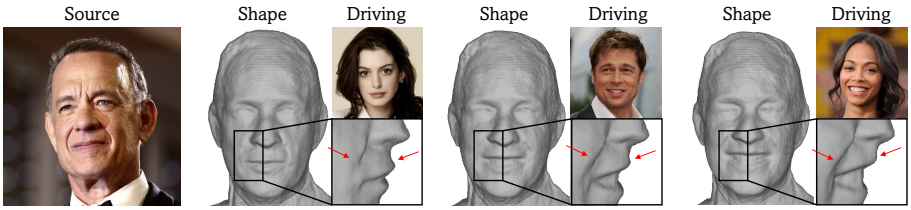

Fig. II: Our 2D motion embedding, even though it is learned from soley 2D videos, effectively captures reasonable 3D geometry changes.

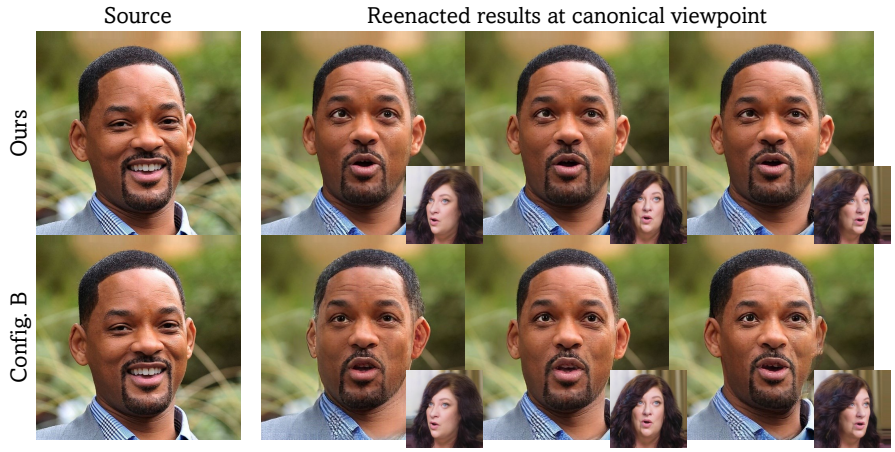

Fig. III: By utilizing multi-view training data, we effectively remove pose-related information from our 2D motion embedding, achieving enhanced 3D consistency while preserving the original capability to represent detailed motion.

to be described using existing 3DMMs. As illustrated in Fig. I, our data-driven motion embedding excels at modeling the nuanced mouth movements that occur during speech. In contrast, 3DMMs struggle to accurately represent the relative positions between lips and teeth.

Additionally, even though PD-FGC is a 2D motion embedding, it well records 3D geometry information. Figure II shows that our motion embedding effectively captures subtle 3D geometry changes around cheeks and lips given different 2D driving images. We contend that this is because most 3D geometry clues can be inferred from 2D appearances.

Furthermore, although the initial PD-FGC representation includes pose-related information due to its 2D nature, our learning approach using multi-view training data enforces 3D consistency while preserving detailed expression information. Figure III demonstrates that our method with multi-view supervision retains the ability to reconstruct subtle expressions compared to config. B using monocular video supervision, while also providing improved 3D consistency.

E User Study

We provide the 20 test cases used for the user study in Fig. IV to X and the original voting results in Tab. I and II.

F Limitations and Future Works

Despite the high-quality synthesis results compared to previous approaches, our method still has several limitations.

First, we observe that when rendering the synthesized results at novel views, the unseen regions in the original source image usually have fewer details compared to the visible regions. We conjecture that this is due to the deterministic property of our 4D head synthesizer, where it tends to predict an "average" expectation of all possible results. We believe integrating the current learning framework with a stochastic generative model (*e.g.*, a Diffusion Model [8]) can alleviate this issue.

Second, our current motion embedding from [9] is learned using video data from Voxceleb2 [3] which may lack extreme expressions. As a result, our method may not well capture exaggerated facial movements such as puffing out the cheeks or poking out the tongue. However, since our learning framework does not require an end-to-end training with the motion encoder, it is possible to retrain the motion embedding using more expressive data or replace it with more advanced motion embedding for better expression imitation in the future.

Finally, as a learning-based approach, our method may produce inferior results for out-of-distribution data and can have slightly different performances for identities of different races. We plan to collect training data of higher diversity and coverage to alleviate this problem. Besides, extending our approach to handle upper-body or full-body avatar synthesis is also an interesting future direction to be explored.

G Ethics Consideration

The proposed method is intended for virtual communication and entertainment. However, a misuse of it for deceptive content creation can be harmful and should be strictly prohibited. Currently, the synthesized results contain certain visual artifacts which could help with deepfake detection.

References

1. Bergman, A., Kellnhofer, P., Yifan, W., Chan, E., Lindell, D., Wetzstein, G.: Generative neural articulated radiance fields. Advances in Neural Information Processing Systems **35**, 19900–19916 (2022)
2. Chan, E.R., Lin, C.Z., Chan, M.A., Nagano, K., Pan, B., De Mello, S., Gallo, O., Guibas, L., Tremblay, J., Khamis, S., et al.: Efficient geometry-aware 3d generative adversarial networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
3. Chung, J.S., Nagrani, A., Zisserman, A.: Voxceleb2: Deep speaker recognition. In: INTERSPEECH (2018)
4. Deng, Y., Wang, D., Ren, X., Chen, X., Wang, B.: Learning one-shot 4d head avatar synthesis using synthetic data. arXiv preprint arXiv:2311.18729 (2023)
5. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4401–4410 (2019)
6. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8110–8119 (2020)
7. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (2015)
8. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
9. Wang, D., Deng, Y., Yin, Z., Shum, H.Y., Wang, B.: Progressive disentangled representation learning for fine-grained controllable talking head synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17979–17989 (2023)
10. Xie, L., Wang, X., Zhang, H., Dong, C., Shan, Y.: Vfhq: A high-quality dataset and benchmark for video face super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 657–666 (2022)
11. Yu, J., Zhu, H., Jiang, L., Loy, C.C., Cai, W., Wu, W.: Celebv-text: A large-scale facial text-video dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14805–14814 (2023)

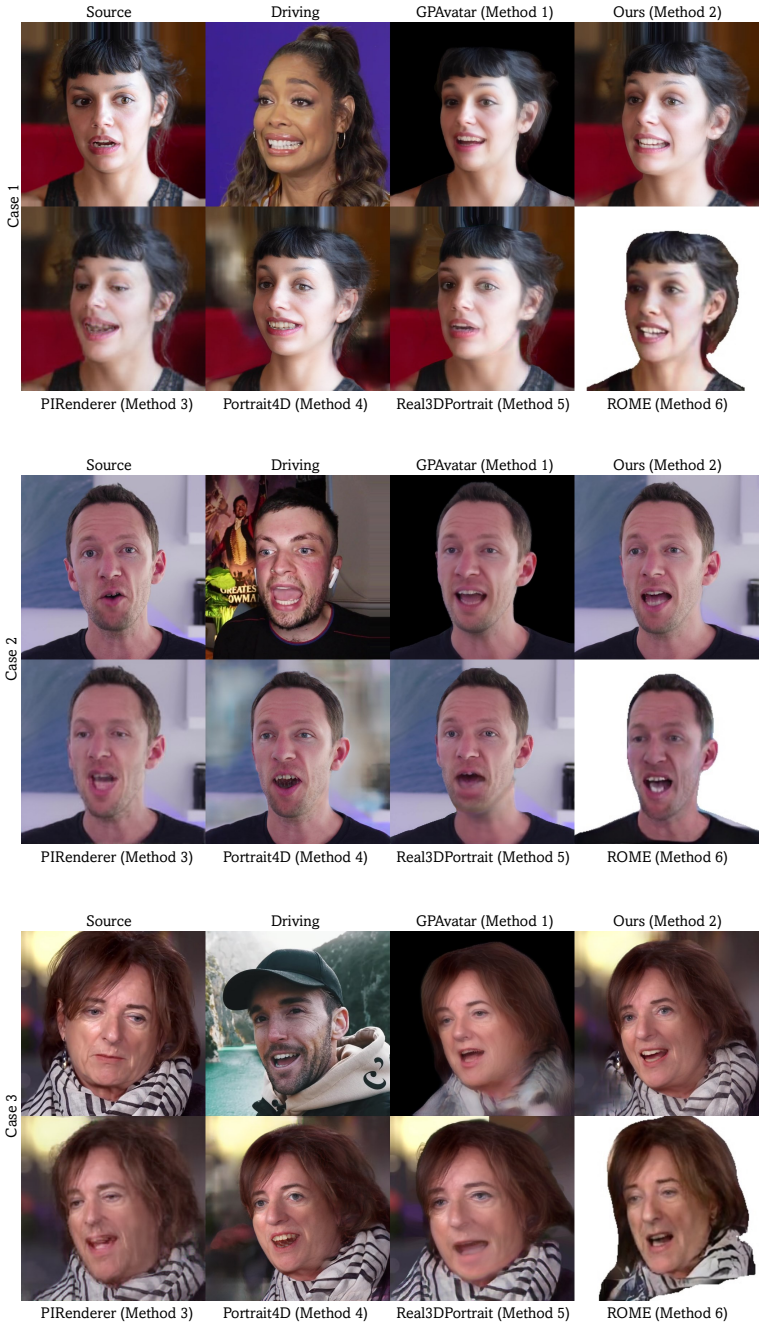

Fig. IV: User study cases.

Fig. V: User study cases.

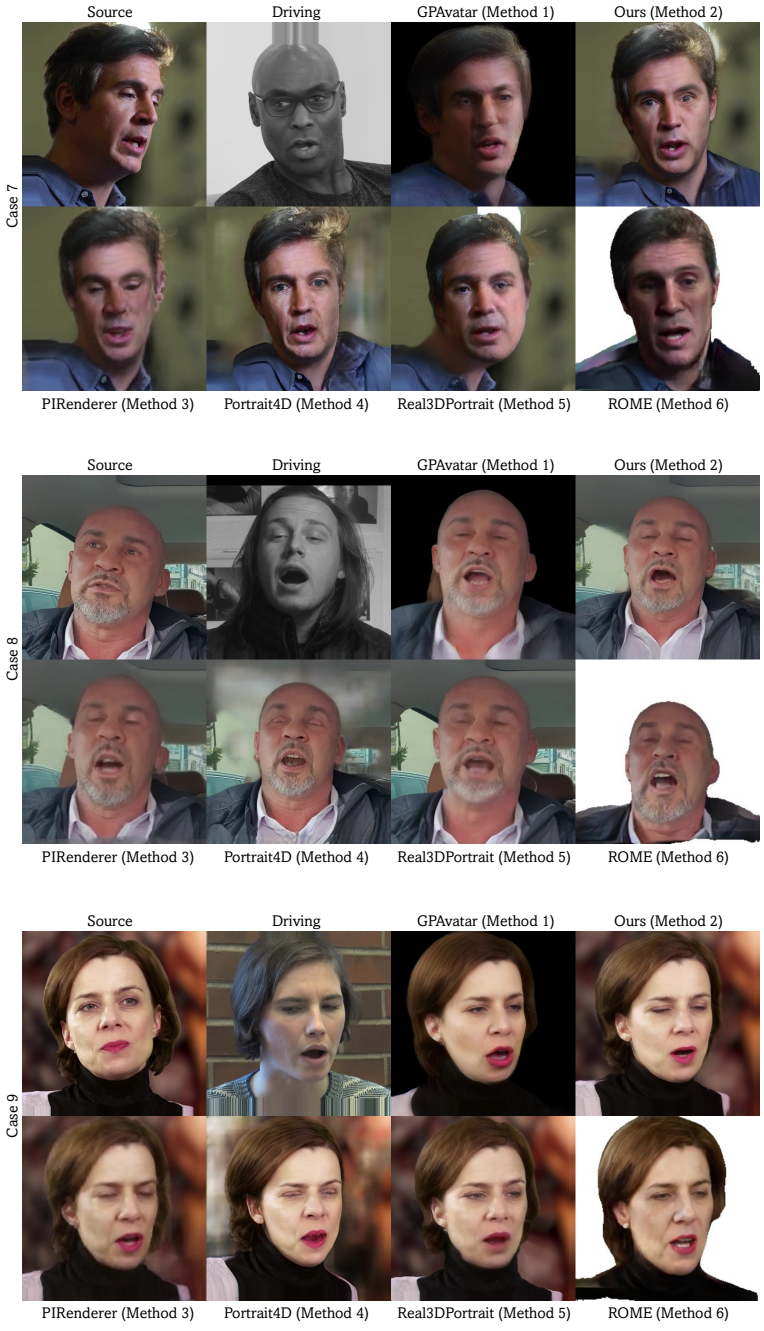

Fig. VI: User study cases.

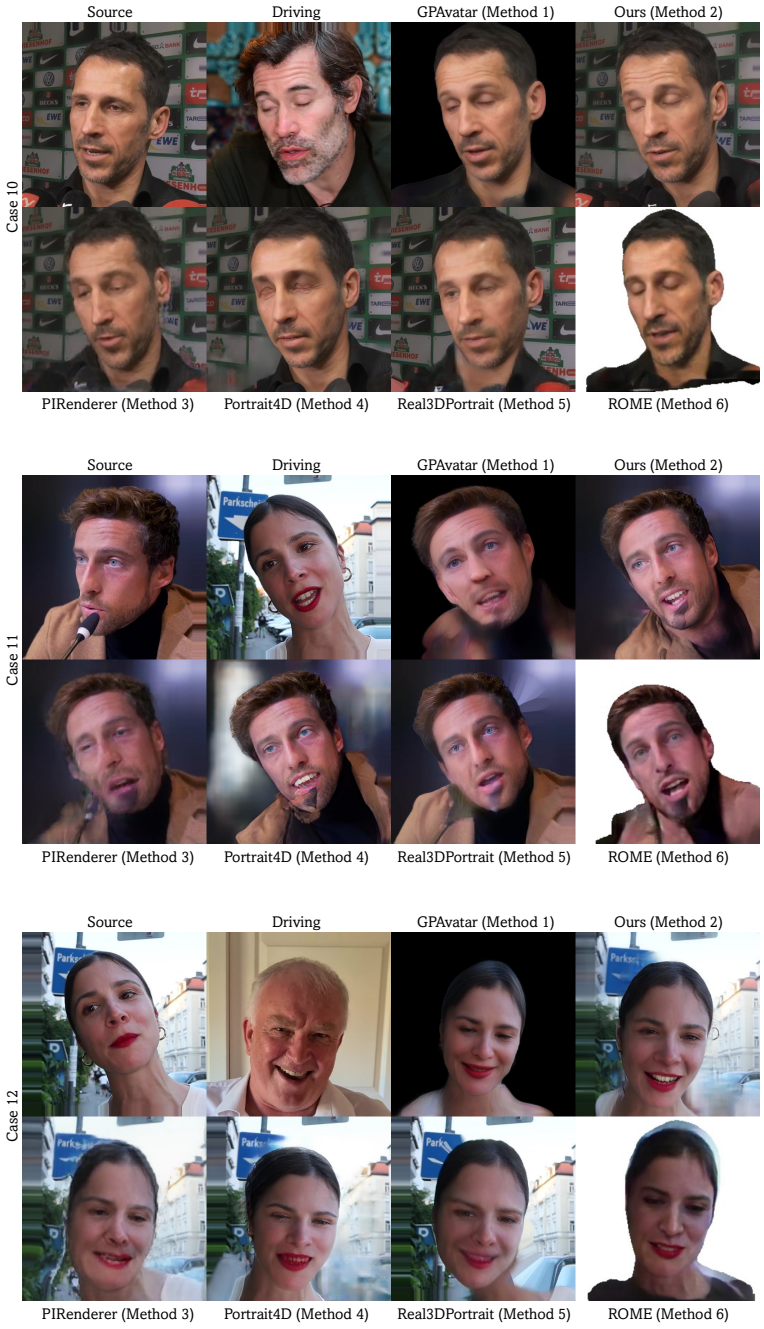

Fig. VII: User study cases.

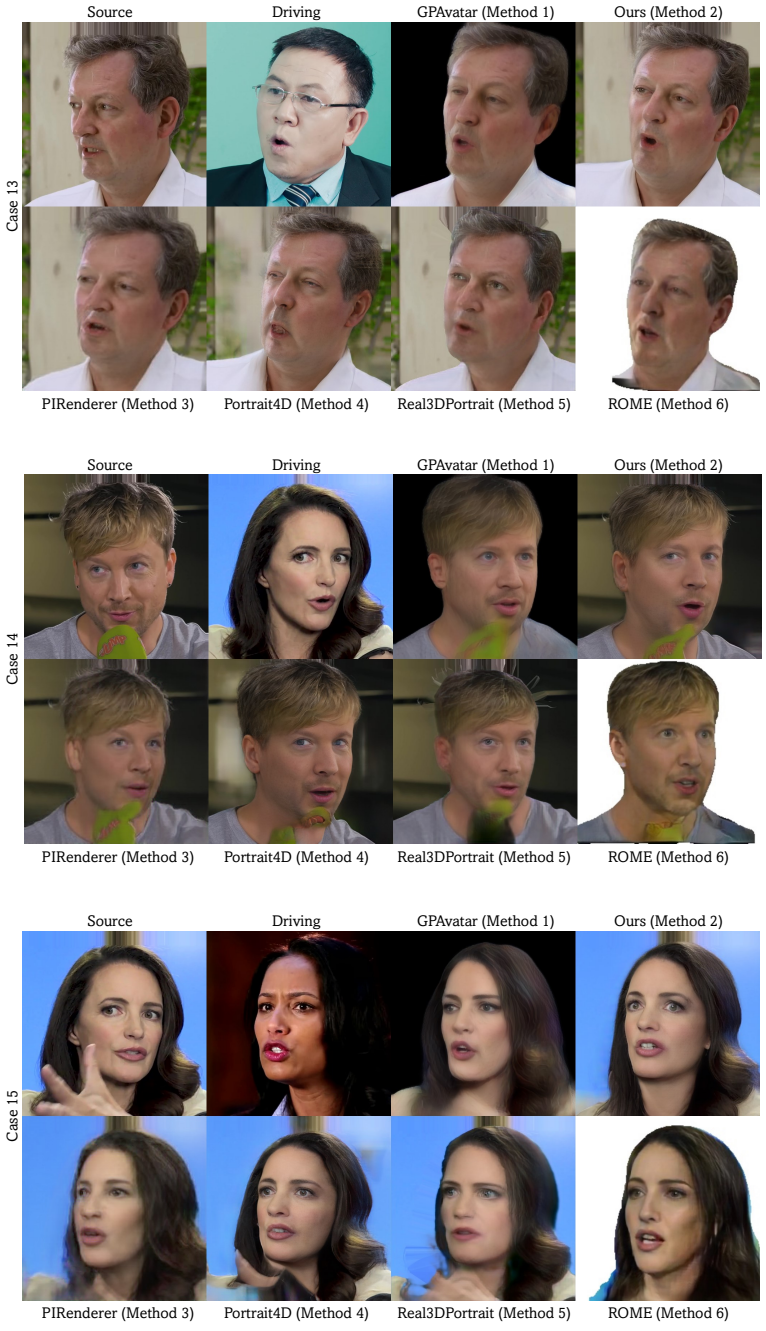

Fig. VIII: User study cases.

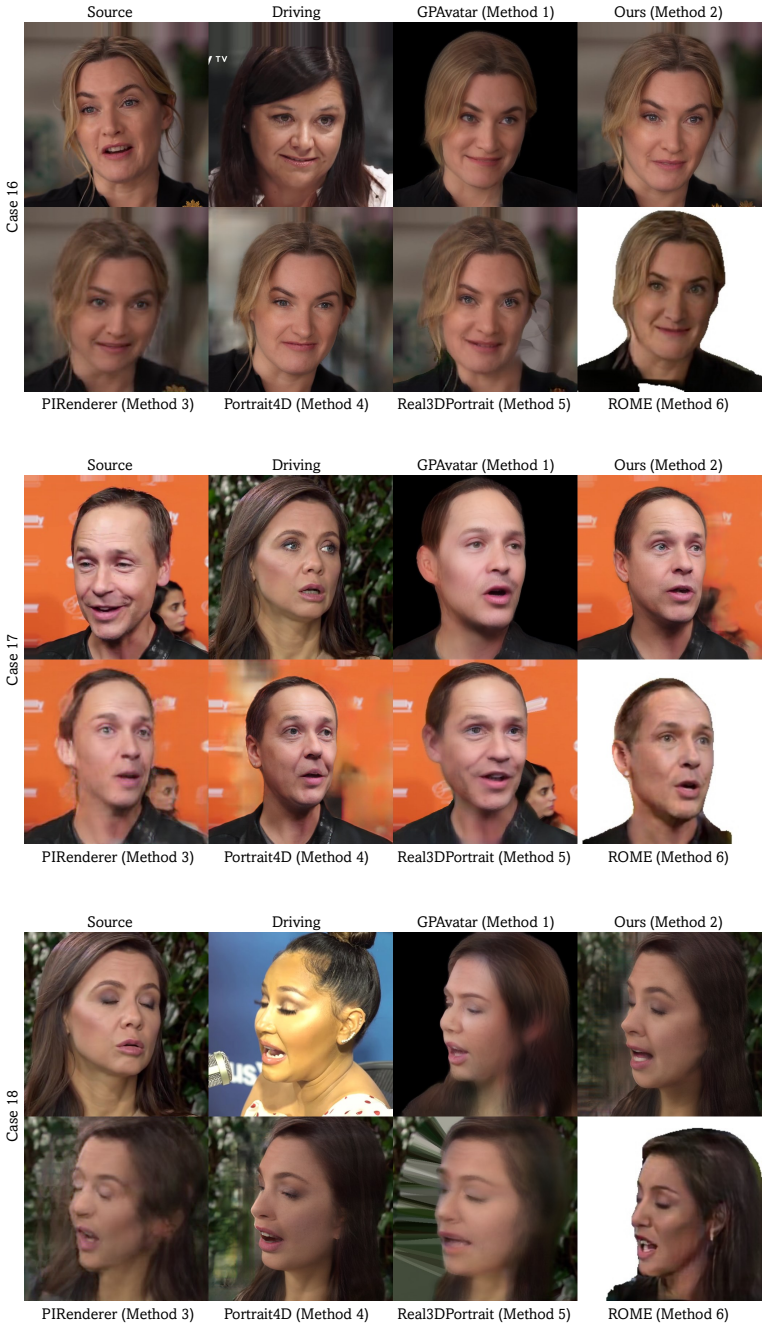

Fig. IX: User study cases.

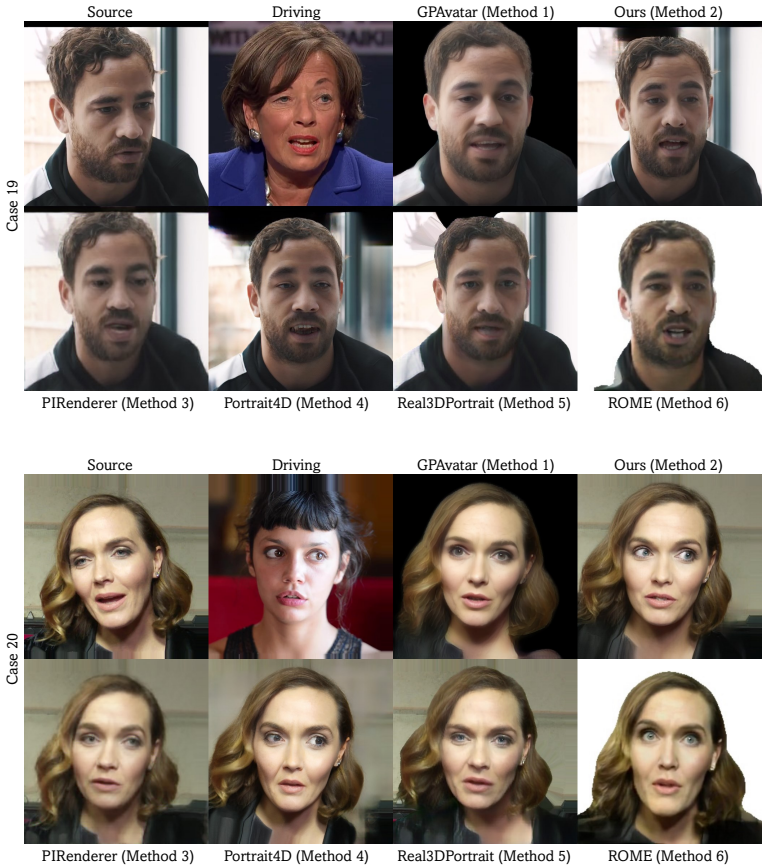

Fig. X: User study cases.

Table I: User study results on expression similarity. Each row shows the choice of one participant.

case1	case2	case3	case4	case5	case6	case7	case8	case9	case10	case11	case12	case13	case14	case15	case16	case17	case18	case19	case20
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
2	1	2	2	1	2	2	2	1	1	2	2	1	2	2	1	2	2	2	2
2	2	2	2	1	2	2	2	1	2	4	2	2	4	2	1	2	2	2	2
4	6	2	1	2	2	2	2	6	1	2	2	2	4	5	3	2	2	6	2
2	2	2	2	1	2	2	2	1	1	2	2	2	2	1	1	2	2	1	2
2	2	2	2	2	2	2	1	1	2	2	2	2	4	2	1	2	2	2	2
2	6	2	4	1	2	2	2	5	2	2	2	2	4	2	1	2	2	1	2
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	1	2	2	1	2
2	6	2	6	6	2	4	1	1	1	6	2	2	4	1	3	1	2	1	2
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
2	2	2	2	2	2	2	2	1	2	2	2	2	2	2	1	2	2	2	2
2	1	2	1	1	2	2	2	1	1	2	2	2	2	2	1	2	2	2	2
2	2	2	4	1	2	2	2	2	2	2	1	2	2	1	1	2	6	2	2
4	1	4	2	1	2	4	5	5	4										
5	5	2	4	3	2	4	2	6	1	1	5	2	4	1	3	6	2	1	6
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
2	1	2	2	1	2	2	2	2	2	2	2	2	2	2	1	6	2	2	2
2	2	4	2	1	2	2	5	1	2	2	2	2	5	1	2	2	2	2	2
2	1	2	2	2	2	2	2	1	2	2	2	2	2	2	1	2	2	4	2
2	1	2	2	3	6	2	2	5	6	2	2	2	4	6	3	2	2	2	2
4	2	2	2	2	1	2	2	1	2	2	2	2	4	2	2	2	2	2	2
2	2	4	2	2	2	2	2	1	1	2	2	2	4	4	4	4	2	2	2

Table II: User study results on image quality. Each row shows the choice of one participant.

case1	case2	case3	case4	case5	case6	case7	case8	case9	case10	case11	case12	case13	case14	case15	case16	case17	case18	case19	case20
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
2	2	4	2	2	2	2	2	2	2	2	1	2	1	2	4	1	2	2	2
2	2	2	2	2	2	2	2	1	2	2	2	2	2	2	2	2	2	2	2
2	4	4	2	2	2	2	2	5	2	2	2	4	4	2	2	2	4	4	2
4	2	4	2	2	4	2	2	1	2	2	2	2	4	2	2	2	2	2	2
2	2	2	2	2	2	2	1	1	2	2	2	2	4	2	1	1	2	2	2
1	2	2	4	2	2	2	2	1	2	2	2	2	4	1	2	2	2	2	1
4	4	4	4	4	4	2	4	4	3	2	2	2	4	4	4	2	4	2	2
2	2	2	2	2	2	2	2	2	2	2	2	2	4	1	2	4	4	2	2
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
2	2	2	2	2	2	2	2	1	2	2	2	2	2	2	1	2	2	2	2
2	2	2	2	2	2	2	2	2	2	2	2	2	4	2	2	2	2	2	2
2	2	2	4	2	1	2	2	2	2	2	2	2	4	4	1	2	6	2	2
4	4	4	4	4	2	4	4	4	1										
2	2	2	2	2	2	2	2	5	3	2	2	2	2	2	1	2	2	1	2
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	6	2	2	2
2	2	4	2	2	2	2	2	1	1	2	2	2	2	1	2	1	2	2	2
2	2	2	2	2	2	2	2	4	2	2	2	2	2	2	2	2	2	2	2
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	4	2
2	2	4	2	2	2	2	2	1	1	2	2	2	4	2	4	4	4	2	2