FaceVerse – Supplementary Material

A. Implement Detail

All the 60K face models from base datasets are used to build the base model, and we use detailed training set which contains 2,310 high-resolution face models from 110 volunteers (each performs 21 specific expressions) to train the detailed face generators, leaving 378 models from 18 volunteers for evaluation. We also sample several Asian face images from the FFHQ dataset[1] to show the performance of our monocular 3D face fitting pipeline. We follow the basic training schedule of Style-GAN during the training process of our detail generator G_{detail} and expression refinement generator G_{exp}, each of which contains an additional encoder input with the conditional image and a normal discriminator. The input and output resolution of the generators is set to 1024×1024. And we choose 80 shape parameters, 120 texture parameters and 64 expression parameters for the base face fitting.

B. Additional Experiments

Comparison with 3DMM methods. We first explain the quantitative comparison of 3DMM methods in detail, as shown in Fig. 11 of the main text. We conduct a quantitative comparison with the state-of-the-art East Asian facial parametric models proposed by FaceScape and Hifi3DFace, as well as the BFM, on 3D scans from our testing set, which contains 378 models from 18 people. The models are fixed in the length of 200mm. We fit the parametric models to 3D scans by an optimization algorithm, where the shape and expression parameters are optimized by

$$L_{ICP} = \lambda_{lms} L_{lms} + \lambda_{near} L_{nearest} \tag{1}$$

where L_{lms} denotes the L2 loss of ground-truth landmarks and the pre-defined corresponding landmarks and $L_{nearest}$ denotes the L2 loss of the model points and the corresponding nearest points of the ground-truth scans. Note that the nearest points are searched iteratively and we set the overall iteration five times. Only the frontal area of BFM, FaceScape and Hifi3DFace is used for error calculation and our base model show better performance than FaceScape, Hifi3DFace and BFM both visually and quantitatively. Since our detailed model need additional texture input, we only use the base model to conduct a fair comparison. In the 3D fitting process, we use 120 shape parameters and 64 expression parameters for our base PCA model, 300 shape parameters and 52 expression parameters for the FaceScape bilinear model, 500 shape parameters and 199 expression parameters for Hifi3DFace PCA model, 80 shape parameters and 64 expression parameters for BFM PCA model.

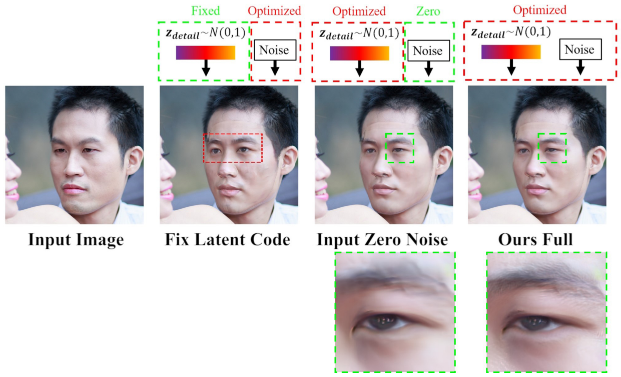

Figure 1. The evaluation of input latent code and injected noise.

Evaluation of latent code and noise input. In order to better understand the mechanics of our model, we conduct an experiment on the input latent code and the injected noise. Our detail model needs conditional UV maps from the base model, latent code input to the mapping network and noise injected to the model weights. As shown in Fig. 1, input with the same base UV maps, we set up three sets of tests during the single-image fitting process: input with fixed latent code and optimizable noise, input with zero noise and optimizable latent code, input with both optimizable noise and latent code. Results in Fig. 1 show that the latent code will affect the detailed shape of facial features and the noise will help to generate tiny details like hair and brows. Besides, the experiment of changing one of noise and latent code while fixing the other is also presented in our video.

C. Detailed Network Architecture

The detailed architecture of our conditional StyleGAN generator is presented in Fig. 2, the input image is a 6-channel UV map with the resolution of 1024×1024. The unlabeled modules (orange and blue rectangles) represent two convolutional layers. The output channel and resolution are also labeled in the convolutional layers. The multi-scale features from the encoder are input to the decoder by the skip connections. The "UP" module represents the "To-RGB" module of the original StyleGANv2. The mapping network has 8 fully connected layers with 512 channels. The style input and noise injection module are the same with the original StyleGANv2. The output image is a 6-channel UV map (detailed generator) or a 3-channel UV map (expression refinement generator). We use the same structure with the original StyleGANv2 except the 6-channel input for our discriminator and the additional normal discriminator.

[1] https://github.com/NVlabs/ffhq-dataset

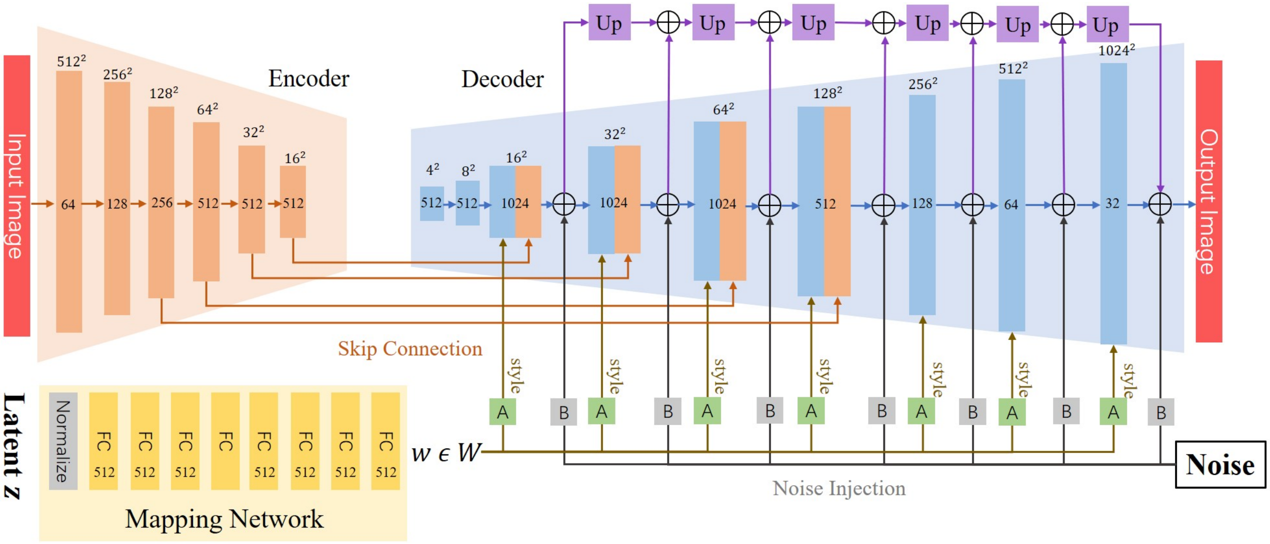

Figure 2. The detailed architecture of our conditional StyleGAN generator.