3D Talking Face with Personalized Pose Dynamics (Supplementary Document)

APPENDIX A
DETAILS OF POSEGAN NETWORK

Here, we discuss the more technical details of our proposed PoseGAN network. First, we define the following two terms:

- Initial pose: the first frame of head pose sequence is the initial pose $\mathbf{p} \in \mathbb{R}^6$, which includes Euler angles (pitch θ_x, yaw θ_y, roll θ_z) in radians and a 3D translation vector $\mathbf{t}$ in millimeters.
- Pose sequence: this represents the head pose sequence $\mathbf{p_s} \in \mathbb{R}^{256\times 6}$ in 256 frames.

A.1 U-net Structure

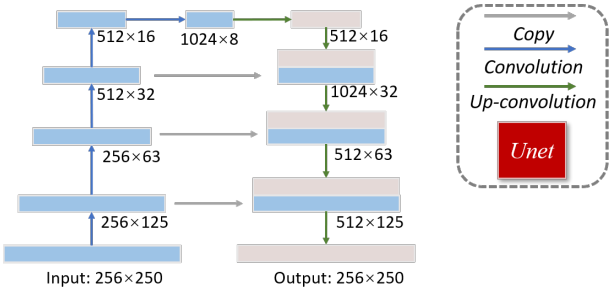

Fig. 1. The network structure of U-net used in our PoseGAN.

We employ the 1D U-net structure [1] in our network, which includes five downsampling layers and five upsampling layers, as shown in Fig. 1. Each convolution layer was followed by batch normalization [2] and ReLU [3].

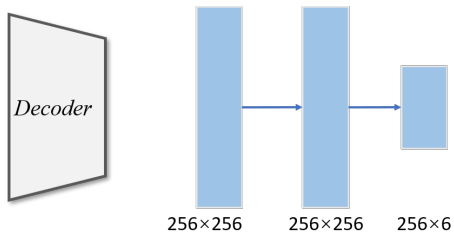

Fig. 2. The decoder structure used to generate the head pose sequence.

A.2 Decoder Structure

Three convolution layers were used to generate the head pose sequence in the decoder structure, as shown in Fig. 2. The first two layers are both followed by batch normalization and ReLU.

APPENDIX B
DETAILS OF PGFACE NETWORK

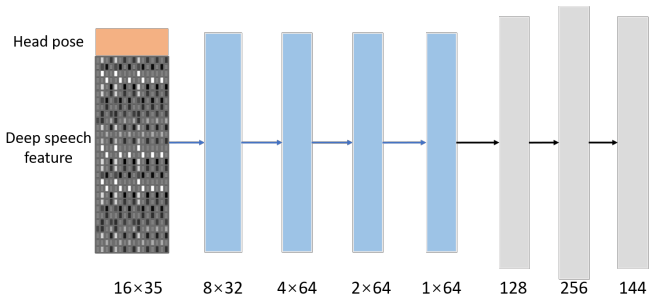

Fig. 3. Network structure of our PGFace module.

As shown in Fig. 3, we use a 16-frame window to compose the input for PGFace, which includes the head poses $\mathbf{p} \in \mathbb{R}^{6\times 16}$ concatenated with the deep speech feature $\mathbf{s} \in \mathbb{R}^{29\times 16}$. The first four convolution layers were all followed by batch normalization and ReLU. Then, three fully connected layers were employed to generate the face shape parameters $[\alpha_{\text{id}}, \alpha_{\text{exp}}]$.

REFERENCES

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, 2015, pp. 234–241.
[2] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.
[3] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 2011, pp. 315–323.