StyleLipSync: Style-based Personalized Lip-sync Video Generation Supplementary Material

Taekyung Ki[1*] Dongchan Min[2*]
[1]AITRICS, [2]Graduate School of AI, KAIST
tkki@aitrics.com, alsehdcks95@kaist.ac.kr
https://stylelipsync.github.io

Appendix A. Model Architectures

We provide the detailed encoder architectures in Figure 5, and MaLS in Figure 1, respectively. Note that we set $L = 7$ for the number of decoder blocks and use $2L = 14$ MaLS modules to generate a video of 256×256 resolution.

Encoders. The reference encoder $\mathbf{E}_{ref}$ consists of a stack of ResidualBlocks and a fully connected layer. The face encoder $\mathbf{E}_{face}$, on the other hands, consists of non-residual DownBlocks, which outputs 6 levels of spatial features maps, each of which are injected into the decoder $\mathbf{G}$ through the proposed SaMFs. We use the audio encoder architecture in [8] as our audio encoder $\mathbf{E}_{aud}$.

Moving-average based Latent Smoothing (MaLS). We use 14 MaLS, each of which consists of a stack of the weighted moving-average, and 1D convolutions. We compute the l-th moving-average $\overline{w}^l_t$ as follows:

$$\overline{w}^l_t = \sum_{\tau=t-1}^{t+1} \alpha_\tau \cdot w^l_\tau, \tag{1}$$

where $t \in [1, 2, \cdots, T]$, $\alpha_{t-1} = 0.25, \alpha_t = 0.5$, and $\alpha_{t+1} = 0.25$. Conv1Ds in the main paper is a stack of 1D convolutions of kernel size 3 with the stride of size 1 and reflection padding of size 1 as shown in Figure 1. To compute $\overline{w}^l_1$ and $\overline{w}^l_T$, we assign $w^l_0 = w^l_1$ and $w^l_{T+1} = w^l_T$, respectively. Since the weights $\alpha_{t-1:t+1}$ are shared at every time step t, averaged style codes $\overline{w}^l_{1:T}$ are calculated at once regardless of the frame length T. We use different Conv1Ds for each l-th MaLS block.

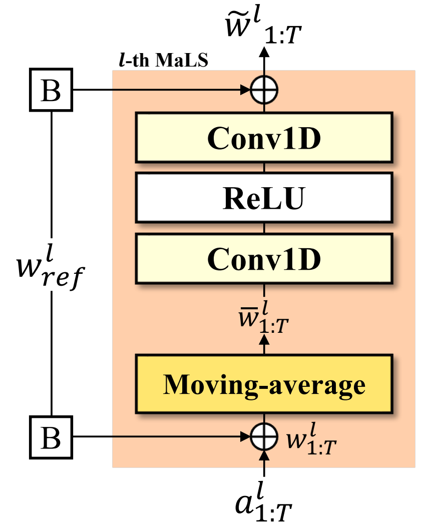


Figure 1. Detailed architecture of the l-th moving-average based latent smoothing (MaLS) module. $\boxed{\text{B}}$ is a broadcasting operator.

Appendix B. Additional Experiments

Reconstruction Results. We compute the CSIM [3] on the reconstruction task in Voxceleb2 test set to support the results the main paper. In Table 1, we report CSIM with the lip-sync metrics (LSE-D and LSE-C [5]). CSIM scores of all models are low compared to the results of the cross-id experiment, as Voxceleb2 [2] inherits more dynamic head pose than HDTF [7]. StyleLipSync can generate more accurate lip-sync video using the proposed pose-aware masking, therefore achieving the best CSIM score.

Table 1. Quantitative comparison of reconstruction of Voxceleb2 test data. The best score for each metric is in **bold**.

Method	Voxceleb2 (Reconstruction)		
	Image	Lip-Sync	
	CSIM ↑	LSE-D ↓	LSE-C ↑
Wav2Lip$_{96\times96}$ [5]	0.533	6.999	**8.329**
ATVG$_{128\times128}$ [1]	0.241	8.821	5.421
MakeItTalk$_{256\times256}$ [9]	0.634	10.895	3.624
PC-AVS$_{224\times224}$ [8]	0.405	7.278	7.699
Ours$_{256\times256}$	**0.789**	**6.628**	8.056

Unseen Face Adaptation. We plot the CSIM [3] scores in Figure 2 to further support the necessity of the sync regularizer $\mathcal{R}_{sync}$. In all cases of generator tuning, CSIM scores are higher than the zero-shot case and comparable

[*] Equal contribution.

with each other since we minimize the perceptual distance between the reference and generated frame [6]. In addition to these visual adaptation, the sync regularizer $\mathcal{R}_{sync}$ enforces maintaining the lip-sync generality as described in the main paper. Note that we do not incorporate ID loss [3] to our training objective. We experimentally found that ID loss (and ℓ^2 loss) is very sensitive in each image instance, so it leads video-level artifacts such as flicker. Therefore, we simplify the training objective where the LPIPS [6] is the only image-level loss.

Our adaptation method requires about 20-25 minutes (3-4k steps) for fine-tuning a video of 60 seconds.

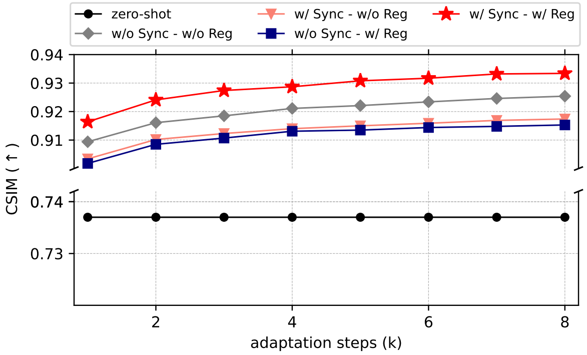


Figure 2. Cosine similarity between face embedding (CSIM) [3] tendency of the adaptation method.

Visualization of SaMFs. In Figure 3, we visualize the unsupervisedly predicted masks S_t^l and their overlay results. We resize and overlay the gray scaled inverse masks $1 - S_t^l$ onto the generated result. The SaMFs learn the attention masks with low scores (blue) on the mouth region throughout all resolutions, which helps the model to improve the lip fidelity as illustrated in the main paper.

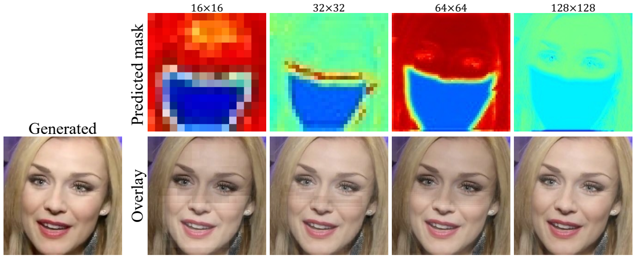
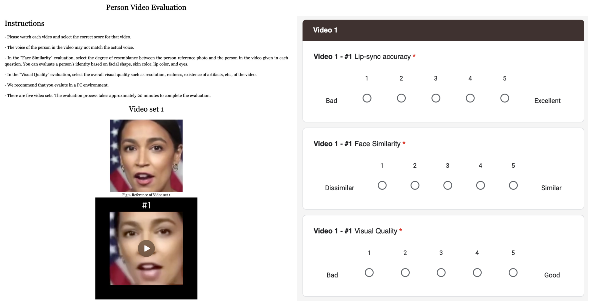


Figure 3. Visualization of SaMFs at each resolution. Please refer to our project page for visualization of the resolution less than 16 × 16.

Details of User Study. We provide an example of user studies in Figure 11. The studies conducted with Five videos in total. Five videos generated by each model were used in random order. Four videos from HDTF [7] are 10 seconds, and the rest from YouTube is 5 seconds.



Person Video Evaluation

Instructions

- Please watch each video and select the correct score for that video.

- The voice of the person in the video may not match the actual voice.

- In the "Face Similarity" evaluation, select the degree of resemblance between the person reference photo and the person in the video given in each question. You can evaluate a person's identity based on facial shape, skin color, lip color, and eyes.

- In the "Visual Quality" evaluation, select the overall visual quality such as resolution, realness, existence of artifacts, etc., of the video.

- We recommend that you evalute in a PC environment.

- There are five video sets. The evaluation process takes approximately 20 minutes to complete the evaluation.

Video set 1

Fig 1. Reference of Video set 1

#1

Video 1

Video 1 - #1 Lip-sync accuracy *

	1	2	3	4	5	
Bad	○	○	○	○	○	Excellent

Video 1 - #1 Face Similarity *

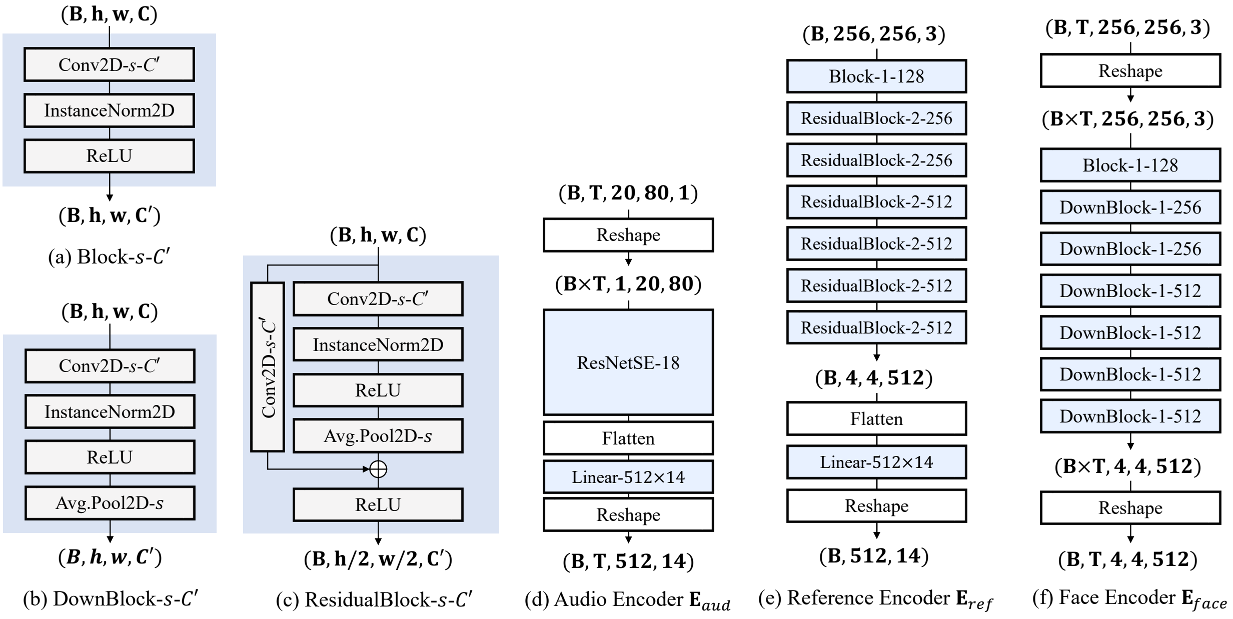
	1	2	3	4	5	
Dissimilar	○	○	○	○	○	Similar

Video 1 - #1 Visual Quality *

	1	2	3	4	5	
Bad	○	○	○	○	○	Good

Figure 4. Example of a questionnaire used in the user study.

Video Results. We additionally provide the video results of our work and other baselines, which includes the videos of reconstruction, cross-id, and ablation results. Please refer to our project page at the front of the paper.



(a) Block-s-C'

(b) DownBlock-s-C'

(c) ResidualBlock-s-C'

(d) Audio Encoder $\mathbf{E}_{aud}$

(e) Reference Encoder $\mathbf{E}_{ref}$

(f) Face Encoder $\mathbf{E}_{face}$

Figure 5. Detailed encoder architectures. $(\mathbf{B}, \mathbf{T}, \mathbf{h}, \mathbf{w}, \mathbf{C})$ is a 5D tensor of the batch size $\mathbf{B}$, the number of frames $\mathbf{T}$, the height $\mathbf{h}$, the width $\mathbf{w}$, and the number of channels $\mathbf{C}$. Similarly, $(\mathbf{B}, \mathbf{h}, \mathbf{w}, \mathbf{C})$ is a 4D tensor excluding the axis of the number of frames in $(\mathbf{B}, \mathbf{T}, \mathbf{h}, \mathbf{w}, \mathbf{C})$. 'Conv2D-$s$-$C'$' means a 2D convolution of 3×3 kernel with stride (s, s), padding (s, s), and output channels C'. 'Linear-C' means a fully-connected layer of the output with C nodes. 'Avg.Pool2D-s' is a 2D average pooling of $s \times s$ kernel with stride (s, s). We employ the audio encoder architecture used in [8, 4] as our audio encoder.

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.

[2] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. In *INTERSPEECH*, 2018.

[3] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[4] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018.

[5] 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.

[6] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.

[7] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3661–3670, 2021.

[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.

[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.