Supplementary Materials

1. Network Structure Details

Style-specific audio-to-animation generator G^{ani}. The structural details of G^{ani} is shown in Figure 1.

- **(a)-(e)**: the 5 basic blocks in G^{ani} with the setting of kernel size, stride and padding.
- **(f)**: the mapping from $\{I^{ref} \in R^{3\times512\times512}, f^{audio} \in R^{T\times15}\}$ to $\hat{f}^{audio}_{ref} \in R^{T\times256}$.
- **(g)**:the translation from $\hat{f}^{audio}_{ref}$ to $\{p^{mou} \in R^{T\times28}, p^{ebro} \in R^{T\times5}, p^{hed} \in R^{T\times5}\}$.
- **(h)**:the setting of channel size of each block.

Animation discriminator. In our experiment, D^{mou}, D^{ebro} and D^{hed} share the same structure. The structural details are illustrated in Figure 2.

3DMM. In our 3DMM, M_0 and $\{V^s_i\}^{60}_{i=1}$ are preserved as in LSFM[1]. We disentangle the mouth and eyebrow movements, and sculpture 28 mouth bases and 5 eyebrow bases as our $\{V^e_j\}^{33}_{j=1}$. The details of eyebrow and mouth bases are shown in Figure 3 and Figure 4 respectively.

Flow-guided video generator G^{vid}. In G^{vid}, the structure of Houglass network is preserved as the original paper[2]. Three separate layers, after the Houglass network, with the setting of kernel size = 7, stride = 1 and padding = 3, are employed to compute M^m, g and M^f. The structure of encoder-decoder is shown in Figure 5.

Discriminator D^{vid}. The structure of D^{vid} is as same as [3]. In this paper, the number of multi-D is set to 1.

2. Implementation Details

In the training stage, G^{ani} and G^{vid} are trained separately. In the audio-to-animation module, f^{audio} is extracted in 100 fps. p^{mou}, p^{ebro} and p^{hed} are up-sample to 100 fps with the linear interpolation. 1000 frames are split as one sample by the temporal window. The slide stride is set to 128. We set $\lambda_{mou} = 100$, $\lambda_{ebro} = \lambda_{hed} = 10$. The batch size is 128. The learning rate is initially set to 0.0005, and stay fixed in the first 50 epoches and linearly decay to 0 within another 50 epoches. It takes about 26 hours to train G^{ani}.

In the animation-to-video module, we set $\lambda_{perc} = \lambda_{FM} = 10$. The batch size is set to 2 with 2 1080ti GPU. In the first 75 epoches, the learning rate is fixed in 0.0004. In the last 75 epoches, the learning rate linearly decay to 0. We take about two weeks to train G^{vid}. Adam optimizer is used in G^{ani} and G^{vid} with default setting.

In the inference stage, we first extract the face shape parameter p^s and initial face animation parameters p^{mou}_{init}, p^{ebro}_{init} and p^{hed}_{init} from I^{ref}. Then, we synthesize the $\{\hat{p}^{mou}, \hat{p}^{ebro}, \hat{p}^{hed}\}$ from $\{I^{ref}, f^{audio}\}$. Finally, we synthesize the video frame-by-frame according to generated animation parameters. Our framework produces videos at about 2 frames per second.

In the experiments, to evaluate our audio-to-animation module and animation-to-video module, 5% of the YAD dataset is randomly selected for testing.

3. Demo Video

A video demo is also included in the supplementary material.

4. Ethical Consideration

We strongly advocate using our technology properly. To prevent the abuse of our method, anyone who employ our method to synthesize fake videos should mark with "fake video". Fortunately, detecting the synthetic and manipulated videos has got much attention and achieved much progress. As part of our responsibility, we are happy to promote the development of detection methodologies by sharing our dataset, source codes for their future research.

References

[1] James Booth, Anastasios Roussos, Allan Ponniah, David Dunaway, and Stefanos Zafeiriou. Large scale 3d morphable models. *International Journal of Computer Vision*, 126(2-4):233–254, 2018.

[2] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016.

[3] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *Advances in Neural Information Processing Systems*, pages 7137–7147, 2019.

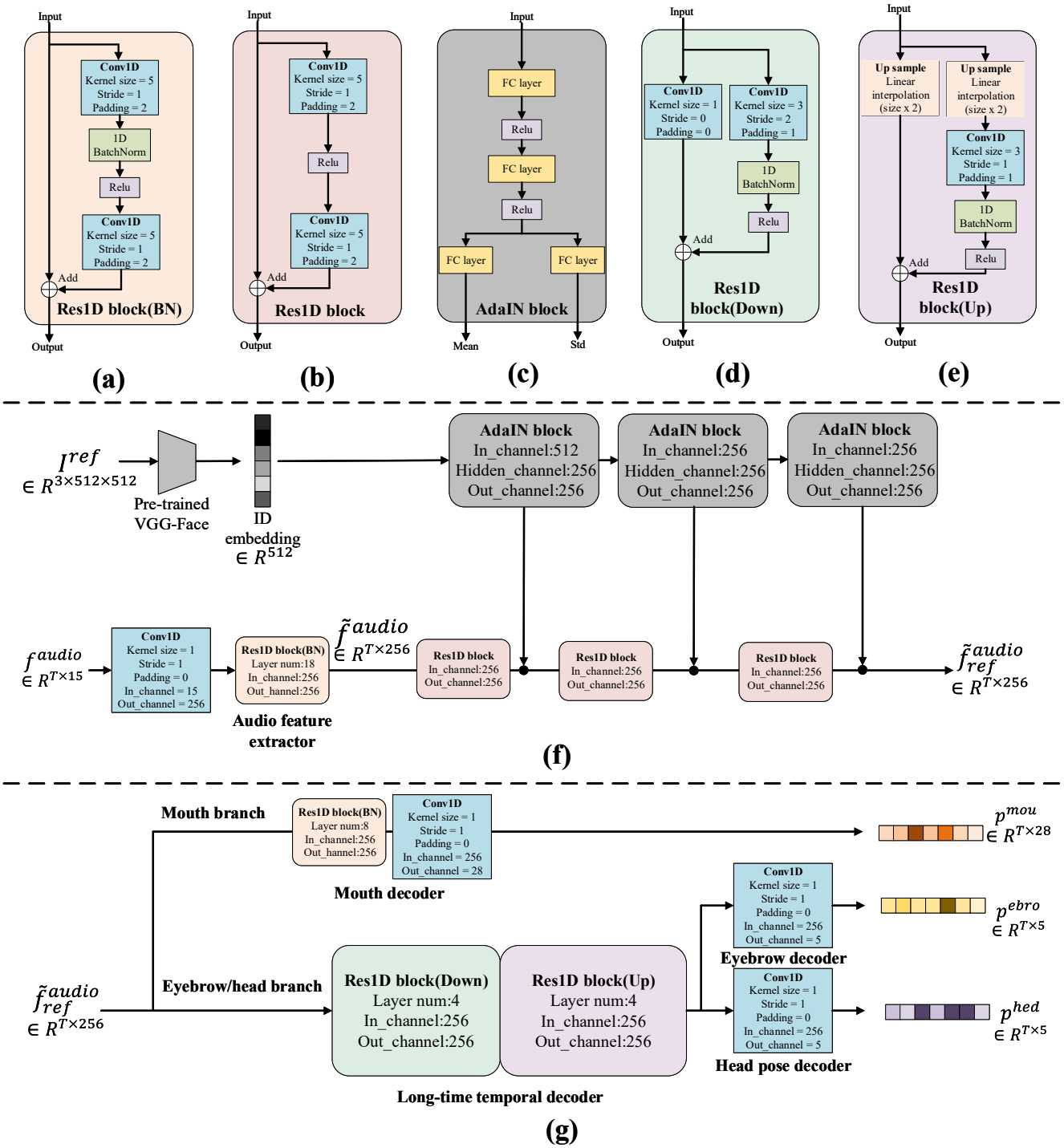

Figure 1. The structural details of style-specific animation generator G^{ani}.

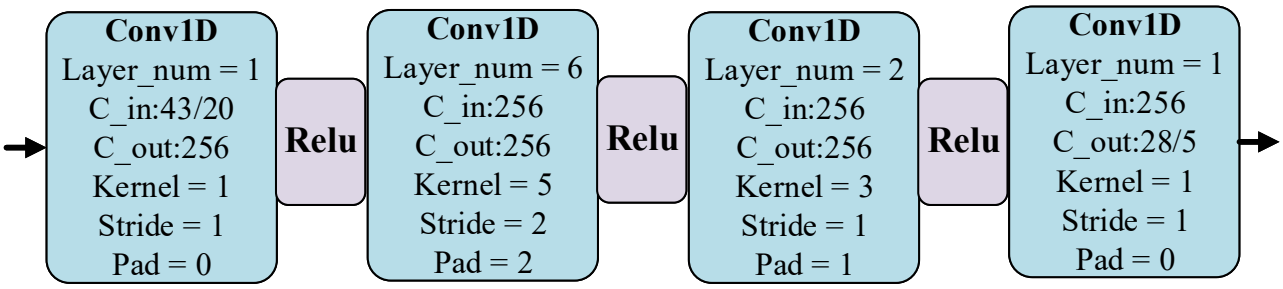

Figure 2. The structural details of animation discriminator D^{mou}, D^{ebro}, D^{hed}.

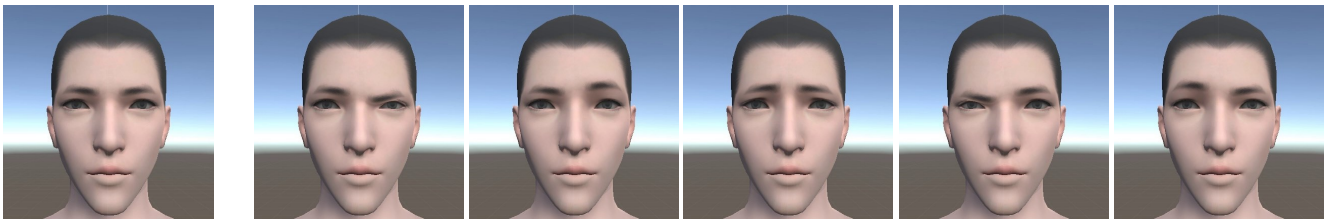

Figure 3. The eyebrow basis in 3DMM.

Figure 4. The mouth basis in 3DMM.

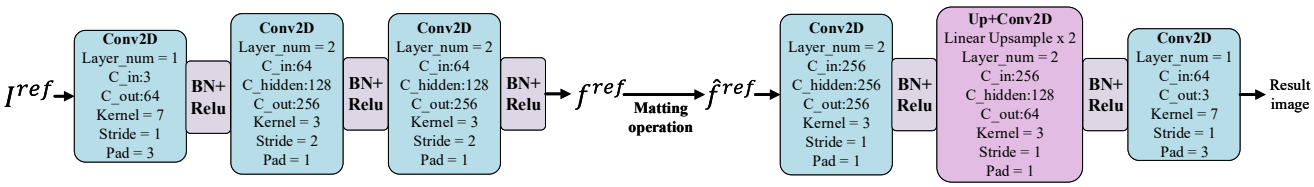

Figure 5. The structural details of encoder-decoder network in G^{vid}.