# StyleAvatar: Real-time Photo-realistic Portrait Avatar from a Single Video


Lizhen Wang
Tsinghua University & NNKosmos Technology
Beijing & Hangzhou, China
wlz18@mails.tsinghua.edu.cn

Xiaochen Zhao
Tsinghua University
Beijing, China
zhaoxc19@mails.tsinghua.edu.cn

Jingxiang Sun
Tsinghua University
Beijing, China
sunjingxiang_stark@126.com

Yuxiang Zhang
Tsinghua University
Beijing, China
yx-z19@mails.tsinghua.edu.cn

Hongwen Zhang
Tsinghua University
Beijing, China
zhanghongwen@tsinghua.edu.cn

Tao Yu
Tsinghua University
Beijing, China
ytrock@mail.tsinghua.edu.cn

Yebin Liu
Tsinghua University
Beijing, China
liuyebin@mail.tsinghua.edu.cn


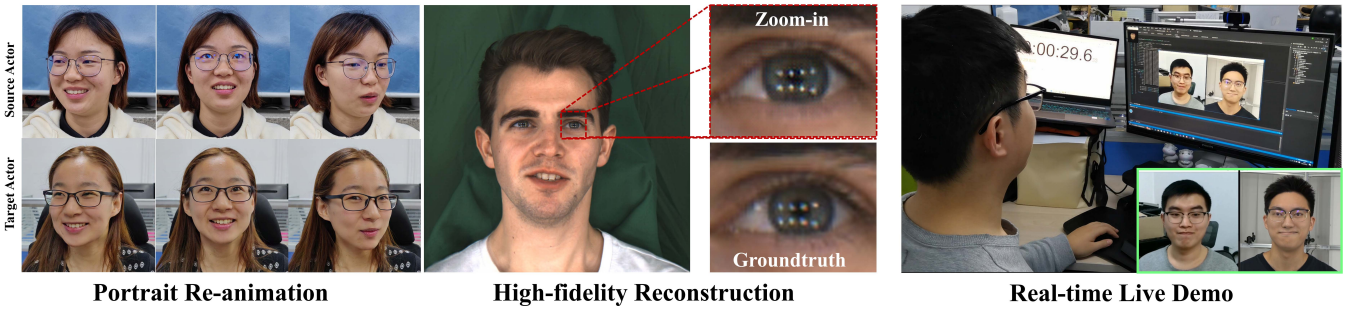


**Figure 1: We present StyleAvatar, a real-time high-fidelity portrait avatar reconstruction method.**


## ABSTRACT

Face reenactment methods attempt to restore and re-animate portrait videos as realistically as possible. Existing methods face a dilemma in quality versus controllability: 2D GAN-based methods achieve higher image quality but suffer in fine-grained control of facial attributes compared with 3D counterparts. In this work, we propose StyleAvatar, a real-time photo-realistic portrait avatar reconstruction method using StyleGAN-based networks, which can generate high-fidelity portrait avatars with faithful expression control. We expand the capabilities of StyleGAN by introducing a compositional representation and a sliding window augmentation method, which enable faster convergence and improve translation generalization. Specifically, we divide the portrait scenes into three parts for adaptive adjustments: facial region, non-facial foreground region, and the background. Besides, our network leverages the best of UNet, StyleGAN and time coding for video learning, which enables high-quality video generation. Furthermore, a sliding window augmentation method together with a pre-training strategy are proposed to improve translation generalization and training performance, respectively. The proposed network can converge within two hours while ensuring high image quality and a forward rendering time of only 20 milliseconds. Furthermore, we propose a real-time live system, which further pushes research into applications. Results and experiments demonstrate the superiority of our method in terms of image quality, full portrait video generation, and real-time re-animation compared to existing facial reenactment methods. Training and inference code for this paper are at https://github.com/LizhenWangT/StyleAvatar.




*SIGGRAPH '23 Conference Proceedings, August 6–10, 2023, Los Angeles, CA, USA*



ACM ISBN 979-8-4007-0159-7/23/08.
https://doi.org/10.1145/3588432.3591517



## CCS CONCEPTS

• **Computing methodologies** → *Motion processing*.

## KEYWORDS

Facial Reenactment, StyleGAN, Video Portraits, Deep Learning, Rendering-to-Video Translation.

**ACM Reference Format:**
Lizhen Wang, Xiaochen Zhao, Jingxiang Sun, Yuxiang Zhang, Hongwen Zhang, Tao Yu, and Yebin Liu. 2023. StyleAvatar: Real-time Photo-realistic Portrait Avatar from a Single Video. In *Special Interest Group on Computer Graphics and InteractiveTechniques Conference Conference Proceedings (SIGGRAPH '23 Conference Proceedings), August 6–10, 2023, Los Angeles, CA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3588432.3591517

## 1 INTRODUCTION

Photo-realistic portrait avatar reconstruction and re-animation is a long standing topic in computer vision and computer graphics, with a wide range of applications from video editing to mixed reality. Recent efforts on portrait avatar based on NeRF [Gafni et al. 2021; Gao et al. 2022; Zheng et al. 2023] or some other 3D representations present [Grassal et al. 2022a; Zheng et al. 2022a,b; Zielonka et al. 2022] have demonstrated that stable 3D avatars can be learned from monocular videos. However, compared to 2D GAN-based methods, most 3D approaches still face limitations in resolution and image quality. Additionally, these methods primarily focus on the rigid facial regions, ignoring long hair, body parts, and background elements. While backgrounds can be overlaid directly, poorly executed overlays often lead to unrealistic results. The ideal portrait avatar should prioritize high-fidelity, fast training, fine-grained control, and real-time efficiency.

Learning 3D head avatars from monocular videos has been a popular topic in recent years. Early works [Gafni et al. 2021; Grassal et al. 2022a; Zheng et al. 2022a] incorporate NeRF into head avatars, achieving promising view consistency. More recent approaches either aim to achieve even better rendering realism [Xu et al. 2023b; Zheng et al. 2022b] or faster training convergence and inference speed [Gao et al. 2022; Xu et al. 2023a; Zielonka et al. 2022] by utilizing more efficient 3D representations [Fang et al. 2022; Müller et al. 2022]. In general, the core idea of 3D methods is to maintain a relatively fixed feature space, such as topology-consistent meshes or a canonical space, to enable each point or voxel to learn certain local features from the video. This strategy leads to greater stability, but also results in smoothed textures due to tracking instability or other factors.

On the other end of the spectrum, benefiting from the powerful StyleGAN [Karras et al. 2021, 2019, 2020], following works [Abdal et al. 2021; Chen et al. 2022; Deng et al. 2020; Härkönen et al. 2020; Shen et al. 2020b; Tewari et al. 2020a,b; Wang et al. 2021b] have drastically improved the semantic editing performance. Some methods [Doukas et al. 2021b; Drobyshev et al. 2022; Khakhulin et al. 2022] can create head avatars from a single image, while others [Sun et al. 2022, 2023; Xiang et al. 2022] can even produce controllable 3D faces with EG3D [Chan et al. 2022]. Nevertheless, these StyleGAN-based methods mostly rely on a highly aligned HD face dataset such as FFHQ, which lacks sufficient variation in facial expressions. Additionally, they can not generate natural head movements in portrait videos.

We propose StyleAvatar, a real-time system for photo-realistic portrait avatar reconstruction using a StyleGAN-based network. Our system is capable of generating a high-fidelity portrait avatar in just three hours. To address the challenges of full photo-realistic portrait video reconstruction, we divide the portrait scenes into three parts: the face, movable body parts (shoulders, neck, and hair), and background. Each part has distinct attributes: the face part provides almost all the moving information through 3DMM, while the background is typically static. The movable body part may contain numerous uncontrollable movements, but we can still learn some trends from facial movements.

To overcome these challenges, we propose StyleAvatar, a real-time system for photo-realistic portrait avatar reconstruction using a StyleGAN-based network. Our system can generate a high-fidelity portrait avatar in just two hours. In order to reconstruct a full portrait video, we divide the video into three parts: facial region, non-facial foreground region (shoulders, neck, hair etc.) and background. Each part has distinct attributes: the facial region can be described by the 3DMM; the non-facial foreground region often exhibits uncontrollable movements, but trends can be learned from facial movements; and the background remains static. To better represent the distinct features of the three parts, we use two StyleGAN generators to generate two static feature maps for the facial region and background, and propose a StyleUNet to generate the non-facial foreground feature map from the input 3DMM rendering. To accelerate the training and inference speed, we use use Neural Textures [Thies et al. 2019] for the facial region during the feature combination stage. Moreover, a sliding window augmentation method is introduced to improve translation generalization and we pre-train the model on a small video dataset to further speed up training. Finally, another StyleUNet is used to generate the final images from the combined feature maps. Our framework is designed to be easily accelerated by TensorRT and OpenGL, with a forward rendering time of only 20 milliseconds, enabling real-time live portrait reenactment. Results and experiments demonstrate that our method outperforms existing facial reenactment methods in terms of image quality, full portrait video generation, and real-time re-animation. Our contributions can be concluded as:

- We introduce a compositional representation that effectively decomposes the facial region, the non-facial foreground region and the background, so that we can make adaptive adjustments according to the characteristics of different regions to increase the stability and the training speed.
- We further propose StyleUNet that leverages the best of UNet, StyleGAN and time coding for video learning, which enables high-quality video generation.
- A sliding window augmentation method together with a pre-training strategy are proposed to improve translation generalization and training performance, respectively.

## 2 RELATED WORKS

*Facial Reenactment Methods.* These methods aim at generating photo-realistic portrait images (including face, hair, neck and even shoulder regions) of a target person given the performance of another person, which is different from face replacement methods [Perov et al. 2020] or face performance capture and animation methods [Li et al. 2012; Weise et al. 2011]. According to the input data, facial reenactment methods can be roughly divided into three categories: multi-view system based methods, single video based methods and single image based methods. Based on multi-view capture systems, recent researches [Lombardi et al. 2018, 2019; Ma et al. 2021; Raj et al. 2021; Wang et al. 2022a; Wei et al. 2019] were able

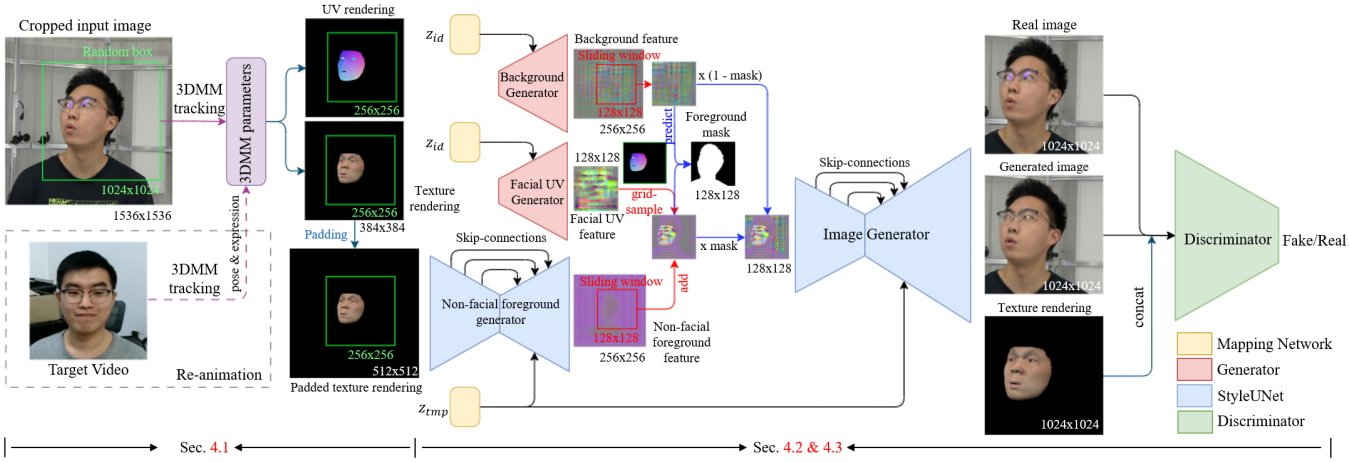


**Figure 2: Our portrait avatar reconstruction and re-animation pipeline consists of three main steps: 3DMM tracking and rendering, feature generation of the facial region, non-facial foreground region, and background, and final image generation from the combined feature map. To achieve this, we utilize two StyleGAN generators, a StyleGAN discriminator, and two StyleUNets. Additionally, we incorporate data augmentation techniques with random boxes on input images and sliding windows on generated feature maps to improve translation generalization.**

to generate facial avatars with impressive subtle details and highly flexible controllability for immersive metric-telepresence. However, difficulties in data acquisition limited the broad applications. On contrast, single image-based methods [Averbuch-Elor et al. 2017; Doukas et al. 2021b; Drobyshev et al. 2022; Geng et al. 2018; Hong et al. 2022; Kowalski et al. 2020; Liu et al. 2001; Mallya et al. 2022; Nagano et al. 2018; Olszewski et al. 2017; Siarohin et al. 2019; Vlasic et al. 2005; Yin et al. 2022] were most easy to capture and could produce photo-realistic facial reenactment results. However, shapes and details may not be consistent when animating to large poses and expressions especially for those regions that have not been well covered in the single input image, not to say the dynamic facial details. Single video based methods [Doukas et al. 2021a; Garrido et al. 2014; Koujan et al. 2020; Suwajanakorn et al. 2017; Thies et al. 2015, 2016] showed more stable facial reenactment results. Among them, [Kim et al. 2018] presented impressive full head reenactment and interactive editing results based on an image2image translation framework. Recent methods [Gafni et al. 2021; Gao et al. 2022; Wang et al. 2021a; Xu et al. 2023a,b; Zielonka et al. 2022] showed state-of-the-art reenactment results with a parameter-controlled neural radiance field. Other methods [Cao et al. 2022; Garbin et al. 2022; Grassal et al. 2022b; Zheng et al. 2022a,b] improved stability and texture quality by utilizing 3D representations like meshes or point clouds. However, existing single-video-based facial reenactment methods still face challenges in recovering elaborate details such as hairs, freckles, and even skin pores.

*StyleGAN-based Facial Image Generation and Editing Methods.* These methods can produce high-resolution and photo-realistic facial images [Karras et al. 2021, 2019, 2020] even under semantic editing operations [Abdal et al. 2021; Alaluf et al. 2021; Chen et al. 2022; Deng et al. 2020; Ghosh et al. 2020; Härkönen et al. 2020; Jang et al. 2021; Ren et al. 2021; Richardson et al. 2021; Shen et al. 2020a,b; Shi et al. 2022; Tewari et al. 2020a,b; Tov et al. 2021; Wang et al. 2021b]. These works conducted semantic modifications on the generated images of StyleGAN by decoupling the input latent space. Moreover, Ghosh *et, al.* [Ghosh et al. 2020] and Shoshan *et, al.* [Shoshan et al. 2021] combined the StyleGAN latent space with the semantic input, such as 3DMM parameters [Blanz and Vetter 1999]. SofGAN [Chen et al. 2022] used semantic segmentation maps to condition the image generator and achieves 3D aware generation using 3D scans additionally for training. These methods provide meaningful control over the shape, pose, hair and style of the generated photo-realistic facial images, but fine-grained expression modifications such as blink remain unavailable. More importantly, these methods heavily rely on a pre-trained StyleGAN latent space, making it difficult for them to maintain temporal stability and consistency of details during the editing process.

## 3 METHOD

As shown in Fig. 2, input with a monocular portrait video, we first perform 3DMM tracking (Sec. 3.1) to generate synthetic renderings with both predicted texture and UV coordinate vertex colors. Next, in the feature generation stage (Sec. 3.2), we divide a portrait feature map into three parts: a static facial feature map generated in UV space generated by a StyleGAN generator; a non-facial foreground feature map generated from the input texture rendering by a StyleUNet; a static background feature map generated by another StyleGAN generator. Then, we employ Neural Textures to extract a facial feature map from the UV space, and introduce a sliding window data augmentation during the combination of feature maps to better utilize information from the entire video. Finally, another StyleUNet is used to generate images from the combined feature maps and a StyleGAN discriminator is introduced for adversarial learning. Note that only the two StyleUNets are computed in the inference stage, and re-animation can be achieved by replacing the input expression and pose parameters from another video.

### 3.1 Data Processing

We first perfrom 3DMM tracking on the input monocular portrait video to generate pixel-aligned 3DMM renderings for subsequent training. We choose to use FaceVerse [Wang et al. 2022b] due to its rich shape and expression bases and we add separate eyeballs. A texture rendering with predicted texture is used for non-facial foreground feature generation and another UV rendering with UV coordinate vertex colors is used for the Neural Textures of the facial region. To supervise the training of mask prediction in the feature combination stage, we generate foreground masks using Robust Video Matting [Lin et al. 2021].

For the 3DMM tracking algorithm, we need to solve for shape coefficients $\theta_{shape}$, expression coefficients $\theta_{expression}$, texture coefficients $\theta_{texture}$, translation $t$, scale $s$, and rotations of the head and two eyeballs $R_1, R_2, R_3$. To improve efficiency, we directly solve for the analytical solutions of these parameters from facial landmarks $K_{tgt}$ detected by MediaPipe[1]. Specifically, we utilize the following energy functions:

$$\arg\min_{R_x,t,s} \|R_x K_{src} + t - K_{tgt}/s\|_2 \tag{1}$$

$$\arg\min_{\delta\theta_{shape},\delta\theta_{exp}} \|R_1(K_{src}+B_{shape}\delta\theta_{shape}+B_{exp}\delta\theta_{exp})+t-K_{tgt}/s\|_2 \tag{2}$$

$$\arg\min_{\delta\theta_{tex}} \|T_{src} + B_{tex}\delta\theta_{tex} - T_{tgt}\|_2 \tag{3}$$

where $Ksrc$ represents the corresponding landmarks on FaceVerse, while $Bshape, B_{exp}, B_{tex}$ represent the shape, expression, and texture bases. To obtain the final parameters, we solve Eq. 1 and Eq. 2 iteratively. Eq. 1 provides us with the rotation, translation, and scale parameters, whereas Eq. 2 gives us the shape and expression coefficients of FaceVerse. Additionally, we use Eq. 3 to solve for the texture coefficients. Eq. 2 and Eq. 2 can be solved by LDLT decomposition, while Eq. 1 can be solved by SVD decomposition. Since we require the predicted shape to be consistent for the same person, and the predicted texture is not critical for our framework, we only solve for the shape and texture coefficients in the first frame.

To generate the UV rendering, we use UV coordinate vertex colors, where the red and green channel pixel values correspond to the x and y position in the UV coordinate. To generate the texture rendering, we utilize the predicted texture and render it with a pre-defined fixed lighting in order to accentuate facial expression changes. Furthermore, in order to better utilize information from the entire video, we propose a sliding window data augmentation approach. Specifically, we enlarge the crop box of the portrait region by 1.5 times, allowing for the preservation of more upper body and background information, as shown in Fig. 2.

### 3.2 Feature Generation and Combination

In the feature generation stage, to create high-quality avatars, we divide the portrait region into three parts and employ StyleGAN-based networks to generate corresponding feature maps. Specifically, we use two StyleGAN generators to generate static feature maps for the facial region and background. The facial feature maps are generated in UV space, and the input identity latent code $z_{id}$ is designed for pre-training, allowing us to generate different feature maps for different videos. For the non-facial foreground region, we propose a StyleUNet to generate the pixel-aligned feature map from the input 3DMM rendering. We believe that plausible movements of this part can be predicted from the movements of 3DMM. To accommodate uncontrollable changes like hair motion, we use the input temporal latent code $z_{tmp}$ as an additional input. We follow Bahmani et al. [Bahmani et al. 2022] in using positional embedding to map person identities and timestamps to higher dimensions, so that they can be input into our network.

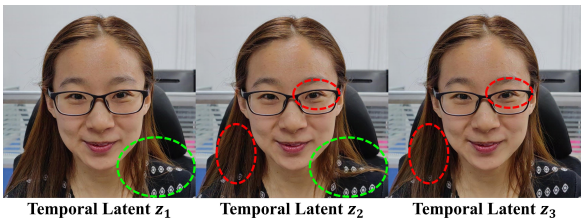


**Figure 3: The information stored in our temporal latent space allows for changes in details such as hair changes when inputting the same rendering with different temporal latent code $z_{tmp}$.**

In the feature combination stage, we combine the three feature maps to generate the final image. For the facial region, the 3DMM already provides the basic geometry, so we adopt the Neural Textures proposed in Deferred Neural Rendering (DNR) [Thies et al. 2019] to generate a facial feature map by sampling from the facial UV feature map using the UV rendering. As for the background and non-facial foreground regions, we crop the generated feature maps using sliding windows to ensure the pixel-aligned relationship between the input feature maps and the output images. To generate dynamic facial changes and protruding shapes such as glasses, we directly add the facial feature map and the non-facial foreground feature map. To ensure that the foreground feature map is in front of the background feature map, we first use a supervised convolutional layer to predict a foreground mask from the foreground and background feature maps, and then combine these two feature maps using the predicted mask. Finally, we use another StyleUNet to generate the final image from the combined feature map and the networks are trained adversarially with a StyleGAN discriminator. We also incorporate the temporal latent code $z_{tmp}$ as an additional input to our StyleUNet. As depicted in Fig. 3, varying $z_{tmp}$ values result in time-related changes such as hair movement. The use of a mapping network and $z_{tmp}$ helps prevent overly smoothed details by incorporating time-related but uncontrollable changes into the latent space. Note that only the two StyleUNets will be computed during the inference stage.

To overcome the challenge of handling free translational motions in 2D avatars, we propose a sliding window data augmentation method to better utilize the information from the entire input video. As demonstrated in Fig. 4a, the use of a simple UNet-based 2D avatar often leads to noticeable visual artifacts, such as image tearing, particularly in cases of excessive head translations. This can be attributed to two factors: firstly, UNets are known to be highly sensitive to image translations; secondly, the central area of the

[1]https://github.com/google/mediapipe

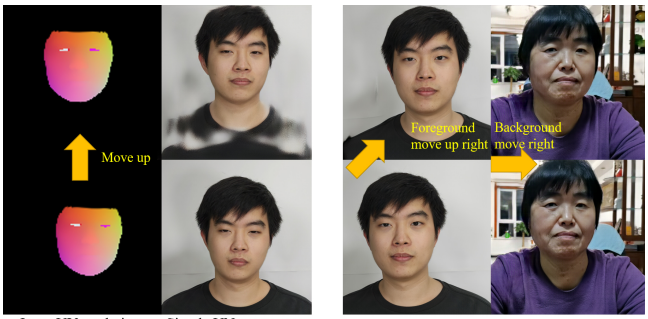


**Figure 4: Free movement in a large range may tear the generated image using a simple UNet. By introducing the sliding windows our avatar can move in a larger range.**

input portrait video tends to have more pronounced pose and expression changes, whereas the border area shows less variation, leading to a reduced ability to generalize translations. To enhance the translation generalization ability, we first employ random sample boxes on both the input images and their corresponding 3DMM renderings. To ensure pixel-aligned features in the non-facial foreground region, we pad the sampled input rendering to a higher resolution (256 to 512). Then, to extract pixel-aligned background and non-facial foreground regions from the generated feature maps, we utilize two sliding windows. The position of these sliding windows is determined by the input random sample boxes. Finally, we concatenate the pixel-aligned texture renderings with the generated images to create fake input images for our discriminator, and concatenate the texture renderings with the input images inside the sample boxes to create real input images for our discriminator. As shown in Fig. 4b, with the aid of the proposed data augmentation method, our approach is capable of addressing large head translations and can generate both foreground and background movements by employing different sliding windows during feature combination.

## 3.3 Network Structure and Loss Functions

As shown in Fig. 2, our framework consists of five networks: two feature generators, two StyleUNets, and a discriminator. To meet real-time requirements and speed up training, we employ wavelet transform in all networks similar to SWAGAN [Gal et al. 2021], which enables us to replace a 1024×1024×3 image with a 512×512×12 representation. Apart from this, the remaining parts of our generator and discriminator follow StyleGAN2 architecture. The StyleUNet uses an encoder-decoder structure, designed for image-to-image translation, with additional input latent code and noise. Similar to UNet [Ronneberger et al. 2015], our encoder extracts multi-scale features that are sent to the decoder via skip-connections. We use a mapping network, the modulated convolution and the temporal latent code to introduce additional variation possibilities for the network, preventing it from smoothing certain features. For example, the uncontrolled hair motions can be accommodated by the time-related latent code. However, this approach cannot guarantee the complete stability of the hair. We use the 64-dimensional latent code as the input of mapping networks in both StyleUNets and generators. As described in [Karras et al. 2021], the noise injection layers of StyleGAN2 can lead to texture sticking artifacts. To mitigate this effect, we use UV renderings to map the noise from UV space to the face regions, and a fixed noise is used for the static background. To incorporate facial priors into the discriminator, we concatenate the texture rendering with the output image, which serves as the input to the discriminator.

In terms of loss functions, as direct supervision is feasible for our task, we utilize common L1 loss and perceptual loss with a VGG19 during the training process. Additionally, we incorporate an L1 loss for the foreground mask. We also include GAN loss for adversarial learning. Our loss functions can be formulated as

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_{percep} + \mathcal{L}_{mask} + \mathcal{L}_{GAN} \tag{4}$$

**Table 1: Quantitative comparisons with one-shot avatar methods.**

| Error Metric | SSIM ↑ | PSNR ↑ | FID ↓ |
|---|---|---|---|
| DaGAN | 0.79 | 22.0 | 73.2 |
| Next3D-single | 0.81 | 24.6 | 24.4 |
| Next3D-refine | 0.79 | 23.5 | 21.4 |
| Ours | **0.87** | **27.1** | **12.2** |

## 3.4 Pre-training and Live System

To expedite training convergence, we use 6 videos cropped from 4K videos for pre-training. As discussed in Sec. 3.2, we use distinct identity latent codes $z_{id}$ for each video. Despite the limited dataset, pre-training has proven effective, as demonstrated in Sec. 4.2. Note that we assume a fixed video length for $z_{tmp}$ of each video and select a fixed $z_{tmp}$ during the inference stage.

To bring our work closer to practical applications, we present a real-time live system consisting of three main steps: 3DMM tracking, OpenGL rendering for input 3DMM renderings, and the two StyleUNets which have been converted to TensorRT models. Our system can run at 35 fps (28 ms per frame) using a 16-bit TensorRT model on a PC with one RTX 3090 GPU, and requires approximately 4 GB of GPU memory during the inference stage. The 16-bit model takes an average of 20 milliseconds GPU time, while the original PyTorch model takes an average of 31 milliseconds. To generate a realistic facial avatar, we need approximately two minutes of video footage featuring a range of head poses and facial expressions. The videos used in this paper include the Obama video courtesy of the White House (public domain), a video provided by IMAvatar, and videos from the MEAD dataset [Wang et al. 2020]. We obtained consent from all actors featured in the remaining videos.

# 4 EXPERIMENTS

## 4.1 Comparisons

We first compare our method to state-of-the-art one-shot facial methods based on the GAN structure to demonstrate our network's ability to achieve fine-grained control of expressions and facial details. For comparison, we select two representative approaches: DaGAN [Hong et al. 2022] as a representation of 2D one-shot avatars

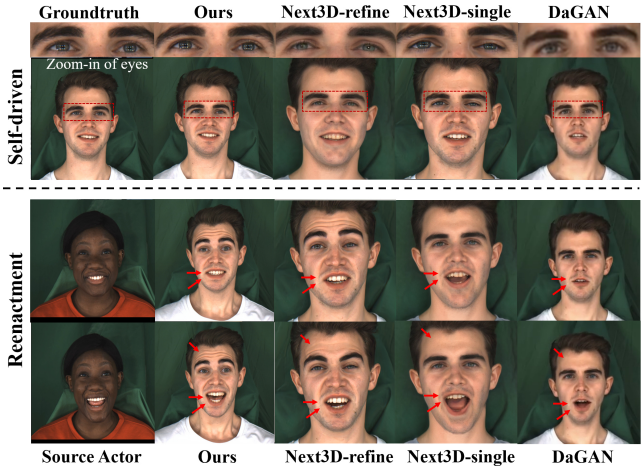


**Figure 5: Self-driven animation and reenactment comparisons with one-shot portrait avatar methods.**

**Table 2: Quantitative comparisons with video-based avatar methods.**

| | Case 1 | | | Case 2 | | |
|---|---|---|---|---|---|---|
| Error Metric | SSIM ↑ | PSNR ↑ | FID ↓ | SSIM ↑ | PSNR ↑ | FID ↓ |
| DVP | 0.80 | 21.6 | 31.3 | 0.85 | 24.2 | 25.9 |
| NeRFace | 0.77 | 18.8 | 36.4 | 0.87 | 22.2 | 50.0 |
| NHA | 0.73 | 16.0 | 83.5 | 0.86 | 19.5 | 62.9 |
| IMAvatar | 0.78 | 19.0 | 59.7 | **0.89** | 25.6 | 58.7 |
| Ours | **0.87** | **25.6** | **15.1** | 0.87 | **26.2** | **13.2** |

and Next3D [Sun et al. 2023] as a representation of 3D one-shot methods. To ensure a fair experiment, we fine-tune Next3D on the monocular video used for the comparison for 24 hours, denoted as "Next3D-refine", while "Next3D-single" refers to the original pre-trained model. As shown in Fig. 5 and Tab. 1, our method outperforms DaGAN and both versions of Next3D in both image quality and control of facial attributes. Although Next3D-refine improves fidelity after fine-tuning on the video, it still cannot achieve fine-grained control of facial expressions. These results suggest that video-based training is still necessary for high-fidelity portrait avatars, and our network structure shows better performance in video-based training. Note that in order to obtain the values presented in Tab. 1, we have first aligned the faces, cropped the images, removed the background, and resized them to a resolution of $512 \times 512$. The values presented in Tab. 1 are calculated based on self-driving images generated from the testing set, which is another video in the MEAD dataset, and the corresponding ground-truth images.

We compare our method with state-of-the-art video-based facial reenactment methods, including Deep Video Portrait (DVP) [Kim et al. 2018], NeRFace [Gafni et al. 2021], IMAvatar [Zheng et al. 2022a], and Neural Head Avatar (NHA) [Grassal et al. 2022a], all of which are trained on monocular videos. We have partitioned 80% of the video into a training set and the remaining 20% into

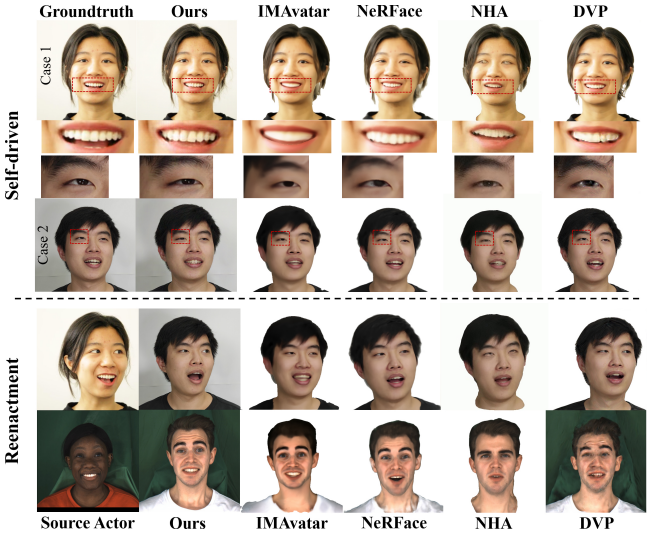


**Figure 6: The comparisons with video-based facial reenactment methods on self-driven re-animation and faical reenactment. Our method can generate more realistic details (e.g. teeth, light points in eyes).**

a testing set for evaluation. Additionally, we have removed the background and resized the images to a resolution of $512 \times 512$. The values presented in Tab. 2 are calculated based on the self-driving images and their corresponding ground-truth images in the testing set. We also perform self-driven re-animation and reenactment for these methods, as shown in Fig. 6. While the 3D methods show stable head geometry, they fail to produce high-fidelity texture details such as hair, teeth, and pupils in the output renderings. By incorporating StyleGAN-based networks and using our data augmentation, our method achieves higher image quality than the existing methods and can preserve more details such as light points in the eyes. As shown in Tab. 2, our method achieves significantly better quantitative results, particularly in the FID metric, indicating that our method generates higher-quality images.

## 4.2 Ablation Study

In our comparisons, we have demonstrated the powerful image generation capabilities of our network, which benefit from the StyleGAN-based StyleUNet structure. However, it should be noted that a simple StyleUNet may not be able to achieve translation generalization, and may require a long time to converge. In the ablation study, we use the term "DVP (Ours)" to refer to a simple UNet input with 3DMM texture rendering, similar to the DVP structure, and "Single-StyleUNet" to refer to a StyleUNet input with 3DMM texture rendering. "DNR (Ours)" represents a comparison with Deferred Neural Rendering [Thies et al. 2019], which only uses the Neural Textures of the facial region as the input of a UNet. Additionally, we include "Nearest Sample" in our training set and "Ours w/o NT", which refers to our method without the Neural Textures. As shown in Fig. 7 and Tab. 3, without our data augmentation and video decomposition, even though "Single-StyleUNet" can still generate high-fidelity images, it is unable to prevent image tearing during

**Table 3: Quantitative comparisons of our ablation study.**

| Error Metric | SSIM ↑ | PSNR ↑ | FID ↓ |
|---|---|---|---|
| DVP (Ours) | 0.85 | 24.2 | 25.9 |
| Single-StyleUNet | 0.87 | 26.2 | 15.05 |
| DNR (Ours) | **0.89** | 26.1 | 27.2 |
| Nearest Sample | 0.81 | 22.1 | 18.3 |
| Ours w/o NT | 0.88 | **26.6** | 15.10 |
| Ours | 0.87 | 26.2 | **13.2** |

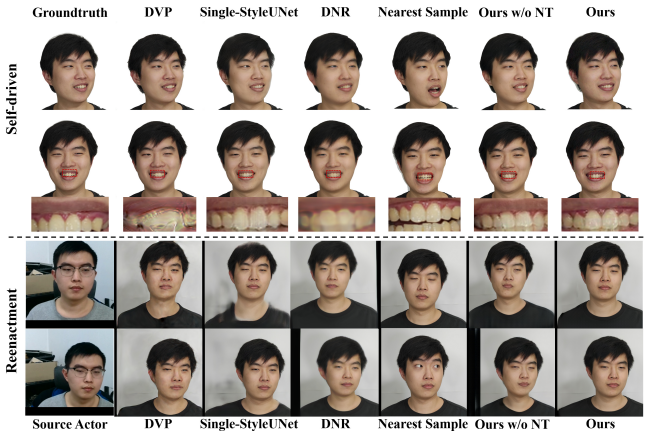


**Figure 7: Ablation study of DVP, DNR, our "single-StyleUNet", "nearest sample", "ours w/o NT" and our full method. Our full method is superior in image quality and translation generalization.**

significant head movements. Both "DVP (Ours)" and "DNR (Ours)" exhibit a noticeable decrease in image quality, highlighting the importance of StyleUNet in generating high-fidelity images. "Ours w/o NT" can still produce high-quality images, suggesting that the Neural Textures primarily contribute to speeding up the training convergence. Note that the cases 1 and 2 presented in Tab. 3 have been marked in the self-driven portion of Fig. 7.

To validate the generalization ability of our method, we present the "Nearest Sample" in Fig. 7 and calculate the average nearest translation, rotation, and expression parameters for both self-driven and reenactment cases, as shown in Tab. 4. It should be noted that we first normalize all parameters using the standard deviation calculated from the training set. The images in Fig. 7 are selected with equal weight given to translation, rotation, and expression parameters, but the values in Tab. 4 are calculated separately for each parameter. The results demonstrate that our method can achieve larger extrapolation on translation while performing similarly to other state-of-the-art methods on expression and rotation. Additionally, our method can only handle rotations up to approximately 30 degrees due to the challenges that large rotations present for face tracking.

As illustrated in Fig.8, due to the video decomposition and Neural Textures, our method achieves significantly faster training speed

**Table 4: The average nearest parameters are presented for both the testing set and the reenactment case. Note that all parameters are normalized by the standard deviation calculated in the training set.**

| Parameter | Translation ↑ | Rotation ↑ | Expression ↓ |
|---|---|---|---|
| Testing Set | 0.052 | 0.058 | 0.316 |
| Reenactment | 2.002 | 0.054 | 0.294 |


**Figure 8: Generated images of different network structures during the training stage. Our method can achieve significantly faster training speed.**

even without pre-training. The fourth column of "Ours w/o pretrain" can learn the basic information of the background and facial region faster than the third column. Only the remaining regions such as shoulders, hair, and small parts such as eyeglasses and teeth require further training. Comparing "Ours" with "Ours w/o pretrain" demonstrates the effectiveness of pre-training. Note that the video in Fig.8 was not used in pre-training.

## 4.3 Training and Rendering Efficiency

As shown in Tab. 5, our method is significantly faster in training and rendering, while also generating images at a higher resolution. This can be attributed to our use of video decomposition, pre-training and a simple network structure that can be easily accelerated by TensorRT. Additionally, we use the vertex color of 3DMM points to represent their corresponding UV coordinates, allowing for real-time rendering using OpenGL. It should be noted that after just two hours of training, our model has already achieved a visually pleasing result, with only the teeth region requiring further training (approximately another 4 hours) to reach convergence.

# 5 DISCUSSION AND CONCLUSION

*Limitations.* The proposed StyleAvatar outperforms state-of-the-art facial reenactment methods, but still has some limitations. First, our image-to-image translation network is limited by the quality

**Table 5: Training time, rendering time and image resolution of each method. Compared to the existing methods, our method is significantly faster in training and is able to render images at a higher resolution in real-time.**

| Method | Training time (hour) | Rendering time per frame (s) | Resolution |
|---|---|---|---|
| IMAvatar | 48 | 100 | 512 |
| NeRFace | 36 | 4 | 512 |
| NHA | 13 | 0.06 | 512 |
| Ours | 2 | 0.028 | 1024 |

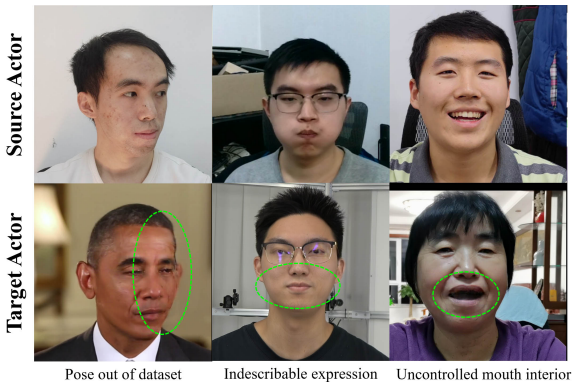


**Figure 9: Failure cases arise due to poses outside of the training dataset, expressions that cannot be modeled by the parametric model, and uncontrollable mouth interior.**

and variation of the training dataset. As a result, we cannot generate rotations and expressions that differ significantly from the training dataset, as shown in the first column of Fig. 9. Second, the input renderings are generated by a 3DMM tracking algorithm. However, the tracked 3DMM is not capable of accurately describing detailed expressions, leading to inaccurate expression control, as shown in the second column of Fig. 9. Additionally, the mouth interior is not constrained, resulting in a lack of realism during the reenactment stage, with the inside of the mouth sometimes appearing blurred.

*Potential Social Impact.* Our method enables a digital portrait copy that can be reenacted by another portrait video. Therefore, given a portrait video of a specific person, it can be used to generate fake portrait videos, which needs to be addressed carefully before deploying the technology.

*Conclusion.* In this paper, we have presented StyleAvatar, a real-time photo-realistic portrait avatar generated from a single video. We have proposed a novel StyleGAN-based framework that can generate a full portrait video, including the shoulders and background, with high image quality. The unique video decomposition and the sliding window data augmentation enable us to achieve faster convergence and more natural movements. Additionally, our proposed live system allows the learned facial avatar to be re-animated by other subjects in real-time. Our extensive results and comprehensive experiments demonstrate that our method outperforms state-of-the-art methods for single-video-based facial avatar reconstruction and reenactment. We believe that our framework will inspire future research on facial reenactment, and our real-time live system has promising potential applications for related tasks.

## ACKNOWLEDGMENTS

This paper is supported by National Key R&D Program of China (2022YFF0902200), the NSFC project No.62125107, No.61827805 and No.62171255, and Guoqiang Institute of Tsinghua University (No.2021GQG0001).

## REFERENCES

Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka. 2021. StyleFlow: Attribute-conditioned exploration of StyleGAN-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics (TOG)* 40, 3 (2021), 1–21.

Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. 2021. Restyle: A residual-based StyleGAN encoder via iterative refinement. In *IEEE/CVF International Conference on Computer Vision (ICCV)*. 6711–6720.

Hadar Averbuch-Elor, Daniel Cohen-Or, Johannes Kopf, and Michael F Cohen. 2017. Bringing portraits to life. *ACM transactions on graphics (TOG)* 36, 6 (2017), 1–13.

Sherwin Bahmani, Jeong Joon Park, Despoina Paschalidou, Hao Tang, Gordon Wetzstein, Leonidas Guibas, Luc Van Gool, and Radu Timofte. 2022. 3d-aware video generation. *arXiv preprint arXiv:2206.14797* (2022).

Volker Blanz and Thomas Vetter. 1999. A Morphable Model for the Synthesis of 3D Faces. In *ACM SIGGRAPH*. ACM, 187–194.

Chen Cao, Tomas Simon, Jin Kyu Kim, Gabe Schwartz, Michael Zollhoefer, Shun-Suke Saito, Stephen Lombardi, Shih-En Wei, Danielle Belko, Shoou-I Yu, et al. 2022. Authentic volumetric avatars from a phone scan. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 1–19.

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. 2022. Efficient geometry-aware 3D generative adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 16123–16133.

Anpei Chen, Ruiyang Liu, Ling Xie, Zhang Chen, Hao Su, and Jingyi Yu. 2022. SofGAN: A portrait image generator with dynamic styling. *ACM Transactions on Graphics (TOG)* 41, 1 (2022), 1–26.

Yu Deng, Jiaolong Yang, Dong Chen, Fang Wen, and Xin Tong. 2020. Disentangled and controllable face image generation via 3D imitative-contrastive learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5154–5163.

Michail Christos Doukas, Mohammad Rami Koujan, Viktoriia Sharmanska, Anastasios Roussos, and Stefanos Zafeiriou. 2021a. Head2Head++: Deep facial attributes re-targeting. *IEEE Transactions on Biometrics, Behavior, and Identity Science* 3, 1 (2021), 31–43.

Michail Christos Doukas, Stefanos Zafeiriou, and Viktoriia Sharmanska. 2021b. HeadGAN: One-shot neural head synthesis and editing. In *IEEE/CVF International Conference on Computer Vision (ICCV)*. 14398–14407.

Nikita Drobyshev, Jenya Chelishev, Taras Khakhulin, Aleksei Ivakhnenko, Victor Lempitsky, and Egor Zakharov. 2022. MegaPortraits: One-shot megapixel neural head avatars. *arXiv preprint arXiv:2207.07621* (2022).

Jiemin Fang, Taoran Yi, Xinggang Wang, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Matthias Nießner, and Qi Tian. 2022. Fast dynamic radiance fields with time-aware neural voxels. In *SIGGRAPH Asia 2022 Conference Papers*. 1–9.

Guy Gafni, Justus Thies, Michael Zollhofer, and Matthias Nießner. 2021. Dynamic neural radiance fields for monocular 4D facial avatar reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8649–8658.

Rinon Gal, Dana Cohen Hochberg, Amit Bermano, and Daniel Cohen-Or. 2021. SWAGAN: A style-based wavelet-driven generative model. *ACM Transactions on Graphics (TOG)* 40, 4 (2021), 1–11.

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

Stephan J Garbin, Marek Kowalski, Virginia Estellers, Stanislaw Szymanowicz, Shideh Rezaeifar, Jingjing Shen, Matthew Johnson, and Julien Valentin. 2022. VolTeMorph: Realtime, Controllable and Generalisable Animation of Volumetric Representations. *arXiv preprint arXiv:2208.00949* (2022).

Pablo Garrido, Levi Valgaerts, Ole Rehmsen, Thorsten Thormahlen, Patrick Perez, and Christian Theobalt. 2014. Automatic face reenactment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4217–4224.

Jiahao Geng, Tianjia Shao, Youyi Zheng, Yanlin Weng, and Kun Zhou. 2018. Warp-guided GANs for single-photo facial animation. *ACM Transactions on Graphics (TOG)* 37, 6 (2018), 1–12.

Partha Ghosh, Pravir Singh Gupta, Roy Uziel, Anurag Ranjan, Michael J Black, and Timo Bolkart. 2020. GIF: Generative interpretable faces. In *International Conference on 3D Vision (3DV)*. IEEE, 868–878.

Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. 2022a. Neural head avatars from monocular RGB videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 18653–18664.

Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. 2022b. Neural head avatars from monocular RGB videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 18653–18664.

Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. 2020. GANSpace: Discovering interpretable GAN controls. *Advances in Neural Information Processing Systems* 33 (2020), 9841–9850.

Fa-Ting Hong, Longhao Zhang, Li Shen, and Dan Xu. 2022. Depth-aware generative adversarial network for talking head video generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 3397–3406.

Wonjong Jang, Gwangjin Ju, Yucheol Jung, Jiaolong Yang, Xin Tong, and Seungyong Lee. 2021. StyleCariGAN: caricature generation via StyleGAN feature map modulation. *ACM Transactions on Graphics (TOG)* 40, 4 (2021), 1–16.

Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2021. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems* 34 (2021), 852–863.

Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4401–4410.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of StyleGAN. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8110–8119.

Taras Khakhulin, Vanessa Sklyarova, Victor Lempitsky, and Egor Zakharov. 2022. Realistic one-shot mesh-based head avatars. In *European Conference of Computer vision (ECCV)*. Springer, 345–362.

Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Niessner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. 2018. Deep video portraits. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 1–14.

Mohammad Rami Koujan, Michail Christos Doukas, Anastasios Roussos, and Stefanos Zafeiriou. 2020. Head2head: Video-based neural head synthesis. In *IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 16–23.

Marek Kowalski, Stephan J Garbin, Virginia Estellers, Tadas Baltrušaitis, Matthew Johnson, and Jamie Shotton. 2020. Config: Controllable neural face image generation. In *European Conference on Computer Vision (ECCV)*. Springer, 299–315.

Kai Li, Feng Xu, Jue Wang, Qionghai Dai, and Yebin Liu. 2012. A data-driven approach for facial expression synthesis in video. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 57–64.

Shanchuan Lin, Linjie Yang, Imran Saleemi, and Soumyadip Sengupta. 2021. Robust High-Resolution Video Matting with Temporal Guidance. *arXiv preprint arXiv:2108.11515* (2021).

Zicheng Liu, Ying Shan, and Zhengyou Zhang. 2001. Expressive expression mapping with ratio images. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*. 271–276.

Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. 2018. Deep appearance models for face rendering. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 1–13.

Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. 2019. Neural volumes: Learning dynamic renderable volumes from images. *arXiv preprint arXiv:1906.07751* (2019).

Shugao Ma, Tomas Simon, Jason Saragih, Dawei Wang, Yuecheng Li, Fernando De La Torre, and Yaser Sheikh. 2021. Pixel codec avatars. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 64–73.

Arun Mallya, Ting-Chun Wang, and Ming-Yu Liu. 2022. Implicit Warping for Animation with Image Sets. *arXiv preprint arXiv:2210.01794* (2022).

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. 2022. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 1–15.

Koki Nagano, Jaewoo Seo, Jun Xing, Lingyu Wei, Zimo Li, Shunsuke Saito, Aviral Agarwal, Jens Fursund, Hao Li, Richard Roberts, et al. 2018. paGAN: real-time avatars using dynamic textures. *ACM Transactions on Graphics (TOG)* 37, 6 (2018), 258–1.

Kyle Olszewski, Zimo Li, Chao Yang, Yi Zhou, Ronald Yu, Zeng Huang, Sitao Xiang, Shunsuke Saito, Pushmeet Kohli, and Hao Li. 2017. Realistic dynamic facial textures from a single image using GANs. In *IEEE International Conference on Computer Vision (ICCV)*. 5429–5438.

Ivan Perov, Daiheng Gao, Nikolay Chervoniy, Kunlin Liu, Sugasa Marangonda, Chris Umé, Mr Dpfks, Carl Shift Facenheim, Luis RP, Jian Jiang, et al. 2020. DeepFaceLab: Integrated, flexible and extensible face-swapping framework. *arXiv preprint arXiv:2005.05535* (2020).

Amit Raj, Michael Zollhoefer, Tomas Simon, Jason Saragih, Shunsuke Saito, James Hays, and Stephen Lombardi. 2021. PVA: Pixel-aligned volumetric avatars. *arXiv preprint arXiv:2101.02697* (2021).

Yurui Ren, Ge Li, Yuanqi Chen, Thomas H Li, and Shan Liu. 2021. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *IEEE/CVF International Conference on Computer Vision (ICCV)*. 13759–13768.

Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. 2021. Encoding in style: a StyleGAN encoder for image-to-image translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2287–2296.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 234–241.

Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. 2020a. Interpreting the latent space of GANs for semantic face editing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 9243–9252.

Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. 2020b. InterFaceGAN: Interpreting the disentangled face representation learned by GANs. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 44, 4 (2020), 2004–2018.

Yichun Shi, Xiao Yang, Yangyue Wan, and Xiaohui Shen. 2022. SemanticStyleGAN: Learning compositional generative priors for controllable image synthesis and editing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 11254–11264.

Alon Shoshan, Nadav Bhonker, Igor Kviatkovsky, and Gerard Medioni. 2021. GAN-Control: Explicitly controllable GANs. In *IEEE/CVF International Conference on Computer Vision (ICCV)*. 14083–14093.

Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019. First order motion model for image animation. *Advances in Neural Information Processing Systems* 32 (2019).

Jingxiang Sun, Xuan Wang, Yichun Shi, Lizhen Wang, Jue Wang, and Yebin Liu. 2022. IDE-3D: Interactive disentangled editing for High-Resolution 3D-Aware portrait synthesis. *ACM Transactions on Graphics (TOG)* 41, 6 (2022), 1–10.

Jingxiang Sun, Xuan Wang, Lizhen Wang, Xiaoyu Li, Yong Zhang, Hongwen Zhang, and Yebin Liu. 2023. Next3D: Generative Neural Texture Rasterization for 3D-Aware Head Avatars. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. 2017. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 1–13.

Ayush Tewari, Mohamed Elgharib, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhöfer, and Christian Theobalt. 2020a. PIE: Portrait image embedding for semantic control. *ACM Transactions on Graphics (TOG)* 39, 6 (2020), 1–14.

Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. 2020b. StyleRig: Rigging StyleGAN for 3D control over portrait images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 6142–6151.

Justus Thies, Michael Zollhöfer, and Matthias Nießner. 2019. Deferred neural rendering: Image synthesis using neural textures. *Acm Transactions on Graphics (TOG)* 38, 4 (2019), 1–12.

Justus Thies, Michael Zollhöfer, Matthias Nießner, Levi Valgaerts, Marc Stamminger, and Christian Theobalt. 2015. Real-time expression transfer for facial reenactment. *ACM Transactions on Graphics (TOG)* 34, 6 (2015), 183–1.

Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2016. Face2Face: Real-time face capture and reenactment of RGB videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2387–2395.

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

Daniel Vlasic, Matthew Brand, Hanspeter Pfister, and Jovan Popović. 2005. Face transfer with multilinear models. *ACM Transactions on Graphics (TOG)* 24, 3 (2005), 426–433.

Daoye Wang, Prashanth Chandran, Gaspard Zoss, Derek Bradley, and Paulo Gotardo. 2022a. MoRF: Morphable radiance fields for multiview neural head modeling. In *ACM SIGGRAPH 2022 Conference Proceedings*. 1–9.

Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. 2020. MEAD: A large-scale audio-visual dataset for emotional talking-face generation. In *European Conference on Computer Vision (ECCV)*. Springer, 700–717.

Lizhen Wang, Zhiyuan Chen, Tao Yu, Chenguang Ma, Liang Li, and Yebin Liu. 2022b. FaceVerse: a fine-grained and detail-controllable 3D face morphable model from a hybrid dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 20333–20342.

Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. 2021b. Towards Real-World blind face restoration with generative facial prior. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 9168–9178.

Ziyan Wang, Timur Bagautdinov, Stephen Lombardi, Tomas Simon, Jason Saragih, Jessica Hodgins, and Michael Zollhofer. 2021a. Learning Compositional Radiance Fields of Dynamic Human Heads. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5704–5713.

Shih-En Wei, Jason Saragih, Tomas Simon, Adam W Harley, Stephen Lombardi, Michal Perdoch, Alexander Hypes, Dawei Wang, Hernan Badino, and Yaser Sheikh. 2019.

VR facial animation via multiview image translation. *ACM Transactions on Graphics (TOG)* 38, 4 (2019), 1–16.

Thibaut Weise, Sofien Bouaziz, Hao Li, and Mark Pauly. 2011. Realtime performance-based facial animation. *ACM transactions on graphics (TOG)* 30, 4 (2011), 1–10.

Jianfeng Xiang, Jiaolong Yang, Yu Deng, and Xin Tong. 2022. GRAM-HD: 3D-Consistent image generation at high resolution with generative radiance manifolds. *arXiv preprint arXiv:2206.07255* (2022).

Yuelang Xu, Lizhen Wang, Xiaochen Zhao, Hongwen Zhang, and Yebin Liu. 2023a. AvatarMAV: Fast 3D Head Avatar Reconstruction Using Motion-Aware Neural Voxels. In *ACM SIGGRAPH 2023 Conference Proceedings*.

Yuelang Xu, Hongwen Zhang, Lizhen Wang, Xiaochen Zhao, Han Huang, Guojun Qi, and Yebin Liu. 2023b. LatentAvatar: Learning Latent Expression Code for Expressive Neural Head Avatar. In *ACM SIGGRAPH 2023 Conference Proceedings*.

Fei Yin, Yong Zhang, Xiaodong Cun, Mingdeng Cao, Yanbo Fan, Xuan Wang, Qingyan Bai, Baoyuan Wu, Jue Wang, and Yujiu Yang. 2022. StyleHEAT: One-shot high-resolution editable talking face generation via pre-trained StyleGAN. In *European Conference on Computer Vision (ECCV)*. Springer, 85–101.

Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C Bühler, Xu Chen, Michael J Black, and Otmar Hilliges. 2022a. IM Avatar: Implicit morphable head avatars from videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 13545–13555.

Yufeng Zheng, Wang Yifan, Gordon Wetzstein, Michael J Black, and Otmar Hilliges. 2022b. PointAvatar: Deformable Point-based Head Avatars from Videos. *arXiv preprint arXiv:2212.08377* (2022).

Zerong Zheng, Xiaochen Zhao, Hongwen Zhang, Boning Liu, and Yebin Liu. 2023. AvatarReX: Real-time Expressive Full-body Avatars. *ACM Transactions on Graphics (TOG)* 42, 4 (2023), 1–19. https://doi.org/10.1145/3592101

Wojciech Zielonka, Timo Bolkart, and Justus Thies. 2022. Instant Volumetric Head Avatars. *arXiv preprint arXiv:2211.12499* (2022).