DiffPortrait3D: Controllable Diffusion for Zero-Shot Portrait View Synthesis

Yuming Gu[1,2], Hongyi Xu[2], You Xie[2], Guoxian Song[2], Yichun Shi[2],
Di Chang[1,2],Jing Yang[1], Linjie Luo[2]
[1]University of Southern California, [2]ByteDance Inc.
https://freedomgu.github.io/DiffPortrait3D
{yuminggu,dichang,jyang010}@usc.edu
{hongyixu,you.xie,guoxian.song,yichun.shi,linjie.luo}@bytedance.com

Abstract

We present DiffPortrait3D, a conditional diffusion model that is capable of synthesizing 3D-consistent photo-realistic novel views from as few as a single in-the-wild portrait. Specifically, given a single RGB input, we aim to synthesize plausible but consistent facial details rendered from novel camera views with retained both identity and facial expression. In lieu of time-consuming optimization and fine-tuning, our zero-shot method generalizes well to arbitrary face portraits with unposed camera views, extreme facial expressions, and diverse artistic depictions. At its core, we leverage the generative prior of 2D diffusion models pre-trained on large-scale image datasets as our rendering backbone, while the denoising is guided with disentangled attentive control of appearance and camera pose. To achieve this, we first inject the appearance context from the reference image into the self-attention layers of the frozen UNets. The rendering view is then manipulated with a novel conditional control module that interprets the camera pose by watching a condition image of a crossed subject from the same view. Furthermore, we insert a trainable cross-view attention module to enhance view consistency, which is further strengthened with a novel 3D-aware noise generation process during inference. We demonstrate state-of-the-art results both qualitatively and quantitatively on our challenging in-the-wild and multi-view benchmarks.

Figure 1. Given a single portrait as reference (left), DiffPortrait3D is adept at producing high-fidelity and 3d-consistent novel view synthesis (right). Notably, without any finetuning, DiffPortrait3D is universally effective across a diverse range of facial portraits, encompassing, but not limited to, faces with exaggerated expressions, wide camera views, and artistic depictions.

1. Introduction

Faithfully reconstructing the 3d appearance of human faces from a single 2D unconstrained portrait is a long-standing goal for computer vision, with a wide range of downstream applications in visual effects, digital avatars, 3D animation, and many others. In this work, we challenge ourselves to synthesize *high-fidelity consistent* novel views from as few as a single portrait, with *high resemblance* to the inputs in both individual appearance, expression and background content. Notably to the best of our knowledge, we are the first *zero-shot* novel portrait synthesis work that supports versatile facial appearances and backgrounds, exaggerated expressions, wide views, and a plethora of artist styles.

Long-range portrait view synthesis from sparse inputs requires a generative prior to hallucinating plausible scene features that are unobserved in the inputs. Recently, 3D aware generative adversarial network (GAN) [2, 5, 6, 11, 16, 35, 42, 53, 54] demonstrated striking quality and multi-

view-consistent image synthesis, by integrating 3D neural representations [34, 52] with style-based image generation [15, 24, 25]. Thereafter a line of work [3, 29, 39, 45, 55] has explored either optimization-based or encoder-based approaches to carefully invert the image into the latent or feature embedding of 3D GANs, and then synthesize novel views with 3D-aware generative priors. Nevertheless, almost all existing 3D-aware GANs are trained on limited image datasets. Hence when it comes to much more wild and nuanced portraits with large domain gap with the training distributions, GANs tend to struggle in faithfully depicting the 3D faces, resulting in loss of resemblance, corrupted geometry, or blurry extrapolation (see Figure 3, 4).

With the recent advent of text-to-image diffusion models [20, 40, 43, 44], we have witnessed unprecedented diversity and stability in image synthesis exhibited by large diffusion models pre-trained on billions of images, such as Imagen [41] and Stable Diffusion (SD) [1]. We therefore aim to capitalize on the generative power of production-ready diffusion models (SD in our work), for the task of portrait view synthesis. However, unlike previous 3D GAN-inversion works, simply inverting the reference image into a generative noise or a textual description does not naturally lift the image into a 3D scene, and it struggles to retain consistent appearances when deviating from the reference view. The introduction of ControlNet [57] enhances the controllability of Stable Diffusion by injecting localized spatial conditions. However, it remains unclear how to achieve appearance-disentangled view control such as in the paradigm of ControlNet. Moreover, without inherent 3D representation, the direct application of existing 2D image diffusion models to long-range animated view synthesis results in severe flickering artifacts.

In this work, we propose *DiffPortrait3D*, a novel zero-shot approach that lifts 2D diffusion model for synthesizing 3D consistent novel views from as few as a single portrait. Our key insight is to decompose the task into explicitly disentangled control of appearance and camera view. Specifically, we first utilize a trainable copy of the SD UNets to derive semantic appearance context from the reference image and then provide layer-by-layer contextual guidance to the self-attention modules of a locked SD network. This allows us to preserve the capability of the large diffusion models while generating images with retained reference characteristics regardless of the rendering views. On top of that, we further achieve view control by adding camera pose attention to the locked UNet decoder as done in ControlNet [57]. By design, the camera pose attention is intelligently extracted from an RGB portrait image of a proxy subject captured at the same view, to minimize appearance leakage from the condition image (e.g., shape and expression from landmarks). Additionally, to alleviate flickering artifacts when animating the views, we adopt a cross-view attention module as used in many video diffusion models [17, 21]. This ensures the unobservable region is completed in a consistent fashion. View consistency is further enhanced during inference with a novel 3D-aware noise generation process.

With the locked parameters of Stable Diffusion, we fine-tuned our control modules in stages with multi-view synthetic dataset by PanoHead [2] and real-image Nersemble dataset [27]. Our method demonstrates native generalization capability to in-the-wild portraits without run-time fine-tuning. We extensively evaluate our framework on a few challenging benchmarks. DiffPortrait3D outperforms prior methods both quantitatively and qualitatively in terms of visual quality, resemblance, and view consistency. The contributions of our work can be summarized as:

- A novel zero-shot view synthesis method that extends 2D Stable Diffusion for generating 3d consistent novel views given as little as a single portrait.
- We demonstrate compelling fine-tuning-free novel view synthesis results given a single unconstrained portrait, regardless of its appearance, expression, pose, and style.
- Explicitly disentangled control of appearance and camera view, enabling effective camera control with preserved identity and expression.
- Long-range 3D view consistency with a cross-view attention module and 3D-aware noise generation.

Our code and model will be available for research purposes.

2. Related Works

Our study focuses on the application of 2D diffusion models for zero-shot portrait novel view synthesis (NVS). Within this context, we undertake an extensive survey of progress in techniques related to novel view synthesis, categorized into regression-based and generative approaches.

Regression based NVS. Facial NVS is attainable through the use of explicit parametric geometry priors, as demonstrated by 3D Morphable Models (3DMM) [13, 36, 47, 50, 59]. However, the limited parametric space of 3DMM poses challenges in faithfully depicting diverse facial expressions. Recent strides in Neural Radiance Fields (NeRF) [16, 22, 34, 56] have yielded high-fidelity results in novel view synthesis. Notably in the realm of portrait NVS, FDNeRF [56] constructed a NeRF model that integrates aligned features from inputs to generate novel view portraits. Nevertheless, achieving photo-realistic 3D-aware novel views with such models typically necessitates the availability of dense calibrated images.

Generative NVS with GAN GANs [15] employ adversarial learning to synthesize images that faithfully capture the distribution of the training dataset. Previous studies have demonstrated the effectiveness of 2D GANs in portrait manipulation, employing techniques such as latent space exploration [8] and exemplar image utilization [23, 51].

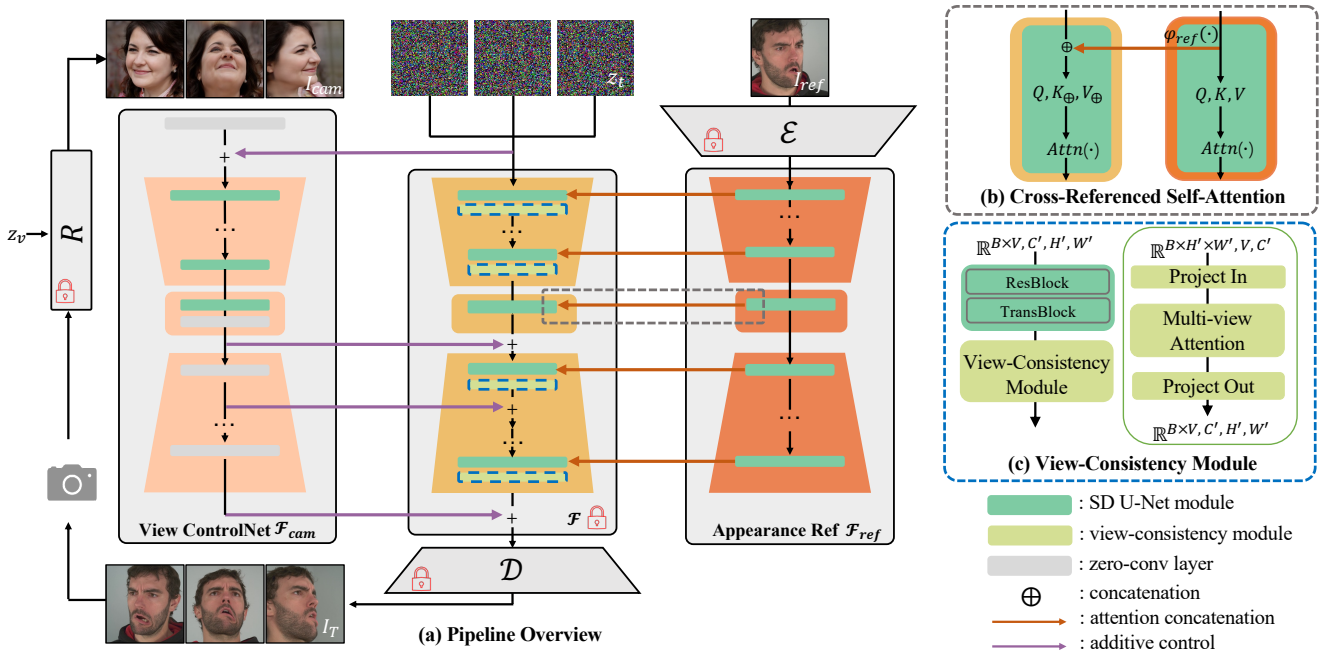

Figure 2. (a) Overview of our DiffPortrait3D framework. Given a single reference image I_{ref}, we aim to synthesize its novel views as I_T at camera perspectives aligned with condition images I_{cam}. We leverage a pre-trained LDM $\mathcal{F}$ as our image synthesis backbone (middle), where its self-attention layers cross query the appearance context from I_{ref} via our appearance reference module $\mathcal{F}_{ref}$ (right). Our view control module (left) $\mathcal{F}_{cam}$ derives additive view condition from I_{cam} and exerts on $\mathcal{F}$. Additionally, we plug in view consistency modules (dotted rectangles, middle) to $\mathcal{F}$ to enhance multi-view coherence. During training, the images I_{cam} are rendered using an off-the-shelf 3D GAN renderer R, where its camera perspectives are aligned with I_T. (b) The intermediate spatial features $\varphi(\cdot)$ sourced from I_{ref} are concatenated into the corresponding self-attention blocks in $\mathcal{F}$. (c) An attention mechanism is employed across the multi-view dimensions by our view-consistency module.

Nevertheless, the absence of inherent 3D representations in these 2D GANs presents a challenge in maintaining 3D consistency for the task of NVS.

Recent advancements on 3D aware GANs [2, 5, 6, 11, 16, 35, 42, 53, 54], built upon foundations of 2D GANs, have demonstrated striking quality and multi-view-consistent image synthesis. These methodologies typically leverage StyleGAN2 [26] as a fundamental component, incorporating it with differential rendering and diverse 3D representations, such as signed distance functions as in StyleSDF [35] and tri-plane representations used by EG3D [6]. Thereafter a line of work [3, 29, 39, 45, 55] has explored either optimization-based or encoder-based approaches to carefully invert the image into the latent or feature embedding of 3D GANs, and then synthesize novel views with 3D-aware generative priors. It is noteworthy, however, that these methods heavily depend on a pre-trained 3D GAN generator and exhibit limitations in their capacity to generate unposed portraits with in-the-wild expressions, styles, and camera views.

Diffusion Model based NVS In lieu of directly confronting the intricacies of learning a 3D diffusion model, recent research endeavors have embraced an alternative strategy, harnessing powerful 2D diffusion models to improve the processes of 3D modeling and novel view synthesis. DreamFusion [38] pioneered this strategy by distilling a 2D text-to-image generation model for fine-tuning a NeRF model. GENVS [7] introduced a diffusion-based model explicitly tailored for 3D-aware generative novel view synthesis from a single input image. Their methodology involves modeling samples from the potential rendering distribution, effectively mitigating ambiguity and generating plausible novel views through the utilization of diffusion processes. Recent noteworthy study, Zero-1-to-3 [31, 32] utilizes a stable diffusion model to capture geometric priors derived from an extensive synthetic dataset, yielding high-quality predictions. Moreover, Consistent123 [30], a case-aware approach, utilizes Zero-1-to-3 as 3D prior for the initial structural representation before generating high texture fidelity. However, it is crucial to note that these approaches primarily concentrate on general objects, resulting in a diminished quality when applied to portrait synthesis.

3. Methods

Given as few as a single RGB portrait image, denoted as I_{ref}, captured from any camera perspective, we aim to syn-

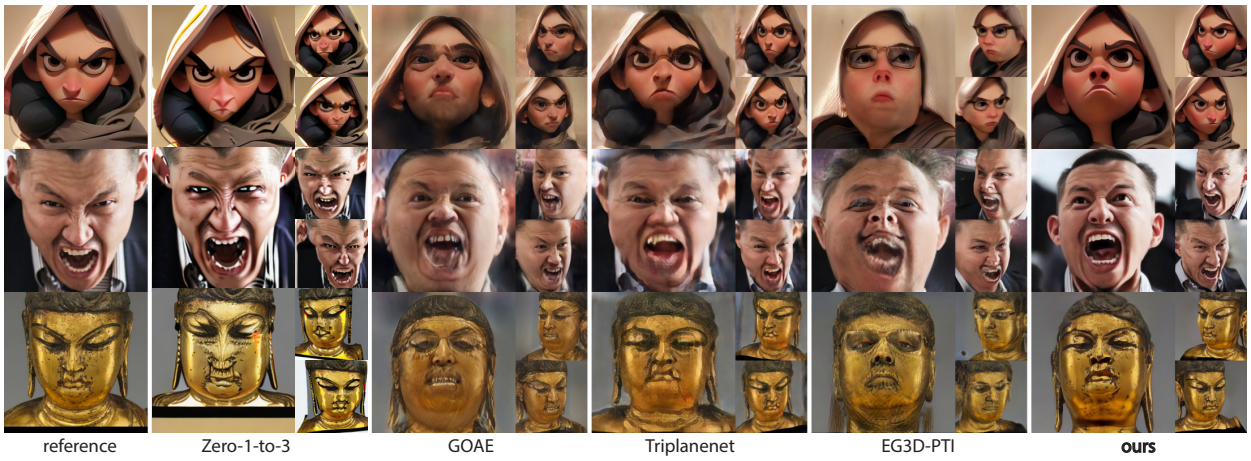

Figure 3. **Qualitative comparison of novel view synthesis on in-the-wild images.** Compared to the baselines, our method shows superior generalization capability to novel view synthesis of wild portraits with unseen appearances, expressions and styles, even without any reliance on fine-tuning.

thesize a new image I_T at a novel query view as indicated by a condition image I_{cam}. The synthesized image I_T should retain the expression and appearance of the foreground individual as well as the background context as in I_{ref}, while follows the rendering view of I_{cam}. Note that I_{cam} and I_{ref} could be of a completely different identity.

Our proposed approach, DiffPortrait3D, leverages a latent diffusion model (LDM) as the backbone of our rendering framework, as depicted in Figure 2 (a) (Section 3.1). We then introduce an auxiliary appearance control branch (Section 3.2) to exert layer-by-layer guidance with local structures and textures from reference images I_{ref}. To enable effective camera control with I_{cam}, our view control module, designed in a fashion of ControlNet [57], implicitly derives camera pose from I_{cam} and inject to the diffusion process as an additive condition (Section 3.3). Lastly we discuss about enhancing view consistency with our integrated multi-view attentions, and noise generation with 3D awareness at inference (Section 3.4).

3.1. Preliminaries

Latent Diffusion Models. Diffusion models [20, 43, 44] are generative models designed to synthesize desired data samples from Gaussian noise via removing noises iteratively. Latent diffusion models [40] are a class of diffusion models that operates in the encoded latent space of an autoencoder $\mathcal{D}(\mathcal{E}(\cdot))$, where $\mathcal{E}$ and $\mathcal{D}$ denotes the encoder and decoder respectively. Specifically, given an image I and the text condition c_{text}, the encoded image latent $z_0 = \mathcal{E}(I)$ is diffused T time steps into a Gaussian-distributed $z_T \sim \mathcal{N}(0,1)$. The model is then trained to learn the reverse denoising process with the objective,

$$L_{ldm} = \mathbb{E}_{z_0, c_{text}, t, \epsilon\sim\mathcal{N}(0,1)}\left[\left\|\epsilon - \epsilon_\theta\left(z_t, c_{text}, t\right)\right\|_2^2\right], \quad (1)$$

The ϵ_θ is formulated as a trainable U-Net architecture with layers of intervened convolutions (ResBlock) and self-/cross-attentions (TransBlock). In this paper, we build our network as a plug-and-play module to the recent state-of-the-art text-to-image latent diffusion model, Stable Diffusion [1].

ControlNet. As introduced by [57], ControlNet effectively enhances latent diffusion models with spatially localized, task-specific image conditions. As its core, it replicates the original Stable Diffusion as a trainable side path, and adds additional "zero convolution" layers. The extra conditions outputted from the "zero convolution" layers are then added to the skipped connections of the SD-UNets. Let c_p be the extra condition, the noise prediction of U-Net with ControlNet then becomes $\epsilon_\theta\left(z_t, c_{text}, c_p, t\right)$.

3.2. Appearance Reference Module

In order to synthesize a novel view of I_{ref} with LDM, one could try to condition the denoising with an "inverted" text condition c_{text} [28]. However, providing a precise textual description of I_{ref} for LDM to comprehensively recover all its components is often a challenging undertaking. Alternatively, one could also condition ϵ_θ on I_{ref} directly as a ControlNet. Such a design, however, tend to generate images predominantly influenced by the camera pose in I_{ref}. Inspired by [4, 49], we opt for integrating appearance attributes of the reference image I_{ref} into the UNet backbone as cross-referenced self-attentions. Note that to eliminate the harmful influence of inaccurate text description, we set c_{text} empty and use the reference image I_{ref} as the only source of appearance.

To illustrate our appearance reference module, let us denote the pretrained LDM as $\mathcal{F}$, where its self-attention is

Figure 4. **Qualitative comparison of novel view synthesis on NeRSemble [27].** Our method achieves effective view control for novel synthesis with the best perceptual quality and retained identity and expression, even for portraits with exaggerated expressions and under substantial change of camera view for synthesis.

calculated as

$$Attn(\cdot) = softmax(\frac{QK^T}{\sqrt{d}}) \cdot V \tag{2}$$

$$Q = W_Q \cdot \varphi(z_t), K = W_K \cdot \varphi(z_t), V = W_V \cdot \varphi(z_t), \tag{3}$$

where Q, K, V are the query, key, and value features projected from the spatial features $\varphi(z_t)$ with corresponding projection matrices respectively.

To guide the denoising process with I_{ref}, we adapt the self-attention mechanism within $\mathcal{F}$ such that it is able to cross query the correlated local contents and textures from $\mathcal{E}(I_{ref})$, in addition to its own spatial features. Specifically we replicate $\mathcal{F}$ into a trainable counterpart $\mathcal{F}_{ref}$ with $\varphi_{ref}(\cdot)$ serving as intermediate representations within the UNet architecture. As depicted in Figure 2 (b), we then modify the vanilla self-attention in $\mathcal{F}$ in a way that the spatial context $\varphi_{ref}(\mathcal{E}(I_{ref}))$ in the appearance branch $\mathcal{F}_{ref}$ is cross-queried layer by layer as,

$$\begin{aligned} K_{\oplus} &= W_K \cdot (\varphi(z_t) \oplus \varphi_{ref}(\mathcal{E}(I_{ref}))), \\ V_{\oplus} &= W_V \cdot (\varphi(z_t) \oplus \varphi_{ref}(\mathcal{E}(I_{ref}))), \end{aligned} \tag{4}$$

where $\oplus$ denotes concatenation. Note that we do not apply noise to I_{ref}, ensuring meticulous transfer of referenced structure and appearance attributes into the novel portrait synthesis. We lock the parameters of SD-UNet $\mathcal{F}$, and train our appearance reference module $\mathcal{F}_{ref}$ with paired multi-view images.

Notable, when more reference images are available, e.g., in some multi-view capture settings, our appearance reference module can be easily extended by concatenating multiple appearance contexts as

$$\varphi(z_t) \oplus \varphi_{ref}(\mathcal{E}(I^1_{ref})) \oplus ... \oplus \varphi_{ref}(\mathcal{E}(I^n_{ref})). \tag{5}$$

Our trained module is capable of seamlessly integrating the multi-view appearance clues into 3D-consistent appearance context (Figure 8).

3.3. View Control Module

In this stage, we aim to attain control over the synthesis viewpoint without influencing either the derived appearance attributes by $\mathcal{F}_{ref}$ or the synthesis capability of a pre-trained LDM $\mathcal{F}$. This naturally leads to the paradigm of ControlNet [57] where the additional view control is connected via "zero convolution" layers of a trainable LDM copy, with both $\mathcal{F}_{ref}$ and $\mathcal{F}$ locked. Here we denote our view control module as $\mathcal{F}_{cam}$, to be trained with multi-view images. One straightforward design of $\mathcal{F}_{cam}$ would be to employ the spatial feature maps extracted from the ground-truth target images as image conditions, such as landmarks, segmentation, or edges. We note that such "ground-truth" condition images are not available during inference and therefore the view is typically manipulated with images of a different identity. However, we argue that such condition images contain entangled semantic appearance information, such as shape and expression, which is likely to be passed along with the camera pose to $\mathcal{F}$. Herein, appearance leakage from the view condition image will be reflected on the novel view synthesis during inference. This artifact is more pronounced when I_{ref} and I_{cam} exhibit distinct appearance features.

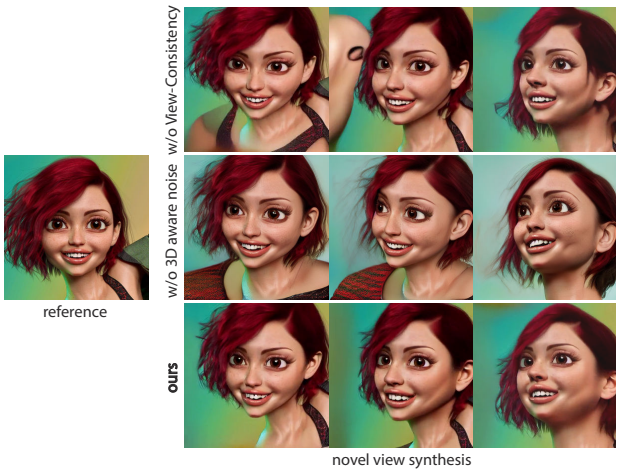

Figure 5. **Ablation on view consistency.** Excessive background variation and slight shading change across multiple novel views are observable without our view-consistency module. Our 3D-aware noise, compared to random Gaussian noise, helps maintain structural coherence during view animation.

Instead, we utilize a portrait image from a distinct random identity as the view condition, and generate novel-view images that mirror the head pose as in the condition portrait I_{cam}. Our design unifies the view manipulation setting in training and inference, and facilitates the natural disentanglement of view and appearance control. However, training ControlNet for cross-identity view control requires paired images at a identical view, and obtaining such data pairs is typically unfeasible in real-world capture settings. To address this hurdle, we leverage off-the-shelf 3D GAN renders $\mathcal{R}(v, z_v)$, as exemplified in prior works [2, 6], to generate synthetic pose images I_{cam}. Here the v denotes the camera parameters calibrated from the target image and z_v is a random Gaussian noise input to the 3D GAN. Since I_{cam} and I_{ref} possess substantial difference in expression and appearance, our view control module is therefore instructed to derive camera pose from I_{cam} only. Moreover, by design, the camera pose is directly interpreted by our view control module, allowing us to mimic the rendering view simply with an RGB image. This largely eases the cumbersomeness in feature processing of I_{cam}, e.g., landmarks detection or semantic parsing, which could be unreliable with heavy occlusion or under wide views.

3.4. View Consistency Module

To this end, we have facilitated the generation of a novel-view portrait via the seamless combination of an appearance reference module, a view ControlNet and a pre-trained LDM. Nevertheless, achieving consistency in features across various views poses a significant challenge as many explanations exist for the unobservable region. Inspired by AnimateDiff [17], we introduce a view consistency module that incorporates cross-view attention within a batch of views. Such a module employ an attention mechanism along the dimension of views to establish feature correlation among the multiple novel view synthesis. Similar to AnimateDiff, we integrate these view consistency modules into the up- and down-sampling blocks of the LDM $\mathcal{F}$, as depicted in Figure 2 (c). However, we note that such frame-wise modules were originally proposed for temporal coherence and as motion prior, trained with sequential video frames. In contrast, the animated view motion is purely defined by the sequence of I_{cam}. Therefore, we trained our view-consistency modules with batches of randomly shuffled views, permitting the modules to focus on cross-view attentions in lieu of motion distribution.

Figure 6. **Reconstruction**. DiffPortrait3D demonstrates meticulous reconstruction of referenced appearance, even with side views and 3D cartoon styles, substantially outperforming the baseline methods.

As illustrated in Figure 2 (c), we train our view-consistency modules in groups of multi-view condition images $\{I_{cam}\}$ with a shape of (B, V, C, H, W), where B and V are the batch size and the number of views, while C, H, W denote the number of image channels, image height and width respectively. We note that the appearance within each batch of generation is referenced from the same image I_{ref}. Inside $\mathcal{F}$, we reshape the input to ResBlocks and TransBlocks as $(B \times V, C', H', W')$, where C', H', W' represent the latent feature channel, height and width respectively. Following the operations of self- and cross-attention, we then transform the layer input into a shape of $(B \times H' \times W', V, C')$, performing view-wise attention within the view consistency modules.

3D-aware inference. It has been empirically observed that the image layout is formed in the early denoising steps. Therefore instead of denoising from multiple random Gaussian noises, structural and textural consistency is likely to be enhanced when synthesizing multiple novel views by initiating the denoising process from "3D-consistent" noise samples. We propose an efficient two-stage process to generate noise samples with 3D awareness. On our multi-view image dataset, we first trained a 3D-convolution based NVS model with inclusion of 3D feature field and neural feature rendering (please refer to the supplementary paper for details). We employ this NVS model to provide a proxy syn-

	Ours	Eg3D-PTI	GOAE	Triplanenet
POSE ↓	-/0.0068/-	-/**0.0021**/-	-/0.0022/-	-/0.0134/-
LPIPS ↓	**0.02/0.23/0.02**	0.21/0.26/0.36	0.12/0.28/0.21	0.10/0.39/0.17
SSIM ↑	**0.92/0.62/0.93**	0.66/0.59/0.46	0.72/0.57/0.54	0.76/0.50/0.63
DIST ↓	**0.06/0.21/0.04**	0.21/0.24/0.27	0.15/0.25/0.20	0.15/0.29/0.19
ID ↑	**0.95/0.70/0.92**	0.15/0.28/0.12	0.55/0.39/0.54	0.70/0.45/0.70
FID ↓	**7.65/27.4/11.7**	33.13/56.2/95.0	54.10/84.8/92.0	63.54/112.1/88.0

(a)

	w/o $\mathcal{F}_{ref}$ finetuning	I_{ref} unaligned	I_{ref} aligned
LPIPS ↓	0.55	0.28	**0.27**
SSIM ↑	0.47	**0.68**	**0.68**
DIST ↓	0.43	0.19	**0.18**
ID ↑	0.21	**0.70**	**0.70**
FID ↓	99.09	25.77	**25.37**

(b)

Table 1. (a) Quantitative comparison of our method and GAN-based baselines, showing numerical results of reconstruction/novel view synthesis of NeRSemble [27], and reconstruction of in-the-wild test images(from left to right). For a fair comparison to our baselines, the evaluation is performed at the resolution of 256×256. (b) Ablation study of our method without finetuning appearance reference module, with unaligned reference images, and with aligned reference images, evaluated on NeRSemble at the resolution of 512×512.

thesis $\tilde{I}$ at the target novel view, which is typically blurry but 3D consistent. We then diffuse the latent feature $\mathcal{E}(\tilde{I})$ with 1000 time steps into a Gaussian noise as the input to the LDM. In essence, the two-step generated noise still contains some image layout semantics in a very coarse grain and in practice, enhanced view consistency is observed in our task as demonstrated in Figure 5.

4. Experiments

Dataset and Training. Our model was trained in three stages on our multi-view image dataset as an image reconstruction task. That being said, both the appearance reference image I_{ref} and the target image I_T are sourced from the same identity but with different views, whereas I_{cam} is synthesized with EG3D [6] using a random latent Gaussian noise and the calibrated camera parameters of I_T. We lock the parameters of the SD-UNet $\mathcal{F}$ during the whole training stage. In the first stage, we train all the parameters of our appearance reference module $\mathcal{F}_{ref}$ without any camera guidance. Next we freeze the weights of $\mathcal{F}_{ref}$, and train our view control module $\mathcal{F}_{cam}$ with paired I_{cam}. Lastly the view consistency module, performing cross-view attentions among 8 views at once, is trained with the rest modules frozen. All training was conducted on 6 Nvidia A100 GPUs at a learning rate of 10^{-5}, with 16 images processed in each step. During inference, we empirically set 100 steps for DDIM denoising [43] and unconditional guidance scale [19] as 3 for a good balance of quality and speed.

We trained our modules on a hybrid dataset comprised of photo-realistic multi-view images NeRSemble [27] and synthetic ones by PanoHead [2]. NeRSemble dataset consists of high-resolution videos of 220 subjects performing a wide range of dynamic expressions, captured from 16 calibrated synchronized cameras. We sampled 2000 pairs of multi-view frames from NeRSemble for training, where 1 randomly-selected view is used for appearance reference and 8 other views as targets. Given the scarcity of available camera views and the background variation, we augmented our training dataset with another 2000 pairs of multi-view images synthesized via PanoHead [2]. For evaluation, we used another unseen 500 multi-view pairs from NeRSemble, and 360 single-view internet-collected in-the-wild portraits [12, 33, 37], containing a wide variation in appearance, expression, camera perspective, and style. We note that for training, all the images are cropped and aligned as in EG3D [6] whereas we do not perform image alignment during inference (unless explicitly stated for comparison to GAN-based methods). For testing on both datasets, the novel camera views are all manipulated with EG3D renderings.

4.1. Qualitative Evaluations

Given a single reference portrait, our method demonstrates high-fidelity and 3D-consistent novel view synthesis at a resolution of 512×512, as illustrated in Figure 1. While only being trained on aligned real portrait images, our method shows superior generalization capability to novel identities, styles, expressions and views. This is largely credited to the preservation of the generative prior of pre-trained LDM by our design. As evidenced in Figure 4, our view control module is also able to effectively control the synthesis view. Compared to the ground truth (second column, Figure 4), our novel portraits are highly plausible but with some noticeable identity differences. This is due to the limited visual appearance clue in the single reference image, and the problem can be largely alleviated with additional references (please refer to Figure 8 and the supplementary paper for visual results).

We extensively compare to a few state-of-the-art novel portrait synthesis works on both image reconstruction (Figure 6) and novel view synthesis (Figure 3, 4): GOAE [55],TriPlaneNet [3], Pivot Tuning (EG3D-PTI) [6] and Zero-1-to-3 [32]. GOAE [55] and TriPlaneNet [3] designed an effective image encoder for EG3D [6], whereas

EG3D-PTI runs latent code optimization and finetunes the weights of EG3D per image. We did not compare to Live3D Portrait [46] given unavailable implementation and model. Zero-1-to-3 [32] leverages Stable Diffusion but was trained on 3D object dataset Objaverse [9]. While not required by our method, we cropped and aligned the test images as in EG3D. Nevertheless, our method outperforms substantially over the prior work in terms of both perceptual quality, and preservation of identity and expression. Notably all 3D GAN-based baselines fail to reconstruct side views (Figure 6), exaggerated expressions (Figure 4), or out-of-domain styles (Figure 3), whereas Zero-1-to-3 synthesizes novel portraits with very limited perceptual quality.

4.2. Quantitative Evaluations

We evaluate methods for single-view novel portrait synthesis on 4 main aspects. We use LPIPS↓ [58], DISTS↓ [14], SSIM↑ [48] for evaluation of 2D image reconstruction, ID↑ [10] for identity consistency, FID↓ [18] for perceptual quality, and POSE ↓ for camera view control accuracy. Notably, to evaluate reconstruction fairly, we estimate camera parameters from the ground-truth target image and uses the EG3D renderings as condition I_{cam}. The error in camera estimation could result in some image misalignment and therefore we mainly rely on perceptual metrics LPIPS and DISTS for reconstruction evaluation. The identity similarity is calculated between the synthesized and reference image by calculating the cosine similarity of the face embeddings with a pretrained face recognition module [10].

Table 1a shows the numerical comparison on reconstruction of NeRSemble and in-the-wild test images, and novel view synthesis of NeRSemble respectively. On all image metrics, our method shows our method is superior than all prior work by a large margin, demonstrating the most compelling image quality. Our pose reconstruction is slightly worse than the baseline. However, we argue that this is largely due to the camera misalignment between the ground truth and the condition EG3D rendering.

4.3. Ablations

We ablate the efficacy of the individual component with extensive ablation experiments for noval view synthesis on NeRSemble test set. As illustrated in Figure 5, we demonstrate the necessity of our view consistency module and 3D-aware noise in maintaining appearance coherence cross multiple views. Without them, substantial variations are observed, especially on the unobserved region of the reference image, when altering the camera views. The weights of our appearance reference module is initiated from a copy of SD-UNet which should be already able to derive local appearance context from the reference image. However, as evidenced by Table 1b and Figure 7, significant improvements are achieved by our finetuning on multi-view images. We

Figure 7. Fine-tuning appearance reference module helps better retain the spatial features from the reference image.

Figure 8. Our method seamlessly supports multiple reference images as input, and the novel view synthesis quality is progressively enhanced with more references.

reason that the necessity of finetuning is due to the removal of cross attention from text. Lastly unlike many GAN-based methods that requires the reference image to be aligned, our model supports free-form portraits as inputs without quality degeneration even though the model was trained on camera-aligned multi-view images. This is numerically shown in Table 1b where aligning the reference image (as in EG3D) only leads to neglectable differences.

5. Discussion

Conclusion. We presented *DiffPortrait3D*, a novel conditional diffusion model that is capable of generating consistent novel portraits from sparse input views. By design, our framework seamlessly cross-references the key characteristics from the input images and effectively adds camera pose control into the latent diffusion process, modulated with enhanced consistency across views. Trained only with a few thousand of synthetic and real multi-view images, our model successfully showcases compelling novel portrait synthesis results, regardless of appearances, expressions, camera perspectives, and styles. This is largely credited to our explicitly disentangled control of appearance and view within both model design and training, without harming the generalization capability of large pretrained diffusion models. We believe that our framework opens up possibilities for accessible 3D reconstruction and visualization from a single picture.

References

[1] Stability AI. Stable diffusion v1.5 model card. *https://huggingface.co/runwayml/stable-diffusion-v1-5*, 2022. 2, 4

[2] Sizhe An, Hongyi Xu, Yichun Shi, Guoxian Song, Umit Y. Ogras, and Linjie Luo. Panohead: Geometry-aware 3d full-head synthesis in 360deg. In *CVPR*, pages 20950–20959, 2023. 1, 2, 3, 6, 7

[3] Ananta R Bhattarai, Matthias Nießner, and Artem Sevastopolsky. Triplanenet: An encoder for eg3d inversion. *arXiv preprint arXiv:2303.13497*, 2023. 2, 3, 7

[4] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. *arXiv preprint arXiv:2304.08465*, 2023. 4

[5] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5799–5809, 2021. 1, 3

[6] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In *CVPR*, 2022. 1, 3, 6, 7

[7] Eric R. Chan, Koki Nagano, Matthew A. Chan, Alexander W. Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. GeNVS: Generative novel view synthesis with 3D-aware diffusion models. In *arXiv*, 2023. 3

[8] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets, 2016. 2

[9] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 8

[10] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, pages 4690–4699, 2019. 8

[11] Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. Gram: Generative radiance manifolds for 3d-aware image generation. *CVPR*, 2022. 1, 3

[12] DeviantArt. deviantart. *https://www.deviantart.com*, 2024. 7

[13] Abdallah Dib, Cedric Thebault, Junghyun Ahn, Philippe-Henri Gosselin, Christian Theobalt, and Louis Chevallier. Towards high fidelity monocular face reconstruction with rich reflectance using self-supervised learning and ray tracing, 2021. 2

[14] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *CoRR*, abs/2004.07728, 2020. 8

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 2

[16] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. *CVPR*, 2022. 1, 2, 3

[17] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 2, 6

[18] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 8

[19] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 7

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020. 2, 4

[21] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv:2204.03458*, 2022. 2

[22] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. Headnerf: A real-time nerf-based parametric head model. In *CVPR*, pages 20374–20384, 2022. 2

[23] Omer Kafri, Or Patashnik, Yuval Alaluf, and Daniel Cohen-Or. Stylefusion: A generative model for disentangling spatial segments, 2021. 2

[24] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 2

[25] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *CVPR*, 2020. 2

[26] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 3

[27] Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. Nersemble: Multi-view radiance field reconstruction of human heads. *ACM Transactions on Graphics*, 2023. 2, 5, 7

[28] Dongxu Li, Junnan Li, and Steven CH Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *arXiv preprint arXiv:2305.14720*, 2023. 4

[29] Connor Z Lin, David B Lindell, Eric R Chan, and Gordon Wetzstein. 3d gan inversion for controllable portrait image animation. *arXiv preprint arXiv:2203.13441*, 2022. 2, 3

[30] Yukang Lin, Haonan Han, Chaoqun Gong, Zunnan Xu, Yachao Zhang, and Xiu Li. Consistent123: One image to highly consistent 3d asset using case-aware diffusion priors, 2023. 3

[31] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization, 2023. 3

[32] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object, 2023. 3, 7, 8

[33] Midjourney. midjourney. *https://www.midjourney.com*, 2024. 7

[34] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 2

[35] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. Stylesdf: High-resolution 3d-consistent image and geometry generation. *CVPR*, 2022. 1, 3

[36] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *2009 sixth IEEE international conference on advanced video and signal based surveillance*, pages 296–301. Ieee, 2009. 2

[37] Pexels. pexels. *https://www.pexels.com/*, 2024. 7

[38] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv*, 2022. 3

[39] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Transactions on Graphics*, 42(1):1–13, 2022. 2, 3

[40] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 2, 4

[41] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS*, 35:36479–36494, 2022. 2

[42] Ivan Skorokhodov, Sergey Tulyakov, Yiqun Wang, and Peter Wonka. Epigraf: Rethinking training of 3d gans. *arXiv preprint arXiv:2206.10535*, 2022. 1, 3

[43] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2, 4, 7

[44] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 2, 4

[45] Alex Trevithick, Matthew Chan, Michael Stengel, Eric R. Chan, Chao Liu, Zhiding Yu, Sameh Khamis, Manmohan Chandraker, Ravi Ramamoorthi, and Koki Nagano. Real-time radiance fields for single-image portrait view synthesis. In *SIGGRAPH*, 2023. 2, 3

[46] Alex Trevithick, Matthew Chan, Michael Stengel, Eric R. Chan, Chao Liu, Zhiding Yu, Sameh Khamis, Manmohan Chandraker, Ravi Ramamoorthi, and Koki Nagano. Real-time radiance fields for single-image portrait view synthesis, 2023. 8

[47] Satoshi Tsutsui, Weijia Mao, Sijing Lin, Yunyi Zhu, Murong Ma, and Mike Zheng Shou. Novel view synthesis for high-fidelity headshot scenes. *arXiv preprint arXiv:2205.15595*, 2022. 2

[48] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 8

[49] Haohan Weng, Tianyu Yang, Jianan Wang, Yu Li, Tong Zhang, CL Chen, and Lei Zhang. Consistent123: Improve consistency for one image to 3d object synthesis. *arXiv preprint arXiv:2310.08092*, 2023. 4

[50] Fanzi Wu, Linchao Bao, Yajing Chen, Yonggen Ling, Yibing Song, Songnan Li, King Ngi Ngan, and Wei Liu. Mvf-net: Multi-view 3d face morphable model regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 959–968, 2019. 2

[51] Sitao Xiang, Yuming Gu, Pengda Xiang, Mingming He, Koki Nagano, Haiwei Chen, and Hao Li. One-shot identity-preserving portrait reenactment. *arXiv preprint arXiv:2004.12452*, 2020. 2

[52] Yiheng Xie, Towaki Takikawa, Shunsuke Saito, Or Litany, Shiqin Yan, Numair Khan, Federico Tombari, James Tompkin, Vincent Sitzmann, and Srinath Sridhar. Neural fields in visual computing and beyond. In *Computer Graphics Forum*, pages 641–676. Wiley Online Library, 2022. 2

[53] Xudong Xu, Xingang Pan, Dahua Lin, and Bo Dai. Generative occupancy fields for 3d surface-aware image synthesis. *NeurIPS*, 34:20683–20695, 2021. 1, 3

[54] Yang Xue, Yuheng Li, Krishna Kumar Singh, and Yong Jae Lee. Giraffe hd: A high-resolution 3d-aware generative model. In *CVPR*, 2022. 1, 3

[55] Ziyang Yuan, Yiming Zhu, Yu Li, Hongyu Liu, and Chun Yuan. Make encoder great again in 3d gan inversion through geometry and occlusion-aware encoding. *arXiv preprint arXiv:2303.12326*, 2023. 2, 3, 7

[56] Jingbo Zhang, Xiaoyu Li, Ziyu Wan, Can Wang, and Jing Liao. Fdnerf: Few-shot dynamic neural radiance fields for face reconstruction and expression editing, 2022. 2

[57] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023. 2, 4, 5

[58] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 8

[59] Yiyu Zhuang, Hao Zhu, Xusen Sun, and Xun Cao. Mofanerf: Morphable facial neural radiance field. In *European Conference on Computer Vision*, pages 268–285. Springer, 2022. 2