ToonifyGB: StyleGAN-based Gaussian Blendshapes for 3D Stylized Head Avatars

Rui-Yang Ju*
National Taiwan University

Sheng-Yen Huang†
National Taiwan University

Yi-Ping Hung‡
National Taiwan University

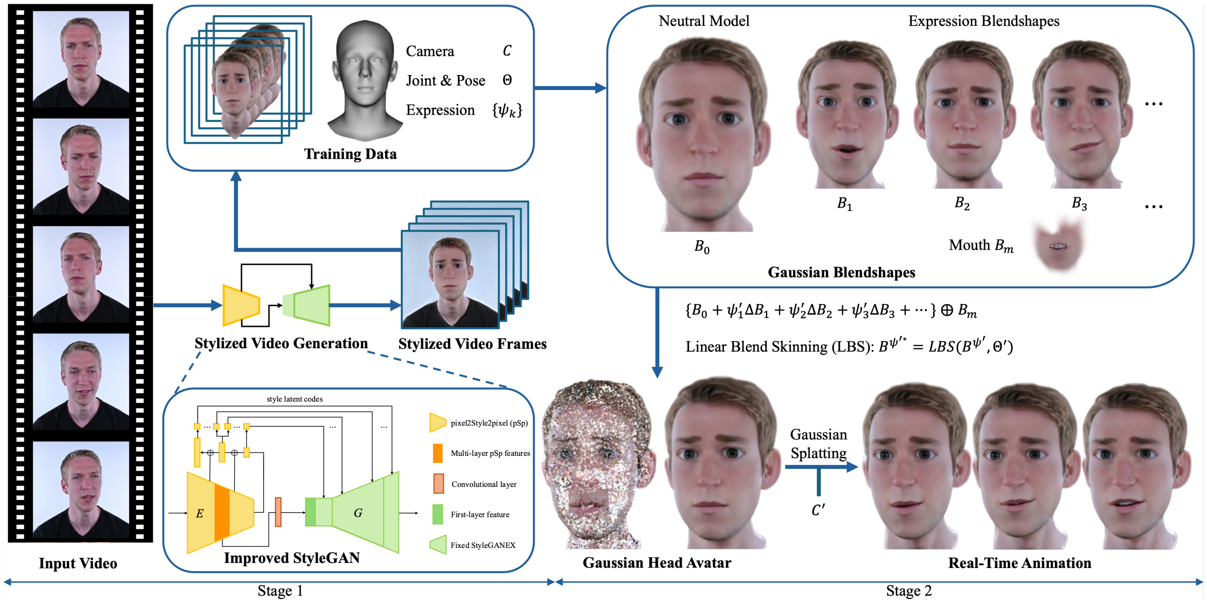

Figure 1: **ToonifyGB**: We propose an efficient two-stage framework that adopts an improved StyleGAN to generate stylized head videos from input video frames (Stage 1), and synthesizes the corresponding 3D avatars using Gaussian blendshapes (Stage 2).

Abstract

Recent advances in 3D Gaussian blendshapes have enabled real-time reconstruction of animatable, photo-realistic head avatars from monocular videos. Toonify, a StyleGAN-based method, is widely used for facial image stylization. To extend Toonify to the synthesis of diverse stylized 3D head avatars using Gaussian blendshapes, we propose an efficient two-stage framework, ToonifyGB, for applications in VR/AR environments. In Stage 1, our method generates stylized videos, and in Stage 2, it synthesizes the corresponding 3D stylized head avatars using Gaussian blendshapes. We evaluate ToonifyGB on the INSTA dataset with two representative styles: Arcane and Pixar. The implementation code is available at `https://ruiyangju.github.io/ToonifyGB`.

Index Terms: Toonify, StyleGAN, Gaussian Splatting, Gaussian Blendshapes, 3D Stylized Head Avatar, Facial Animation.

1 Introduction

Recent advances in 3D head reconstruction have enabled people to appear in virtual and augmented reality (VR/AR) environments using photo-realistic head avatars for telepresence and social interaction. In these scenarios, users often prefer stylized avatars that protect their privacy, match the aesthetics of virtual worlds, and still preserve personal identity and expressiveness.

Toonify [3], a StyleGAN [1]-based facial stylization method, shows the potential to translate real portraits into stylized images. However, existing Toonify-based methods mainly focus on 2D image editing, and do not provide 3D stylized head avatars suitable for use in VR/AR applications. In addition, recent methods in 3D head reconstruction have mainly targeted photo-realistic avatars. In particular, 3D Gaussian Blendshapes (3DGB) [2] integrate blendshape modeling with Gaussian splatting, achieving real-time rendering and state-of-the-art (SOTA) performance in head reconstruction from monocular videos. While such methods are attractive for telepresence, they do not support stylized 3D avatars that match the artistic styles commonly used in VR/AR environments.

To bridge this gap, we propose ToonifyGB, an efficient two-stage framework that synthesizes and animates 3D stylized head avatars from monocular videos by combining StyleGAN-based video stylization with 3D Gaussian blendshapes.

2 Method

Given a monocular input video, ToonifyGB performs frame-by-frame stylization to generate the corresponding stylized frames. Conventional methods typically require face alignment, cropping, and editing before compositing the stylized results back into the original frames. This process often introduces visual discontinuities at the seams, resulting in noticeable jitter in the output video. To address this issue, we adopt an improved StyleGAN model based on StyleGANEX [4] in Stage 1, enabling stable stylized video generation at the original resolution, as shown in Fig. 1.

To prepare the training data for Stage 2, we follow 3DGB [2] and use the facial tracker to compute FLAME meshes (50k Gaussians for the neutral model and 14k for the mouth interior), which include a neutral head model and a set of expression blendshapes. This process provides training camera parameters C, joint and pose parameters Θ, and expression coefficients $\{\psi_k\}$ for each frame. In addition to enabling facial expression control, the FLAME model based on Principal Component Analysis (PCA) provides joint and pose parameters for controlling head, eyeball, eyelid, and jaw movements.

*e-mail: jryjry1094791442@gmail.com
†e-mail: d12944001@csie.ntu.edu.tw
‡e-mail: hung@csie.ntu.edu.tw

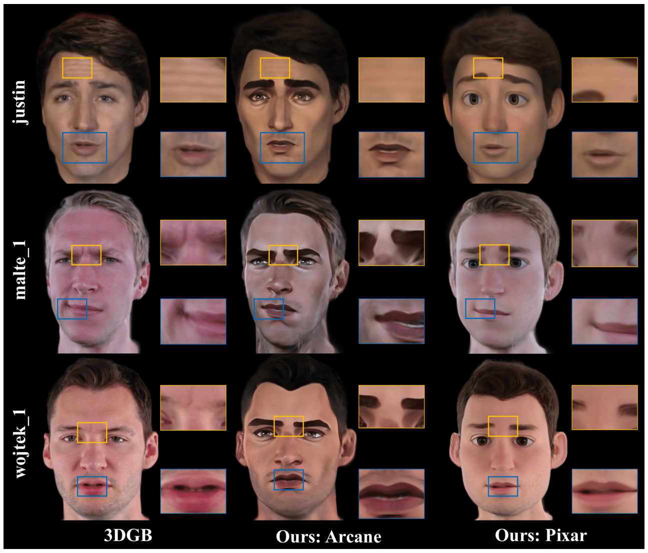

Figure 2: **Qualitative Comparison:** We present 3D head avatars synthesized by the baseline method 3DGB [2] and our method.

Table 1: **Time Comparison:** We record the training time (in minutes) and rendering speed (in fps) of different methods on GPU RTX 4090.

Method	Training Time			Rendering
	Stage 1	Stage 2	Total	Speed
3DGB [2]	–	44.5	44.5	136.8
Ours (Arcane)	4.0	44.7	48.7	136.0
Ours (Pixar)	4.1	43.2	47.3	132.5

We then apply Linear Blend Skinning (LBS) to transform the Gaussian model using the extracted joint and pose parameters. The transformed Gaussian model is rendered in real-time as a 3D stylized head avatar via Gaussian Splatting. Finally, by integrating the rendering camera parameters C', we enable novel-view 3D stylized head avatar rendering and animation.

3 Experiments

To evaluate the effectiveness of ToonifyGB, we compare it with the SOTA baseline [2] on six videos from the INSTA dataset [5].

Training Time and Rendering Speed: We measure the training time (in minutes) and rendering speed (in fps) of both our method and 3DGB [2] for animatable head reconstruction. The comparison results are summarized in Tab. 1. Although our method integrates stylization into 3D head avatars (i.e., additional training time in Stage 1), the total training time remains comparable to that of 3DGB [2]. Furthermore, the rendering speed of ToonifyGB for both stylization styles is similar to that of the baseline [2].

Qualitative Comparison: A qualitative comparison with 3DGB [2] is presented in Fig. 2. Our method effectively captures and preserves high-frequency details in the stylized videos, including forehead wrinkles and fine mouth-region deformations. Compared to the SOTA baseline method, ToonifyGB can synthesize 3D stylized head avatars with comparable quality and detail.

Ablation Study on Source Videos for Driving Animation: To demonstrate the importance of the generated stylized video in driving the animation, we compare animations driven by the original input video (real face) and by our generated stylized video, as shown in Fig. 3. It can be observed that using the original input video (real face) as the driving source often leads to unsatisfactory results, especially around the mouth region. This error occurs due to significant differences in expression blendshapes between the real and stylized domains. These results highlight the importance of the stylized driving videos generated in Stage 1 of our framework.

Ablation Study on Face Alignment and Cropping: We further compare stylized 3D head avatars (using the “justin” data) synthesized from videos processed by our method against those generated from videos preprocessed with face alignment and cropping. The resulting avatars are evaluated using the standard metrics used in 3DGB [2]. As shown in Tab. 2, our method outperforms the conventional method with face alignment and cropping across all metrics. This demonstrates that our method, based on the 3DGB framework, effectively eliminates jitter during video generation, enabling higher-quality synthesis of 3D stylized head avatar animations.

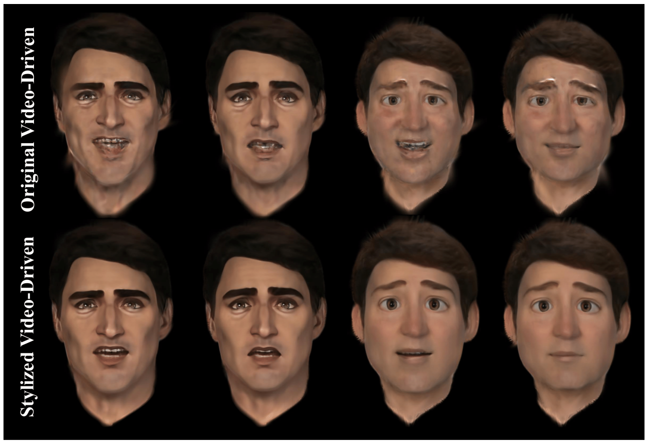

Figure 3: **Ablation Study:** We present 3D stylized head avatar animation driven by the original input videos and our generated videos.

Table 2: **Ablation Study:** We compare 3D head avatars synthesized from different input videos, where one is generated by our method and the other uses face alignment and cropping as preprocessing.

Method	PSNR↑	SSIM↑	LPIPS↓
Face Align & Crop	32.23	0.9387	0.1587
Ours	33.27	0.9645	0.0796

4 Conclusion

We presented ToonifyGB, an efficient two-stage framework that synthesizes and animates stylized 3D head avatars from monocular videos by combining StyleGAN-based video stylization with 3D Gaussian blendshapes. Our method produces stable stylized videos, learns expressive 3D blendshape models, and renders high-quality stylized avatars in real-time. Beyond technical improvements, ToonifyGB provides a practical solution for creating identity-preserving, stylized avatars that align with the aesthetics and privacy requirements of VR/AR environments. As ToonifyGB supports real-time 3D stylized head avatar rendering and animation, it can be integrated into VR/AR telepresence, social VR platforms, virtual characters, and immersive content creation systems.

Acknowledgments

This work was supported in part by the National Science and Technology Council (NSTC), Taiwan, under Grants NSTC 114-2221-E-002-001- and NSTC 114-2420-H-002-010-.

References

[1] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. *CVPR*, 2019. 1

[2] S. Ma, Y. Weng, T. Shao, and K. Zhou. 3D gaussian blendshapes for head avatar animation. *ACM SIGGRAPH*, 2024. 1, 2

[3] J. N. Pinkney and D. Adler. Resolution dependent gan interpolation for controllable image synthesis between domains. *NeurIPS WS*, 2020. 1

[4] S. Yang, L. Jiang, Z. Liu, and C. C. Loy. Styleganex: Stylegan-based manipulation beyond cropped aligned faces. *ICCV*, 2023. 1

[5] W. Zielonka, T. Bolkart, and J. Thies. Instant volumetric head avatars. *CVPR*, 2023. 2