Supplementary Materials for "High-fidelity Facial Avatar Reconstruction from Monocular Video with Generative Priors"

Yunpeng Bai[1]*, Yanbo Fan[2]†, Xuan Wang[3], Yong Zhang[2], Jingxiang Sun[4], Chun Yuan[1,5]†, Ying Shan[2]

[1] Tsinghua Shenzhen International Graduate School,
[2]Tencent AI Lab, [3]Ant Group,[4]Tsinghua University, [5]Peng Cheng Laboratory

1. Implementation Details

Here we present the implementation details of the encoder for RGB frames, 3DMM expression coefficients, and audio features, respectively. **1)** for the input of RGB frames, the encoder $\boldsymbol{f}$ is composed of 6 ResBlocks and 4 FC layers; **2)** for the input of 3DMM expression coefficients, the encoder $\boldsymbol{f}_e$ is realized by a 6-layer MLP network; **3)** for the input of audio features, we follow AD-NeRF [1] for the realization of $\boldsymbol{f}_a$. Specifically, the audio features will first be sent to a 1-D convolutional layer to get the latent code of each frame. Then, these latent codes will be filtered in time by an attention module with five 1-D convolutional layers and softmax activation. Finally, the temporally filtered latent code is passed through an FC layer to obtain the coefficient $\boldsymbol{\alpha}$.

2. Evaluations on More Complex Situations

We test our method under several extreme situations.

Complex lighting conditions. We test our method on a video with complex lighting conditions, as shown in Figure 1. Note that we don't consider the decoupling of the light. And the lighting does not change with the change of viewpoint.

Figure 1. Results under complex lighting.

Partial occlusions. We test our method on images with partial occlusions. When the occluded part is not crucial, our method can still reconstruct results similar to the ground truth (second line in Figure 2). When some crucial parts are occluded, such as the mouth, our results are different from the ground truth but still reasonable (first line in Figure 2).

*Work done during an internship at Tencent AI Lab.
†Corresponding Authors.

Large pose/expressions. We train our model on videos without large poses/expressions and then test our method on extreme cases. The results are shown in Figure 4, denoted as *Ours few*. The reconstructed results only have mild expressions compared to the ground truth. However, our whole model (denoted as *Ours whole*) has good generalization for large poses and expressions.

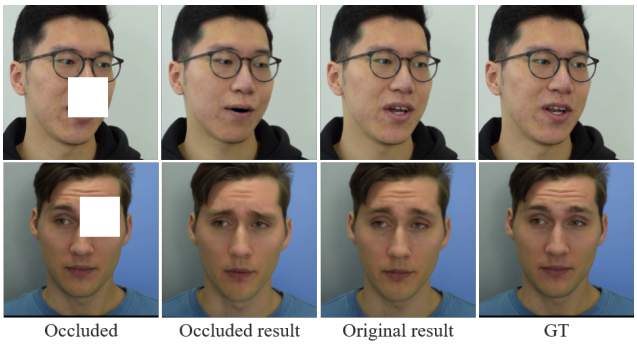

Figure 2. Results under partial occlusion.

3. Transferability of the Basis

As mentioned in the paper, the basis used in our method can be shared between different modalities. To verify this, we train encoders for different modalities on the same basis, denoted as *Shared basis*. The comparative results are shown in Figure 3, where the second column refers to the results that were trained with the specific basis for each modality. The results demonstrate that sharing the basis between different modalities is feasible and does not affect performance.

4. Ablation Studies of the Two-Stage Training Scheme

1) We show the performance **without fintuning the pretrain generator**, for both PTI and our methods, denoted as *PTI w/o finetune* and *Ours w/o finetune*, respectively. The

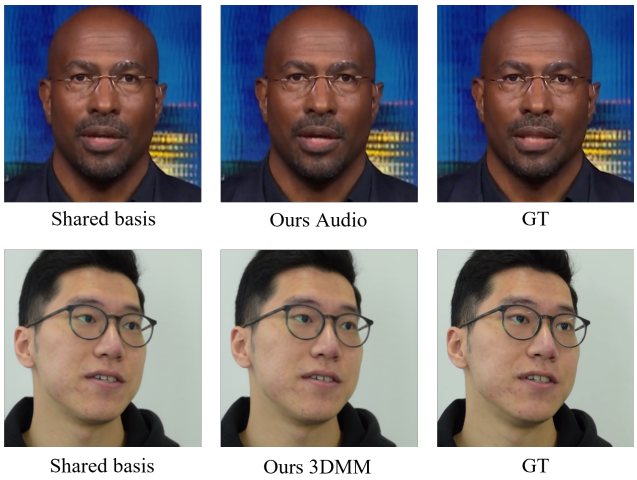

Figure 3. Transferability of the basis.

results are shown in Figure 4. Without finetuning, the results obtained are significantly distorted compared to the ground truth. **2)** If the **generator is also finetuned at the beginning of training**, some grid artifacts may appear in the results, affecting the overall performance. The results are shown in Figure 4, denoted as *Finetune at beginning*.

References

[1] Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5784–5794, 2021. 1

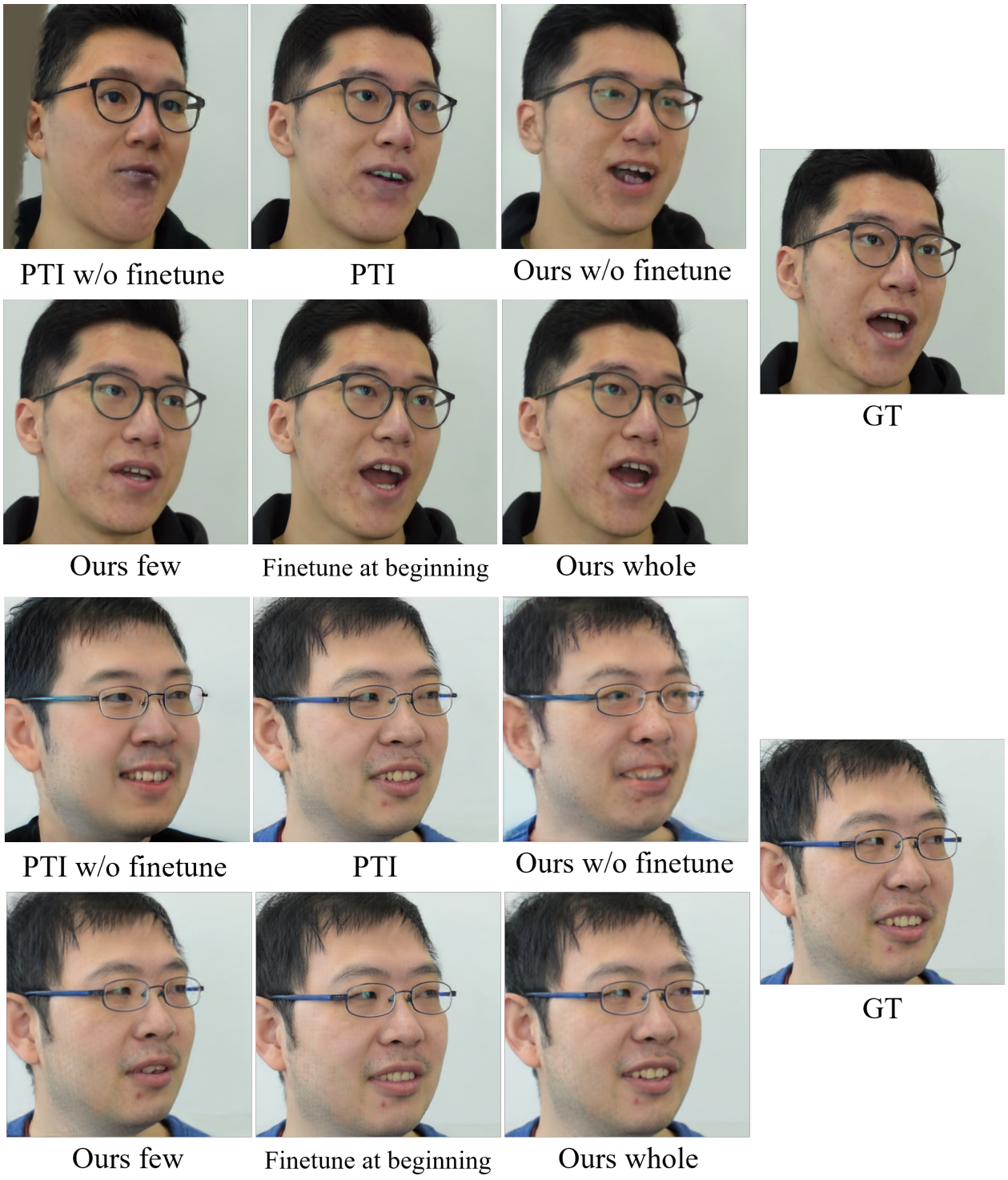

Figure 4. Comparison results for studies of the two-stage training scheme and generalization for large poses/expressions.