

# Emo-Avatar: Efficient Monocular Video Style Avatar through Texture Rendering

Pixin Liu\*  
 University of Rochester  
 USA  
 pliu23@u.rochester.edu

Hang Hua  
 University of Rochester  
 USA  
 hhua2@cs.rochester.edu

Jiebo Luo  
 University of Rochester  
 USA  
 jluo@cs.rochester.edu

Luchuan Song\*  
 University of Rochester  
 USA  
 slc0826@mail.ustc.edu.cn

Yunlong Tang  
 University of Rochester  
 USA  
 yunlong.tang@rochester.edu

Chenliang Xu  
 University of Rochester  
 USA  
 chenliang.xu@rochester.edu

Daoan Zhang  
 University of Rochester  
 USA  
 daoan.zhang@rochester.edu

Huaijin Tu  
 Georgia Institute of Technology  
 USA  
 htu35@gatech.edu



**Figure 1:** We introduce Emo-Avatar, a method efficiently designed for high-fidelity animatable portrait reconstruction and editing from a short monocular video with one-shot reference image in 5 minutes. Our method maintains clear textures, consistent motion, and coherent style in both self-driving and cross-driving settings.

## ABSTRACT

Artistic video portrait generation is a significant and sought-after task in the fields of computer graphics and vision. While various methods have been developed that integrate NeRFs or StyleGANs with instructional editing models for creating and editing drivable portraits, these approaches face several challenges. They often rely heavily on large datasets, require extensive customization processes, and frequently result in reduced image quality. To address the above problems, we propose the Efficient Monotonic Video Style Avatar (Emo-Avatar) through deferred neural rendering that enhances StyleGAN's capacity for producing dynamic, drivable portrait videos. We first delved into whether pre-trained StyleGAN can encode video portraits. It has been shown that StyleGAN can encode aligned, multi-view animatable portrait videos. However, for portrait videos containing upper body elements, StyleGAN struggles to accurately reconstruct these unaligned avatars. To fully leverage the facial information from the pre-trained StyleGAN for portrait rendering. We proposed a two-stage deferred neural rendering pipeline. In the first stage, we utilize few-shot PTI initialization to initialize the StyleGAN generator through several extreme poses sampled from the video to capture the consistent representation of aligned faces from the target portrait. In the second stage, we propose a Laplacian pyramid for high-frequency texture sampling from UV maps deformed by dynamic flow of expression for motion-aware texture prior integration to provide torso features to enhance StyleGAN's ability to generate complete and upper body for portrait video rendering. Emo-Avatar reduces style customization time from hours to merely 5 minutes compared with existing methods. In addition, Emo-Avatar requires only a single reference image for editing and employs region-aware contrastive learning with semantic invariant CLIP guidance, ensuring consistent high-resolution output and identity preservation. Through both quantitative and qualitative assessments, Emo-Avatar demonstrates superior performance over existing methods in terms of training efficiency, rendering quality and editability in self- and cross-reenactment.

## CCS CONCEPTS

- Computing methodologies → Animation; Motion processing.

## KEYWORDS

StyleGAN, Facial Reenactment, Video Portraits, Face Editing

## 1 INTRODUCTION

In recent years, artistic video portraits have gained widespread prominence across various domains like social media, art, cinema, and advertising, underscoring their cultural and commercial value. Traditionally, creating these portraits involves a skill-intensive, laborious process, posing challenges in scalability and replication. To address this, advancements in computer vision and graphics are increasingly focusing on automating their generation.

StyleGAN-based methods like StyleGANEX [41] and VToonify [40] have leveraged the pre-trained latent style distribution for style transfer and semantic editing to address these challenges. However,

these image translation-centric methods are not extendable to novel view synthesis and rely heavily on extensive collections of stylized images for fine-tuning. This reliance poses substantial challenges in compiling large datasets with consistent styling. Furthermore, StyleGAN's limitation in effectively separating facial motion from identity features often compromises the personalization and quality of portraits, with the subject's identity unintentionally shifting during the style transfer process. Other prevalent approaches employ monocular videos to train models for assimilating individual identity features, augmented by CLIP [26], Instruct-pix2pix (Ip2p) [1], or Score Distillation Sampling [25] (SDS) for style editing guidance. However, these methods encounter significant limitations: they are not pre-trained on diverse datasets, struggle to maintain explicit control over style variations and necessitate prolonged fine-tuning periods. Particularly, diffusion-based approaches like Ip2p and SDS often demonstrate inconsistency and are susceptible to progressive image blurring, especially when the desired styles deviate substantially from the original image.

To tackle these challenges, we introduce Emo-Avatar, a novel approach that is efficient in terms of parameters, space, and time for the synthesis and editing of drivable portraits. Emo-Avatar minimally modifies pre-trained StyleGAN via Few-shot PTI Initialization to optimize the target portrait's coarse template style code and integrate the target portrait style into the latent distribution. Then Emo-Avatar employs Neural Texture [34], integrating UV mapping and facial expression data for consistent identity and motion-aware texture representation. Finally, We utilize a StyleGAN to integrate coarse template code and motion-aware texture-prior for deferred neural rendering of the drivable portrait. During the editing phase, our approach only requires **one** reference image containing the target style for fine-tuning. This is in contrast to previous Ip2p-based editing methods, which iteratively update the entire dataset. Specifically, we incorporate a novel Region-aware Style Contrastive Loss, utilizing semantic invariant CLIP guidance, to mitigate the editing inconsistencies and overfitting on the reference image. This approach achieves high-resolution instructional editing while maintaining the original identity. Our novel method enables rapid and precise style customization, significantly reducing the style fine-tuning time to just 5 minutes, a substantial improvement from the typical 4-5 hours required by previous methods.

We validate the efficacy of our approach with both quantitative and qualitative experiments, benchmarking it against existing methods for drivable portrait generation and video translation. Our key contributions are outlined as follows:

- We pioneer the integration of Neural Texture with StyleGAN, creating a straightforward but potent pipeline for generating and editing personalized drivable style portraits.
- Our approach employs parameter-efficient fine-tuning and explores an optimal skip connection that maintains StyleGAN's generation and domain transferability. This significantly speeds up the portrait generation training process to 1 hour and the editing fine-tuning to just 5 minutes.
- Our region-aware CLIP-based contrastive learning approach prevents blurring caused by inconsistencies during editing and over-fitting on the reference image, ensuring high-resolution output during instructional editing.

<sup>\*</sup>Both authors contributed equally to this research.

## 2 RELATED WORK

**Driveable Portrait Rendering.** Recent advancements in 2D-based animation draw inspiration from classical rendering techniques, employing morphable model reconstruction, forward rendering with optimized textures [6, 35], and image synthesis through deep neural network [12, 14]. Noteworthy among these approaches is Deep Video Portraits [15], which utilizes rendered correspondence maps in conjunction with image-to-image translation networks to produce highly realistic imagery. In contrast to dense conditioning inputs or rendered feature maps, some methods focus on rendered facial landmarks [31, 45]. First-Order Motion Model [31] is a data-driven approach that separates appearance and motion in videos, enabling the transfer of motion from a source video to a target image.

Other recent works [4, 16, 27] rely on semantic features from 3D templates [3] to control facial motion. Such as Head2Head and Head2Head++ [4, 16] encodes Normalized Mean Face Coordinates (NMFC) features with gaze image to animate the target portrait. The StyleHEAT [44] optimizes latent codes through inversion and leverages audio features to learn motion for target movement generation. The Style Avatar [37] combines UV and texture rendering with two StyleUNETs, achieving significant improvements in facial synthesis quality.

Compared with 2D-based methods above, reconstructing 3D head portraits from monocular videos is a complex task. Mainstream approaches have predominantly focused on creating mesh-based portraits using head templates from video [3, 24]. To address dynamic elements like hair gazes, and teeth, Neural Head portrait [8] employs neural networks to dynamically enhance the texture and geometry of head portraits, utilizing the FLAME model [17]. However, challenges such as blurred textures due to geometric inaccuracies persist. Innovative solutions like IMAvatar [48] have shifted towards using implicit geometry and texture models to overcome the limitations of mesh templates. This approach has been further refined in PointAvatar [49], which combines explicit point clouds with implicit representations to enhance image quality.

**Controllable Video Editing.** In the dynamic realm of controllable video editing, the field has witnessed substantial progress. This advancement is manifested through a spectrum of innovations, encompassing cutting-edge editing techniques as exemplified [30, 47], and [42], as well as the ingenious fusion of linguistic, musical, and visual elements, notably presented in the works [38, 39]. The domain is further augmented by strides in multimodal interactions and the art of summarization, as elucidated by [20, 33], complemented by exploratory strides into the realms of content modification and the nuanced editing within latent spaces, detailed in the studies by [11, 19]. Central to these burgeoning developments are StyleGAN [14] and its cognate methodologies, which have proven to be instrumental. The synergistic integration with CLIP [26] propels StyleGAN-Nada [5], heralding a novel paradigm where textual inputs elegantly steer the editing narrative. Anchoring these multi-faceted advancements are diffusion models, notably the pioneering work [10], which consistently serve as a cornerstone, driving the relentless evolution and refinement of video editing technologies. Diffusion models [10] have been extensively investigated in the context of instruction-guided video generation tasks. Diffusion

models can generate high-fidelity images by iteratively denoising data, starting from a mixture of pure and noisy samples. It enables low-resource and barrier-free style transfer, making it an attractive option for a broad range of applications. Recent advancements have seen numerous works elevating 2D diffusion processes to 3D, applying these processes extensively in the realm of 3D editing. Broadly, these works can be categorized into two types. The first type [18, 50], exemplified by Dreamfusion’s [25] introduction of SDS loss, involves feeding the noised rendering of the current 3D model, along with other conditions, into a 2D diffusion model. The scores generated by the diffusion model then guide the direction of model updates. The second type [9, 29, 32] focuses on conducting 2D editing based on given prompts for the multiview rendering of a 3D model. This approach creates a multi-view 2D image dataset, which is then utilized as a training target to guide the 3D model.

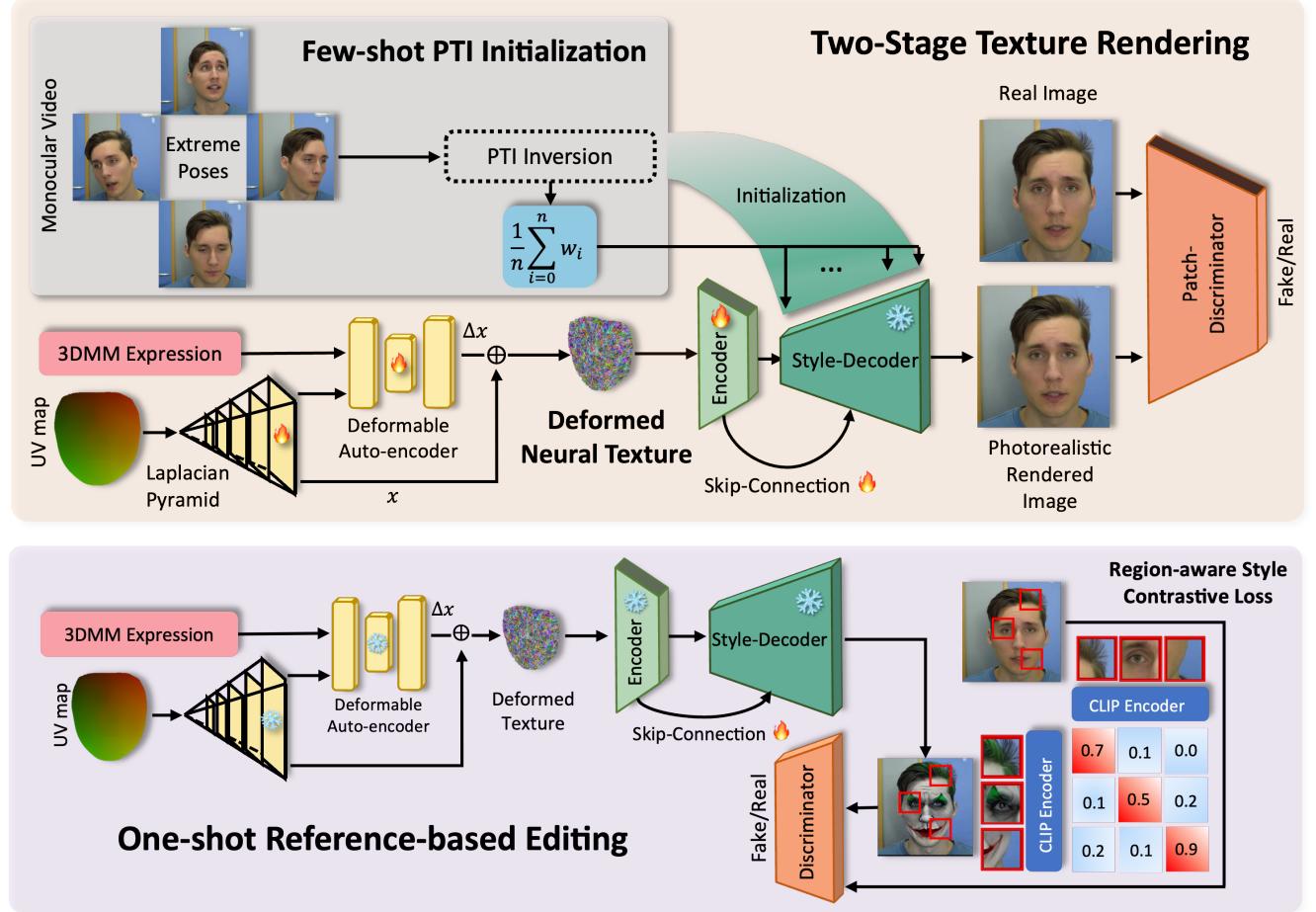
## 3 EFFICIENT STYLEGAN-BASED TEXTURE RENDERING.

Generating high-fidelity animatable portrait videos and achieving personalized editing is a critical task in computer vision. To achieve this goal, a robust facial neural representation model is essential. In this work, we leverage the powerful StyleGAN [14] to boost the performance of neural texture rendering with stylization, it has demonstrated impressive results in various applications involving facial analysis. The StyleAGN has end-to-end differentiability and the ability to control picture style, so it is suitable for our field. However, it is unable to be directly extended for unaligned portrait generation. To address this issue, we propose an effective two-stage method that, while preserving the pre-trained generalization capabilities of StyleGAN, integrates an unaligned motion-aware texture prior into StyleGAN. This enables the pre-trained generative power of StyleGAN to be generalized to deferred drivable portrait rendering and editing tasks.

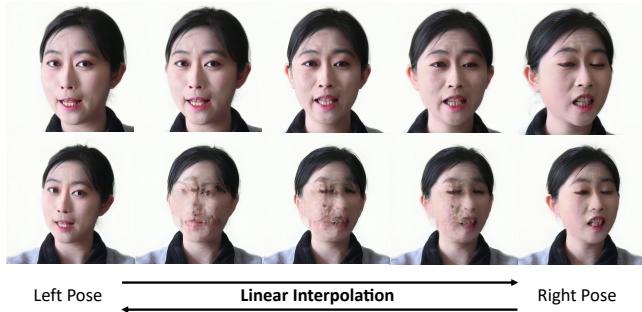
### 3.1 Preliminaries of StyleGAN

Before introducing the Emo-Avatar framework, we evaluated StyleGAN’s ability to generate animatable video portraits, which involves capturing varying expressions, continuous facial motions, and cohesive upper body movement during head rotations. Unlike the aligned images in the pre-trained FFHQ dataset, animatable portraits are often unaligned and captured in diverse settings, with a variety of head positions and orientations.

To assess StyleGAN’s effectiveness, we applied the GAN inversion method on both aligned and unaligned portraits, comparing the rendering results. This was crucial to determine if StyleGAN could accurately represent a dynamic portrait video. Our evaluation focused on frames showing extreme left and right head poses from videos as inputs for GAN inversion. This approach tested StyleGAN’s limits in rendering realistic, continuous motion and its ability to capture the nuanced changes in facial orientation and expression. The insights gained from this assessment were instrumental in shaping the Emo-Avatar framework, enhancing our understanding of the capabilities and limitations of StyleGAN in animatable portrait generation.



**Figure 2: Overview of Emo-Avatar.** The pipeline steps include portrait video generation and reference based. The Two-stage Texture Rendering is shown in the upper part, we initialize the generator though few-shot PTI initialization and utilize Deformed Neural Texture to provide motion-aware texture prior information though skip connection to StyleGAN for portrait video generation; In the lower part, we finetune the skip-connection module within the pre-trained encoder and Style-Decoder for updating the stylization parameters. The direction of stylization guidance focuses on each image patches.



**Figure 3: Interpolation of GAN inversion:** Latent code interpolation between extreme pose parameters along the x-axis for aligned (upper) and unaligned (lower) video portraits.

In Figure 3, the linear interpolation of latent codes for extreme poses is presented in two rows: the first for aligned and the second for unaligned inversion. With aligned inversion, interpolating

between two style codes yields images that maintain texture quality and exhibit consistent, smooth transitions in facial expressions and poses. This demonstrates StyleGAN's capability in handling aligned facial data. In contrast, the unaligned inversion results reveal StyleGAN's limitations. When processing unaligned faces, particularly in the animatable portrait domain, the model struggles, leading to blurred images. This blurring highlights its difficulty in accurately reconstructing the complex, varied aspects of unaligned faces, including nuanced head movements and expressions. This comparison underlines a key finding: while pre-trained StyleGAN is effective for aligned facial portraits, it falls short in encoding complete portraits with upper body information, unable to capture the full range of portrait dynamics. Additionally, from the first line in Figure 3, we observe that even though only two images from extreme poses in the left and right directions are used for GAN inversion, StyleGAN is still capable of rendering relatively good intermediate images when interpolating the latent codes. This

suggests that after GAN Inversion, the latent space encoded in StyleGAN remains continuous, motion-aware, and can be effectively sampled. Therefore, we can sample a small number of images from the video to perform GAN inversion, thereby obtaining the video's neural representation model.

### 3.2 Two-Stage Texture Rendering

In the previous section, we find that the pretrained StyleGAN hardly to handle with wild animatable portraits. To address this issue, we designed a Two-Stage Texture Rendering method to progressively encode animatable portrait information into StyleGAN's latent space. In the first stage, we adopt Few-shot PTI Initialization to encode only the facial information from the video into StyleGAN. In the second stage, we further encode dynamic facial expressions from 3DMM and the relative positioning information of the upper body and face from the UV Map into StyleGAN, enabling it to generate complete video portraits.

**3.2.1 Few-shot PTI Initialization.** To better utilize the information in the pre-trained StyleGAN, we propose the Multi-view PTI Initialization to tailor a StyleGAN based on facial information. In the previous section, we discovered through experiments that using only a few images of extreme angles to invert StyleGAN can effectively represent intermediate head rotations. Therefore, in our proposed Multi-view PTI Initialization, we sampled only *four* images of extreme poses as inputs, representing the top, bottom, left, and right views. This strategy not only speeds up the rendering process but also ensures the quality of the generated video.

Specifically, for the PTI initialization, we employ PTI inversion [28] to encode personalized avatar information from the video into StyleGAN as the initialization latent space. The PTI inversion can embed the target portrait within StyleGAN's latent distribution with subtly modifying the original model parameters. Specifically, We first fix StyleGAN's parameters and optimize  $w$  to minimize the discrepancy between the generated and target images, indicating that  $w$  closely aligns with our target in the latent space. Subsequently, we fix  $w$  and fine-tune StyleGAN to enhance the similarity of the generated image to the target at this specific  $w$ . Hence, by using only four images of extreme poses, PTI inversion can encode the entire video's information into a neural representation based on StyleGAN.

**3.2.2 Deformed Neural Texture.** The goal of this module is to provide motion-aware texture prior information and the torso features to the generator. The UV maps define the facial locations and contain high-frequency texture features. Thus, applying UV-based rendering can easily control the facial motions and achieve high-quality portrait rendering. In our experimental setup, we follow the inspiration from the MipMaps to build the Laplacian Pyramid with 12 distinct layers of texture maps. While the texture maps are sampled from static UVs corresponding to each frame in the video, they do not contain dynamic facial expressions. To capture the dynamic flow of facial expressions, we employ a sliding window technique to track the expression parameters from two frames preceding and succeeding the current frame (thus encompassing a total of five frames) extracted from the 3D Morphable Model (3DMM). The objective here is to capture the dynamic continuity of facial

expressions, which we refer to as the "motion flow". To capture this motion flow, we first map expression parameters to a latent vector using a 3-layer MLP. This process aggregates temporal information. We then construct an auto-encoder, comprising a 5-layer convolutional encoder and decoder, designed for multi-scale feature extraction. The auto-encoder requires the source texture and driving facial expressions as inputs, producing the desired flow fields for texture deformation. The motion parameters are integrated into each convolutional layer through adaptive instance normalization (AdaIN) [13] defined as follows:

$$\text{AdaIN}(\text{tex}, \text{exp}) = M_s(\text{exp}) \frac{\text{tex} - \mu(\text{tex})}{\sigma(\text{tex})} + M_b(\text{exp}),$$

where  $\mu(\cdot)$  and  $\sigma(\cdot)$  denote the average and variance operations, respectively.  $M_s$  and  $M_b$  are used to estimate the adapted mean and bias value according to the target motion. Each feature map  $\text{tex}$  is normalized, then scaled and biased using corresponding scalar components. This encoding process estimates the deformation of the original texture map, which we then integrate with the original texture. By incorporating these motion-aware features, we enable StyleGAN to render neural portraits with both facial expressions and movements accurately captured and represented.

**3.2.3 Training strategy of Texture Rendering.** For neural texture synthesis, we employ both  $L_1$  loss and LPISP loss, which are instrumental in aligning synthesized images with ground truth, particularly in terms of texture and perceptual fidelity. We focus on the first three channels of the 12-channel hierarchical texture maps  $I_{\text{tex}}^{(1:3)}$ , comparing them against RGB ground-truth images  $I_{gt}$ . During training, we specifically mask the UV region to concentrate on this area only. For StyleGAN-based texture rendering, we perform fine-tuning on entire ground-truth images to derive  $I_{nt}$ .

$$\mathcal{L}_{rgb} = ||I_{\text{tex}}^{(1:3)} - I_{gt}||_1 + ||I_{nt} - I_{gt}||_1 + \text{LPISP}(I_{nt}, I_{gt}). \quad (1)$$

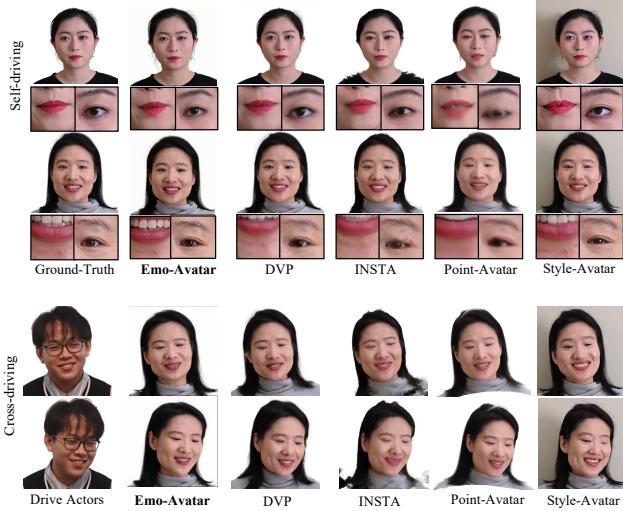
To further enhance image fidelity, we introduce a patch-ware conditional discriminator, using UV maps as a condition to compare patches of generated images with ground truth. This method, employing conditional adversarial loss [22], significantly improves the detailed matching of generated and real images:

$$\begin{aligned} \mathcal{L}_{cGAN}(G, D_p) = & \mathbb{E}_{I_{gt}, uv} [\log D_p(I_{gt}, uv)] + \\ & \mathbb{E}_{uv} [\log (1 - D_p(uv, I_{nt}))] \end{aligned} \quad (2)$$

where a patch discriminator  $D_p$  aims to distinguish between  $\{(I_{gt}, uv)\}$  and  $\{(I_{nt}, uv)\}$ , and  $(\cdot, \cdot)$  denotes a concatenation operation.

### 3.3 One-shot Reference-based Editing

To efficiently conduct editing while retaining more pre-trained information, in this phase we only train the skip connections and freeze all other modules in the generator. For the ablation study of this part, please refer to the Appendix for additional information. We then leverage the discriminator in Stage 1 to be fine-tuned on reference images to provide guidance to the model as well. However, simply relying on a single reference image is easy to lead to a significant over-fitting problem, causing the model to be unable to construct stylized images for other poses.



**Figure 4: We compare our methods with the current SOTA portrait video avatar rendering techniques. Our model outperforms all other methods in details such as eyes and teeth.**

Existing methods that employ iterative dataset updates for editing with diffusion models face a significant issue: the resulting images often become progressively blurred. This occurs because the diffusion models tend to lose the original high-frequency information during the editing process, and the updated datasets lack consistency in the style edits applied. To counter these challenges, we introduce a patch-wise contrastive loss (PatchNCE). Our approach utilizes contrastive learning on the embedded features of the model to better maintain local spatial information and details. We use pre-trained CLIP models to extract feature embeddings from generated images  $I_{nt}$  by Emo-Avatar and reference images. We randomly crop patches in these images and embed them with the CLIP encoder  $E$ . We then aim to bring ‘positive’ patches, cropped from the same position in both images, closer together in the feature space, while distancing ‘negative’ patches, cropped from different positions. Let  $v$  denote the embedded query patch from  $I_{ref}$ . Let  $v^+$  and  $\{v_i^-\}_{i=1}^N$  be the embedded positive patch and  $N$  negative patches from  $I_{nt}$ , respectively. The patch-wise loss can be written as:

$$L_{\text{patch}} = -\log \left[ \frac{\exp(v \cdot v^+)}{\exp(v \cdot v^+) + \sum_{i=1}^N \exp(v \cdot v_i^-)} \right]. \quad (3)$$

## 4 EXPERIMENT

### 4.1 Experimental Setup

**Datasets and Preprocess with Head Tracking.** Our method takes a monocular video as input to generate the volumetric portrait video. To demonstrate the highest quality achievable by each method, we use self-recorded videos rather than publicly available internet videos, which often undergo quality degradation due to compression. We employ Nikon Z7 camera with a manual fixed-focus lens to record 4 actors, with each actor being filmed in a 5-minute video at 4K resolution and 25 FPS. For the monocular



**Figure 5: We compare our methods with the current SOTA video editing methods. The comparison demonstrates the quality of the generated portraits is significantly better than others.**

Methods	F-LMD↓	PSNR↑	SD↓	LPIPS↓
	Quantitative results			
PointAvatar [49]	2.74	24.39	6.20	0.086
INSTA [51]	2.81	25.97	7.73	0.094
DVP [15]	2.93	25.32	5.25	0.086
Style-Avatar [37]	2.64	<b>27.83</b>	4.78	0.075
Emo-Avatar	<b>2.57</b>	26.73	<b>3.50</b>	<b>0.058</b>

**Table 1: Quantitative comparison with the baseline methods for self-reenactment.**

Methods	F-LMD↓	BIQ↓ ( $\times 10^{-2}$ )	SD↓ ( $\times 10^{-2}$ )	CLIP-D↑ ( $\rightarrow 1$ )	CLIP-D $^D$ ↑ ( $\rightarrow 1$ )
	Quantitative results				
Ip2p	24.11	0.917	0.225	0.272	0.627
TokenFlow	18.14	0.823	0.175	<b>0.281</b>	0.822
RAV	9.42	0.835	0.221	0.254	0.792
GT	–	<b>0.805</b>	–	0.013	0.004
Our	<b>3.194</b>	<b>0.822</b>	<b>0.171</b>	0.263	<b>0.914</b>

**Table 2: Quantitative comparison for video portrait editing with the baseline methods. The SD metric employs image pairs that belong to different styles (original captured images and artistic images).**

head tracking, we follow Faceverse [36] to extract the UV map and facial expression parameters in videos.

**Implementation Details.** We implement our model with PyTorch and a single A6000 GPU. For stage 1, we train the model in 1 hours. For stage 2, we randomly select one reference image and obtain its edited version through Instruct-pixel2pixel for one-shot editing. It takes 5 minutes to achieve personalized instructional editing.

### 4.2 Qualitative Evaluation

We present a qualitative evaluation aimed at elucidating the discernible distinctions among different methods. We perform our experiments on the same set of facial/head poses as the template input, which is not involved in training.

**Evaluation Metrics.** We evaluate the effectiveness of our method based on four aspects. (1) F-LMD [2]: The distance on the whole face to measure the differences in head pose and facial expression. (2) The Sharpness Difference (SD) [21]: It is used to evaluate the sharpness difference between the source and stylized images, which is implemented by the pixel-level difference in rows and columns. We anticipate that a pair with good SD can exhibit equivalent changes in gradient (e.g. shadows). (3) Image Spatial Quality: we adopt the PSNR to measure the overall image quality, the Learned Perceptual Image Patch Similarity (LPIPS) [46] for the details. (4) Blind/Referenceless Image Spatial Quality Evaluator (BIQ) [23]: The no-reference spatial domain image quality assessment method, is used to assess the quality of edited images directly.(5) Text-Image Consistency (CLIP-D, CLIP-D<sup>v</sup>) [9]: The CLIP-D is the embedding between text and image to calculate cosine similarity for each pair via CLIP direction and CLIP-D<sup>v</sup> is to assess temporal consistency across the entire video, as defined in the official report of in2n.

**Comparison Setting.** Our evaluation of the method’s efficacy in portrait generation and editing encompasses two experimental settings. We first assess Emo-Avatar’s proficiency in both self-reenactment and cross-reenactment tasks using self-recorded datasets captured via camera. We compare our method with the state-of-the-art methods, including: (1) Point-Avatar [49], (2) Instant-Avatar (IN-STA) [51], (3) Deep Video Portrait (DVP) [15], (4) Style-Avatar [37]. To ensure a rigorous test, we randomly selected 500 frames from each video across both datasets, excluding any frames used in training, for our analysis. For assessing video portrait editing capabilities, we conducted a comparative study with leading state-of-the-art video editing techniques: (1) InstructPix2Pix(ip2p) [1]: we use ip2p for image-level editing. (2)TokenFlow [7]: Due to TokenFlow’s GPU memory requirements increasing with the length of the video, to ensure a fair comparison, we limit the video segments used in our evaluation to 3 seconds, consisting of approximately 75 frames. (3) Rerender-A-Video (RAV) [43]: We followed the released data preprocessing and rendering steps on our monotonic video.

**Evaluation Results.** The average of quantitative results on 4 datasets are summarized in Tab 1 and Tab 2, for video portrait rendering and editing respectively. For portrait rendering, According to the results shown in the upper of Table 2, our method outperforms the others in terms of text-image consistency (CLIP-D<sup>v</sup>) and motion synchronization (F-LMD). Even though there is a slight disadvantage in CLIP-D compared to Ip2p, they are seriously affected by consistency (CLIP-D<sup>v</sup>), in which the editing is applied to the image level, and temporal consistency is not guaranteed. Our method also achieves the best on the BIQ and SD.

### 4.3 Quantitative Results

**Evaluation Results.** In our comparative analysis using a variety of frames, our method distinctly outperforms others by achieving a seamless artistic blend with the source, coupled with the retention of high-quality and consistent facial poses. Specifically, Ip2p excels in editing capabilities and maintaining a strong correlation between text and image. However, its performance is hindered by an inability to uniformly edit throughout an entire video sequence, which is reflected in its lower CLIP-D score. In contrast, RAV exhibits significant visual artifacts, especially around the mouth area

Methods	Editing-Quality↑ Video-Realness↑ Motion-Acc↑		
	User-Study		
RAV [43]	1.40	2.05	3.46
Ip2p [51]	1.07	1.81	3.38
TokenFlow [7]	4.03	3.22	3.86
<b>Emo-Avatar</b>	<b>4.53</b>	<b>4.05</b>	<b>4.71</b>

**Table 3: The MOS score for human evaluation. Each one comes from a 5-point Likertscale (1-Bad, 2-Poor, 3-Fair, 4-Good, 5-Excellent). The closer to 5 the better, we bold the best results.**

Methods	Stage 1					Stage 2		
	LPIPS↓	PSNR↑	SD↓	Time↓	CLIP-D↑	CLIP-D <sup>v</sup> ↑	Time↓	
NE-Avatar	0.085	25.63	0.225	120	0.143	0.627	30	
Unaligned-Inv	0.145	24.57	0.254	90	0.142	0.792	5	
Our w/o $\mathcal{L}_{patch}$	<b>0.058</b>	0.835	0.221	85	0.254	0.792	5	
Ours + in2n	-	-	-	-	0.224	0.792	180	
<b>Our</b>	<b>0.058</b>	<b>26.73</b>	<b>0.171</b>	60	0.263	<b>0.914</b>	5	

**Table 4: Quantitative comparison of our ablation studies.**

— characterized by unclear teeth and a general facial blur, as particularly noted in the “Bronze Statue” scenario. TokenFlow, while attempting to create realistic avatars, often diverges from the intended outcome, resulting in cartoon-like faces inconsistent with the specified prompt. Our model, leveraging the robust foundation established in its initial training phase, consistently replicates the target portrait across various views. During the editing process, it demonstrates an exceptional capacity to conform to the nuances of the reference image. Our pipeline maintains consistency and quality significantly elevates our model above all baseline comparisons.

**User Study.** A user study involving 12 participants and 25 video clips assessed the visual quality of generated portraits. We presented one video clip with the original clip and its corresponding edited result at a time and asked participants to respond to three statements: (1) “How do you like the editing quality in the video?”, (2) “How real is the video?” and (3) “Do you think the motion is accurate and consistent?”. The video clips are shown in a random order, and each video clip is shown exactly once to assess the first impression of participants. The results are shown in Tab. 3. Our method scored significantly higher in Editing-Quality (4.53 to 4.03), video quality (4.05 to 3.22), and motion consistency (4.71 to 4.86), demonstrating fewer artifacts and more consistent performance.

### 4.4 Ablation Study

**Effectiveness of Few-shot PTI Initialization.** The key idea of our Emo-Avatar is to utilize Few-shot PTI Initialization to embed portrait information into StyleGAN’s latent distribution to allow StyleGAN to represent intermediate head rotated avatars. We explore the following setting: (1) randomly select one image (2) randomly select multiple images (4) use images present extreme poses for initialization. As shown in Fig 6, randomly selecting a single image for initialization will lead to blurred generation for side views. However, by selecting the 4 extreme extrinsic parameters along the



**Figure 6: Few-shot PTI initialization from extreme poses helps the generator to capture the side views.**

x and y-axis, the initialization will be capable to consistently representing extreme views. In addition, we also explore the performance by utilizing extreme poses for few-shot PTI initialization, which demonstrates a lower image quality and longer time for training in the first stage.

**Effectiveness of Pre-trained StyleGAN Latent Space.** To explore whether our great rendering capability and fast training come from retaining StyleGAN’s latent distribution, we explored the following two settings. We build the ablation baseline named Non-Efficient Avatar(NE-Avatar). In this case, the model architecture of the whole pipeline is the same as our method; however, for NE-Avatar, we do not freeze StyleGAN’s parameters but train together with other modules during texture rendering. As shown in Tab 4, retaining StyleGAN’s latent distribution significantly reduced the time for editing from 30 minutes to 5 minutes.

## 5 CONCLUSION

In conclusion, our Emo-Avatar represents a significant advancement in efficient avatar generation. Combining StyleGAN with neural texture rendering, we achieve avatars that are not only motion-consistent but also quickly adaptable in style. Our method strikes a balance between natural, fluid movements and a range of stylistic options, validated through extensive testing. This approach not only showcases technical robustness but also has practical applications in various digital interaction scenarios. We anticipate that our work will inspire future innovations, particularly in integrating StyleGAN with Large Language Models (LLMs), potentially transforming the landscape of virtual communication and interactive experiences.

## REFERENCES

- [1] Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18392–18402.
- [2] Lele Chen, Zhiheng Li, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. 2018. Lip movements generation at a glance. In *Proceedings of the European conference on computer vision (ECCV)*. 520–535.
- [3] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. 2019. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 0–0.
- [4] Michail Christos Doukas, Mohammad Rami Koujan, Viktoria Sharmancka, Anastasios Roussos, and Stefanos Zafeiriou. 2021. Head2head++: Deep facial attributes re-targeting. *IEEE Transactions on Biometrics, Behavior, and Identity Science* 3, 1 (2021), 31–43.
- [5] Rinot Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. 2021. StyleGAN-NADA: CLIP-Guided Domain Adaptation of Image Generators. *arXiv:2108.00946 [cs.CV]*
- [6] Pablo Garrido, Levi Valgaerts, Ole Rehmsen, Thorsten Thormahlen, Patrick Perez, and Christian Theobalt. 2014. Automatic face reenactment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4217–4224.
- [7] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. 2023. TokenFlow: Consistent Diffusion Features for Consistent Video Editing. *arXiv preprint arXiv:2307.10373* (2023).
- [8] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. 2022. Neural head avatars from monocular rgb videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18653–18664.
- [9] Ayaan Haque, Mathew Tancik, Alexei A Efros, Aleksander Holynski, and Angjoo Kanazawa. 2023. Instruct-nerf2nerf: Editing 3d scenes with instructions. *arXiv preprint arXiv:2303.12789* (2023).
- [10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.
- [11] Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A Smith, and Jiebo Luo. 2022. Promptcap: Prompt-guided task-aware image captioning. *arXiv preprint arXiv:2211.09699* (2022).
- [12] He Huang, Philip S Yu, and Changhu Wang. 2018. An introduction to image synthesis with generative adversarial nets. *arXiv preprint arXiv:1803.04469* (2018).
- [13] Xun Huang and Serge Belongie. 2017. Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization. *arXiv:1703.06868 [cs.CV]*
- [14] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4401–4410.
- [15] 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.
- [16] Mohammad Rami Koujan, Michail Christos Doukas, Anastasios Roussos, and Stefanos Zafeiriou. 2020. Head2head: Video-based neural head synthesis. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. IEEE, 16–23.
- [17] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)* 36, 6 (2017), 194:1–194:17. <https://doi.org/10.1145/3130800.3130813>
- [18] Yuhan Li, Yishun Dou, Yue Shi, Yu Lei, Xuanhong Chen, Yi Zhang, Peng Zhou, and Bingbing Ni. 2023. Focaldreamer: Text-driven 3d editing via focal-fusion assembly. *arXiv preprint arXiv:2308.10608* (2023).
- [19] David Chuan-En Lin, Fabian Caba Heilbron, Joon-Young Lee, Oliver Wang, and Nikolas Martelaro. 2022. VideoMap: Video Editing in Latent Space. *arXiv preprint arXiv:2211.12492* (2022).
- [20] Jingyang Lin, Hang Hua, Ming Chen, Yikang Li, Jenhao Hsiao, Chiuman Ho, and Jiebo Luo. 2023. VideoXum: Cross-modal Visual and Textural Summarization of Videos. *arXiv preprint arXiv:2303.12060* (2023).
- [21] Michael Mathieu, Camille Couprie, and Yann LeCun. 2015. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440* (2015).
- [22] Mehdi Mirza and Simon Osindero. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784* (2014).
- [23] Anish Mittal, Anush K Moorthy, and Alan C Bovik. 2011. Blind/referenceless image spatial quality evaluator. In *2011 conference record of the forty fifth asilomar conference on signals, systems and computers (ASILOMAR)*. IEEE, 723–727.
- [24] 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 Trans. Graph.* 37, 6 (2018), 258–1.
- [25] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. 2022. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988* (2022).

- [26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.
- [27] Yurui Ren, Ge Li, Yuanqi Chen, Thomas H Li, and Shan Liu. 2021. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13759–13768.
- [28] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. 2021. Pivotal Tuning for Latent-based Editing of Real Images. *ACM Trans. Graph.* (2021).
- [29] Ruizhi Shao, Jingxiang Sun, Cheng Peng, Zerong Zheng, Boyao Zhou, Hongwen Zhang, and Yebin Liu. 2023. Control4D: Dynamic Portrait Editing by Learning 4D GAN from 2D Diffusion-based Editor. *arXiv preprint arXiv:2305.20082* (2023).
- [30] Chaehun Shin, Heeseung Kim, Che Hyun Lee, Sang-gil Lee, and Sungroh Yoon. 2023. Edit-A-Video: Single Video Editing with Object-Aware Consistency. *arXiv preprint arXiv:2303.07945* (2023).
- [31] Aliaksandr Siarohin, Stéphanie Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019. First Order Motion Model for Image Animation. In *Conference on Neural Information Processing Systems (NeurIPS)*.
- [32] Yuqi Sun, Reian He, Weimin Tan, and Bo Yan. 2023. Instruct-NeuralTalker: Editing Audio-Driven Talking Radiance Fields with Instructions. *arXiv preprint arXiv:2306.10813* (2023).
- [33] Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jia An, Jingyang Lin, Rongyi Zhu, et al. 2023. Video Understanding with Large Language Models: A Survey. *arXiv preprint arXiv:2312.17432* (2023).
- [34] 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.
- [35] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2016. Facevr: Real-time facial reenactment and eye gaze control in virtual reality. *arXiv preprint arXiv:1610.03151* (2016).
- [36] Lizhen Wang, Zhiyu Chen, Tao Yu, Chenguang Ma, Liang Li, and Yebin Liu. 2022. FaceVerse: a Fine-grained and Detail-controllable 3D Face Morphable Model from a Hybrid Dataset. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2022)*.
- [37] 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. *arXiv preprint arXiv:2305.00942* (2023).
- [38] Teng Wang, Jinru Zhang, Junjie Fei, Yixiao Ge, Hao Zheng, Yunlong Tang, Zhe Li, Mingqi Gao, Shanshan Zhao, Ying Shan, et al. 2023. Caption anything: Interactive image description with diverse multimodal controls. *arXiv preprint arXiv:2305.02677* (2023).
- [39] Siting Xu, Yunlong Tang, and Feng Zheng. 2023. Launchpadgt: Language model as music visualization designer on launchpad. *arXiv preprint arXiv:2307.04827* (2023).
- [40] Shuai Yang, Liming Jiang, Ziwei Liu, and Chen Change Loy. 2022. Vtoonify: Controllable high-resolution portrait video style transfer. *ACM Transactions on Graphics (TOG)* 41, 6 (2022), 1–15.
- [41] Shuai Yang, Liming Jiang, Ziwei Liu, and Chen Change Loy. 2023. StyleGANEX: StyleGAN-Based Manipulation Beyond Cropped Aligned Faces. *arXiv preprint arXiv:2303.06146* (2023).
- [42] Shuzhou Yang, Chong Mou, Jiwen Yu, Yuhuan Wang, Xiandong Meng, and Jian Zhang. 2023. Neural Video Fields Editing. *arXiv preprint arXiv:2312.08882* (2023).
- [43] Shuai Yang, Yifan Zhou, Ziwei Liu, , and Chen Change Loy. 2023. Rerender A Video: Zero-Shot Text-Guided Video-to-Video Translation. In *ACM SIGGRAPH Asia Conference Proceedings*.
- [44] 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*. Springer, 85–101.
- [45] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. 2019. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE/CVF international conference on computer vision*. 9459–9468.
- [46] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 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.
- [47] Sharon Zhang, Jiaju Ma, Jiajun Wu, Daniel Ritchie, and Maneesh Agrawala. 2023. Editing Motion Graphics Video via Motion Vectorization and Transformation. *ACM Transactions on Graphics (TOG)* 42, 6 (2023), 1–13.
- [48] Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C Bühlér, Xu Chen, Michael J Black, and Otmar Hilliges. 2022. Im avatar: Implicit morphable head avatars from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13545–13555.
- [49] Yufeng Zheng, Wang Yifan, Gordon Wetzstein, Michael J. Black, and Otmar Hilliges. 2023. PointAvatar: Deformable Point-Based Head Avatars From Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 21057–21067.
- [50] Jingyu Zhuang, Chen Wang, Liang Lin, Lingjie Liu, and Guanbin Li. 2023. Dreameditor: Text-driven 3d scene editing with neural fields. In *SIGGRAPH Asia 2023 Conference Papers*. 1–10.
- [51] Wojciech Zienolka, Timo Bolkart, and Justus Thies. 2023. Instant volumetric head avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4574–4584.