–Supplementary Material–
Accurate 3D Face Reconstruction with Facial Component Tokens

Tianke Zhang[1,2*] Xuangeng Chu[2] Yunfei Liu[2] Lijian Lin[2] Zhendong Yang[1,2]
Zhengzhuo Xu[1,2] Chengkun Cao[2] Fei Yu[3] Changyin Zhou[3] Chun Yuan[1†] Yu Li[2†]
[1]Tsinghua Shenzhen International Graduate School
[2]International Digital Economy Academy (IDEA) [3]Vistring Inc.

Overview

This supplementary material provides additional details and additional results that are not included in the main paper due to page limitations. The list of items included are:

- Ablation study on backbone in Sec. A.
- More face tracking results in Sec. B.
- Summarizing the difference of existing methods in Sec. C.
- More qualitative comparison results with existing methods in Sec. D.
- Video demo. (separate file: *video-demo.mp4*)

A. Ablation Study on Backbone Structure

We have tested some variants of TokenFace by changing the backbone ViT structure and listed the comparison in Table S-1. As same as the ablations studies in the main paper, we use validation set of the NoW dataset [6] to evaluate the reconstruction performance. We can see that, generally, using larger ViT can lead to lower reconstruction error. Our method is also compatible with ViT accelaration method like ToMe [1] module, which improves the training and inference speed of the model by aggregating similar tokens together through a matching algorithm. In our main paper, we use TokenFace (ViT Base) to report the results.

B. Face Tracking Results

As extension of the face-tracking comparison experiments in our main paper, we conduct two additional tests, which are presented in Fig. S-1. TokenFace produce more accurate estimation than ResNet method and TokenFace with temporal modeling (TokenFace-T) can further improve the performance.

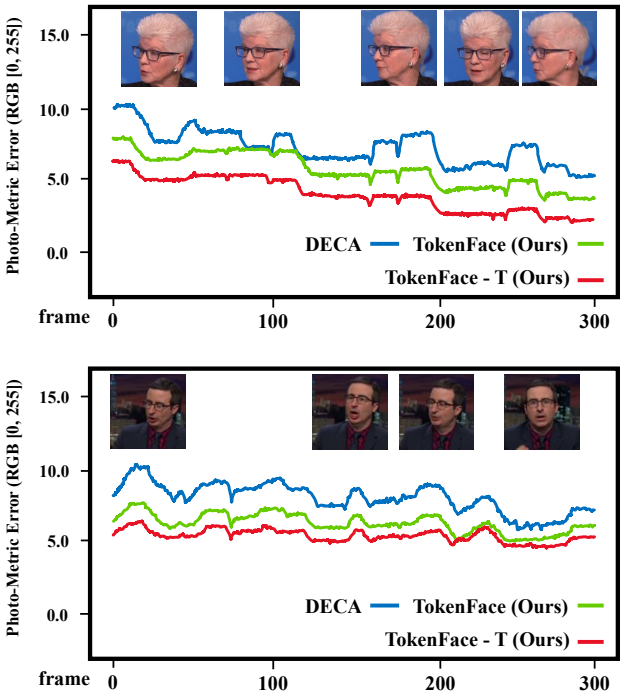

Figure S-1. **Visualization of face tracking error on a video clip**. We calculate the photometric error. TokenFace–T denotes TokenFace with the temporal module.

C. More Comparisons with other methods

We summarize the difference of training dataset types and reconstructed properties of different methods in Table S-2. We also list their reconstruction performance again in the table. As can be seen, we are the only method which utilize hybrid 2D and 3D dataset and get the best performance. The latest method MICA [7] use 3D metrical data but can only predict shape parameters.

Model Name	Params / M	FLOPs / G	Calculation Speed / fps	Reconstruction Error		
				Median	Mean	Std
ResNet-50	47.54	9.19	53.7	1.19	1.47	1.25
TokenFace (ViT Tiny)	31.67	7.08	58.2	1.12	1.39	1.17
TokenFace (ViT Small)	52.24	9.17	42.9	0.97	1.28	1.02
TokenFace (ViT Base)	126.82	21.65	14.5	0.79	0.99	0.85
TokenFace (ViT Base) + ToMe [1]	119.93	19.37	17.7	0.81	1.01	0.87

Table S-1. **Ablation study on backbone structure**. It can be seen that the change of backbone brings an increase in the number of parameters and computation, but the structure still outperforms ResNet50 under the same conditions, and at the same time, the ToMe module plays an important role in accelerating the reconstruction.

Method	Training Data		Reconstructed Param.	Reconstruction Error		
	2D	3D		Median	Mean	Std
Deep3DFaceRecon[Pytorch] [2]	✓		S, E, J, C, T, L	1.20	1.52	1.28
3DDFA-v2 [4]	✓		S, E, J, C, T, L	1.30	1.85	1.41
DECA [3]	✓		S, E, J, C, T, L	1.18	1.46	1.25
MICA [7]		✓	S	0.90	1.11	0.92
TokenFace (Ours)	✓	✓	S, E, J, C, T, L	0.79	0.99	0.85

Table S-2. **Comparison with the existing 3D face reconstruction methods**. The six facial parameters are **S** (shape), **E** (expression), **J** (jaw pose), **C** (camera pose), **T** (texture), and **L** (lighting).

D. Qualitative Comparisons

More visual comparison of the 3D reconstruction results of different methods can be found in Figure S-2. These examples are from IMDB-WIKI [5], a dataset from film and TV containing wider range of different human races, ages, poses, expressions, and occlusions than other face dataset. Despite the increased complexity of this dataset, TokenFace still produces consistently accurate reconstruction results, demonstrating remarkable generalization ability.

References

[1] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*, 2022. 1, 2

[2] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2, 3

[3] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (ToG)*, 40(4):1–13, 2021. 2, 3

[4] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3d dense face alignment. In *European Conference on Computer Vision*, pages 152–168. Springer, 2020. 2, 3

[5] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Dex: Deep expectation of apparent age from a single image. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 10–15, 2015. 2, 3

[6] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael Black. Learning to regress 3D face shape and expression from an image without 3D supervision. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 7763–7772, June 2019. 1

[7] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. In *European Conference on Computer Vision (ECCV)*. Springer International Publishing, Oct. 2022. 1, 2

Figure S-2. **Visual comparison of 3D face reconstruction quality of ours and some other representative methods in IMDB-WIKI Datset[5]**. From top to bottom are input image, Deep3DFaceRecon [2], 3DDFAv2 [4], DECA [3], and TokenFace (Ours).