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

Rui-Yang Ju[1] Sheng-Yen Huang[1] Yi-Ping Hung[1]
[1]Graduate Institute of Networking and Multimedia, National Taiwan University, Taiwan
jryjry1094791442@gmail.com, d12944001@csie.ntu.edu.tw, hung@csie.ntu.edu.tw
https://ruiyangju.github.io/ToonifyGB

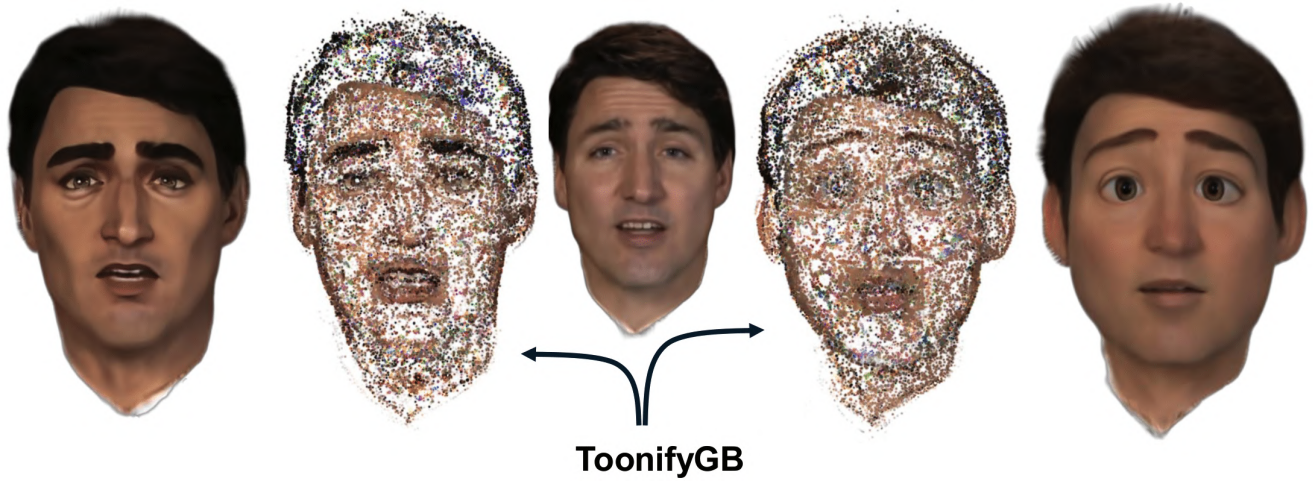

Figure 1. **ToonifyGB**: We propose an efficient two-stage framework that employs an improved StyleGAN to generate stylized head videos from input video frames and synthesize the corresponding 3D avatars using Gaussian blendshapes. Our method supports real-time synthesis of stylized avatar animations (with 50k Gaussians for the neutral model and 14k Gaussians for the mouth interior) in diverse styles such as Arcane and Pixar.

Abstract

The introduction of 3D Gaussian blendshapes has enabled the real-time reconstruction of animatable head avatars from monocular video. Toonify, a StyleGAN-based method, has become widely used for facial image stylization. To extend Toonify for synthesizing diverse stylized 3D head avatars using Gaussian blendshapes, we propose an efficient two-stage framework, ToonifyGB. In Stage 1 (stylized video generation), we adopt an improved StyleGAN to generate the stylized video from the input video frames, which overcomes the limitation of cropping aligned faces at a fixed resolution as preprocessing for normal StyleGAN. This process provides a more stable stylized video, which enables Gaussian blendshapes to better capture the high-frequency details of the video frames, facilitating the synthesis of high-quality animations in the next stage. In Stage 2 (Gaussian blendshapes synthesis), our method learns a stylized neutral head model and a set of expression blendshapes from the generated stylized video. By combining the neutral head model with expression blendshapes, ToonifyGB can efficiently render stylized avatars with arbitrary expressions. We validate the effectiveness of ToonifyGB on benchmark datasets using two representative styles: Arcane and Pixar.

1. Introduction

With the advancement of 3D head reconstruction technologies, individuals can now personalize unique avatars for telepresence and virtual/augmented reality applications, which serve as a crucial foundation for the rise of the metaverse. Considering user preferences and privacy concerns, the creation of stylized avatars has become an important research topic. Toonify [36], a StyleGAN-based method, was designed for 2D facial image stylization, presenting

the potential of translating real portraits into stylized 2D images. While such methods focus on 2D images, recent advances in 3D head reconstruction have mainly targeted photo-realistic avatars. In contrast, stylized 3D head avatars emphasize personal identity and the faithful transfer of artistic styles.

Blendshapes are an efficient facial animation representation that synthesize continuous and high-quality expressions by blending a set of 3D meshes, each corresponding to a specific facial expression. These facial shapes are synthesized by linearly blending the basis meshes using weighting coefficients. With the introduction of Neural Radiance Fields (NeRF) [31], Gao *et al.* [11] and Zheng *et al.* [57] incorporated the blendshape concept into NeRF, enabling avatar animation through a group of NeRF blendshapes that are linearly blended. Furthermore, the recently proposed 3D Gaussian Splatting (3DGS) [22] significantly improved rendering efficiency and achieved higher-quality head reconstruction, outperforming NeRF in both speed and quality. Building on this, 3D Gaussian Blendshapes (3DGB) [29] successfully integrated blendshapes with Gaussian splatting, achieving real-time rendering and state-of-the-art performance in head reconstruction.

In contrast to previous works focused on photo-realistic 3D head avatar reconstruction, we propose ToonifyGB, a two-stage framework for synthesizing and animating 3D stylized head avatars. Given monocular video frames, Stage 1 adopts an improved StyleGAN to generate a more stable and less jittery stylized video, without requiring fixed resolution or pre-aligned face cropping. In Stage 2, we build upon 3DGB to learn a neutral head model and a set of expression blendshapes, each represented as 3D Gaussians. Finally, by incorporating a facial tracker [59], ToonifyGB uses the tracked motion parameters to animate 3D stylized head avatars.

The contributions of this work are as follows:

- We propose ToonifyGB, an efficient two-stage framework that synthesizes 3D stylized head avatars from monocular videos using Gaussian blendshapes, supporting diverse styles with real-time animation.
- We demonstrate that reducing per-frame jitter in the generated video enables Gaussian blendshapes to better capture high-frequency details, thereby improving the quality of 3D stylized head avatar animations.
- To the best of our knowledge, this work is the first to synthesize 3D stylized head avatars using Gaussian blendshapes.

2. Related Work

2.1. StyleGAN and Toonify

StyleGAN [18, 19] has been widely used to generate realistic facial images across diverse styles. Inversion of StyleGAN enables projecting real facial images into its latent space, allowing subsequent edits such as adding glasses or changing hairstyles or age [1, 35]. To enhance inversion efficiency, methods such as pSp [38] and e4e [40] employ encoders to directly project target faces into their corresponding latent codes. However, these methods often struggle to reconstruct fine image details, resulting in unsatisfactory reconstruction quality. To address these limitations, ReStyle [4] and HFGI [42] improve reconstruction fidelity by respectively predicting latent code residuals and correcting intermediate features. Nevertheless, these methods remain limited to aligned and cropped facial images for effective editing and reconstruction.

Recently, researchers [10, 15, 33, 36, 49] have explored the use of StyleGAN for target-domain image generation through transfer learning. Among these works, Toonify [36] fine-tunes the trained generator to blend realistic textures with toonified facial structures. In addition to image editing, StyleGAN has also been widely applied to video editing. Related studies have focused on enhancing video editing performance by employing temporal correlations in low-dimensional latent codes [9], disentangling identity from facial attributes [52], incorporating sketch-based branches [28], and tuning the generator to maintain temporal consistency [41]. However, these methods typically rely on face alignment and cropping as preprocessing. Although StyleGAN3 [20] was introduced to support unaligned face inputs, a subsequent study [5] has shown that it struggles to encode facial features effectively without preprocessing, often resulting in structural artifacts. To overcome these limitations, methods such as VToonify [50] and StyleGANEX [51] have been proposed to directly process videos beyond pre-aligned cropping. Nevertheless, these methods remain limited to 2D representations and have yet to be extended to 3D applications.

2.2. 3D Head Avatar

Since the introduction of NeRF [31], implicit representation-based methods [7, 13, 47, 53, 57] for head reconstruction have achieved remarkable progress. 3DGS [22] has obtained a significant breakthrough in 3D reconstruction, further advancing the development of downstream applications such as 3D head modeling. Although several Gaussian-based methods [3, 8, 24, 29, 37, 45, 48] have demonstrated high-quality head reconstruction and impressive rendering performance, they typically focus on photo-realistic avatars, with relatively limited exploration of avatar stylization. Stylized head avatars, characterized by geometric abstraction and artistic expression, differ significantly from the photo-realistic avatars synthesized by the aforementioned methods.

Pre-trained 3D GANs [44] enable high-quality generation, making 3D head stylization possible. Although fine-

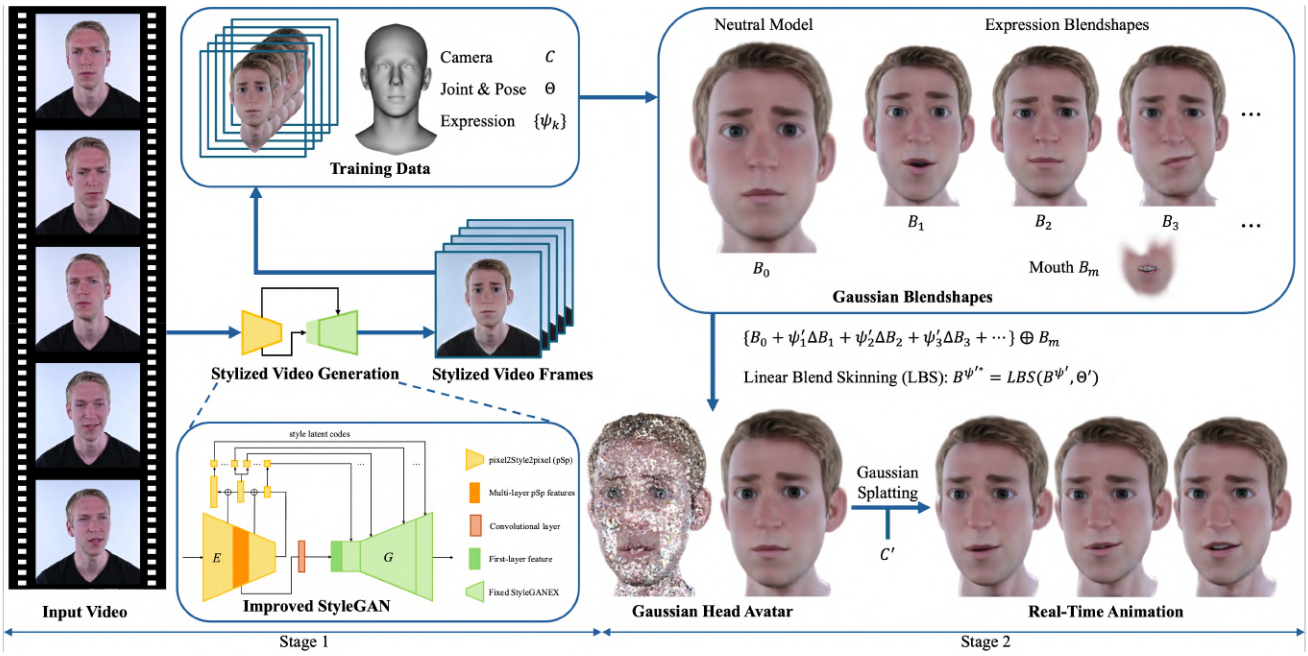

Figure 2. **Pipeline:** Our ToonifyGB framework consists of two stages: Stage 1 involves the generation of stylized videos, and Stage 2 focuses on the synthesis of 3D stylized head avatars using Gaussian blendshapes.

tuning 3D generators for geometric and texture-based stylization has proven effective [2, 17, 25, 34, 43, 54], performing independent fine-tuning for each new style remains costly. Toonify3D [16] addressed this limitation by predicting facial surface normals using the proposed StyleNormal, enabling direct face stylization without additional fine-tuning. Similarly, DeformToon3D [55] introduced StyleField to predict conditional 3D deformations, aligning NeRF representations in real space with style space to achieve geometric stylization and obviate per-style fine-tuning. However, Toonify3D suffers from limited data diversity, and DeformToon3D cannot support novel-view animations, which limits their application scenarios.

3. Method

3.1. ToonifyGB Framework

Given a monocular video input, ToonifyGB applies frame-by-frame stylization to generate the corresponding stylized frames. For inputs such as live streams or selfie videos, the face often occupies only a small portion of each frame, while the rest includes the hairstyle and upper body. Traditional methods [18, 19] typically require face alignment, cropping, and editing before synthesizing the results back into the original frame. 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 [51] in Stage 1, enabling stable stylized video generation at the original resolution, as shown in Figure 2.

To prepare the training data for Stage 2, we follow the method in [29, 60], using the facial tracker from [59] to compute FLAME [26] meshes, including a neutral head model and a set of expression blendshapes. This process also provides camera parameters C, joint and pose parameters Θ, and expression coefficients $\{\psi_k\}$ for each frame. In addition to enabling facial expressions control, the FLAME model based on Principal Component Analysis (PCA) provides joint and pose parameters for controlling head, eyeball, eyelid, and jaw movements. As shown in Figure 2, we apply Linear Blend Skinning (LBS) to transform the Gaussian model based on the extracted joint and pose parameters. The transformation is defined as:

$$B^{\psi *} = LBS(B^{\psi}, \Theta). \tag{1}$$

The transformed Gaussian model is then rendered in real-time as a 3D stylized head avatar using Gaussian Splatting. Finally, by integrating the camera parameters, we enable novel-view rendering and animation.

3.2. Stylized Video Generation

As shown in StyleGANEX [51], manipulating feature maps at different layers of StyleGAN leads to different spatial effects in the generated faces. Specifically, while shifting or rotating the feature maps in deeper layers (i.e., Layer 7) produces consistent global transformations, similar operations in shallow layers (i.e., Layer 1) fail to preserve facial structure due to the low spatial resolution of the 4×4 feature

Figure 3. **Visualization of stylized video generation results** in "Arcane" and "Pixar" styles on the INSTA [60] and NeRFBlendShape [11] datasets, covering both male and female subjects.

map, causing blurring and loss of detail. To address this limitation, we adopt StyleGANEX [51], an enhanced variant of StyleGAN2 [18], which increases the spatial resolution of shallow feature maps (Layers 1–7) to 32×32. This improvement enables finer control over facial geometry and enhances the generation quality for unaligned faces.

Our specific architectural improvements of the generator are as follows. First, we replace the constant 4×4 input of the first layer with a variable feature map of resolution $1/32$ of the final output, enabling support for arbitrary input sizes. Then, we replace the standard convolutions in the shallow layers with dilated convolutions to enlarge the receptive field. Finally, we remove all upsample operations before the eighth layer, ensuring that the seven shallow layers maintain the same 32×32 resolution.

These architectural improvements effectively address the limitations beyond pre-aligned cropping. As shown in Figure 3, our method consistently generates high-quality stylized head videos across diverse styles, regardless of gender.

3.3. Gaussian Blendshapes Synthesis

We represent all Gaussian head avatars using 3D Gaussians. As described in 3DGS [22], each Gaussian has some basic properties including Gaussian center μ, scale s, color c, opacity α, and rotation q. Based on 3DGB [29], our Gaussian blendshape representation consists of a neutral model B_0 and a set of n expression blendshapes $B_1, B_2, \ldots, B_n$. Each Gaussian in the neutral model B_0 has a set of blend weights w to control joint and pose. In addition, each Gaussian in an expression blendshape B_k corresponds one-to-one to a Gaussian in the neutral model B_0. The difference between B_k and B_0 is defined as the difference in their corresponding Gaussian properties: $\Delta B_k = B_k - B_0$. The expression of Gaussian head avatar B^ψ can be computed as follows:

$$B^\psi = B_0 + \sum_{k=1}^{n} \psi_k \, \Delta B_k \tag{2}$$

where ψ_k denotes the expression coefficients. Here, B^ψ represents the untransformed expression model, and the final posed Gaussian model, obtained via Linear Blend Skinning (LBS), is defined in Equation 1.

Since the FLAME meshes and blendshape models do not include interior mouth components such as teeth, we adopt the method of 3DGB [29] by defining a separate set of Gaussians for the mouth B_m. The properties of these mouth Gaussians are not affected by expression changes, they only move with the jaw joint in the FLAME model. The mouth Gaussians for the head avatar, B_m^*, are computed via linear blend skinning (LBS) as:

$$B_m^* = LBS(B_m, \Theta). \tag{3}$$

The transformed Gaussian model $(B^{\psi*}, B_m^*)$ is rendered into a complete 3D head avatar using real-time Gaussian Splatting, with the overall pipeline shown in Figure 2.

3.4. Loss Function

We adopt the loss function from 3DGB [29], and define the total loss as follows:

$$L = \lambda_1 L_{rgb} + \lambda_2 L_\alpha + \lambda_3 L_{reg}, \tag{4}$$

where the default weights of λ_1, λ_2 and λ_3 are set to 1, 10, 100, respectively.

The RGB loss L_{rgb} encourages the rendered image to resemble the target video frame in both color and structure. It is computed as a weighted combination of an L_1 loss and a differentiable Structural Similarity (D-SSIM) loss:

$$L_{rgb} = \lambda_{rgb} L_1 + (1 - \lambda_{rgb}) L_{D-SSIM}, \tag{5}$$

where the default weight λ_{rgb} is set to 0.2.

The opacity loss L_α penalizes opacity values outside the head mask. For each frame i, we compute the accumulated opacity image I_α^i and the corresponding head mask M_h^i, and average the error over F frames:

$$L_\alpha = \frac{1}{F} \sum_{i=1}^{F} \frac{1}{|P|} \sum_{p \in P} \left(I_\alpha^i(p) - M_h^i(p) \right)^2. \tag{6}$$

The regularization loss L_{reg} constrains the mouth Gaussians to remain within a predefined cylindrical volume V. Let $\{\mathbf{x}_i\}_{i=1}^N$ denote the centers of Gaussians located in the

Figure 4. **Visualization of stylized video generation results:** We present details of the real head from the input video, and the "Arcane" stylized head generated by our method. From left to right, the results for the video samples "bala" and "wojtek_1" are shown.

Table 1. **Video durations and inference times:** Duration (in seconds) of the input videos, and inference time (in seconds) of our method.

Video samples	justin	malte_1	nf_01	bala	wojtek_1	person_0004
Duration	98	130	130	159	137	60
Inference	221	260	213	342	275	108

mouth region. To penalize points outside the volume, we employ a signed distance function $SDF(\mathbf{x}_i, V)$, and define the loss as follows:

$$L_{reg} = \frac{1}{N}\sum_{i=1}^{N}\left(\max\left(SDF(\mathbf{x}_i, V), 0\right)\right)^2, \qquad (7)$$

where N is the number of mouth Gaussians.

4. Experiments

4.1. Baselines

Due to the current lack of methods for synthesizing 3D stylized head avatars using Gaussian blendshapes, we compare our method against the following state-of-the-art methods for photo-realistic 3D head avatar synthesis: INSTA [60], PointAvatar [58], FLARE [6], SplattingAvatar [39], FlashAvatar [45], and 3DGB [29]. Notably, 3DGB shares a similar architecture with ours but focuses on photo-realistic avatar synthesis and does not support the synthesis of diverse stylized avatars.

4.2. Dataset

We evaluate both our method and state-of-the-art photo-realistic avatar synthesis methods using six videos from the INSTA [60] dataset. Each video is cropped and resized to 512×512 resolution, with sequence lengths ranging from 1,000 to 4,000 frames. Following the method of 3DGB [29], we retain the final 350 frames of each video for testing. Both 3DGB [29] and our method apply the same preprocessing pipeline [59, 60], including background removal and FLAME parameter extraction.

Table 2. **Quantitative comparison of video stabilization:** We compare the original input (OI), the aligned input (AI), our "Arcane" (OA), and the aligned "Arcane" (AA) videos.

Video Samples		justin	malte_1	nf_01	bala	wojtek_1	person_0004
ITF↑	OI	37.78	38.47	31.82	37.73	39.02	37.17
	AI	32.45	28.84	26.49	27.97	29.09	34.80
	OA	35.80	34.51	29.36	36.01	36.77	33.82
	AA	31.35	26.04	25.31	26.43	28.31	30.84
ISI↑	OI	0.9685	0.9709	0.9361	0.9614	0.9651	0.9277
	AI	0.9066	0.9276	0.8918	0.8995	0.9126	0.9270
	OA	0.9700	0.9643	0.9382	0.9685	0.9670	0.9532
	AA	0.9034	0.8965	0.8887	0.8963	0.9030	0.9036

4.3. Evaluation Metrics

We employ two metrics to evaluate video stabilization: Inter-frame Transformation Fidelity (ITF) [30, 32, 46] and Inter-frame Similarity Index (ISI) [12, 14]. ITF measures the inter-frame Peak Signal-to-Noise Ratio (PSNR) in dB based on the mean squared error. The intuitive idea of ITF is that a more stable video (i.e., less jittery) will have greater similarity between adjacent frames compared to an unstable version of the same video. ISI computes the average Structural Similarity (SSIM) between adjacent frames across the video. Higher ISI values indicate greater perceptual similarity between frames, leading to improved visual comfort for viewers.

Table 3. **Quantitative comparison of 3D head avatars:** We evaluate our method and state-of-the-art methods on the INSTA [60] dataset. In each metric group, the best value is highlighted in **bold**, and the second-best is underlined.

Method	justin		malte_1		nf_01		bala		wojtek_1		person_0004	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
INSTA [60]	31.66	0.9591	27.44	0.9159	26.45	0.8937	29.53	0.8896	31.36	0.9452	25.44	0.8478
PointAvatar [58]	30.40	0.9373	24.98	0.8853	25.25	0.8919	27.88	0.8658	28.82	0.9192	23.29	0.8576
FLARE [6]	29.10	0.9363	25.93	0.8973	25.97	0.9027	27.20	0.8761	27.84	0.9216	25.53	0.9015
SplattingAvatar [39]	30.93	0.9482	27.66	0.9243	27.08	0.9202	32.14	0.9272	29.54	0.9400	26.49	0.9075
FlashAvatar [45]	32.16	0.9611	27.45	0.9326	28.02	0.9326	30.27	0.8494	32.02	0.9509	25.49	0.8996
3DGB [29]	32.63	0.9643	28.65	**0.9432**	28.06	0.9340	33.29	0.9457	**32.57**	**0.9623**	23.66	0.8449
Ours (Arcane)	33.12	0.9628	**29.55**	0.9360	28.33	0.9288	**33.39**	**0.9488**	30.56	0.9436	**28.76**	**0.9110**
Ours (Pixar)	**33.42**	**0.9662**	27.01	0.9375	**28.34**	**0.9341**	30.84	0.9337	31.14	0.9583	23.16	0.8338

Table 4. **Performance comparison:** We record the training time (in minutes) and the rendering speed (in fps) of 3DGB and our method in both "Arcane" (A) and "Pixar" (P) styles.

Video Samples		justin	malte_1	nf_01	bala	wojtek_1	person_0004
Train↓	3DGB	41	44	44	**44**	49	45
	Ours (A)	**40**	45	44	45	50	**44**
	Ours (P)	43	**40**	**43**	**44**	**45**	**44**
Render↑	3DGB	**143**	**142**	130	134	138	**134**
	Ours (A)	140	**142**	**131**	**135**	**140**	128
	Ours (P)	141	133	128	132	134	127

For 3D head avatar synthesis, we evaluate the performance of our method and state-of-the-art methods for photo-realistic avatar synthesis using standard evaluation metrics [29, 56], including Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM). In addition, we record the training time (in minutes) and the rendering speed (in frames per second, fps) for each method. In the ablation study, we additionally adopt the Learned Perceptual Image Patch Similarity (LPIPS) metric to better capture perceptual differences between the synthesized avatars and the ground truth.

4.4. Implementation Details

To ensure a fair performance comparison, the training and testing of all methods are performed on a single RTX 4090 GPU. Our methods are implemented in Python using the PyTorch framework.

For 2D stylized video generation, we use the pre-trained models provided by StyleGANEX [51]. For training the 3D stylized head avatars, we employ the Adam optimizer [23], setting the initial learning rates of the Gaussian properties $\{\mathbf{x}_k, \alpha_k, \mathbf{s}_k, \mathbf{q}_k, S_{H_k}\}$ to 3.2×10^{-7}, 5×10^{-5}, 5×10^{-4}, 1×10^{-4}, and 1.25×10^{-3}, respectively. Following 3DGB [29], the initial number of sampled Gaussians is 50k for the neutral head model and 14k for the mouth interior.

4.5. Quantitative Comparison

4.5.1. Video Stabilization

We adopt an improved StyleGAN model to generate six videos in the "Arcane" style. The durations of the videos and their corresponding inference times are summarized in Table 1. All input videos have a resolution of 512×512 pixels, and inference is performed on a single NVIDIA RTX 4090 GPU. For video durations ranging from 60 to 160 seconds, the generation times span approximately 100 to 350 seconds.

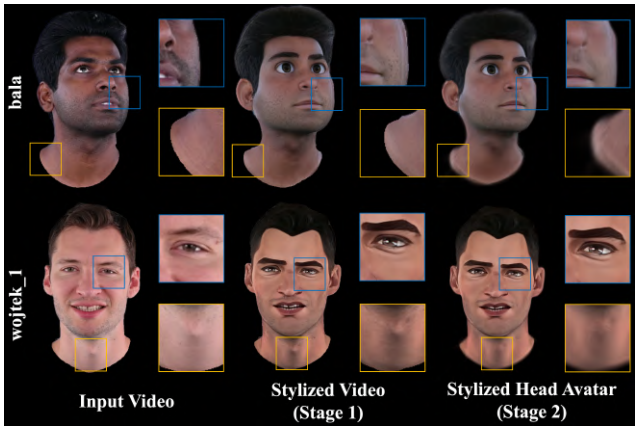

Figure 5. **Qualitative comparison of each stage:** We present the input video head frames, the corresponding stylized videos, and 3D head avatars synthesized by our method.

To evaluate the impact of preprocessing, we apply a standard face alignment technique based on a facial keypoint predictor [21] to the input videos. We compare the original input videos (Original Input, OI) with their aligned counterparts (Aligned Input, AI). Likewise, we compare the "Arcane" style outputs generated from unaligned inputs (Ours Arcane, OA) with those generated from aligned inputs (Aligned Arcane, AA).

As shown in Table 2, both the Inter-frame Transformation Fidelity (ITF) and Inter-frame Similarity Index (ISI) scores for AI are consistently lower than those for OI. Similarly, AA exhibits lower ITF and ISI scores compared to OA. These results suggest that applying face alignment and cropping prior to frame-by-frame generation (i.e., AI and AA) tends to introduce greater temporal instability, resulting in more jittery outputs.

Table 5. **User preference study:** We conduct user preference studies on the "Arcane" and "Pixar" styles, where users rate their preferences on a scale from 1 to 5, with higher scores indicating greater satisfaction, across three evaluation criteria: Style Consistency, Identity Preservation, and Overall Quality. The highest percentage is highlighted in **bold**.

Style	Arcane			Pixar		
Evaluation Criteria	Style Consistency	Identity Preservation	Overall Quality	Style Consistency	Identity Preservation	Overall Quality
1: Very Dissatisfied	0.0%	2.8%	0.6%	2.8%	5.6%	2.2%
2: Dissatisfied	5.0%	10.0%	6.1%	5.0%	6.7%	5.0%
3: Neutral	11.1%	15.6%	15.0%	13.3%	24.4%	20.0%
4: Satisfied	32.2%	**39.4%**	36.7%	29.4%	**36.7%**	**37.8%**
5: Very Satisfied	**51.7%**	32.2%	**41.7%**	**49.4%**	26.7%	35.0%

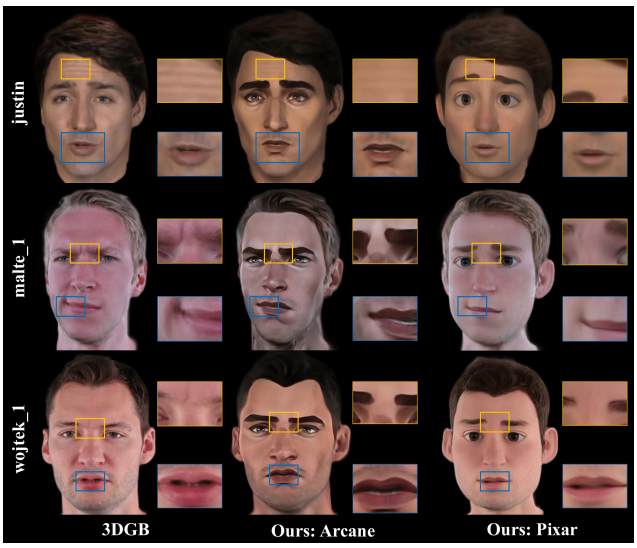

Figure 6. **Qualitative comparison of baseline and ours:** We present 3D head avatars using Gaussian blendshapes synthesized by 3DGB [29] and our method.

4.5.2. 3D Head Avatar

We evaluate our method and state-of-the-art methods using standard metrics for animatable head reconstruction. The quantitative results are presented in Table 3, and the training and rendering times for both the baseline methods and ours are reported in Table 4. With the additional integration of stylization, our method achieves performance comparable to the state-of-the-art on the PSNR and SSIM metrics in most cases, and even outperforms them on certain data. Specifically, our method outperforms all other methods on synthesizing the "Arcane" style for the "bala" and "person_0004" data, as well as the "Pixar" style for the "justin" and "nf_01" data.

In addition, although our method integrates stylization into 3D head avatars, its training and rendering times remain comparable to those of the method of 3DGB [29]. In certain cases, our method is even more efficient in both training and rendering. Combined with the additional time required for video generation (as shown in Table 1), the overall time cost of our method remains acceptable.

4.6. Qualitative Comparison

We present the original video head frames, the corresponding stylized video frames generated by our method, and the 3D stylized head avatars synthesized using Gaussian blendshapes. The qualitative comparison is shown in Figure 5. The examples are selected from the "bala" dataset in the "Pixar" style and the "wojtek_1" dataset in the "Arcane" style.

In the stylized video, the "bala" data exhibits artifacts along the side edge of the head. We attribute this to the latent space distribution learned by StyleGAN, which tends to produce striped artifacts when the viewing angle falls outside the distribution covered by the training data. Notably, these artifacts are not present in the corresponding 3D stylized head avatars rendered by our method. Furthermore, the 3D stylized head avatars successfully preserve fine details from the stylized videos, such as the mole near the eye in the "wojtek_1" dataset. However, since the 3D avatar synthesis mainly focuses on the facial region, the neck area is typically blurred, as observed in both cases. This blurring leads to the lower quantitative performance, since the neck region is included in the evaluation.

The qualitative comparison with 3DGB [29] is presented in Figures 6. Our method effectively captures and preserves high-frequency details in the stylized videos. Compared to the state-of-the-art method, ToonifyGB can synthesize 3D stylized head avatars with comparable quality and detail.

4.7. Visualization

To better demonstrate the visual quality of our generated videos, we present several examples in Figure 3, and select two representative videos for detailed comparison in Figure 4. Specifically, we show real head frames from the "bala" and "wojtek_1" videos, as well as the corresponding heads of generated videos in the "Arcane" style.

The results demonstrate that key facial features, such as the beard, mouth shape, and even small details like the black mole above the eye in the lower right image, are well preserved after the stylization process. These details highlight the excellent performance of our method in terms of detail preservation and identity consistency.

Table 6. **Ablation study on face alignment and cropping:** We compare 3D head avatars synthesized from different input videos: one generated by our method, and the other using face alignment and cropping as prprocessing.

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

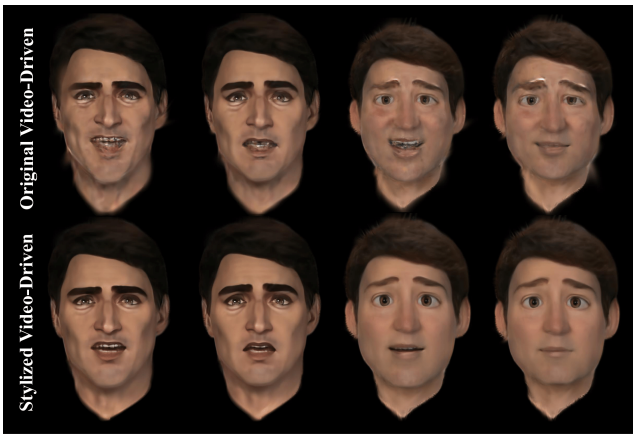

Figure 7. **Ablation study on the effect of different driving videos:** We present 3D stylized head avatar animation driven by the original input videos and our generated videos.

4.8. User Study

To more effectively evaluate the 3D stylized head avatars synthesized by our method, we conduct a user preference study. We collect 180 votes for both the "Arcane" and "Pixar" styles, respectively, with users rating their preferences using Likert scales [27] across three evaluation criteria: Style Consistency, Identity Preservation, and Overall Quality. Each criterion is assessed using a five-point scale: 1 for Very Dissatisfied, 2 for Dissatisfied, 3 for Neutral, 4 for Satisfied, and 5 for Very Satisfied.

Specifically, Style Consistency evaluates how well the stylized output aligns with the defining characteristics of the style; Identity Preservation measures whether the avatar retains the unique features and identity of the original character after stylization; and Overall Quality provides a comprehensive assessment of the visual appeal and overall quality of the synthesized avatar.

The results are presented in Table 5, indicating that most users show great satisfaction with the "Arcane" style, particularly in terms of Style Consistency (51.7% of users are very satisfied) and Overall Quality (41.7% of users are very satisfied). Although the rating for Identity Preservation is slightly lower, the average score remains favorable.

In addition, the "Pixar" style is also favored by users, particularly in Style Consistency (49.4% of users are very satisfied). For Identity Preservation and Overall Quality, the majority of users (over 60%) indicate satisfaction with our 3D stylized head avatars.

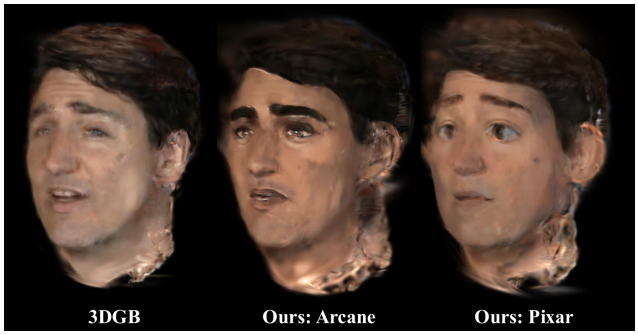

Figure 8. **Limitation:** We present side-view renderings synthesized by 3DGB [29] and our method.

4.9. Ablation Study

4.9.1. Face Alignment and Cropping

We compare 3D stylized 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 PSNR, SSIM, and Learned Perceptual Image Patch Similarity (LPIPS). As shown in Table 6, our method outperforms the traditional method with face alignment and cropping across all evaluation metrics. This demonstrates that our method effectively eliminates jitter during video generation, enabling higher-quality synthesis of 3D stylized head animations.

4.9.2. Source Videos for Driving Animation

Compared with the architecture of 3DGB [29] that synthesizes 3D photo-realistic head avatars, our framework includes an additional Stage 1 to generate the stylized video. To demonstrate the importance of the generated stylized video in driving the animation, we compare the results of using the original input video (real face) versus our generated stylized video as the driving source, as shown in Figure 7.

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 videos generated by Stage 1 of our framework. Therefore, we recommend using the generated stylized videos, rather than the original input videos, as the driving source for 3D stylized head avatar animation.

5. Limitation

Our method struggles to render side views of 3D stylized head avatars when the training data (i.e., input video) lacks side-view representations of the real head. As shown in Figure 8, we present side-view renderings synthesized by both 3DGB [29] and our method, and this limitation is also observed in the state-of-the-art methods. In fact, existing NeRF-based and Gaussian-based methods have yet to effectively address this issue. Rendering novel views from single-view training data remains an open problem for future research. Two directions to address this limitation include employing 2D GANs to synthesize videos with side views as additional training data, and enhancing the generalization ability of our model.

6. Conclusion

We propose a novel two-stage framework, named ToonifyGB, which utilizes Gaussian blendshapes to synthesize head animations in diverse styles from monocular videos. In Stage 1, the proposed method adopts an improved StyleGAN-based model to generate stylized videos without requiring face alignment or cropping as preprocessing. This results in more temporally stable outputs, providing a reliable foundation for high-quality 3D head avatar animation synthesis. Stage 2 focuses on constructing 3D stylized head avatars using Gaussian blendshapes, enabling fine-grained expression modeling and satisfactory animation. Our method supports real-time generation of stylized avatar animations in popular styles such as "Arcane" and "Pixar".

For future work, we plan to integrate motion capture technologies to enable real-time expression control of 3D stylized avatars. Specifically, we aim to explore more efficient approaches for obtaining real-time expression parameters, bypassing the complexity of traditional PCA inversion. This direction is expected to further broaden the applicability of ToonifyGB in virtual character interaction and personalized avatar generation.

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] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4432–4441, 2019. 2

[2] Rameen Abdal, Hsin-Ying Lee, Peihao Zhu, Menglei Chai, Aliaksandr Siarohin, Peter Wonka, and Sergey Tulyakov. 3davatargan: Bridging domains for personalized editable avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4552–4562, 2023. 3

[3] Rameen Abdal, Wang Yifan, Zifan Shi, Yinghao Xu, Ryan Po, Zhengfei Kuang, Qifeng Chen, Dit-Yan Yeung, and Gordon Wetzstein. Gaussian shell maps for efficient 3d human generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9441–9451, 2024. 2

[4] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Restyle: A residual-based stylegan encoder via iterative refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6711–6720, 2021. 2

[5] Yuval Alaluf, Or Patashnik, Zongze Wu, Asif Zamir, Eli Shechtman, Dani Lischinski, and Daniel Cohen-Or. Third time's the charm? image and video editing with stylegan3. In *European Conference on Computer Vision*, pages 204–220, 2022. 2

[6] Shrisha Bharadwaj, Yufeng Zheng, Otmar Hilliges, Michael J Black, and Victoria Fernandez-Abrevaya. Flare: Fast learning of animatable and relightable mesh avatars. *ACM Transactions on Graphics*, 42(6):15, 2023. 5, 6

[7] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022. 2

[8] Yufan Chen, Lizhen Wang, Qijing Li, Hongjiang Xiao, Shengping Zhang, Hongxun Yao, and Yebin Liu. Monogaussianavatar: Monocular gaussian point-based head avatar. In *ACM SIGGRAPH Conference Papers*, pages 1–9, 2024. 2

[9] Gereon Fox, Ayush Tewari, Mohamed Elgharib, and Christian Theobalt. Stylevideogan: A temporal generative model using a pretrained stylegan. In *British Machine Vision Conference*, 2021. 2

[10] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics*, 41(4):1–13, 2022. 2

[11] Xuan Gao, Chenglai Zhong, Jun Xiang, Yang Hong, Yudong Guo, and Juyong Zhang. Reconstructing personalized semantic facial nerf models from monocular video. *ACM Transactions on Graphics*, 41(6):1–12, 2022. 2, 4, 12

[12] Wilko Guilluy, Azeddine Beghdadi, and Laurent Oudre. A performance evaluation framework for video stabilization methods. In *European Workshop on Visual Information Processing*, pages 1–6, 2018. 5

[13] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. Headnerf: A real-time nerf-based parametric head model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20374–20384, 2022. 2

[14] Jerin Geo James, Devansh Jain, and Ajit Rajwade. Globalflownet: Video stabilization using deep distilled global motion estimates. In *Proceedings of the IEEE/CVF Winter Con-*

ference on Applications of Computer Vision, pages 5078–5087, 2023. 5

[15] Wonjong Jang, Gwangjin Ju, Yucheol Jung, Jiaolong Yang, Xin Tong, and Seungyong Lee. Stylecarigan: caricature generation via stylegan feature map modulation. *ACM Transactions On Graphics*, 40(4):1–16, 2021. 2

[16] Wonjong Jang, Yucheol Jung, Hyomin Kim, Gwangjin Ju, Chaewon Son, Jooeun Son, and Seungyong Lee. Toonify3d: Stylegan-based 3d stylized face generator. In *ACM SIGGRAPH Conference Papers*, pages 1–11, 2024. 3

[17] Wonjoon Jin, Nuri Ryu, Geonung Kim, Seung-Hwan Baek, and Sunghyun Cho. Dr. 3d: Adapting 3d gans to artistic drawings. In *SIGGRAPH Asia Conference Papers*, pages 1–8, 2022. 3

[18] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 2, 3, 4

[19] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020. 2, 3

[20] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in neural information processing systems*, 34:852–863, 2021. 2

[21] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1867–1874, 2014. 6

[22] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4):139–1, 2023. 2, 4

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. 6

[24] Tobias Kirschstein, Simon Giebenhain, Jiapeng Tang, Markos Georgopoulos, and Matthias Nießner. Gghead: Fast and generalizable 3d gaussian heads. In *SIGGRAPH Asia Conference Papers*, pages 1–11, 2024. 2

[25] Yushi Lan, Xuyi Meng, Shuai Yang, Chen Change Loy, and Bo Dai. Self-supervised geometry-aware encoder for style-based 3d gan inversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20940–20949, 2023. 3

[26] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Transactions on Graphics*, 36(6):194–1, 2017. 3

[27] Rensis Likert. A technique for the measurement of attitudes. *Archives of psychology*, 1932. 8

[28] Feng-Lin Liu, Shu-Yu Chen, Yu-Kun Lai, Chunpeng Li, Yue-Ren Jiang, Hongbo Fu, and Lin Gao. Deepfacevideoediting: Sketch-based deep editing of face videos. *ACM Transactions on Graphics*, 41(4):1–16, 2022. 2

[29] Shengjie Ma, Yanlin Weng, Tianjia Shao, and Kun Zhou. 3d gaussian blendshapes for head avatar animation. In *ACM SIGGRAPH Conference Papers*, pages 1–10, 2024. 2, 3, 4, 5, 6, 7, 8, 9, 13

[30] Lucio Marcenaro, Gianni Vernazza, and Carlo S Regazzoni. Image stabilization algorithms for video-surveillance applications. In *Proceedings of the IEEE International Conference on Image Processing*, pages 349–352, 2001. 5

[31] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2

[32] Carlos Morimoto and Rama Chellappa. Evaluation of image stabilization algorithms. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2789–2792, 1998. 5

[33] Utkarsh Ojha, Yijun Li, Jingwan Lu, Alexei A Efros, Yong Jae Lee, Eli Shechtman, and Richard Zhang. Few-shot image generation via cross-domain correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10743–10752, 2021. 2

[34] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. Stylesdf: High-resolution 3d-consistent image and geometry generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13503–13513, 2022. 3

[35] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021. 2

[36] Justin NM Pinkney and Doron Adler. Resolution dependent gan interpolation for controllable image synthesis between domains. In *NeurIPS Workshop on Machine Learning for Creativity and Design*, 2020. 1, 2

[37] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20299–20309, 2024. 2

[38] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2287–2296, 2021. 2

[39] Zhijing Shao, Zhaolong Wang, Zhuang Li, Duotun Wang, Xiangru Lin, Yu Zhang, Mingming Fan, and Zeyu Wang. Splattingavatar: Realistic real-time human avatars with mesh-embedded gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1606–1616, 2024. 5, 6

[40] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics*, 40(4):1–14, 2021. 2

[41] Rotem Tzaban, Ron Mokady, Rinon Gal, Amit Bermano, and Daniel Cohen-Or. Stitch it in time: Gan-based facial editing of real videos. In *SIGGRAPH Asia Conference Papers*, pages 1–9, 2022. 2

[42] Tengfei Wang, Yong Zhang, Yanbo Fan, Jue Wang, and Qifeng Chen. High-fidelity gan inversion for image attribute editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11379–11388, 2022. 2

[43] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, et al. Rodin: A generative model for sculpting 3d digital avatars using diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4563–4573, 2023. 3

[44] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. *Advances in neural information processing systems*, 29, 2016. 2

[45] Jun Xiang, Xuan Gao, Yudong Guo, and Juyong Zhang. Flashavatar: High-fidelity head avatar with efficient gaussian embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1802–1812, 2024. 2, 5, 6

[46] Jie Xu, Hua-wen Chang, Shuo Yang, and Minghui Wang. Fast feature-based video stabilization without accumulative global motion estimation. *IEEE Transactions on Consumer Electronics*, 58(3):993–999, 2012. 5

[47] Yuelang Xu, Lizhen Wang, Xiaochen Zhao, Hongwen Zhang, and Yebin Liu. Avatarmav: Fast 3d head avatar reconstruction using motion-aware neural voxels. In *ACM SIGGRAPH Conference Papers*, pages 1–10, 2023. 2

[48] Yuelang Xu, Benwang Chen, Zhe Li, Hongwen Zhang, Lizhen Wang, Zerong Zheng, and Yebin Liu. Gaussian head avatar: Ultra high-fidelity head avatar via dynamic gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2024. 2

[49] Shuai Yang, Liming Jiang, Ziwei Liu, and Chen Change Loy. Pastiche master: Exemplar-based high-resolution portrait style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7693–7702, 2022. 2

[50] Shuai Yang, Liming Jiang, Ziwei Liu, and Chen Change Loy. Vtoonify: Controllable high-resolution portrait video style transfer. *ACM Transactions on Graphics*, 41(6):1–15, 2022. 2

[51] Shuai Yang, Liming Jiang, Ziwei Liu, and Chen Change Loy. Styleganex: Stylegan-based manipulation beyond cropped aligned faces. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21000–21010, 2023. 2, 3, 4, 6

[52] Xu Yao, Alasdair Newson, Yann Gousseau, and Pierre Hellier. A latent transformer for disentangled face editing in images and videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13789–13798, 2021. 2

[53] Tarun Yenamandra, Ayush Tewari, Florian Bernard, Hans-Peter Seidel, Mohamed Elgharib, Daniel Cremers, and Christian Theobalt. i3dmm: Deep implicit 3d morphable model of human heads. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12803–12813, 2021. 2

[54] Bowen Zhang, Yiji Cheng, Chunyu Wang, Ting Zhang, Jiaolong Yang, Yansong Tang, Feng Zhao, Dong Chen, and Baining Guo. Rodinhd: High-fidelity 3d avatar generation with diffusion models. In *European Conference on Computer Vision*, pages 465–483. Springer, 2024. 3

[55] Junzhe Zhang, Yushi Lan, Shuai Yang, Fangzhou Hong, Quan Wang, Chai Kiat Yeo, Ziwei Liu, and Chen Change Loy. Deformtoon3d: Deformable 3d toonification from neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9110–9120, 2023. 3

[56] 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 Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 6

[57] Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C Bühler, Xu Chen, Michael J Black, and Otmar Hilliges. Im avatar: Implicit morphable head avatars from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13545–13555, 2022. 2

[58] Yufeng Zheng, Wang Yifan, Gordon Wetzstein, Michael J Black, and Otmar Hilliges. Pointavatar: Deformable point-based head avatars from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21057–21067, 2023. 5, 6

[59] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. In *European Conference on Computer Vision*, pages 250–269, 2022. 2, 3, 5

[60] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Instant volumetric head avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4574–4584, 2023. 3, 4, 5, 6, 12

Figure 9. Visualization of stylized video generation results on the videos from the INSTA [60] and NeRFBlendShape [11] datasets.

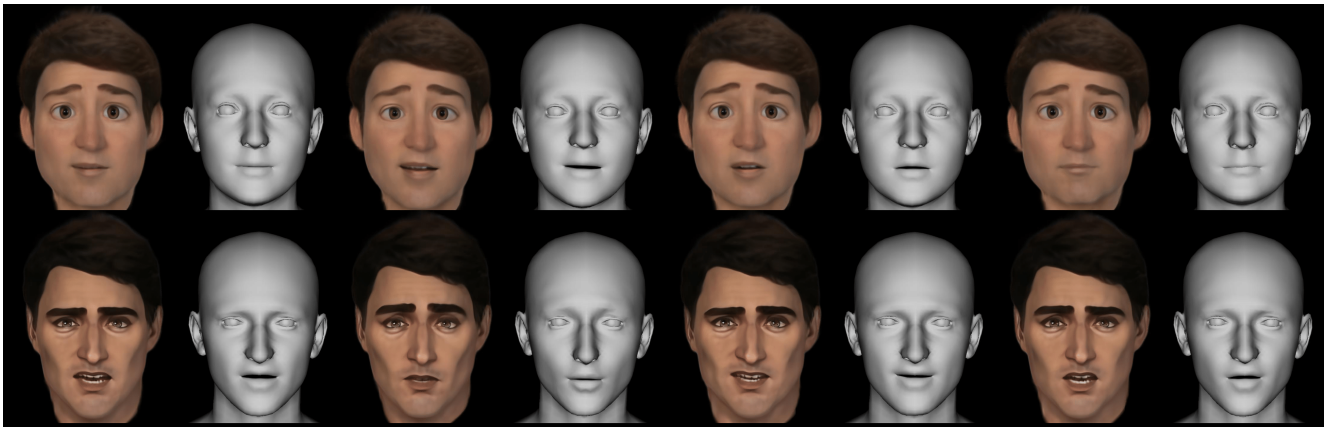

Figure 10. Visualization of the synthesized 3D stylized head avatars. Each avatar closely resembles its corresponding FLAME mesh while capturing the stylized appearance.

Figure 11. **More examples for qualitative comparison:** We present input video head frames, and 3D head avatars using Gaussian blendshapes synthesized by 3DGB [29] and our method.