

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

Rui-Yang Ju¹, Sheng-Yen Huang¹, Yi-Ping Hung¹

¹Graduate Institute of Networking and Multimedia, National Taiwan University
jryjry1094791442@gmail.com, d12944001@csie.ntu.edu.tw, hung@csie.ntu.edu.tw

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.

Introduction

With the advancement of 3D head reconstruction technologies, individuals are now able to personalize unique avatars for telepresence and virtual/augmented reality applications, which serve as an essential foundation for the rise of the metaverse. Considering user preferences and privacy concerns, the creation of stylized avatars is an important topic that deserves further research. Toonify (Pinkney and Adler 2020) is a StyleGAN-based method designed for 2D facial image stylization. In contrast to photo-realistic 3D head avatars, stylized 3D head avatars emphasize the expression of personal identity and the faithful transfer of target styles.

Blendshape is an efficient facial animation representation that synthesizes 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) (Mildenhall et al. 2021), Gao *et al.* (Gao et al. 2022) and Zheng *et al.* (Zheng et al. 2022) incorporated the blendshape concept into NeRF, enabling avatar animation

through the construction of a group of NeRF blendshapes that are linearly blended. Furthermore, the recently proposed 3D Gaussian Splatting (3DGS) (Kerbl et al. 2023) significantly improved rendering efficiency and delivered high-quality head reconstruction, outperforming NeRF in both speed and quality. Based on this, 3D Gaussian Blendshapes (3DGB) (Ma et al. 2024) successfully integrated the blendshape 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 adopt an improved StyleGAN to generate a more stable and less jittery stylized video, without requiring on fixed resolution and pre-aligned face cropping. In Stage 2, we begin with 3DGB to learn a neutral head model and a set of expression blendshapes, each represented as 3D Gaussians. Finally, by incorporating the facial tracker (Zielonka, Bolkart, and Thies 2022), we use the tracked motion parameters to drive ToonifyGB for animating 3D stylized head avatars.

In summary, 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 and enabling 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 based on Gaussian blendshapes.

Related Work

StyleGAN and Toonify

StyleGAN (Karras, Laine, and Aila 2019; Karras et al. 2020) 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 for subsequent edits such as adding glasses or changing hairstyle or age (Abdal, Qin, and Wonka 2019; Patashnik

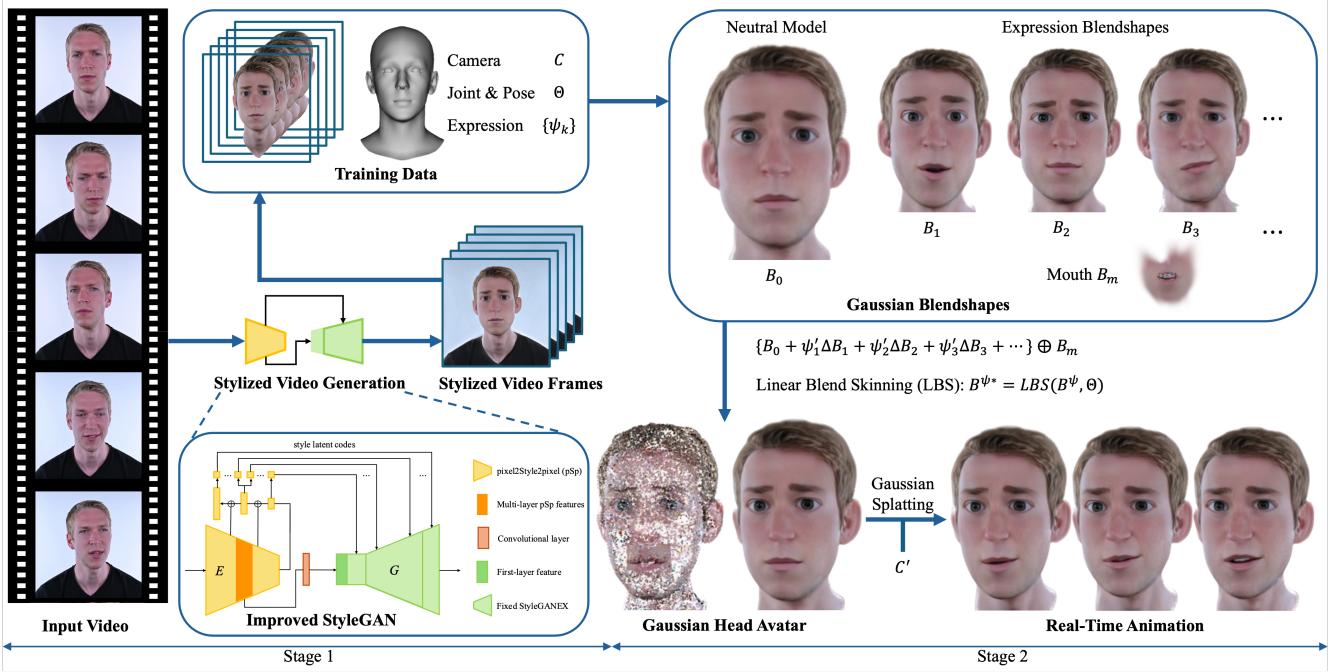


Figure 1: Pipeline: Our ToonifyGB framework consists of two stages: Stage 1 involves the generation of stylized videos, and Stage 2 focuses on the synthesis of 3D stylized head avatars using Gaussian blendshapes.

et al. 2021). To enhance inversion efficiency, methods such as pSp (Richardson et al. 2021) and e4e (Tov et al. 2021) 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 (Alaluf, Patashnik, and Cohen-Or 2021) and HFGI (Wang et al. 2022) improve reconstruction fidelity by respectively predicting latent code residuals and correcting middle layer features. Nevertheless, these methods remain limited to aligned and cropped facial images to achieve effective editing and reconstruction.

Recently, researchers (Pinkney and Adler 2020; Ojha et al. 2021; Jang et al. 2021; Yang et al. 2022a; Gal et al. 2022) have explored using StyleGAN for target-domain image generation through transfer learning. Among these works, Toonify (Pinkney and Adler 2020) 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 (Fox et al. 2021), disentangling identity from facial attributes (Yao et al. 2021), incorporating sketch-based branches (Liu et al. 2022), and tuning the generator to maintain temporal consistency (Tzaban et al. 2022). However, these methods typically rely on face alignment and cropping as preprocessing. Although StyleGAN3 (Karras et al. 2021) was introduced to support unaligned face inputs, a subsequent study (Alaluf et al. 2022) 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 (Yang et al. 2022b) and StyleGANEX (Yang et al. 2023) 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.

3D Head Avatar

Since the introduction of NeRF (Mildenhall et al. 2021), implicit representation-based methods (Yenamandra et al. 2021; Zheng et al. 2022; Hong et al. 2022; Chan et al. 2022; Xu et al. 2023) for head reconstruction have achieved remarkable progress. 3DGS (Kerbl et al. 2023) 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 (Qian et al. 2024; Xu et al. 2024; Chen et al. 2024; Ma et al. 2024; Xiang et al. 2024; Abdal et al. 2024; Kirschstein et al. 2024) 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 (Wu et al. 2016) enable high-quality generation, making 3D head stylization possible. Although fine-tuning 3D generators for geometric and texture-based stylization has proven effective (Or-El et al. 2022;

Jin et al. 2022; Abdal et al. 2023; Lan et al. 2023; Wang et al. 2023; Zhang et al. 2024), performing independent fine-tuning for each new style remains costly. Toonify3D (Jang et al. 2024) addressed this limitation by predicting facial surface normals using the proposed StyleNormal, enabling direct face stylization without additional fine-tuning. Similarly, DeformToon3D (Zhang et al. 2023) 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.

Method

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 (Karras, Laine, and Aila 2019; Karras et al. 2020) 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 (Yang et al. 2023) in Stage 1, enabling stable stylized video generation at the original resolution, as shown in Figure 1.

To prepare the training data for Stage 2, we follow the method in (Zielonka, Bolkart, and Thies 2023; Ma et al. 2024), using the facial tracker from (Zielonka, Bolkart, and Thies 2022) to compute FLAME (Li et al. 2017) 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 1, we apply Linear Blend Skinning (LBS) to transform the Gaussian model based on the extracted joint and pose parameters. The transformation is defined as:

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

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 C , we enable novel-view rendering and animation.

Stylized Video Generation

As shown in StyleGANEX (Yang et al. 2023), 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



Figure 2: **Visualization of stylized video generation results** on the INSTA (Zielonka, Bolkart, and Thies 2023) and NeRFBlendShape (Gao et al. 2022) datasets.

facial structure due to the low spatial resolution of the 4×4 feature map, causing blurring and loss of detail. To address this limitation, we adopt StyleGANEX (Yang et al. 2023), an enhanced variant of StyleGAN2 (Karras, Laine, and Aila 2019), 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 2, our method consistently generates high-quality stylized head videos across diverse styles, regardless of gender.

Gaussian Blendshapes Synthesis

We represent all Gaussian head avatars using 3D Gaussians. As described in 3DGS (Kerbl et al. 2023), each Gaussian has some basic properties including Gaussian center μ , scale s , color c , opacity α , and rotation q . Based on 3DGB (Ma et al. 2024), our Gaussian blendshape representation consists of a neutral model B_0 and a set of n expression blendshapes B_1, B_2, \dots, 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:

$$B^\psi = B_0 + \sum_{k=1}^n \psi_k \Delta B_k \quad (2)$$

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 (Ma et al. 2024) 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:

$$B_m^* = LBS(B_m, \Theta). \quad (3)$$

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 1.

Loss Function

We adopt the loss function from 3DGB (Ma et al. 2024), and define the total loss as follows:

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

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:

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

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:

$$L_\alpha = \frac{1}{F} \sum_{i=1}^F \|I_\alpha^i - M_h^i\|_2. \quad (6)$$

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 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:

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

where N is the number of mouth Gaussians.

Experiments

Baseline and Dataset

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 (Zielonka, Bolkart, and Thies 2023), PointAvatar (Zheng et al. 2023), FLARE (Bharadwaj et al. 2023), SplattingAvatar (Shao et al. 2024), FlashAvatar (Xiang et al. 2024), and 3DGB (Ma et al. 2024). 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.

We evaluate both our method and state-of-the-art photo-realistic avatar synthesis methods using six videos from the INSTA (Zielonka, Bolkart, and Thies 2023) 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 (Ma et al. 2024), we retain the final 350 frames of each video for testing. Both 3DGB (Ma et al. 2024) and our method apply the same preprocessing pipeline (Zielonka, Bolkart, and Thies 2022, 2023), including background removal and FLAME parameter extraction.

Evaluation Metrics

We employ two metrics to evaluate video stabilization: Inter-frame Transformation Fidelity (ITF) (Morimoto and Chellappa 1998; Marcenaro, Vernazza, and Regazzoni 2001; Xu et al. 2012) and Inter-frame Similarity Index (ISI) (Guilley, Beghdadi, and Oudre 2018; James, Jain, and Rajwade 2023). 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.

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 (Zhang et al. 2018; Ma et al. 2024), 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.

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 (Yang et al. 2023). For training the 3D stylized head avatars, we



Figure 3: Visualization of stylized video generation: 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 data samples “bala” and “wojtek_1” are shown.

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

Dataset	justin	malte_1	nf.01	bala	wojtek_1	person_0004
Duration	98	130	130	159	137	60
Inference	221	260	213	342	275	108

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.

Dataset	justin	malte_1	nf.01	bala	wojtek_1	person_0004
ITF↑	OI	37.78	38.47	31.82	37.73	39.02
	AI	32.45	28.84	26.49	27.97	29.09
	OA	35.80	34.51	29.36	36.01	36.77
	AA	31.35	26.04	25.31	26.43	30.84
ISI↑	OI	0.9685	0.9709	0.9361	0.9614	0.9651
	AI	0.9066	0.9276	0.8918	0.8995	0.9126
	OA	0.9700	0.9643	0.9382	0.9685	0.9670
	AA	0.9034	0.8965	0.8887	0.8963	0.9030

employ the Adam optimizer (Kingma and Ba 2015), setting the initial learning rates of the Gaussian properties $\{\mathbf{x}_k, \alpha_k, \mathbf{s}_k, \mathbf{q}_k, SH_k\}$ to $3.2 \times 10^{-7}, 5 \times 10^{-5}, 5 \times 10^{-4}, 1 \times 10^{-4}$, and 1.25×10^{-3} , respectively. Following 3DGB (Ma et al. 2024), the initial number of sampled Gaussians is 50k for the neutral head model and 14k for the mouth interior.

Quantitative Comparison

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.

To evaluate the impact of preprocessing, we apply a standard face alignment technique based on a facial keypoint predictor (Kazemi and Sullivan 2014) 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.

3D Head Avatar We evaluate our method and state-of-the-art methods using standard metrics for animatable head reconstruction, as well as training and rendering times. The quantitative results are presented in Table 3, and the training and rendering times are recorded 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 (Ma et al. 2024). 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.

Table 3: **Quantitative comparison of 3D head avatars:** We evaluate our method and SOTA methods on the INSTA (Zielonka, Bolkart, and Thies 2023) 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↑										
INSTA	31.66	0.9591	27.44	0.9159	26.45	0.8937	29.53	0.8896	<u>31.36</u>	0.9452	25.44	0.8478
PointAvatar	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	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	30.93	0.9482	27.66	0.9243	27.08	0.9202	32.14	0.9272	29.54	0.9400	<u>26.49</u>	<u>0.9075</u>
FlashAvatar	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	32.63	<u>0.9643</u>	<u>28.65</u>	0.9432	28.06	0.9340	<u>33.29</u>	<u>0.9457</u>	32.57	0.9623	23.66	0.8449
Ours (Arcane)	<u>33.12</u>	0.9628	29.55	0.9360	<u>28.33</u>	0.9288	33.39	0.9488	30.56	0.9436	28.76	0.9110
Ours (Pixar)	33.42	0.9662	27.01	<u>0.9375</u>	28.34	0.9341	30.84	0.9337	31.14	<u>0.9583</u>	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.

Method	Metric	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	<u>44</u>
	Ours (P)	43	40	43	44	45	<u>44</u>
Render↑	3DGB	143	142	130	134	138	134
	Ours (A)	140	142	131	135	140	128
	Ours (P)	141	133	128	132	134	127

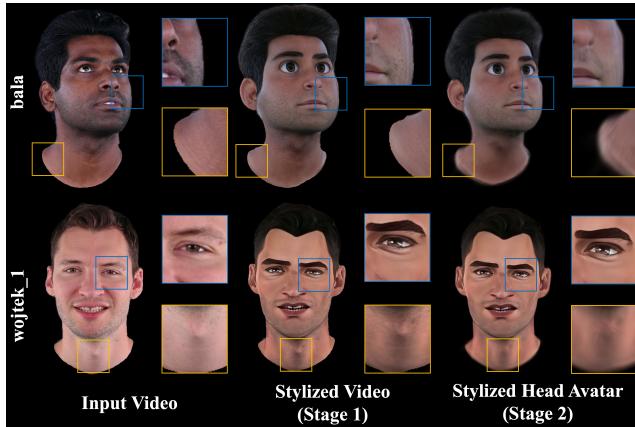


Figure 4: **Qualitative comparison of each stage:** We present the input video head frames, the corresponding stylized video frames generated by our method, and the 3D stylized head avatars synthesized using Gaussian blendshapes. The qualitative comparison is shown in Figure 4. The examples are selected 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

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 4. 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

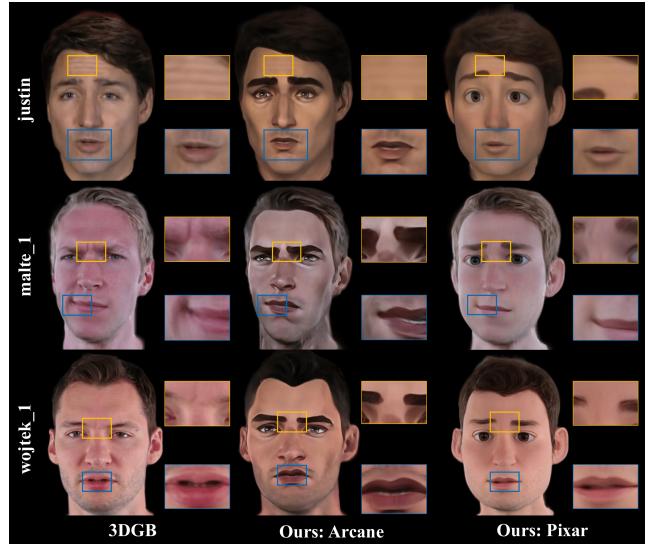


Figure 5: **Qualitative comparison of baseline and ours:** We present 3D head avatars using Gaussian blendshapes synthesized by 3DGB (Ma et al. 2024) and our method.

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 (Ma et al. 2024) is presented in Figures 5. Our method effectively captures and preserves high-frequency details in the stylized videos. Compared to the SOTA method, ToonifyGB can synthesize 3D stylized head avatars with comparable quality and detail.

Visualization

Generated Video Details To better demonstrate the visual quality of our generated videos, we present several examples in Figure 2, and select two representative videos for detailed comparison in Figure 3. 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,

Table 5: 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 preprocessing.

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



Figure 6: 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.

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.

Ablation Study

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 5, 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.

Source Videos for Driving Animation Compared with the architecture of 3DGB (Ma et al. 2024) 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 6.

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

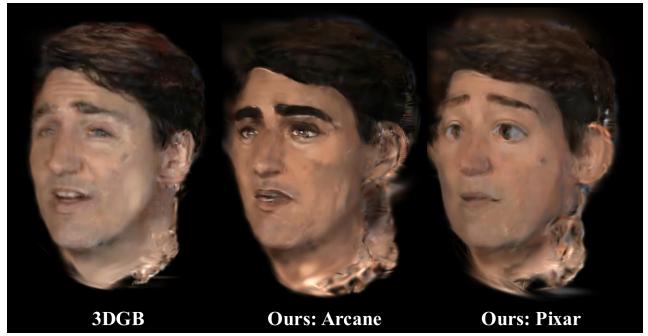


Figure 7: Limitation: We present side-view renderings synthesized by 3DGB (Ma et al. 2024) and our method.

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.

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 7, we present side-view renderings synthesized by both 3DGB (Ma et al. 2024) 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.

Conclusion

We propose a novel two-stage framework, 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. This direction is expected to further broaden the applicability of ToonifyGB in virtual character interaction and personalized avatar generation.

References

- Abdal, R.; Lee, H.-Y.; Zhu, P.; Chai, M.; Siarohin, A.; Wonka, P.; and Tulyakov, S. 2023. 3davatargan: Bridging domains for personalized editable avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4552–4562.
- Abdal, R.; Qin, Y.; and Wonka, P. 2019. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4432–4441.
- Abdal, R.; Yifan, W.; Shi, Z.; Xu, Y.; Po, R.; Kuang, Z.; Chen, Q.; Yeung, D.-Y.; and Wetzstein, G. 2024. Gaussian shell maps for efficient 3d human generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9441–9451.
- Alaluf, Y.; Patashnik, O.; and Cohen-Or, D. 2021. Restyle: A residual-based stylegan encoder via iterative refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6711–6720.
- Alaluf, Y.; Patashnik, O.; Wu, Z.; Zamir, A.; Shechtman, E.; Lischinski, D.; and Cohen-Or, D. 2022. Third time’s the charm? image and video editing with stylegan3. In *European Conference on Computer Vision*, 204–220.
- Bharadwaj, S.; Zheng, Y.; Hilliges, O.; Black, M. J.; and Fernandez-Abreva, V. 2023. Flare: Fast learning of animatable and relightable mesh avatars. *ACM Trans. Graph.*, 42(6): 15.
- Chan, E. R.; Lin, C. Z.; Chan, M. A.; Nagano, K.; Pan, B.; De Mello, S.; Gallo, O.; Guibas, L. J.; Tremblay, J.; Khamis, S.; et al. 2022. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16123–16133.
- Chen, Y.; Wang, L.; Li, Q.; Xiao, H.; Zhang, S.; Yao, H.; and Liu, Y. 2024. Monogaussianavatar: Monocular gaussian point-based head avatar. In *ACM SIGGRAPH 2024 Conference Papers*, 1–9.
- Fox, G.; Tewari, A.; Elgharib, M.; and Theobalt, C. 2021. Stylevideogan: A temporal generative model using a pre-trained stylegan. In *British Machine Vision Conference*.
- Gal, R.; Patashnik, O.; Maron, H.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4): 1–13.
- Gao, X.; Zhong, C.; Xiang, J.; Hong, Y.; Guo, Y.; and Zhang, J. 2022. Reconstructing personalized semantic facial nerf models from monocular video. *ACM Transactions on Graphics (TOG)*, 41(6): 1–12.
- Guilluy, W.; Beghdadi, A.; and Oudre, L. 2018. A performance evaluation framework for video stabilization methods. In *2018 7th European Workshop on Visual Information Processing (EUVIP)*, 1–6.
- Hong, Y.; Peng, B.; Xiao, H.; Liu, L.; and Zhang, J. 2022. Headnerf: A real-time nerf-based parametric head model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20374–20384.
- James, J. G.; Jain, D.; and Rajwade, A. 2023. Globalflownet: Video stabilization using deep distilled global motion estimates. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5078–5087.
- Jang, W.; Ju, G.; Jung, Y.; Yang, J.; Tong, X.; and Lee, S. 2021. Stylecarigan: caricature generation via stylegan feature map modulation. *ACM Transactions On Graphics (TOG)*, 40(4): 1–16.
- Jang, W.; Jung, Y.; Kim, H.; Ju, G.; Son, C.; Son, J.; and Lee, S. 2024. Toonify3D: StyleGAN-based 3D Stylized Face Generator. In *ACM SIGGRAPH 2024 Conference Papers*, 1–11.
- Jin, W.; Ryu, N.; Kim, G.; Baek, S.-H.; and Cho, S. 2022. Dr. 3d: Adapting 3d gans to artistic drawings. In *SIGGRAPH Asia 2022 Conference Papers*, 1–8.
- Karras, T.; Aittala, M.; Laine, S.; Häkkinen, E.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2021. Alias-free generative adversarial networks. *Advances in neural information processing systems*, 34: 852–863.
- Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4401–4410.
- Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8110–8119.
- Kazemi, V.; and Sullivan, J. 2014. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1867–1874.
- Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4): 139–1.
- Kingma, D. P.; and Ba, J. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.
- Kirschstein, T.; Giebenhain, S.; Tang, J.; Georgopoulos, M.; and Nießner, M. 2024. Gghead: Fast and generalizable 3d gaussian heads. In *SIGGRAPH Asia 2024 Conference Papers*, 1–11.
- Lan, Y.; Meng, X.; Yang, S.; Loy, C. C.; and Dai, B. 2023. Self-supervised geometry-aware encoder for style-based 3D GAN inversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20940–20949.
- Li, T.; Bolkart, T.; Black, M. J.; Li, H.; and Romero, J. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Trans. Graph.*, 36(6): 194–1.
- Liu, F.-L.; Chen, S.-Y.; Lai, Y.-K.; Li, C.; Jiang, Y.-R.; Fu, H.; and Gao, L. 2022. Deepfacevideoediting: Sketch-based deep editing of face videos. *ACM Transactions on Graphics (TOG)*, 41(4): 1–16.
- Ma, S.; Weng, Y.; Shao, T.; and Zhou, K. 2024. 3d gaussian blendshapes for head avatar animation. In *ACM SIGGRAPH 2024 Conference Papers*, 1–10.

- Marcenaro, L.; Vernazza, G.; and Regazzoni, C. S. 2001. Image stabilization algorithms for video-surveillance applications. In *Proceedings 2001 International Conference on Image Processing*, volume 1, 349–352.
- Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.
- Morimoto, C.; and Chellappa, R. 1998. Evaluation of image stabilization algorithms. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 5, 2789–2792.
- Ojha, U.; Li, Y.; Lu, J.; Efros, A. A.; Lee, Y. J.; Shechtman, E.; and Zhang, R. 2021. Few-shot image generation via cross-domain correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10743–10752.
- Or-El, R.; Luo, X.; Shan, M.; Shechtman, E.; Park, J. J.; and Kemelmacher-Shlizerman, I. 2022. Stylesdf: High-resolution 3d-consistent image and geometry generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13503–13513.
- Patashnik, O.; Wu, Z.; Shechtman, E.; Cohen-Or, D.; and Lischinski, D. 2021. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2085–2094.
- Pinkney, J. N.; and Adler, D. 2020. Resolution dependent gan interpolation for controllable image synthesis between domains. In *NeurIPS Workshop on Machine Learning for Creativity and Design*.
- Qian, S.; Kirschstein, T.; Schoneveld, L.; Davoli, D.; Giebenhain, S.; and Nießner, M. 2024. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20299–20309.
- Richardson, E.; Alaluf, Y.; Patashnik, O.; Nitzan, Y.; Azar, Y.; Shapiro, S.; and Cohen-Or, D. 2021. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2287–2296.
- Shao, Z.; Wang, Z.; Li, Z.; Wang, D.; Lin, X.; Zhang, Y.; Fan, M.; and Wang, Z. 2024. Splattingavatar: Realistic real-time human avatars with mesh-embedded gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1606–1616.
- Tov, O.; Alaluf, Y.; Nitzan, Y.; Patashnik, O.; and Cohen-Or, D. 2021. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4): 1–14.
- Tzaban, R.; Mokady, R.; Gal, R.; Bermano, A.; and Cohen-Or, D. 2022. Stitch it in time: Gan-based facial editing of real videos. In *SIGGRAPH Asia 2022 Conference Papers*, 1–9.
- Wang, T.; Zhang, B.; Zhang, T.; Gu, S.; Bao, J.; Baltrusaitis, T.; Shen, J.; Chen, D.; Wen, F.; Chen, Q.; et al. 2023. Rodin: A generative model for sculpting 3d digital avatars using diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4563–4573.
- Wang, T.; Zhang, Y.; Fan, Y.; Wang, J.; and Chen, Q. 2022. High-fidelity gan inversion for image attribute editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11379–11388.
- Wu, J.; Zhang, C.; Xue, T.; Freeman, B.; and Tenenbaum, J. 2016. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. *Advances in neural information processing systems*, 29.
- Xiang, J.; Gao, X.; Guo, Y.; and Zhang, J. 2024. Flashavatar: High-fidelity head avatar with efficient gaussian embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1802–1812.
- Xu, J.; Chang, H.-w.; Yang, S.; and Wang, M. 2012. Fast feature-based video stabilization without accumulative global motion estimation. *IEEE Transactions on Consumer Electronics*, 58(3): 993–999.
- Xu, Y.; Chen, B.; Li, Z.; Zhang, H.; Wang, L.; Zheng, Z.; and Liu, Y. 2024. Gaussian head avatar: Ultra high-fidelity head avatar via dynamic gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1931–1941.
- Xu, Y.; Wang, L.; Zhao, X.; Zhang, H.; and Liu, Y. 2023. Avatarmav: Fast 3d head avatar reconstruction using motion-aware neural voxels. In *ACM SIGGRAPH 2023 Conference Papers*, 1–10.
- Yang, S.; Jiang, L.; Liu, Z.; and Loy, C. C. 2022a. Pastiche master: Exemplar-based high-resolution portrait style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7693–7702.
- Yang, S.; Jiang, L.; Liu, Z.; and Loy, C. C. 2022b. Vtoonify: Controllable high-resolution portrait video style transfer. *ACM Transactions on Graphics (TOG)*, 41(6): 1–15.
- Yang, S.; Jiang, L.; Liu, Z.; and Loy, C. C. 2023. Styleganex: Stylegan-based manipulation beyond cropped aligned faces. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 21000–21010.
- Yao, X.; Newson, A.; Gousseau, Y.; and Hellier, P. 2021. A latent transformer for disentangled face editing in images and videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13789–13798.
- Yenamandra, T.; Tewari, A.; Bernard, F.; Seidel, H.-P.; Elgharib, M.; Cremers, D.; and Theobalt, C. 2021. i3dmm: Deep implicit 3d morphable model of human heads. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12803–12813.
- Zhang, B.; Cheng, Y.; Wang, C.; Zhang, T.; Yang, J.; Tang, Y.; Zhao, F.; Chen, D.; and Guo, B. 2024. Rodinhd: High-fidelity 3d avatar generation with diffusion models. In *European Conference on Computer Vision*, 465–483. Springer.
- Zhang, J.; Lan, Y.; Yang, S.; Hong, F.; Wang, Q.; Yeo, C. K.; Liu, Z.; and Loy, C. C. 2023. DeformToon3D: Deformable 3D Toonification from Neural Radiance Fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9110–9120.
- Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as

a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 586–595.

Zheng, Y.; Abrevaya, V. F.; Bühlert, M. C.; Chen, X.; Black, M. J.; and Hilliges, O. 2022. Im avatar: Implicit morphable head avatars from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13545–13555.

Zheng, Y.; Yifan, W.; Wetzstein, G.; Black, M. J.; and Hilliges, O. 2023. Pointavatar: Deformable point-based head avatars from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21057–21067.

Zielonka, W.; Bolkart, T.; and Thies, J. 2022. Towards metrical reconstruction of human faces. In *European Conference on Computer Vision*, 250–269.

Zielonka, W.; Bolkart, T.; and Thies, J. 2023. Instant volumetric head avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4574–4584.