Identity-Preserving Talking Face Generation with Landmark and Appearance Priors (*Supplementary Document*)

Weizhi Zhong[1] Chaowei Fang[2] Yinqi Cai[1] Pengxu Wei[1]
Gangming Zhao[3] Liang Lin[1] Guanbin Li[1]*

[1]Sun Yat-sen University [2]Xidian University [3]The University of Hong Kong

{zhongwzh5, caiyq27}@mail2.sysu.edu.cn, chaoweifang@outlook.com
{weipx3, liguanbin, linlng}@mail.sysu.edu.cn, gangmingzhao@gmail.com

1. Implementation Details

1.1. Network Architectures

Audio-to-Landmark Generator. The landmark generator includes three encoders and L transformer modules. The network architectures of the reference encoder E_r, the pose encoder E_p, and the audio encoder E_a are depicted in Figure 1. In each convolutional layer, 'k', 's', and 'p' refer to kernel size, stride, and padding, respectively. BN stands for batch normalization [3].

Landmark-To-Video Rendering Model. The landmark-to-video rendering stage consists of an alignment module G_a and a translation module G_r. The network architectures are illustrated in Figure 2. For SPADE and AdaIN layers, 'mod' and 'h' denote the channel of the input and the size of hidden variable, respectively.

1.2. Implementation Details

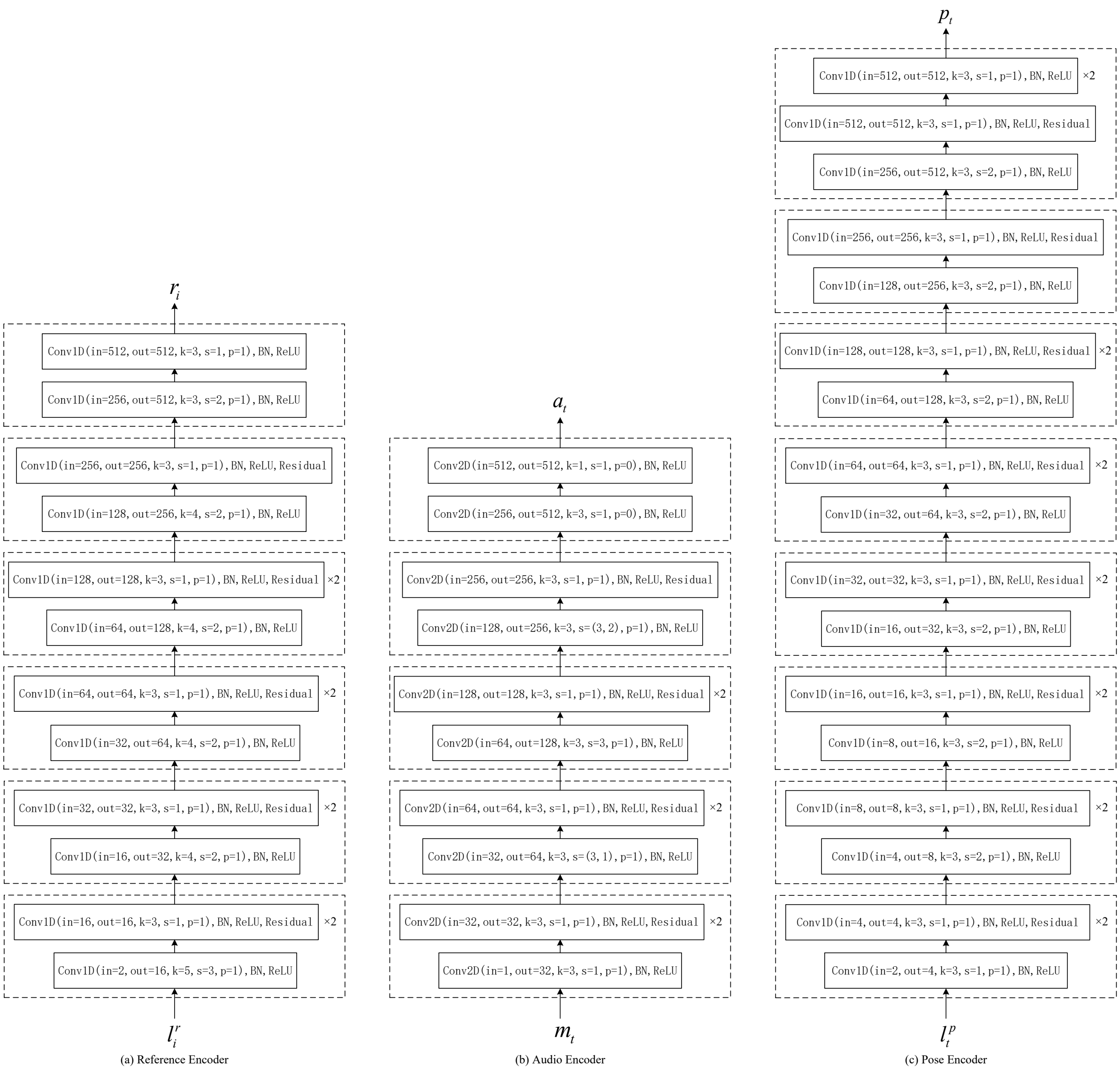
Our method is implemented with PyTorch [6] on Ubuntu 18.04 with RTX 3090 GPUs. In the landmark generator, we use four transformer modules ($L = 4$). Each transformer module has four attention heads ($h = 4$) and feature dimension of 512 ($d = 512$). The sketch is generated by drawing the landmarks and their connections on the image plane. The spatial resolutions of h^1 and h^2 are 128x128 and 64x64, respectively. We adopt Adam [5] with learning rate of 10^{-4} for optimizing network parameters. During inference, we select reference images by ensuring that both open and closed mouths are accessible.

We implemented the LSTM-based variant of our method by replacing the transformer modules with bidirectional LSTM modules [2]. The bidirectional LSTM has 4 layers, an input size of 1536, and a hidden size of 512. To enlarge the predicted landmarks, which take the range of 0 to 1, we multiplied them by 200. Note that 200 is approximately equal to the maximum height of the cropped face image. We then calculated the landmark error as the mean squared error between the predicted landmarks and the ground-truth landmarks.

1.3. Implementation of Other Methods

We conduct experiment on the reconstruction and dubbing settings. In the reconstruction setting, we input the original audio to reconstruct the talking face videos. In the dubbing setting, the input audio comes from another video. For fair comparison, we re-implement existing methods including **EAMM** [4], **PC-AVS** [8], **Wav2Lip** [7], **MakeItTalk** [9], and **ATVGnet** [1] as follows. **EAMM** [4] synthesizes face videos from a source image, a source audio, a pose sequence, and a set of emotion source images. Here, we re-implement it by regarding the first frame of each video as the source image and replacing the emotion source images with original input frames. During the testing process of **PC-AVS** [8], we select the first frame of each video as its identity reference, and the pose source is obtained from the corresponding video frame for each time

[*] Corresponding author is Guanbin Li.


r_i
Conv1D(in=512, out=512, k=3, s=1, p=1), BN, ReLU
Conv1D(in=256, out=512, k=3, s=2, p=1), BN, ReLU
Conv1D(in=256, out=256, k=3, s=1, p=1), BN, ReLU, Residual
Conv1D(in=128, out=256, k=4, s=2, p=1), BN, ReLU
Conv1D(in=128, out=128, k=3, s=1, p=1), BN, ReLU, Residual ×2
Conv1D(in=64, out=128, k=4, s=2, p=1), BN, ReLU
Conv1D(in=64, out=64, k=3, s=1, p=1), BN, ReLU, Residual ×2
Conv1D(in=32, out=64, k=4, s=2, p=1), BN, ReLU
Conv1D(in=32, out=32, k=3, s=1, p=1), BN, ReLU, Residual ×2
Conv1D(in=16, out=32, k=4, s=2, p=1), BN, ReLU
Conv1D(in=16, out=16, k=3, s=1, p=1), BN, ReLU, Residual ×2
Conv1D(in=2, out=16, k=5, s=3, p=1), BN, ReLU
l_i^r
a_t
Conv2D(in=512, out=512, k=1, s=1, p=0), BN, ReLU
Conv2D(in=256, out=512, k=3, s=1, p=0), BN, ReLU
Conv2D(in=256, out=256, k=3, s=1, p=1), BN, ReLU, Residual
Conv2D(in=128, out=256, k=3, s=(3, 2), p=1), BN, ReLU
Conv2D(in=128, out=128, k=3, s=1, p=1), BN, ReLU, Residual ×2
Conv2D(in=64, out=128, k=3, s=3, p=1), BN, ReLU
Conv2D(in=64, out=64, k=3, s=1, p=1), BN, ReLU, Residual ×2
Conv2D(in=32, out=64, k=3, s=(3, 1), p=1), BN, ReLU
Conv2D(in=32, out=32, k=3, s=1, p=1), BN, ReLU, Residual ×2
Conv2D(in=1, out=32, k=3, s=1, p=1), BN, ReLU
m_t
p_t
Conv1D(in=512, out=512, k=3, s=1, p=1), BN, ReLU ×2
Conv1D(in=512, out=512, k=3, s=1, p=1), BN, ReLU, Residual
Conv1D(in=256, out=512, k=3, s=2, p=1), BN, ReLU
Conv1D(in=256, out=256, k=3, s=1, p=1), BN, ReLU, Residual
Conv1D(in=128, out=256, k=3, s=2, p=1), BN, ReLU
Conv1D(in=128, out=128, k=3, s=1, p=1), BN, ReLU, Residual ×2
Conv1D(in=64, out=128, k=3, s=2, p=1), BN, ReLU
Conv1D(in=64, out=64, k=3, s=1, p=1), BN, ReLU, Residual ×2
Conv1D(in=32, out=64, k=3, s=2, p=1), BN, ReLU
Conv1D(in=32, out=32, k=3, s=1, p=1), BN, ReLU, Residual ×2
Conv1D(in=16, out=32, k=3, s=2, p=1), BN, ReLU
Conv1D(in=16, out=16, k=3, s=1, p=1), BN, ReLU, Residual ×2
Conv1D(in=8, out=16, k=3, s=2, p=1), BN, ReLU
Conv1D(in=8, out=8, k=3, s=1, p=1), BN, ReLU, Residual ×2
Conv1D(in=4, out=8, k=3, s=2, p=1), BN, ReLU
Conv1D(in=4, out=4, k=3, s=1, p=1), BN, ReLU, Residual ×2
Conv1D(in=2, out=4, k=3, s=1, p=1), BN, ReLU
l_t^p

(a) Reference Encoder

(b) Audio Encoder

(c) Pose Encoder

Figure 1. Network architectures of reference encoder E_r, audio encoder E_a, and pose encoder E_p in the audio-to-landmark generator.

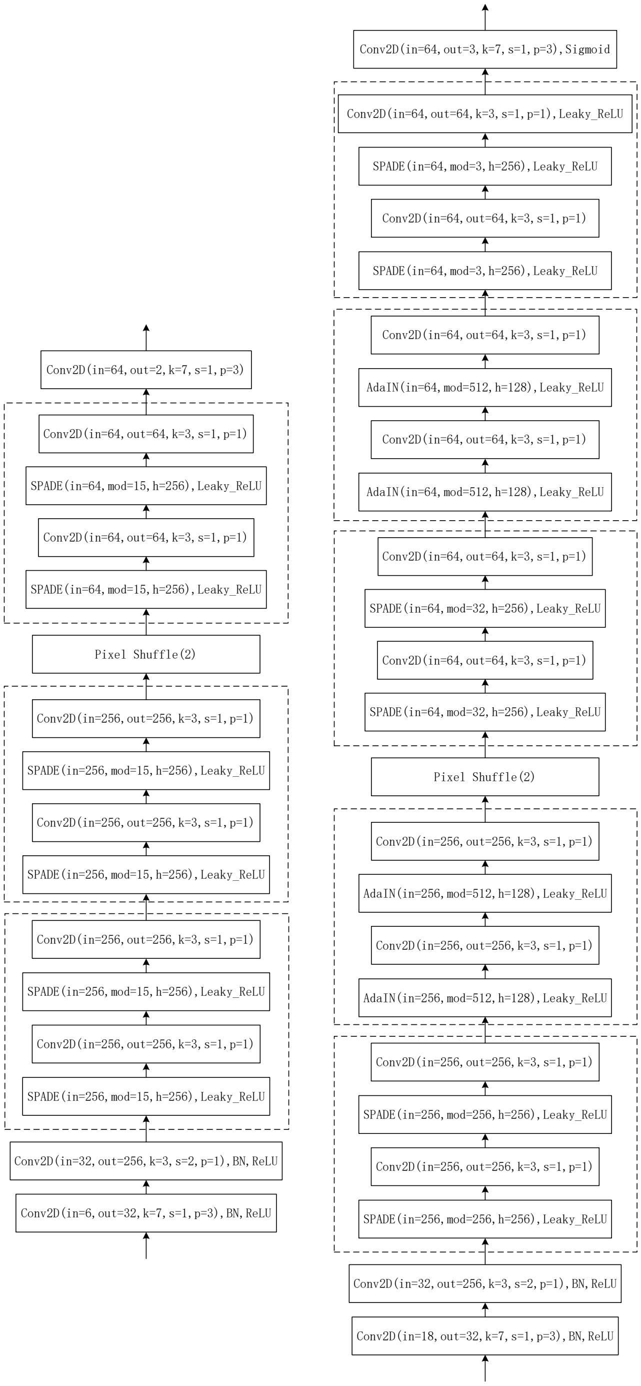
step. For **Wav2Lip** [7], we use the first frame as the reference image in the reconstruction task and regard the corresponding video frame as the reference image at each time step in the dubbing task. For **MakeItTalk** [9] and **ATVGnet** [1], we select the first frame of each video as the identity image.

2. Ethical Discussion.

Audio-driven talking face video generation is significant in extensive real-world applications, but it might be misused for media manipulation or other illegal profits. To combat these malicious behaviors, we will restrict the usage of our code and watermark the generated result. Besides, we are willing to share our synthetic videos with the deepfake detection community to improve their algorithms. We believe that properly using this technology will benefit our daily lives.

3. Complementary Qualitative Comparison

We provide more visualization examples in this section. Moreover, a demo video will be provided in project page: https://github.com/Weizhi-Zhong/IP_LAP.


Conv2D(in=64, out=2, k=7, s=1, p=3)
Conv2D(in=64, out=64, k=3, s=1, p=1)
SPADE(in=64, mod=15, h=256), Leaky_ReLU
Conv2D(in=64, out=64, k=3, s=1, p=1)
SPADE(in=64, mod=15, h=256), Leaky_ReLU
Pixel Shuffle(2)
Conv2D(in=256, out=256, k=3, s=1, p=1)
SPADE(in=256, mod=15, h=256), Leaky_ReLU
Conv2D(in=256, out=256, k=3, s=1, p=1)
SPADE(in=256, mod=15, h=256), Leaky_ReLU
Conv2D(in=256, out=256, k=3, s=1, p=1)
SPADE(in=256, mod=15, h=256), Leaky_ReLU
Conv2D(in=256, out=256, k=3, s=1, p=1)
SPADE(in=256, mod=15, h=256), Leaky_ReLU
Conv2D(in=32, out=256, k=3, s=2, p=1), BN, ReLU
Conv2D(in=6, out=32, k=7, s=1, p=3), BN, ReLU
Conv2D(in=64, out=3, k=7, s=1, p=3), Sigmoid
Conv2D(in=64, out=64, k=3, s=1, p=1), Leaky_ReLU
SPADE(in=64, mod=3, h=256), Leaky_ReLU
Conv2D(in=64, out=64, k=3, s=1, p=1)
SPADE(in=64, mod=3, h=256), Leaky_ReLU
Conv2D(in=64, out=64, k=3, s=1, p=1)
AdaIN(in=64, mod=512, h=128), Leaky_ReLU
Conv2D(in=64, out=64, k=3, s=1, p=1)
AdaIN(in=64, mod=512, h=128), Leaky_ReLU
Conv2D(in=64, out=64, k=3, s=1, p=1)
SPADE(in=64, mod=32, h=256), Leaky_ReLU
Conv2D(in=64, out=64, k=3, s=1, p=1)
SPADE(in=64, mod=32, h=256), Leaky_ReLU
Pixel Shuffle(2)
Conv2D(in=256, out=256, k=3, s=1, p=1)
AdaIN(in=256, mod=512, h=128), Leaky_ReLU
Conv2D(in=256, out=256, k=3, s=1, p=1)
AdaIN(in=256, mod=512, h=128), Leaky_ReLU
Conv2D(in=256, out=256, k=3, s=1, p=1)
SPADE(in=256, mod=256, h=256), Leaky_ReLU
Conv2D(in=256, out=256, k=3, s=1, p=1)
SPADE(in=256, mod=256, h=256), Leaky_ReLU
Conv2D(in=32, out=256, k=3, s=2, p=1), BN, ReLU
Conv2D(in=18, out=32, k=7, s=1, p=3), BN, ReLU

(a) Alignment Module

(b) Translation Module

Figure 2. Network architectures of the alignment module G_a and the translation module G_r in the landmark-to-video rendering model.

3.1. Comparison with Other Methods

We show the qualitative comparisons of our method against state-of-the-art person-generic methods in Figure 3. The input audio corresponds to the word “know” under the reconstruction setting. As can be seen, face images generated by our method are visually closer to the ground truth and have fewer artifacts than results of other methods.

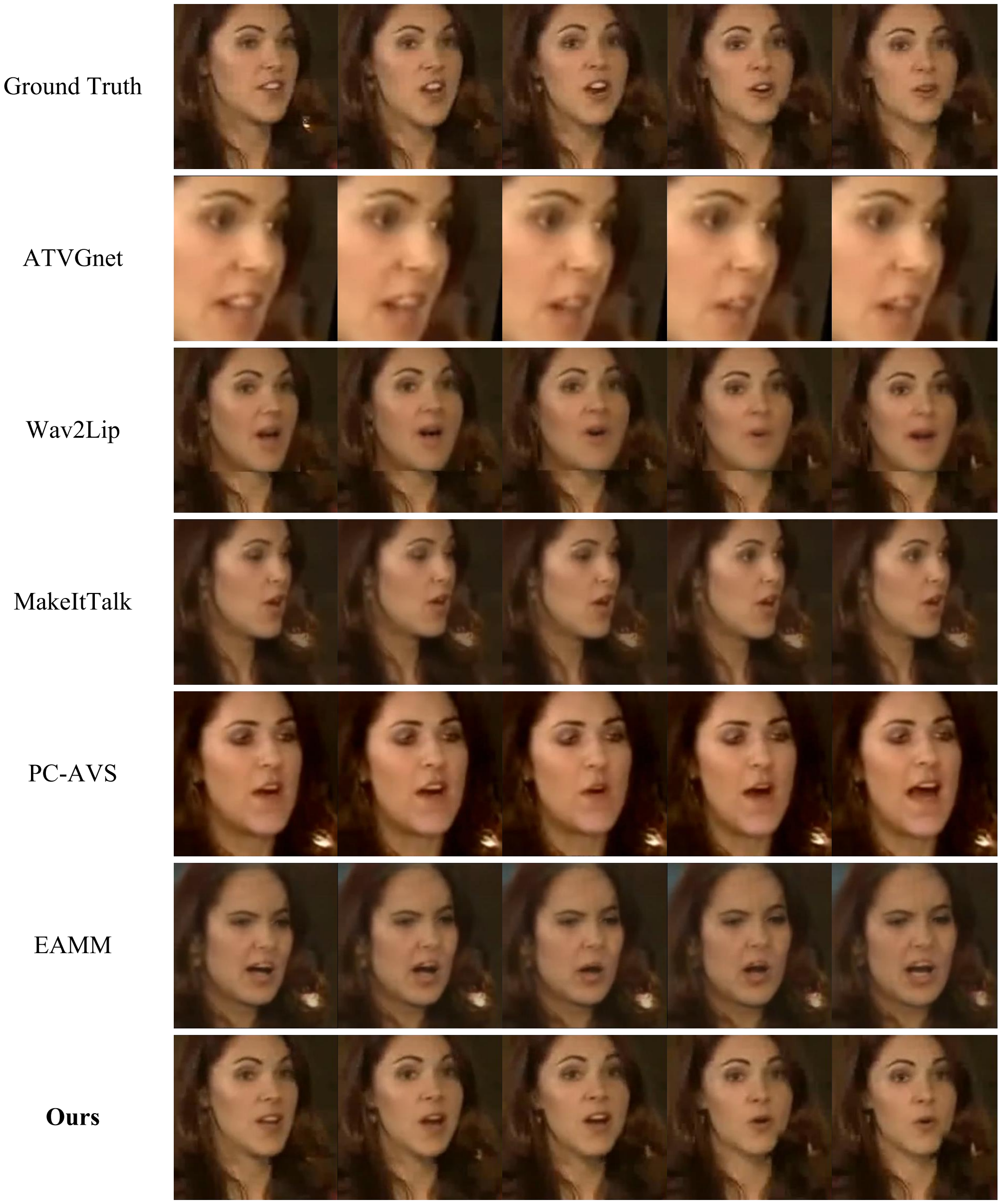


Figure 3. Qualitative comparisons with state-of-the-art person-generic methods. The input audio corresponds to the word "know". Our method produces realistic and lip-synced talking face videos with a better identity preservation effect.

3.2. Effectiveness of Transformer Encoder

In Figure 4, we provide an example for showcasing the superiority of using Transformer over using LSTM. The input audio corresponds to the word "cracked". The results of the LSTM-based method are prone to have smaller mouth openness than the results of the transformer-based method, leading to inconsistent mouth shapes with the ground-truth. The reason is that LSTM is less effective in modeling temporal dependencies and relationship between audio and landmark, deriving less accurate landmarks.

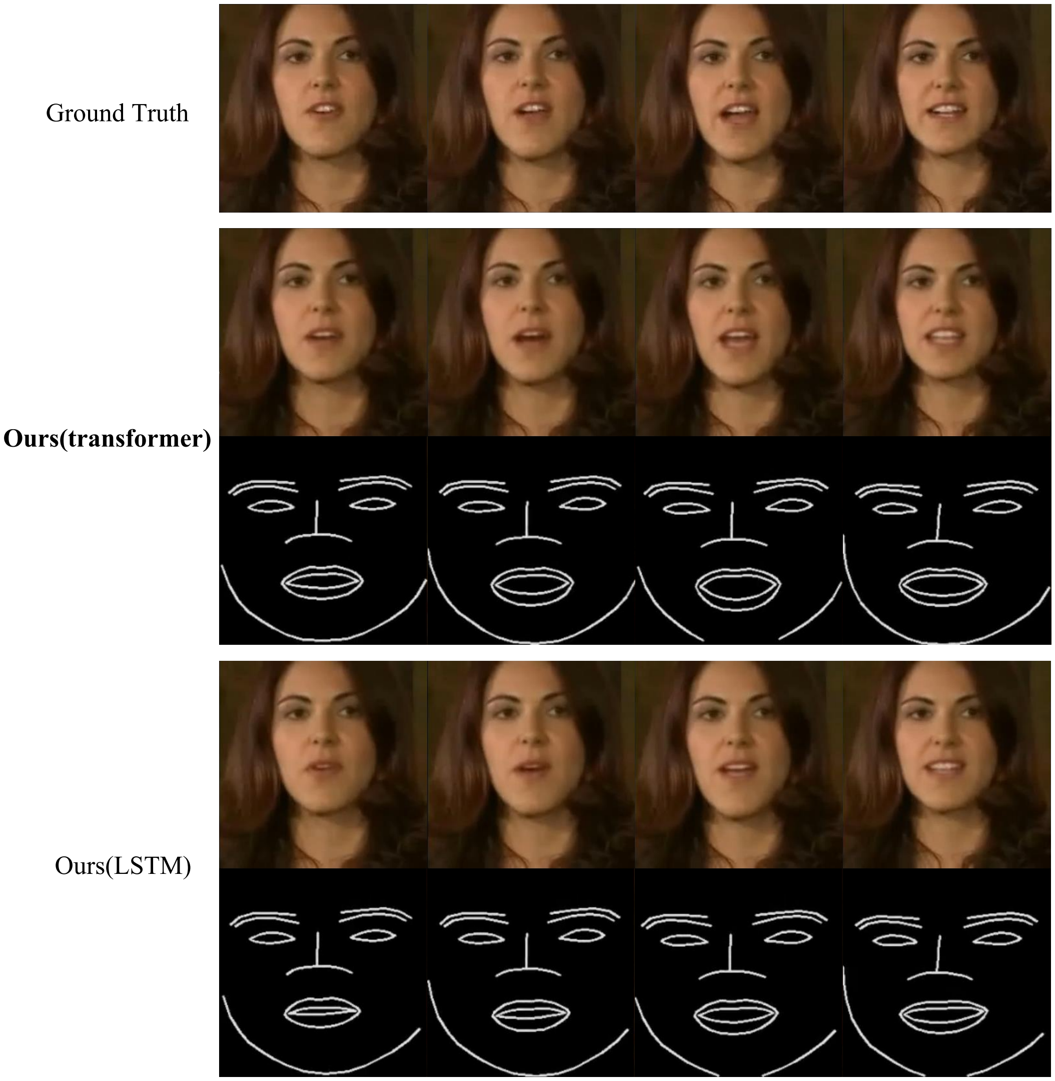


Figure 4. Qualitative comparison between variants of our method using Transformer (middle row) or LSTM (bottom row) to build the audio-to-landmark generator. The ground-truth images are shown in the top row. The input audio corresponds to the word “cracked”. Landmarks generated by the LSTM based method have smaller mouth openness than those generated by the Transformer based method, leading to less accurate mouth shapes.

3.3. Effectiveness of Warping Operation and Audio Feature

We implement variants of our method without using the warping operation or the audio feature guidance during the landmark-to-video rendering stage. Face images produced by them are shown in Figure 5. It can be seen that the visual quality deteriorates without using the warping operation to align reference images with the target head pose and expression (i.e. ‘w/o warping’). Especially, the teeth region becomes more blurry. Besides, without using audio features to enhance the lip synchronization and mouth details in the translation module (i.e. ‘w/o audio’), the lip shape is not as accurate as that in the images generated by the final variant of our method.

3.4. Effectiveness of Multiple Reference Images

In Figure 6, we provide face images generated with 1, 5, or 25 reference images. Using only one reference image produces some artifacts around the mouth. The reason is that one reference image is insufficient to cover all the facial details. Our method produces more realistic results when more reference images are provided.

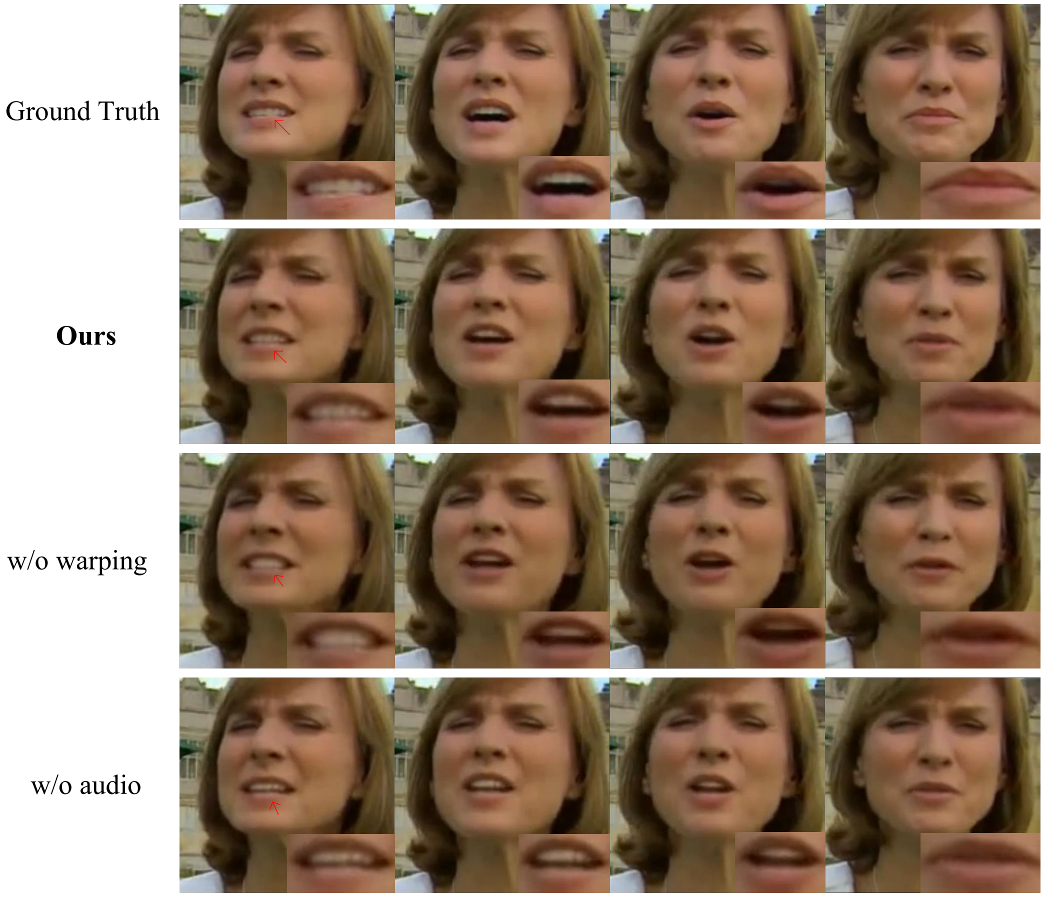


Figure 5. Visualization results of variants of our method without using the warping operation ('w/o warping') or the audio feature guidance ('w/o audio') in the landmark-to-video rendering model. The input audio corresponds to the word "sumptuous". Without using the warping operation to align reference images with the target pose and expression, the generated images are more blurry. Without using audio features in the translation module, the lip shape is not as accurate as that in the images generated by the final variant of our method.

References

[1] Lele Chen, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7832–7841, 2019. 1, 2

[2] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 1

[3] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015. 1

[4] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Qianyi Wu, Wayne Wu, Feng Xu, and Xun Cao. Eamm: One-shot emotional talking face via audio-based emotion-aware motion model. In *ACM SIGGRAPH 2022 Conference Proceedings*, SIGGRAPH '22, 2022. 1

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

[6] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 1

[7] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 484–492, 2020. 1, 2

[8] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4176–4186, 2021. 1

[9] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makelttalk: speaker-aware talking-head animation. *ACM Transactions on Graphics (TOG)*, 39(6):1–15, 2020. 1, 2

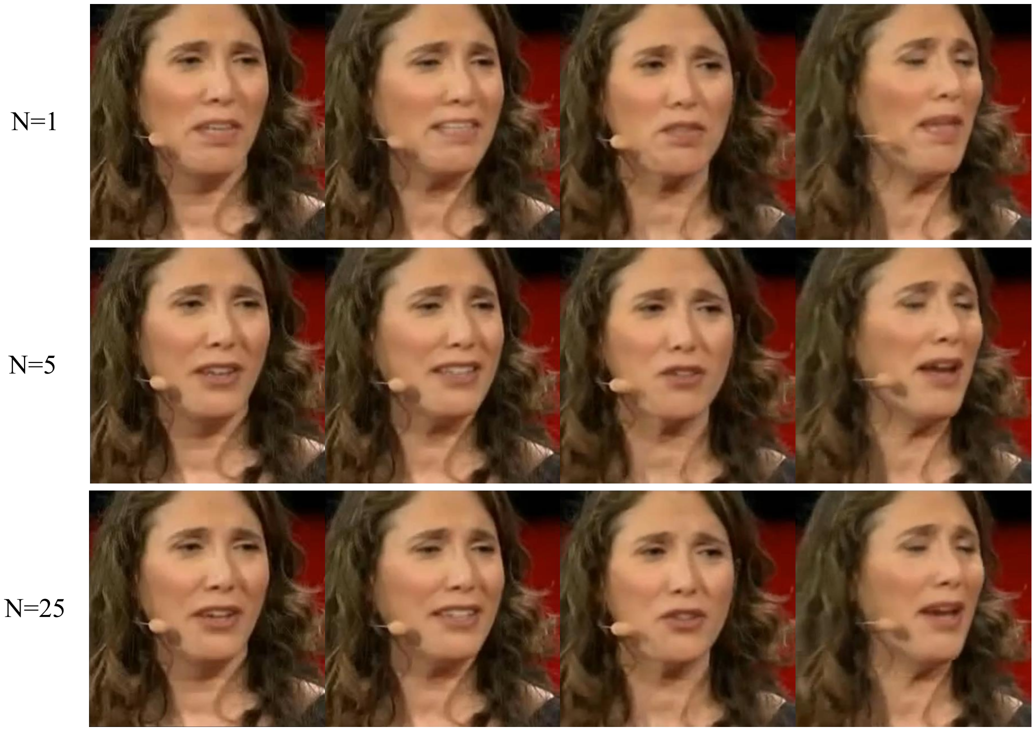


Figure 6. Visualization of face images generated by our method with different numbers of reference images. From top to bottom, 1, 5, or 25 reference images are used, respectively. Using multiple reference images can generate much cleaner results than using only one reference image.