

SegTalker: Segmentation-based Talking Face Generation with Mask-guided Local Editing

Supplementary Material

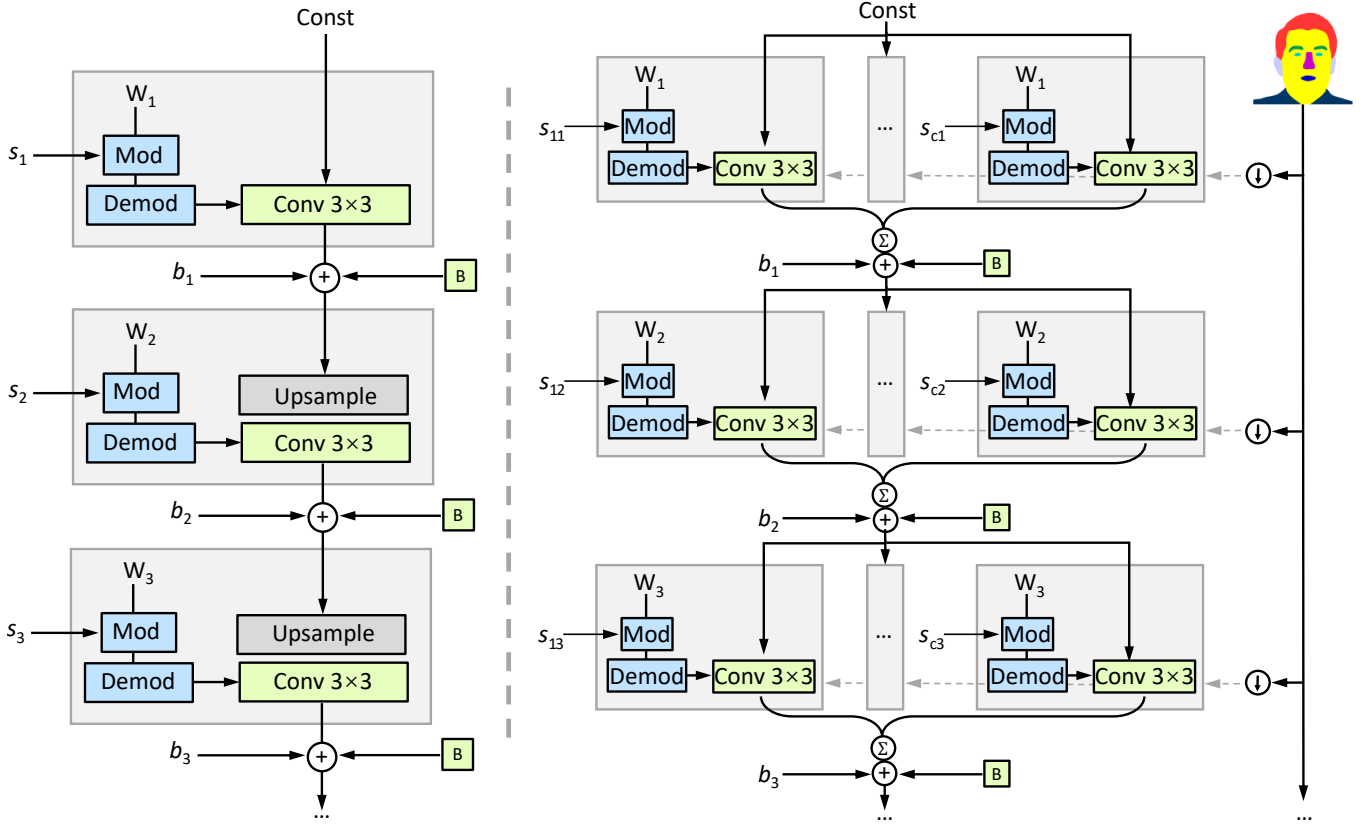


Figure 1: Comparison of the original StyleGAN and the employed mask-guided StyleGAN. Left: Original StyleGAN; Right: Mask-guided StyleGAN.

A MASK-GUIDED GENERATOR DETAILS

We utilize a mask-guided generator to integrate style codes with segmentation for synthesizing video frames, which are illustrated in fig. 1. The original StyleGAN start from a constant feature with a spatial size of 4×4 and consist of a series of style blocks. Each block contains a modulation, a demodulation and a 3×3 convolution layer. a noise broadcast operation B is introduced to improve the diversity of generative images. W and b in each block are denoted as learnable weights. A additional upsampling layer with a factor of 2 is employed between two blocks to increase the resolution of feature maps.

Unlike the global semantic control through style codes in the original StyleGAN, we extract localized style codes corresponding to different semantic sub-regions via masks, thereby enabling localized control. We extend the style block of original styleGAN into a mask-guided style block according to a mask. In particular, we aggregate the intermediate feature maps along with per-region

mask, which are formulated as:

$$F_l = \sum_{j=1}^C (F_{l-1} * W'_{jl}) \circ (\text{Down}(M)_i == j), \{l = 1, 2, \dots, C\} \quad (1)$$

$$W'_{jl} = \text{Demod}(\text{Mod}(W_l, s_{jl})) \quad (2)$$

where F_l and F_{l-1} denote the feature map of layer l and layer $l - 1$, respectively. W'_{jl} represent the scaled kernel weight for the j -th semantic region in the l -th layers. $\text{Down}(\dots)$ function down-samples the mask to align with the input feature map. We follow the same modulation and demodulation as described in the original StyleGAN. In eq. (2), W_l denotes the original kernel weight for l -th layer and s_{jl} denotes the style codes of j -th semantic regions for l -th layer.

B MORE RESULTS

B.1 Talking Segmentation

We present more qualitative audio-driven talking segmentation results, which are shown in fig. 2. Examination results show the generated segmentations effectively delineate distinct facial regions, even elaborating details such as earrings. Furthermore, the generated lip motions exhibit robust synchronization with the reference video.

B.2 Talking Face

We present more qualitative results with state-of-the-art talking face methods: SadTalker, Wav2lip and VideoReTalking, where the results are illustrated in fig. 3. The generated frames of SadTalker exhibit poor visual quality and can not handle the scenarios of heavily variant pose move. The synthesized faces of Wav2lip exhibit worse detail in the lip and teeth regions. Although VideoReTalking can yield visually gratifying results, it exists identity drift and contains artifacts in local regions. Qualitative results demonstrate that our method produces more realistic and high-fidelity results while maintaining rich facial textures and identity details.

B.3 Additional Facial Editing

More facial editing cases can be found in fig. 4. In fig. 4, by simply editing the eye regions of mask, we can manipulate blinking in a controllable manner to synthesize realistic talking face videos.

B.4 Additional Swapping Background

Additional swapping background results are shown in fig. 5. Our model intrinsically disentangles the foreground and background, allowing for seamless background swapping and augmenting the application scenarios of talking faces.

C LIMITATIONS AND DISCUSSIONS

Although our method generates realistic and high-fidelity video which enable facial editing from a audio and a video, there still have some limitations in our framework. Since our approach employ facial editing by simply manipulate mask, currently only some simple attribute such as blinking, eyebrows can be achieved. Future research will concentrate efforts on exploring textual guidance methods to enable global and local attribute editing capabilities for talking face videos.

However, synthesized videos have the potential to be exploited for spreading misinformation, defrauding the public, and infringing on personal privacy. To mitigate this threat, researchers have proposed various techniques such as digital watermarking technology for detecting synthetic media. Future research needs to continue strengthening ethical and legal norms to ensure the healthy development and application of talking face technologies.

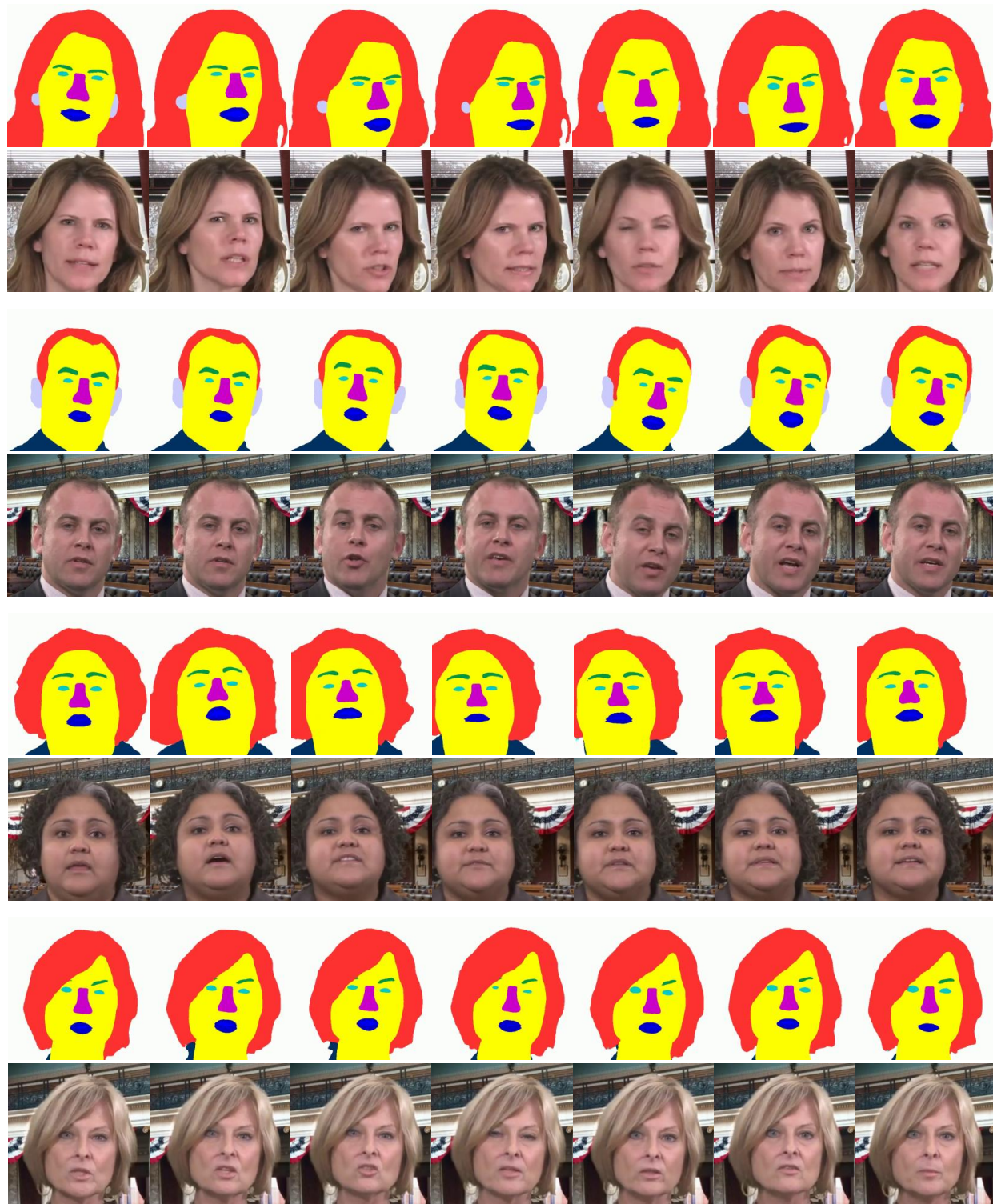


Figure 2: Visualization of synthesized segmentation(row 1, row 3, row5, row7) and real images(row 2, row 4, row6, row8). It can be seen that TSG can produce lip synchronized segmentation.

SegTalker

SadTalker

Wav2lip

Video
ReTalking

...

Figure 3: Additional qualitative comparisons of our results with several state of the art methods for talking face synthesis. In each block, our method is illustrated in the first row and synthesized images of different method (SadTalker, Wav2lip, VideoReTalking) follow the next.



Figure 4: Additional qualitative results of facial editing. Our method produces more high-fidelity results in editing regions while maintaining the details and identity information of other regions.

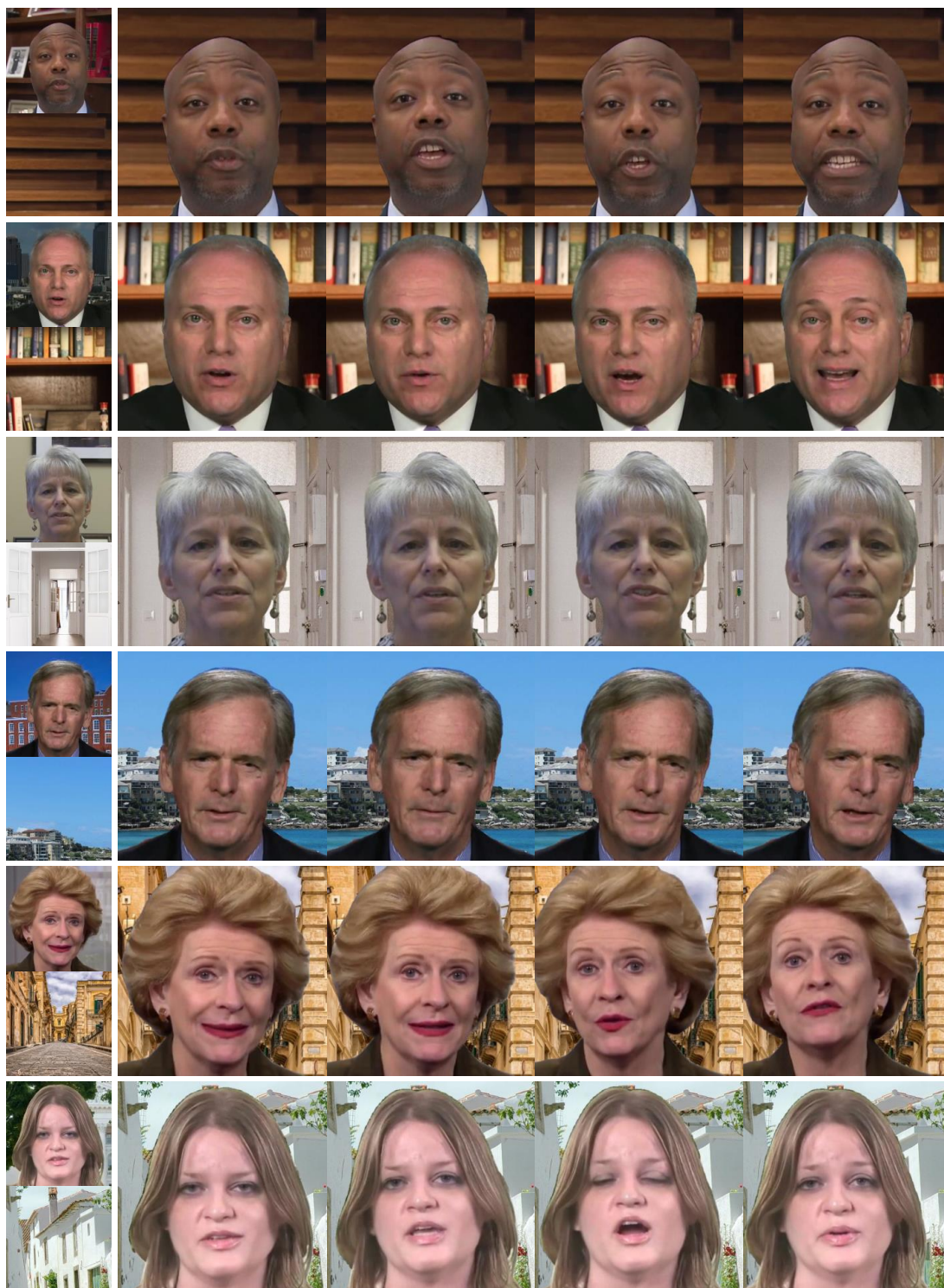


Figure 5: Additional example of Swapping background. Given a video and a background image, our method can produce natural and photo-realistic swapping videos.