**Abstract**

Despite recent advances in syncing lip movements with any audio waves, current methods still struggle to balance generation quality and the model’s generalization ability. Previous studies either require long-term data for training or produce a similar movement pattern on all subjects with low quality. In this paper, we propose StyleSync, an effective framework that enables high-fidelity lip synchronization. We identify that a style-based generator would sufficiently enable such a charming property on both one-shot and few-shot scenarios. Specifically, we design a mask-guided spatial information encoding module that preserves the details of the given face. The mouth shapes are accurately modified by audio through modulated convolutions. Moreover, our design also enables personalized lip-sync by introducing style space and generator refinement on only limited frames. Thus the identity and talking style of a target person could be accurately preserved. Extensive experiments demonstrate the effectiveness of our method in producing high-fidelity results on a variety of scenes. Resources can be found at https://hangz-nju-cuhk.github.io/projects/StyleSync.

---

*Equal contribution.
†Corresponding authors.
1. Introduction

The problem of generating lip-synced videos according to conditional audio is of great importance to the field of digital human creation, audio dubbing, film-making, and entertainment. While the rapid development of this area has been witnessed within recent years, most methods [6, 8, 9, 16, 19, 24, 27, 28, 38, 39, 46, 50, 54, 57–60] focus on generating a whole dynamic talking head. Results created under such settings can hardly be blended into an existing scene. Under real-world scenarios like audio dubbing, one crucial need is to seamlessly alter the mouth or facial area while preserving other parts of the scene unchanged, making these methods non-feasible.

Previous methods take two different paths for achieving seamless mouth modification. A number of studies [38, 44] pursue realistic results on person-specific settings, which require long-term clips for target modeling. Moreover, they rely on prior 3D facial structural information. The uncertainty and errors accumulated in the 3D fitting procedure would greatly influence their performances. On the other hand, it is desired to build models that break the data limitation on more generalized scenes. As a result, a few methods [31, 32, 41] design person-agnostic models without relying on 3D or 2D structural priors. Nevertheless, such a setting is extremely challenging.

In order to produce high-fidelity lip-synced results on any-length videos, two essential challenges need to be addressed. 1) How to efficiently design a powerful generative backbone network that supports both accurate audio information expression and seamless local alternation. Intuitively, the lip-sync quality naturally contradicts the preservation of the original target frame information [32, 59]. 2) How to effectively leverage as much provided information as possible and involve the personalized properties to a generalized model. Though few-shot meta-learning has been proven effective in generating talking heads [5, 7, 56, 60], how to involve such ability into a lip-syncing pipeline has not been explored.

In this paper, we propose a highly concise and comprehensive framework named StyleSync, which produces high-fidelity lip-sync results on both generalized and personalized scenarios. The key is our simple but lip-sync-oriented modifications to style-based generator. Though style-based generators [22, 23] have been leveraged in various talking head generation methods [5, 55, 59], their successes are only partially instructive. They aim at producing the whole head, which leads to unstable background and distortions which are non-acceptable in our scenarios.

By revisiting the details of style-based generators, we identify a few simple but essential modifications that make our framework suitable for lip-syncing. Different from the above methods, we adopt a masked mouth modeling protocol [32, 41] and delicately design a Mask-based Spatial Information Encoding strategy, where both the target and reference frames’ information is encoded into a noise space [52, 53] of the generator according to different masking schemes. While the information on audio dynamics and high-level reference frame is injected into the style-modulated convolution in a similar manner as [25, 59]. In this way, our method can be benefited from the strong generative power of style-based generators and also keeps the advantage of easy implementation and fast training.

Moreover, our network modification enables personalized information preserving (e.g., speaking styles and details of the mouth and jaw). We take inspiration from the recent studies of inverting StyleGAN priors [1, 2, 35, 45] and propose a Personalized Optimization scheme. As audio dubbing is normally performed on speaking videos, our model can make use of only a few seconds of the person’s information and optimize additional person-specific parameters including the $W^+$ and the generator. Extensive experiments show that our framework clearly outperforms previous state of the arts on the one-shot setting by a large margin, and the target-specific optimization further enhances the fidelity of our results.

Our contributions can be summarized as follows: 1) We present the StyleSync framework, which adopts simple but effective modifications including the Mask-based Spatial Information Encoding to a style-based generator. 2) We propose the Personalized Optimization procedure which involves few-shot person-specific optimization into our framework. 3) Extensive experiments demonstrate that our framework can directly produce accurate and high-fidelity one-shot lip-sync results. Moreover, our proposed personalized optimization further improves the generation quality. Our method outperforms previous methods by a clear margin.

2. Related Work

2.1. Audio-Driven Facial Animation

The topic of audio-driven facial animation has long been a research interest in both the computer vision and graphics community. Studies have been carried out on both digital 3D human faces [15, 20, 33, 61] and realistic human heads [14, 17, 42]. We focus on human heads in the real world.

Talking Head Generation. Most studies on lip-syncing target generating the whole head of a talking person [6, 8, 9, 16–18, 25, 26, 40, 42, 54, 58–60]. Specifically, a number of studies leverage structural information such as 2D [9] landmarks, 3D landmarks [60] and 3D meshes [6]. The uncertainty and inaccuracy of such representations would lead to error accumulation during the talking head generation procedure. Most person-specific methods [19, 27, 42] rely on them and produce results with a poor generalization or lip-
sync quality. Recently, NeRF is also used in person-specific modeling [16, 26, 43], but they also perform poorly when driven by audio from a different person.

**Lip-Syncing on Faces.** The other type of study focuses on lip-syncing the mouth part while keeping other information untouched in videos [31, 32, 39, 44]. Our work lies within the same scope. While Thies et al. [44] and Song et al. [38] produce realistic results, they rely on person-specific training on more than 2 minutes of data.

Specifically, Wav2Lip [32] generates person-agnostic lip-sync results that are highly accurate. However, they build the generative model on low-resolution images, leading to blurry results. We identify that it would be easier for the model to learn the mouth motions if less information about the image quality needs to be recovered. Moreover, their methods cannot capture the personalized patterns given the target template video.

### 2.2. Style-based Generator for Faces

StyleGAN models [21–23] have shown great success on image generation tasks, particularly on facial image generation and editing [1, 2, 34, 36, 37, 45].

**StyleGAN Inversion.** StyleGAN inversion is the practice of recovering the latent space of a given image with a pre-trained StyleGAN. Abdal et al. [1, 2] propose to invert images not only using the style space $W$, but also expand it to the $W^+$, which better preserves details. Recent studies propose to learn encoders [3, 34, 45] for better editability and faster optimization results. Furthermore, recent studies try to refine the generator’s parameters through pivotal finetuning [35]. In our work, we take inspiration of previous StyleGAN inversion studies and propose a personalized optimization procedure. We encode a $W^+$ space from reference frames following [45] which enables personalized lip movements learning.

**Applications.** StyleGAN architecture has also been leveraged in face restorations [48, 53] and face swapping [51, 52]. These tasks require preserving the original facial emotion and expressions. On the other hand, style-based generators have also been applied to create facial animations [5, 25, 59], which are highly related to our task. However, they rely on the style vectors in the $W$ or $W^+$ space for controlling both the appearances and motion dynamics. One major drawback of this setting is that $W^+$ space cannot easily account for the spatial consistency of backgrounds, leading to non-realistic results or visible artifacts.

### 3. Methodology

In this section, we introduce the details of our StyleSync framework as depicted in Fig. 2. It is simply designed by leveraging several successes of previous style-based generators. We introduce our modifications to the style-based generator that make a successful backbone in Sec. 3.1 and the training objectives of our generalized model in Sec. 3.2. Specifically, we illustrate our *Personalized Optimization* procedure which further strengthens our results with person-specific optimization in Sec. 3.3.

**Task Formulation.** Our goal is to sync the lip motion of a target person with any provided audio and seamlessly blend it into the original target video. We formulate our training in a typical self-reconstruction manner. The training setting is similar to Wav2Lip [32]. Given a target video $V = \{I_1, \ldots, I_T\}$ with $T$ frames, we mask out the lower half of the face including the jaw and cheeks with a mask $M$. The goal is to recover the whole face with its corresponding audio $a = \{a_1, \ldots, a_T\}$ ($a$ is processed to spectrograms). As no information about the mouth and jaw shape is provided on the masked target frame $I_i^m = (1 - M) \ast I_i$, we leverage a random reference frame $I_{ref} \in V$ during training to provide the desired context.

### 3.1. Modifying Style-based Generator for Lip Sync

**Style-based Generator Overview.** We perform lip-sync-oriented modifications to build a successful backbone for lip sync with the StyleGAN2 [23] architecture. The original StyleGAN2 takes a style vector $w$ as input and uses it to modulate convolution operations in a total of $L$ generative layers during training. During inference, different $w$s can be used at different layers, formulating the $W^+$ space $W = \{w_1, \ldots, w_L\}$. The basic elements of the original StyleGAN2, including the $W^+$ space and the StyleGAN generator $G$’s layers are depicted in blue on Fig. 2. For simplification, certain detailed structures are omitted.

**Mask-based Spatial Information Encoding.** Our goal is to seamlessly blend a lip-synced mouth into the target frame with the assistance of a reference image. As discussed in Sec. 2.2, previous methods leveraging the style-based generator are not directly applicable to our scene, we seek a different way to encode the spatial information on faces. It is proven by recent studies on face restoration [53] and swapping [52] that content-guided feature maps can be attached to StyleGAN layers in a similar status as *noise* without affecting the expressive power of the generative model. These contextual-rich noises $N = \{N_1, \ldots, N_L\}$ could well preserve the original spatial structure and attributes of the encoded visual information. This property provides an option for leveraging the attributes information lying in the reference image. Thus following Wav2Lip [32], we concatenate the masked and reference input together as the visual input of the framework and encode the corresponding feature maps $F = \{F_1, \ldots, F_L\}$. The simplest setting is to regard $N = F$.

However, the goal of generating lip motions needs to create dynamics that are different from the original facial structure but keeps facial identity, which is substantially different from previous methods that aim to keep the structure
fixed. The mouth shapes would be greatly influenced by the reference image $I_{ref}$ when adopting the same protocol. We identify that more than desired facial structure information is infused in the low-level information of $F$ (layers with higher resolutions). As a result, we intuitively mask out the low-level part of the information in the face features. Here we simply define

$$ F_i' = (1 - M) * F_i, \text{for } l > \lfloor \frac{L}{2} \rfloor. \hspace{1cm} (1) $$

The facial information is colored in red on Fig. 2. The $\mathbf{1}$ is an all-one matrix that has the same size as $M$.

**Style Information Encoding.** We follow previous studies [5, 40, 59] to encode both audio dynamics and facial information into the style space ($W$ space) of the style-based generator. The audio feature $f_a$ is encoded from the encoder $E_a$ and the face feature $f_f$ is the bottleneck of the face encoder $E_{face}$, $w = \text{concat}(f_a, f_f)$. As the information in the reference frames is already fused to the generator through spatial noise encoding, the fusing of the face feature is less necessary. Experiments show that the difference is subtle with or without $f_f$, but we still keep this design in accordance with common practice.

**Ingredients for Personalization.** Our above designs are not only simple but also possess external potentials of personalized lip-sync modeling. We leave the details to Sec. 3.3 and briefly discuss the additional ingredients that make our model comprehensive.

Specifically, our extension is inspired by the advances in StyleGAN inversion [1, 2, 35]. It has pointed out that extending $w$ to the $W^+$ space which contains $W = \{w + \Delta w_1, \ldots, w + \Delta w_L\}$ is essential to recovering specific images. Moreover, recent studies [4, 35] also explore tuning the parameters of the generator to improve the network’s fitting ability on a specific target. Thus our generator has the potential of learning limited parameter shifts $\Delta P$ in order to fit a specific person.

These personalized ingredients are depicted in green on Fig. 2 and are not optimized during the generalized training procedure. In order to keep the pipeline consistent across both settings, we set $\Delta W = 0$ and $\Delta P = 0$ in our generalized model.

### 3.2. Backbone Training Objectives

During training, the whole backbone networks take the 6-channel concatenation of $[I^m_t, I_{ref}]$ and the audio clip $a_t$ at the same time step as input. It predicts $I'_t = G(N, W)$, which aims at recovering the unmasked target image $I_t$. In order to keep our design simple, the training objectives are mostly aligned with StyleGAN2 [23] and Wav2Lip [32].

**Reconstruction Loss.** The pixel-level reconstruction loss $\mathcal{L}_{rec}$ is fundamental for training our task. Here we leverage the commonly used $L_1$ loss and perceptual loss [47].

$$ \mathcal{L}_{rec} = ||I'_t - I_t||_1 + \sum_{m=1}^{N_{vgg}} ||VGG_m(I'_t) - VGG_m(I_t)||_1, $$

where $VGG_m$ is the $m$th layer's output of a pre-trained VGG19 network.

**Adversarial Loss.** We directly adopt the same discriminator $D$ from StyleGAN2 [23] for the adversarial training:

$$ \mathcal{L}_{adv} = \min_{G} \max_{D} \mathbb{E}_{I_t} [\log D(I_t)] + \mathbb{E}_{I'_t} [\log (1 - D(I'_t))]. $$

![Figure 2. Our StyleSync framework.](image)
particularly, we initialize the discriminator’s weights with the pre-trained version in StyleGAN2 [23]. this practice produces visibly sharper results.

lip-sync loss. we additionally train a syncnet [12] which consists of a visual encoder \( S_v \) and an audio encoder \( S_a \). it identifies whether visual and audio clips are timely aligned with contrastive loss. when training the generator, we predict 5 consecutive frames \( I^t_{t+4} \) within one batch at each inference step and supervise the training with syncnet’s assistance. the objective is:

\[
\mathcal{L}_{sync} = - \frac{S_v(I^t_{t+4})^T \cdot S_a(a_{t+4})}{\|S_v(I^t_{t+4})\|_2 \|S_a(a_{t+4})\|_2} \tag{4}
\]

the overall loss functions across the generalized training can be written as:

\[
\mathcal{L}_g = \mathcal{L}_{adv} + \lambda_r \mathcal{L}_{rec} + \lambda_s \mathcal{L}_{sync}. \tag{5}
\]

3.3. personalized optimization

after the training of the backbone networks, our method can readily generate high-fidelity lip sync results for arbitrary subject. however, as the model is trained in a generalized manner, the generated mouth motion patterns across different people are basically the same. it has been verified that different identities possess different talking styles [50, 54]. here we pursue to capture such personalized property with the original template video. notably, we focus on the few-shot setting that only less than one minute of the original video is given, which cannot be handled by previous person-specific models [16, 19, 38, 44]. below, we illustrate how we successfully design our personalized optimization module based on the above discussions.

basic learning settings. the basic learning procedure of the personalized optimization is similar to backbone training. on a template video \( \mathbf{V} = \{I_1, \ldots, I_T\} \) with audio \( \mathbf{a} = \{a_1, \ldots, a_T\} \), we randomly select two frames as target \( I_i \) and reference \( I_{i+\Delta} \) for training as one sample. during training, the generalized loss \( \mathcal{L}_g \) continuously supervises the whole procedure.

as the encoder \( E_v \) encodes person-agnostic speech content information, it is fixed during the personalized optimization step. while the face encoder \( E_{face} \) encodes the attribute information of the target person, we do not optimize its parameters on a single person to avoid overfitting.

style space optimization. according to our formulation, the \( \mathbf{W}^+ \) space contains both the lip motion and the high-level facial style information. thus optimizing it would ideally lead to person-specific lip motion pattern. e4e [45] verifies that learning the small displacements around the original \( \mathbf{W} \) space would both extend the inversion quality of the image and enables strong editing ability. thus we propose to learn a set of \( \Delta \mathbf{W} = MLP_{\Delta \mathbf{W}}(f_j) \) in a similar way according to the encoded face feature \( f_j \) with a few layers of \( \text{MLPs} \) \( MLP_{\Delta \mathbf{W}} \).

generator tuning. the generator most accounts for the synthesizing and the blending quality. we allow the parameters \( \mathbf{P} \) of our generator \( G \) to shift a little margin to \( \Delta \mathbf{P} \) with the personalized data we use.

learning objectives. the final learning objectives of the personalized optimization procedure can be summarized as:

\[
\mathcal{L}_p = \mathcal{L}_g + \lambda_p \left( \sum |\Delta \mathbf{W}_j|_2^2 + \sum |\Delta \mathbf{P}_n|_2^2 \right), \tag{6}
\]

where \( \mathbf{W}_j \) denote the \( j \)th parameter of the \( i \)th element in \( \mathbf{W} \). this stands the same for \( \mathbf{P}_n \). we restrict the displacements in a limited step.

4. experiments

datasets. we conduct experiments on two commonly used audio-visual datasets, lrw [11] and voxceleb2 [10]. we follow the datasets’ original train/test split and train our model on a mixture of these two datasets.

- lrw [11] is an audio-visual dataset collected from bbc news for lip reading. it is one of the earliest large-scale audio-visual datasets with high quality. it consists of 1,000 one-second utterances in 500 words.
- voxceleb2 [10] voxceleb [30] and voxceleb2 are large-scale audio-visual datasets collected for the speaker verification task. we re-download around one-fifth of voxceleb2 for training, which are of high quality. the official test set is processed with gpenn [53] for evaluation.

implementation details. we process all videos at 25 fps and align all faces according to pre-detected landmarks at the eyes. all faces are cropped to the size of \( 256 \times 256 \). a same u-shape mask is adopted as shown in fig. 2 to erase the mouth, cheeks and jaws at the target frame. all audios are processed in the same way as wav2lip [32]. in order to keep our design simple and avoid extra hyperparameter tuning, we adopt most settings from the previous studies [23, 32, 53]. specifically, the generator has a total of \( L = 14 \) style-convolution layers. \( \lambda_r \) is empirically set to 10 and all other \( \lambda_s \) are selected to 1. for personalized optimization, we train the person’s data for 5 epochs. long training would not lead to better results. particularly, we find that training the personalized model together with the general training data leads to more stable results, the portion of personal data and general data is set as 1:1 in our experiments.
Comparing Methods. As our model is built based on the person-agnostic setting, we compare our StyleSync framework with three state-of-the-art methods including, MakeitTalk [60], PC-AVS [59] and Wav2Lip [32]. MakeitTalk [60] synthesizes talking head videos with natural head pose. PC-AVS [59] a pose source video to achieve pose-controllable talking face generation. While MakeitTalk and PC-AVS produce a whole talking head, Wav2Lip which focuses on the mouth part is our main competing method. As Wav2Lip is originally trained on a low-resolution with small networks, we carefully reproduce it on the same data and even loss functions as our methods. We denote this model as Wav2Lip-H (High-Quality). For fair comparison, all basic comparisons are carried out on the generalized setting Ours-G.

We also perform personalized optimization on the test set, where each frame lasts less than 10 seconds for both datasets. The results are denoted as Ours-P. We equally perform the generator finetuning on the above methods, and all of them perform worse than their original models, thus the results are not listed below. Additionally, we compare our personalized setting with ADNeRF [16], a recent talking head advance using neural radiance fields [29].

4.1. Quantitative Evaluation

As quantitative evaluations can only be performed on the self-construction setting, we avoid directly leveraging all information within the frame for reconstruction. We uniformly select a same random frame as the reference frame for PC-AVS, Wav2Lip, Wav2Lip-H and ours.

Evaluation Metrics. We follow previous studies [8, 9, 58, 59] to adopt the popularly used SSIM [49], PSNR metrics for evaluating the generation quality, and the landmark distances around the mouth (LMD) and SyncNet’s [12] confidence score to show the lip-sync quality. Please be noted that the SyncNet metric is computed on the officially released version, which is different from our implementation in the loss functions.

Particularly, as the details of the target person cannot be fully recovered, we leverage the ArcFace [13] network and compute the frame-by-frame feature cosine distances $D_{1D}$ for evaluating whether the generated results preserve the identity well.

Evaluation Results. The quantitative experiments are carried out on the test set of LRW [11] and VoxCeleb2 [10] datasets. Please refer to Table 1 for the results. It can be seen that on the LRW dataset, the generalized version
of our StyleSync already outperforms previous methods by a large margin for generation quality. While on the VoxCeleb2 dataset, the gap is less obvious. The reason might be that most training data of our methods are basically frontal view faces selected from LRW, while VoxCeleb2 contains more complicated scenes. On the other hand, after our personalized optimization, the performance of our model advances again. This shows the generated results are clearly more similar to the targets at this stage, which can also be verified by looking into the identity distances.

Meanwhile, our method also achieves comparable performance on the lip-sync metrics on both datasets. Our LMD score is slightly better than the competing methods. As for the SyncNet score, we achieve comparable results on LRW that are closer to the ground truth’s SyncNet score. We argue it is meaningless to refer to the SyncNet scores for verifying the lip-sync quality once the metric has outperformed the ground truth. The $\text{Sync}_{\text{conf}}$ only reflects how well an audio-visual pair fits the learned SyncNet model rather than the true perceptual quality. Thus though generated results might outperform ground truth on the metric, it does not mean better sync quality.

After the personalization, the lip-sync score degrades. We assume that the specificity of the speaking style reduces the mouth opening level. This is also shown in Fig. 5. The woman tends not to open her mouth wide, leading to smaller mouth movements.

### 4.2. Qualitative Evaluation

Subjective evaluation is crucial in identifying the ability of generative models, particularly on videos. We strongly recommend readers to watch our supplementary video at https://hangz-nju-cuhk.github.io/projects/StyleSync.

Two cross-driven examples and their comparisons with the SOTA methods are shown in Fig. 3. We use an conditional audio from an arbitrary person selected from the test set to drive the template. It can be seen that MakeitTalk [60] cannot produce accurate lip movements. Moreover, both PC-AVS and Wav2Lip produce visible artifacts or blurry results. Our reproduced Wav2Lip-H adopts the same training setting as ours, thus generates plausible results for most cases. Nevertheless, our generalized model still produces the most accurate mouth shapes with the highest fidelity.

Additionally, we show a case of personalized optimization on a 50-second video clip in Fig. 1. It can be seen that our method preserves the speaking style of the target person with accurate lip sync. Comparing our personalized results (Ours-P) with AD-NeRF [16] (Fig. 4), it is evident that our method maintains more person-specific details with higher fidelity, even though we use significantly less personal data (10s for tuning vs. 5min for training in [16]).

**User Study.** We also invite 15 participants to conduct a user study for further subjective evaluation and the results are reported in Table 2. Specifically, we randomly select 63 videos from LRW and VoxCeleb2 datasets as test sets and generate videos by using our generalized StyleSync and the comparison methods accordingly. By adopting the commonly used Mean Opinion Scores (MOS) rating protocol, we request all the participants to provide their ratings (from 1 to 5, the higher the better) on three aspects for each generated video: (1) Lip-Sync quality; (2) Generation quality; (3) Video realness.

As shown in Table 2, MakeitTalk [60] achieves the lowest scores in all the aspects due to lacking reasonable lip movements and head pose. PC-AVS [59], Wav2Lip [32] and Wav2Lip-H achieve relatively higher lip-sync score, but only Wav2Lip-H is able to synthesize videos with less blurry textures around the mouth. Overall, our StyleSync

<table>
<thead>
<tr>
<th>Method</th>
<th>LRW</th>
<th></th>
<th>VoxCeleb2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SSIM ↑</td>
<td>PSNR ↑</td>
<td>LMD ↓</td>
<td>$\text{Sync}_{\text{conf}}$ ↑</td>
</tr>
<tr>
<td>MakeitTalk</td>
<td>0.69</td>
<td>29.83</td>
<td>2.75</td>
<td>3.88</td>
</tr>
<tr>
<td>PC-AVS</td>
<td>0.79</td>
<td>30.26</td>
<td>1.84</td>
<td>7.19</td>
</tr>
<tr>
<td>Wav2Lip</td>
<td>0.79</td>
<td>30.54</td>
<td>1.28</td>
<td>7.39</td>
</tr>
<tr>
<td>Wav2Lip-H</td>
<td>0.80</td>
<td>31.38</td>
<td>1.20</td>
<td>7.19</td>
</tr>
<tr>
<td>GT</td>
<td>1.00</td>
<td>100.00</td>
<td>0.00</td>
<td>7.65</td>
</tr>
<tr>
<td>Ours-G</td>
<td>0.85</td>
<td>31.78</td>
<td>1.18</td>
<td>7.25</td>
</tr>
<tr>
<td>Ours-P</td>
<td>0.88</td>
<td>32.66</td>
<td>0.86</td>
<td>6.35</td>
</tr>
</tbody>
</table>

Table 1. Quantitative results on LRW and VoxCeleb. For LMD the lower the better, and the higher the better for other metrics.
and in-kind contribution from the industry partner(s).

Acknowledgement. This study is supported by the Ministry of Education, Singapore, under its MOE AcRF Tier 2 (MOE-T2EP20221-0012), NTU NAP, and under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

Table 2. User study measured by Mean Opinion Scores. The scores are ranged from 1 (worst) to 5 (best).

<table>
<thead>
<tr>
<th>Approach</th>
<th>MOS</th>
<th>Lip-Sync Quality</th>
<th>Generation Quality</th>
<th>Video Realness</th>
</tr>
</thead>
<tbody>
<tr>
<td>MakeItTalk</td>
<td>2.06</td>
<td>3.00</td>
<td>3.49</td>
<td>3.67</td>
</tr>
<tr>
<td>PC-AVS</td>
<td>2.63</td>
<td>2.16</td>
<td>1.87</td>
<td>3.42</td>
</tr>
<tr>
<td>Wav2Lip</td>
<td>1.89</td>
<td>2.16</td>
<td>2.17</td>
<td>2.98</td>
</tr>
<tr>
<td>Ours-G</td>
<td>0.82</td>
<td>31.54</td>
<td>1.15</td>
<td>7.31</td>
</tr>
<tr>
<td>Ours-P</td>
<td>0.80</td>
<td>31.34</td>
<td>1.55</td>
<td>7.41</td>
</tr>
</tbody>
</table>

Table 3. Quantitative ablations results on VoxCeleb2.

<table>
<thead>
<tr>
<th>Method</th>
<th>SSIM ↑</th>
<th>PSNR ↑</th>
<th>LMD ↓</th>
<th>Syncconf ↑</th>
<th>LID ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o mask</td>
<td>0.81</td>
<td>31.20</td>
<td>1.80</td>
<td>7.89</td>
<td>0.90</td>
</tr>
<tr>
<td>w/o sync</td>
<td>0.80</td>
<td>31.34</td>
<td>1.55</td>
<td>7.41</td>
<td>0.90</td>
</tr>
<tr>
<td>Ours-G</td>
<td>0.79</td>
<td>31.00</td>
<td>1.47</td>
<td>8.25</td>
<td>0.90</td>
</tr>
<tr>
<td>P w/o (\Delta W)</td>
<td>0.82</td>
<td>31.51</td>
<td>1.32</td>
<td>7.31</td>
<td>0.92</td>
</tr>
<tr>
<td>Ours-P</td>
<td>0.82</td>
<td>31.54</td>
<td>1.15</td>
<td>7.26</td>
<td>0.93</td>
</tr>
</tbody>
</table>

Figure 5. Ablation study with visual results. Zoom in for details.

4.3. Ablation Study

To further demonstrate the contributions of our novel designs, we perform an ablation study on VoxCeleb2 dataset under both generalized setting and personalized optimization setting (denoted as “Ours-G” and “Ours-P”, respectively). Concretely, we construct two variants for “Ours-G” (denoted as “w/o mask” and “w/o sync”) by removing the masking operation in Eq. 1 and the lip-sync loss during training, respectively. While for “Ours-P”, we form one variant (denoted as “Ours-P w/o \(\Delta W\)” by setting \(\Delta W = 0\)). We have also experimented personalized optimization with fixed generated, however, this would lead to blurry results. This ablation is omitted here.

The quantitative and qualitative results are shown in Table 3 and Fig. 5. As shown in Table 3, we observe that both “w/o mask” and “w/o sync” achieve comparable performance on the image quality metrics but leads to a performance drop on the lip-sync metrics. “w/o mask” achieves worse scores when compared with “Ours-G” in terms of LMD and Sync, which indicates that the proposed masking strategy can obviously alleviate the influence from the reference frames. While “w/o sync” reasonably suffers from lip-sync degradation and the results in Fig. 5 illustrate inconsistent lip movements when compared with the lip-synced video. The results demonstrate the effectiveness of the additional supervision from SyncNet.

It can be seen from Fig. 5 that with the personalized optimization, both the identity and the pattern of the mouth opening on the generated frames become more similar to the original template video. It is also clear that after the personalized training, the mouth opening is less obvious than in the generalized model. We analyze that this is part of the talking style of this target person. This also leads to poorer metric values on SyncNet.

In terms of the comparisons between our personalized model and “Ours-P w/o \(\Delta W\)”, it achieves similar scores on the image quality as well as lip-sync quality. However, the identity score without \(\Delta W\) is slightly degraded, which indicates that it is beneficial to involve \(\Delta W\) into the training procedure for better preserving identity.

5. Conclusion and Discussions

Conclusion. In this paper, we propose StyleSync, which produces high-fidelity lip sync results for both the one-shot and the few-shot settings. We highlight the unique proprieties of our method: 1) Video results produced by our generalized model clearly outperform previous state-of-the-arts. 2) Our model is built upon the success of recent style-based generators with simple modification. It is easy to implement and friendly to train. 3) By involving StyleGAN-inspired personalized optimization procedure, our model can further be improved on a specific person given only a few clips.

Limitations. As our method blends a lip-synced face into an existing video with a fixed mask, the head pose and expressions of the target person cannot be changed. Additionally, under certain extreme cases where the target’s jaw is extremely large, it might get out of our masked area.

Ethical Considerations. Our method could be used to create non-existing talks and speeches, which might be maliciously used. We will issue our core models strictly to research institutions.
References


[31] Se Jin Park, Minski Kim, Joanne Hong, Jeongsoo Choi, and Yong Man Ro. Synctalkface: Talking face generation with precise lip-syncing via audio-lip memory. In AAAI Conference on Artificial Intelligence. Association for the Advancement of Artificial Intelligence, 2022. 2, 3


[34] Elad Richardson, Michael Zollhöfer, Yifeng Ma, Suzhen Wang, Linchong Li, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. ACM Transactions on Graphics (TOG), 2021. 2, 3, 5


[40] Yasheng Sun, Hang Zhou, Ziwei Liu, and Hideki Koike. Speech2talking-face: Inferring and driving a face with synchronized audio-visual representation. In IJCAI, 2021. 2, 4


[54] Ran Yi, Zipeng Ye, Juyong Zhang, Hujun Bao, and Yong-Jin Liu. Audio-driven talking face video generation with

[55] Fei Yin, Yong Zhang, Xiaodong Cun, Mingdeng Cao, Yanbo Fan, Xuan Wang, Qingyan Bai, Baoyuan Wu, Jue Wang, and Yujiu Yang. Styleheat: One-shot high-resolution editable talking face generation via pre-trained stylegan. ECCV, 2022. 2


