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 \( p \in \mathbb{R}^6 \), which includes Euler angles (pitch \( \theta_x \), yaw \( \theta_y \), roll \( \theta_z \)) in radians and a 3D translation vector \( t \) in millimeters.
- **Pose sequence**: this represents the head pose sequence \( p_s \in \mathbb{R}^{256 \times 6} \) in 256 frames.

**A.1 U-net Structure**

![Fig. 1. The network structure of U-net used in our PoseGAN.](image)

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].

**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.](image)

As shown in Fig. 3, we use a 16-frame window to compose the input for PGFace, which includes the head poses \( p \in \mathbb{R}^{6 \times 16} \) concatenated with the deep speech feature \( s \in \mathbb{R}^{29 \times 16} \). 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 \([\alpha_{id}, \alpha_{exp}]\).

**REFERENCES**