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.

Layer	Activation	Output shape
Dense	-	16384
Reshape	-	256x8x8
Conv3x3	LeakyReLU	512x8x8
structure	LeakyReLU	2048x8x8
PixelShuffle	-	512x16x16
Conv3x3	LeakyReLU	2048x16x16
PixelShuffle	-	512x32x32
Conv3x3	LeakyReLU	512x32x32
Conv3x3	LeakyReLU	512x32x32
Conv3x3	LeakyReLU	1024x32x32
PixelShuffle	-	256x64x64
Conv3x3	LeakyReLU	256x64x64
Conv3x3	LeakyReLU	256x64x64
Conv3x3	LeakyReLU	512x64x64
PixelShuffle	-	128x128x128
Conv3x3	LeakyReLU	128x128x128
Conv3x3	LeakyReLU	128x128x128
Conv3x3	LeakyReLU	256x128x128
PixelShuffle	-	64x256x256
Conv3x3	LeakyReLU	64x256x256
Conv3x3	LeakyReLU	64x256x256
Conv1x1	-	3x256x256

Table 1. Decoder architecture. Leaky ReLU activations use a slope of 0.1. Pairs of consecutive convolutions are composed as residual blocks [2].

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

[1] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020. 1

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[3] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[4] Jacek Naruniec, Leonhard Helminger, Christopher Schroers, and Romann M Weber. High-resolution neural face swapping for visual effects. In *Computer Graphics Forum*, volume 39, pages 173–184. Wiley Online Library, 2020. 1

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1