Supplementary Material for
Semi-supervised Speech-driven 3D Facial Animation via Cross-modal Encoding

Anonymous ICCV submission

Paper ID 1586

In this supplementary material, we describe details of the network architecture and the training process of our methods. Additionally, we discuss the limitations of our method and propose potential solutions. A supplementary video is also provided, which presents further qualitative comparisons, ablation studies, and visual results of our approach.

1. Network Architecture

**Speech Encoder.** The speech encoder consists of an audio feature extractor and a multi-layer transformer encoder. The audio feature extractor is initialized with pre-trained wav2vec2.0 [1] weights and generates audio features of dimension 1024. The structure of the multi-layer transformer encoder is adopted from [5], which consists of an input linear layer, a 3-layer transformer encoder, and an output linear layer. The self-attention and feed-forward layers have a hidden size of 512, and 4 attention heads are employed. The input linear layer converts the audio feature into 512-dimensional hidden embedding, while the output linear layer preserves the hidden dimension. The resulting speech encoding has a dimension of 512.

**Visual Encoder.** The shared visual encoder adopts ResNet34 [2] as the backbone, followed by an average pooling layer and a single fully connected layer. The resulting visual encoding has a dimension of 512.

**Decoder.** We borrow the decoder structure from [4], which consists of several upsampling blocks with an upsampling scale of 2. The details are depicted in Table 1.

2. Implementation Details

**Data Batch Organization.** In training phase, each batch of data contains speech, real facial images and synthetic facial images. Specifically, a batch is composed of 20 speech snippets, 20 real facial images, and 14 synthetic facial images, where the speech snippets and real facial images are synchronized. In addition, two neutral expression images are included, one is a real face and the other is a synthetic face.

**Training details.** We resize the facial images to 256x256 and convert them to grayscale images. We use the Adam optimizer [3] with a learning rate of 1e-4. During training, the parameters of the audio feature extractor are fixed. The model is trained for 15000 steps. We evaluate our model using the last checkpoint.

3. Limitations and Solutions

The proposed network consists of only one real face decoder and one synthetic face decoder, which cannot achieve image-to-image translation for multiple identities. For example, generating animations for additional CG characters is infeasible. To overcome this limitation, we can extend...
the network architecture to multiple decoders, with each decoder corresponding to a unique identity. It is worth noting that in order to train this multi-decoder network, images of multiple identities are required for training.

References


