Supplementary Material: Vivify Your Talking Avatar via Zero-Shot Expressive Facial Style Transfer

Liyang Chen[1,3], Zhiyong Wu[1], Runnan Li[3], Weihong Bao[1], Jun Ling[2], Xu Tan[3], Sheng Zhao[3]
[1]Shenzhen International Graduate School, Tsinghua University
[2]Shanghai Jiao Tong University [3]Microsoft
{cly21, bwh21}@mails.tsinghua.edu.cn zywu@sz.tsinghua.edu.cn
lingjun@sjtu.edu.cn {runnan.li, xuta, sheng.zhao}@microsoft.com

In this supplementary material, we will discuss more detail of the proposed method VAST. In Section. 1, the loss function for the variational style enhancer will be explained. In Section. 2, we will present the architecture of the image renderer. In Section. 3, we will introduce how we conduct the mean opinion score (MOS) test.

1. Variational Style Enhancer

One of the key contributions of our method is that we propose the variational style enhancer, which enhances the style space to be highly expressive and meaningful. Without this enhancer, the style space learned by the style encoder and hybrid decoder is flat. The variational style enhancer is based on variational autoencoder [3, 5] and normalizing flow [7]. The training of this enhancer can be considered as the reconstruction of facial expressions, which is conditional on the speech. If we remove the normalizing flow module, the loss function thus becomes:

$$\begin{aligned}\ln p_\theta(\boldsymbol{X} \mid \boldsymbol{A}) \geq\ & \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{X},\boldsymbol{A})}[\ln p_\theta(\boldsymbol{X} \mid \boldsymbol{z}, \boldsymbol{A})] \\ & - \mathrm{KL}(q_\phi(\boldsymbol{z} \mid \boldsymbol{X}, \boldsymbol{A}) \| p(\boldsymbol{z}|\boldsymbol{A})),\end{aligned} \tag{1}$$

where $\boldsymbol{X}$ is the facial expression sequence, and $\boldsymbol{A}$ is the corresponding phonetic posteriorgram (PPG) [6] sequence. Since the latent variable $\boldsymbol{z}$ can be considered to be independent with $\boldsymbol{A}$, Eq. 1 is further defined as

$$\begin{aligned}\ln p_\theta(\boldsymbol{X} \mid \boldsymbol{A}) \geq\ & \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{X},\boldsymbol{A})}[\ln p_\theta(\boldsymbol{X} \mid \boldsymbol{z}, \boldsymbol{A})] \\ & - \mathrm{KL}(q_\phi(\boldsymbol{z} \mid \boldsymbol{X}, \boldsymbol{A}) \| p(\boldsymbol{z})).\end{aligned} \tag{2}$$

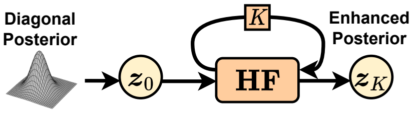

Figure 1. Illustration of the variational style enhancer.

To achieve a more flexible posterior distribution other than a simple diagonal Gaussian, we apply the householder-transformation [7] based normalizing flow to enhance the variational inference [4]. By applying a sequence of invertible mappings $\mathbf{H}^{(k)}, k = 1, \ldots, K$, over the initial variable, we obtain a more valid and flexible probability distribution at the end of this sequence. As shown in Fig. 1, $\boldsymbol{z}^{(k)} = \mathbf{H}^{(k)}(\boldsymbol{z}^{(k-1)})$ and the distribution of $\boldsymbol{z}^{(k)}$ can be transformed from the previous $\boldsymbol{z}^{(k-1)}$:

$$p(\boldsymbol{z}^{(k)}) = p(\boldsymbol{z}^{(k-1)})|\det \frac{\partial {\mathbf{H}^{(k)}}^{-1}}{\partial \boldsymbol{z}^{(k)}}| \tag{3}$$

$$= p(\boldsymbol{z}^{(k-1)})|\det \frac{\partial \mathbf{H}^{(k)}}{\partial \boldsymbol{z}^{(k-1)}}|^{-1}, \tag{4}$$

where $\det$ denotes the Jacobian determinant of the transformation. The density of $\boldsymbol{z}^{(k)}$ is obtained by successively transforming $\boldsymbol{z}^{(0)}$ through a sequence of transformations:

$$\boldsymbol{z}^{(K)} = \mathbf{H}^{(K)} \circ \ldots \circ \mathbf{H}^{(2)} \circ \mathbf{H}^{(1)}(\boldsymbol{z}^{(0)}), \tag{5}$$

$$\ln p_K(\boldsymbol{z}^{(K)}) = \ln p_0(\boldsymbol{z}^{(0)}) - \sum_{k=1}^{K} \ln|\det \frac{\partial \mathbf{H}^{(k)}}{\partial \boldsymbol{z}^{(k-1)}}|. \tag{6}$$

With the enhanced posterior distribution replacing the vanilla diagonal posterior, Eq. 2 thus becomes

$$\begin{aligned}\ln p_\theta(\boldsymbol{X} \mid \boldsymbol{A}) \geq\ & \mathbb{E}_{q_\phi(\boldsymbol{z}^{(0)}|\boldsymbol{X},\boldsymbol{A})}[\ln p_\theta(\boldsymbol{X} \mid \boldsymbol{z}^{(K)}, \boldsymbol{A}) \\ & + \sum_{k=1}^{K} \ln|\det \frac{\partial \mathbf{H}^{(k)}}{\partial \boldsymbol{z}^{(k-1)}}|] \\ & - \mathrm{KL}(q_\phi(\boldsymbol{z}^{(0)} \mid \boldsymbol{X}, \boldsymbol{A}) || p(\boldsymbol{z}^{(K)})),\end{aligned} \tag{7}$$

where the first reconstruction term is formulated with $\boldsymbol{z}^{(K)}$, since we finally sample from the distribution of $\boldsymbol{z}^{(K)}$. The first term can also be $\mathbb{E}_{q_\phi(\boldsymbol{z}^{(0)}|\boldsymbol{X},\boldsymbol{A})}[\ln p_\theta(\boldsymbol{X} \mid \boldsymbol{z}^{(0)}, \boldsymbol{A})]$.

2. Renderer Structure

We adopt the conditional generative network (CGAN) [2] as the basic framework for the renderer. The eroded

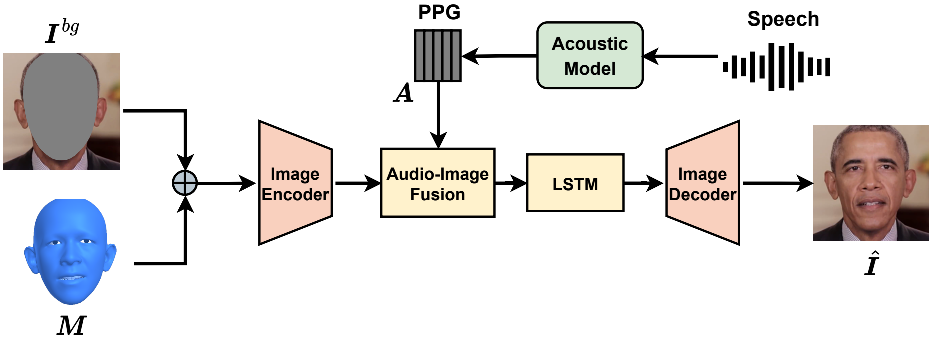

Figure 2. Illustration of the renderer module. The background images and mesh images are concatenated. The images are fed as a sequence of 16 frames for a training sample. The input can be of variable length in the inference stage.

background images $\boldsymbol{I}^{bg}$ and corresponding 3D face representation mesh sequence [8] $\boldsymbol{M}$ are taken as the conditional input. As shown in Fig. 2, the renderer is designed as an encoder-decoder structure. The image encoder is composed of three convolutional layers with stride 2. The encoded bottleneck features contain compact information about face shape, appearance, and image background. To enable the renderer to be perceptual about the relation between speech features and the mouth-region movements and enhance the synthesis accuracy on the mouth shape, the speech features PPG are injected into the bottleneck features. The PPG features and image encoder features are sent into the audio-image fusion module and output the final bottleneck features. This fusion module is constructed with four convolutional layers. To capture the time dependency among the sequence of audio and image features, a long short-term memory (LSTM) [1] module is utilized after the fusion module. Finally, the decoder which is composed of two transposed convolutional layers is employed to output the reconstructed images $\hat{\boldsymbol{I}}$.

3. MOS Guideline

In the authenticity evaluation, expressiveness evaluation and ablation study, we extensively conduct the MOS tests to verify the effectiveness of the proposed method. Three main aspects are taken into account in these tests: speech-lip sync, expressiveness & richness, and overall naturalness. Fifteen judgers participate in these tests. We now list the questions they are asked to evaluate these three aspects.

1. Speech-Lip Sync. How much do the lip movements match the audio? Very good (5) for no wrong lip movements and have nothing different from the ground-truth person talking. Very poor (1) for the lip movements are totally unreasonable and cannot read content from the lips at all.

2. Expressiveness & Richness. How vivid or exaggerated is the avatar presented? Very good (5) for rich and contagious expression on the mouth region. Very poor (1) for unreal and monotonous facial movements.

3. Overall Naturalness. How real and natural is the synthesized video? Very good (5) for high-quality images and natural avatar appearance. Very poor (1) for fake artifacts or blurred images that can be easily observed.

References

[1] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[2] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976, 2017.

[3] Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2022.

[4] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 37, pages 1530–1538, 2015.

[5] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Proceedings of the International Conference on Neural Information Processing Systems (NIPS)*, volume 28, page 3483–3491, 2015.

[6] Lifa Sun, Kun Li, Hao Wang, Shiyin Kang, and Helen Meng. Phonetic posteriorgrams for many-to-one voice conversion without parallel data training. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2016.

[7] Jakub M Tomczak and Max Welling. Improving variational auto-encoders using householder flow. In *Proceedings of the International Conference on Neural Information Processing Systems Workshops*, 2016.

[8] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J. Cashman, and Jamie Shotton. Fake it till you make it: Face analysis in the wild using synthetic data alone. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3681–3691, 2021.