LipFormer: High-fidelity and Generalizable Talking Face Generation with A Pre-learned Facial Codebook

<Supplementary Material >

This is the supplementary material for the paper "High-fidelity and Generalizable Talking Face Generation with A Pre-learned Facial Codebook". In this supplemental document, we introduce our collected YouTubeHQ dataset in detail and provide more implementation details. Then, we provide more experimental results. For more video results, please refer to the attached video file.

A. The YouTubeHQ dataset

We collect this dataset for two main purposes:
- Proving that our method performs better in generating high-quality (HQ) talking face. Existing talking face datasets (e.g. LRS2) usually consist of low-resolution video data, lacking of fine grained facial texture. It is difficult to distinguish which one is better at fine-grained lip and texture generation on such datasets. Compared with LRS2, Table 1 shows that our method outperforms state-of-the-art more significantly on YouTubeHQ, demonstrating that our approach captures more high-fidelity facial details.
- Proving that our method has good generalization (see Figure 8). The model in Figure 8 is trained on FFHQ and YouTubeHQ, tested on YouTubeHQ. We use the audio of B frame to drive the A frame (B and A are from different identities), showing the generalization of our LipFormer.

YouTubeHQ contains over 6400 high-quality short video sequences with audio track (Fig. A). The videos were crawled from YouTube. We filter out videos that have resolution lower than $720p$, retaining only high quality videos and split each video into short clips. Overall, we carefully separate the YouTubeHQ dataset into training and validation sets with 20000 and 1560 clips respectively to ensure no identity overlaps. The average video lengths vary from 4 to 7 seconds and all in 25 fps. The YouTubeHQ dataset includes vastly more variation than LRW [2] and LRS2 [1] in terms of age and mouth shape, and also has much better coverage of identities.

B. Model Settings

In this part, we introduce the model settings as well as the network architecture in detail.

B.1. HQ Codebook

In the stage of HQ codebook learning, for face encoder Enc and decoder Dec, we adopt similar architecture to that of VQGAN [3] with minor modification. We increase the resolution of the model form 256 to 512. For face encoder Enc, we introduce an extra down-sample layer and a ResNet block. Meanwhile, an up-sample layer and a ResNet block are added to the decoder Dec. We set the down sample rate of the face encoder Enc to 32. The dimension of the output encoded feature is $8 \times 16 \times 512$ (*i.e.*, a $256 \times 512 \times 3$ half face input will be encoded into a $8 \times 16 \times 512$ face feature). The input dimension of the decoder is set to $16 \times 16 \times 512$. The codebook size of each codebook ($\mathbb{C}_U$ and $\mathbb{C}_B$) is set to 4096. We set the dimension of each discrete code in the codebooks to 32 and introduce a linear projection from the output of the face encoder to a 32-d code space (*e.g.*, reduced from a 512-d vector to a 32-d vector per code). Also, another linear projection is added before the input of the decoder to project each 32-d code back to a 512-d feature.

B.2. LipFormer

The audio encoder E_{Aud} is a 13-layer convolutional network, which encodes the mel-spectrogram input $\boldsymbol{A}$ to the 512-d audio feature $\boldsymbol{F}_{Aud}$. In the Adaptive Face Warping Module, the keypoints extractor F_e is implemented as a 10-layer convolutional network, followed by a softmax operation, in which the strides are set to 1. For the offsets regressor F_d, we implement it as an U-Net-like network. Specifically, it consists of 2 down-sample layers, 2 up-sample layers and 6 ResNet blocks. For the Transformer Module, it consists of 12 8-head cross attention blocks, each followed by a residual layer, a conditional layer normalization operation and a feed-forward network (FFN). During the training we linearly anneal the temperature of the Gumbel-softmax operation, from $\tau = 1.0$ to $\tau = 0.5$ for iterations 1 to 100, 000 and then kept at $\tau = 0.5$ until training ends.

Figure A. The YouTubeHQ dataset includes a lot of variation in terms of age, ethnicity, viewpoint, and mouth shape.

	LRS2		YouTubeHQ	
Models	PSNR(↑)	SSIM(↑)	PSNR(↑)	SSIM(↑)
Baseline Model	31.613	0.843	28.035	0.749
+ FFHQ Pre-training	32.630	0.873	31.980	0.845
+ Adaptive Warping	32.411	0.865	31.637	0.833
+ FFHQ pre-training & Adaptive Warping	33.497	0.891	33.249	0.876

Table A. Ablation study of FFHQ Pre-training and the Adaptive Face Warping Module.

C. Experimental Results

C.1. More Metrics and Datasets

We add more experimental results on LRW, LRS3 and HDTF datasets. All results are shown in Tab. B. FPS is tested on NVIDIA Tesla V100 GPU. It can be seen our method outperforms others on most metrics. Also, We add LSE-D and LSE-C to evaluate Adaptive Warping(AW) and FFHQ pre-training(pt) effects on lip sync in Tab. C. Both modules have positive effects on the lip sync as they did on the image quality.

User Study is listed in Tab. E.Lip-quality means the quality of mouth region, lip-artifacts are the degree of artifacts.

C.2. Ablation Study

In this part, we further verify the effectiveness of different components. The ablative experiments contain: (1) Baseline Model, where the Adaptive Face Warping Module is removed, and the codebooks are learned without FFHQ dataset (only trained with LRS2 training set); (2) + FFHQ Pre-training, where the Adaptive Face Warping Module is not included, and the codebooks are learned with both LRS2 training set and FFHQ dataset; (3) + Adaptive Warping, where the Adaptive Face Warping Module is utilized for facial texture aligning, and the codebooks are learned without FFHQ dataset; (4) + FFHQ Pre-training & Adaptive Warping, which is the full LipFormer model. All models are trained on LRS2 training set and tested on the evaluation sets of LRS2 and YouTubeHQ respectively. The results in Tab. A verify that the FFHQ Pre-training enables the model

	FPS↑	YouTubeHQ/ LRS2 LSE-D↓	LSE-C↑	FID↓	CPBD↑	LRW/ LRS3/ HDTF PSNR↑	SSIM↑	LMD↓	LSE-D↓	LSE-C↑	FID↓	CPBD↑
ATVG	**36.13**	9.65	4.03	12.87/ 8.04	0.22/ 0.20	31.09/ 27.87/ 24.86	0.77/ 0.71/ 0.71	2.03/ 3.14/ 3.14	7.87/ 9.04/ 9.58	5.71/ 4.40/ 4.22	6.41/ 9.34/ 12.63	0.12/ 0.18/ 0.19
Wav2Lip	32.05	**7.68**	**5.57**	11.15/ 4.78	0.23/ 0.27	32.27/ 30.11/ 26.37	0.87/ 0.83/ 0.77	1.41/ 1.98/ 2.26	**6.62/ 6.67**/ 7.90	**7.15/ 8.90/ 5.23**	2.74/ 4.53/ 10.04	0.15/ 0.27/ 0.21
PC-AVS	4.63	8.31	5.28	12.33/ 9.22	0.21/ 0.21	29.39/ 27.84/ 25.22	0.76/ 0.72/ 0.72	1.61/ 2.99/ 2.51	7.55/ 8.16/ 8.19	6.20/ 5.81/ 4.83	7.04/ 9.83/ 12.82	0.10/ 0.19/ 0.20
LipFormer	9.92	7.71	5.48	**3.93/ 3.76**	**0.29/ 0.29**	**33.83/ 32.93/ 33.26**	**0.90/ 0.87/ 0.87**	**1.26/ 1.38/ 1.34**	6.96/ 6.89/ **7.89**	6.71/ 8.10/ 5.17	**2.38/ 3.79/ 3.85**	**0.18/ 0.28/ 0.29**

Table B. We add 1) LRW,LRS3,HDTF, 2) missing metrics for LRS2,YouTubeHQ, 3) FPS. SyncTalkFace is ignored for code unavailable.

Variants	LSE-D↓	LSE-C↑
w/o AW	7.91	5.36
w/o FFHQ pt	8.15	5.24
LipFormer	7.71	5.48

Table C. Lip-sync metrics.

n	PSNR↑	SSIM↑	LSE-D↓	LSE-C↑
2048	32.86	0.86	7.76	5.40
4096	**33.25**	**0.88**	**7.71**	**5.48**
8192	31.98	0.84	7.94	5.29

Table D. Ablation of codebook size.

Metrics	Wav2Lip	LipFormer
lip-sync↑	**3.24**	2.74
lip-quality↑	1.12	**2.97**
lip-artifacts↓	3.88	**2.09**

Table E. User Study.

Figure B

to better generalize to unseen identities. The Adaptive Face Warping Module further facilitates image quality.

Multiple results of LipFormer with or without the Adaptive Face Warping Module are displayed in Fig. B. It shows that LipFormer without the Adaptive Face Warping Module is able to generate results that are close to the target frame. But sometimes its output dose not preserve the pose and texture in the target, when the references have quite different poses (see in row 4). Fortunately, our Adaptive Face Warping Module is able to recover them.

The ablative experiment of codebook size is also conducted. We show the results in Tab. D. Note that performance improves with increasing n(2048∼4096) due to better expressivity of codebook, but larger codebook(n=8192) may contain redundant elements, leading to ambiguity in lip-codes predictions.

References

[1] Triantafyllos Afouras, Joon Son Chung, Andrew W. Senior, Oriol Vinyals, and Andrew Zisserman. Deep audio-visual speech recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019. 1

[2] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In *Asian Conf. Comput. Vis.*, 2016. 1

[3] Björn Ommer, Patrick Esser, Robin Rombach, Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 1