3D-Aware Video Generation

Anonymous authors
Paper under double-blind review

Abstract

Generative models have emerged as an essential building block for many image synthesis and editing tasks. Recent advances in this field have also enabled high-quality 3D or video content to be generated that exhibits either multi-view or temporal consistency. With our work, we explore 4D generative adversarial networks (GANs) that learn unconditional generation of 3D-aware videos. By combining neural implicit representations with time-aware discriminator, we develop a GAN framework that synthesizes 3D video supervised only with monocular videos. We show that our method learns a rich embedding of decomposable 3D structures and motions that enables new visual effects of spatio-temporal renderings while producing imagery with quality comparable to that of existing 3D or video GANs.

Figure 1: **3D-Aware video generation.** We show multiple frames and viewpoints of two 3D videos, generated using our model trained on the FaceForensics dataset (Rössler et al., 2019). Our 4D GAN generates 3D content of high quality while permitting control of time and camera extrinsics. Video results can be viewed from our supplementary html.

1 Introduction

Recent advances in generative adversarial networks (GANs) (Goodfellow et al., 2014) have led to artificial synthesis of photorealistic images (Karras et al., 2019; 2020; 2021). These methods have been extended to enable unconditional generation of high-quality videos (Chen et al., 2021a; Yu et al., 2022) and multi-view-consistent 3D scenes (Gu et al., 2022; Chan et al., 2022; Or-El et al., 2022a). However, despite important applications in visual effects, computer vision, and other fields, no generative model has been demonstrated that is successful in synthesizing 3D videos to date.

We propose the first 4D GAN that learns to generate multi-view-consistent video data from single-view videos. For this purpose, we develop a 3D-aware video generator to synthesize 3D content that is animated with learned motion priors, and permits viewpoint manipulations. Two key elements of our framework are a time-conditioned 4D generator that leverages emerging neural implicit scene representations (Park et al., 2019; Mescheder et al., 2019; Mildenhall et al., 2020) and a time-aware video discriminator. Our generator takes as input two latent code vectors for 3D identity and motion, respectively, and it outputs a 4D neural fields that can be queried continuously at any spatio-temporal $xyzt$ coordinate. The generated 4D fields can be used to render realistic video frames from arbitrary camera viewpoints. To train the 4D GAN, we use a discriminator that takes two randomly sampled video frames from the generator (or from real videos) along with their time differences to score the realism of the motions. Our model is trained with an adversarial loss where the generator is encouraged, by the discriminator, to render realistic videos across all sampled camera viewpoints.

Our contributions are following: i) We introduce the first 4D GAN which generates 3D-aware videos supervised only from single-view 2D videos. ii) We develop a framework combining implicit fields with a time-aware discriminator, that can be continuously rendered for any $xyzt$ coordinate. iii) We evaluate the effectiveness of our approach on various video datasets. We show that the trained 4D GAN is able to synthesize plausible videos that allows viewpoint changes, whose visual and motion qualities are competitive against the state-of-the-art 2D video GANs' outputs.

2 Related Work

In this section, we discuss the most relevant literature on image and video synthesis as well as neural implicit representations in the context of 2D and 3D content generation. See appendix for a more complete list of references.

GAN-based Image Synthesis Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) have demonstrated impressive results on multiple synthesis tasks such as image generation (Brock et al., 2019; Karras et al., 2019; 2020), image editing (Wang et al., 2018; Shen et al., 2020; Ling et al., 2021) and image-to-image translation (Isola et al., 2017; Zhu et al., 2017; Choi et al., 2018). To allow for increased controllability, during the image synthesis process, several recent works have proposed to disentangle the underlying factors of variation (Reed et al., 2014; Chen et al., 2016; Lee et al., 2020; Shoshan et al., 2021) or rely on pre-defined templates (Tewari et al., 2020a;b). However, since most of these methods operate on 2D images, they often lack physically-sound control in terms of viewpoint manipulation. In this work, we advocate modelling both the image and the video generation process in 3D in order to ensure controllable generations.

Neural Implicit Representations Neural Implicit Representation (NIR) (Mescheder et al., 2019; Park et al., 2019; Chen & Zhang, 2019) have been extensively employed in various generation tasks due to their continuous, efficient, and differentiable nature. These tasks include 3D reconstruction of objects and scenes, novel-view synthesis of static and dynamic scenes, inverse graphics, and video representations (Jiang et al., 2020; Chibane et al., 2020; Sitzmann et al., 2020; Barron et al., 2021; Sajjadi et al., 2022; Park et al., 2021a; Li et al., 2021; Niemeyer et al., 2020; Chen et al., 2021a). Among the most widely used NIRs are Neural Radiance Fields (NeRFs) (Mildenhall et al., 2020) that combine NIRs with volumetric rendering to enforce 3D consistency while performing novel view synthesis. In this work, we employ a generative variant of NeRF (Gu et al., 2022) and combine it with a time-aware discriminator to learn a generative model of videos from unstructured videos.

Closely related to our work are recent approaches that try to control the motion and the pose of scenes (Lin et al., 2022; Liu et al., 2021a; Zhang et al., 2021; Ren et al., 2021). In particular, they focus on transferring or controlling the motion of their target objects instead of automatically generating plausible motions. They often use networks overfitted to a single reconstructed scene (Chen et al., 2021b; Zhang et al., 2021) or rely on pre-defined templates of human faces or bodies (Liu et al., 2021a; Ren et al., 2021). In our work, we build a 4D generative model that can automatically generate diverse 3D content along with its plausible motion without using any pre-defined templates.

3D-Aware Image Generations Another line of research investigates how 3D representations can be incorporated in generative settings for improving the image quality (Park et al., 2017; Nguyen-Phuoc et al., 2018) and increasing the control over various aspects of the image formation process (Gadelha et al., 2017; Chan et al., 2021; Henderson & Ferrari, 2019). Towards this goal, several works (Henzler et al., 2019; Nguyen-Phuoc et al., 2019; 2020) proposed to train 3D-aware GANs from a set of unstructured images using voxel-based representations. However, due to the low voxel resolution and the inconsistent view-controls stemming from the use of pseudo-3D structures that rely on non-physically-based 2D-3D conversions, these methods tend to generate images with artifacts and struggle to generalize in real-world scenarios. More recent approaches rely on volume rendering to generate 3D objects (Schwarz et al., 2020; Chan et al., 2021; Niemeyer & Geiger, 2021a). Similarly, Zhou et al. (2021), Chan et al. (2021), and DeVries et al. (2021) explored the idea of combining NeRF with GANs for designing 3D-aware image generators. Likewise, StyleSDF

(Or-El et al., 2022a) and StyleNeRF (Gu et al., 2022) proposed to combine an MLP-based volume renderer with a style-based generator (Karras et al., 2020) to produce high-resolution 3D-aware images. Deng et al. (2022) explored learning a generative radiance field on 2D manifolds and Chan et al. (2022) introduced a 3D-aware architecture that exploits both implicit and explicit representations. To improve view-consistency, EpiGRAF (Skorokhodov et al., 2022b) proposes patch-based training to discard the 2D upsampling network, while VoxGRAF (Schwarz et al., 2022) uses sparse voxel grids for efficient rendering without a superresolution module. AniFaceGAN (Wu et al., 2022) proposes an animatable 3D-aware GAN for face animation generation supervised by image collections.. In contrast to this line of research that focuses primarily on 3D-aware image generation, we are interested in *3D-aware video generation*. In particular, we build on top of StyleNeRF (Gu et al., 2022) to allow control on the 3D camera pose during the video synthesis. To the best of our knowledge, this is the first work towards 3D-aware video generation trained from unstructured 2D data.

GAN-based Video Synthesis Inspired by the success of GANs and adversarial training on photorealistic image generation, researchers shifted their attention to various video synthesis tasks (Tulyakov et al., 2018; Holynski et al., 2021; Tian et al., 2021). Several works pose the video synthesis as an autoregressive video prediction task and seek to generate future frames conditioned on the previous using either recurrent (Kalchbrenner et al., 2017; Walker et al., 2021) or attention-based (Rakhimov et al., 2020; Weissenborn et al., 2020; Yan et al., 2021) models. Other approaches (Saito et al., 2017; Tulyakov et al., 2018; Aich et al., 2020) tried to disentangle the motion from the image generation during the video synthesis process. To facilitate generating high-quality frames, Tian et al. (2021) and Fox et al. (2021) employed a pre-trained image generator of Karras et al. (2020). Recently, LongVideoGAN (Brooks et al., 2022) has investigated synthesizing longer videos of more complex datasets. Closely related to our method are the recent work of DIGAN (Yu et al., 2022) and StlyeGAN-V (Skorokhodov et al., 2022a) that generate videos at continuous time steps, without conditioning on previous frames. DIGAN (Yu et al., 2022) employs an NIR-based image generator (Skorokhodov et al., 2021) for learning continuous videos and introduces two discriminators: the first discriminates the realism of each frame and the second operates on image pairs and seeks to determine the realism of the motion. Similarly, StyleGAN-V (Skorokhodov et al., 2022a) employs a style-based GAN (Karras et al., 2020) and a single discriminator that operates on sparsely sampled frames. In contrast to Yu et al. (2022) and Skorokhodov et al. (2022a), we focus on *3D-aware video generation*. In particular, we build on top of StyleNeRF (Gu et al., 2022) and DIGAN (Yu et al., 2022) and demonstrate the ability of our model to render high quality videos from diverse viewpoint angles. Note that this task is not possible for prior works that do not explicitly model the image formation process in 3D.

3 Method

The two main components of our 4D GAN framework are a time-conditioned neural scene generator and a time-aware discriminator. Our generator networks take as input two independent noise vectors, $\boldsymbol{m}$ and $\boldsymbol{z}$, that respectively modulate the motion and the content of the 4D fields. To render an image at a specific time step t, we sample the camera extrinsics according to dataset-dependent distribution and conduct volume rendering through the time-conditioned radiance and density fields. Our time-aware discriminator measures the realism of a pair of frames, given their time difference, to promote plausible 3D video generation. The overview of our pipeline can be found in Fig. 2.

3.1 Time-Conditioned Implicit Fields

We build on top of existing coordinate-based neural generators (Gu et al., 2022; Or-El et al., 2022a) to model continuous implicit fields using a Multi-layer Perceptron (MLP) that outputs a density $\sigma(\boldsymbol{x}, t; \boldsymbol{z}, \boldsymbol{m})$ and appearance feature $\boldsymbol{f}(\boldsymbol{x}, t, \boldsymbol{d}; \boldsymbol{z}, \boldsymbol{m})$ for a given spatio-temporal query $(\boldsymbol{x}, t)$, and view direction $\boldsymbol{d}$. Here, $\boldsymbol{z}$ and $\boldsymbol{m}$ are 3D content and motion latent vectors, respectively.

We process a motion latent vector $\boldsymbol{m}$ sampled from a unit sphere with an MLP conditioned on time step $t \in [0, 1]$, to obtain a final motion vector $\boldsymbol{n}(\boldsymbol{m}, t)$ via multiplicative conditioning:

$$\boldsymbol{n}(\boldsymbol{m}, t) = \psi^3 \circ \psi^2 \circ (t \cdot (\psi^1(\boldsymbol{m}))), \tag{1}$$

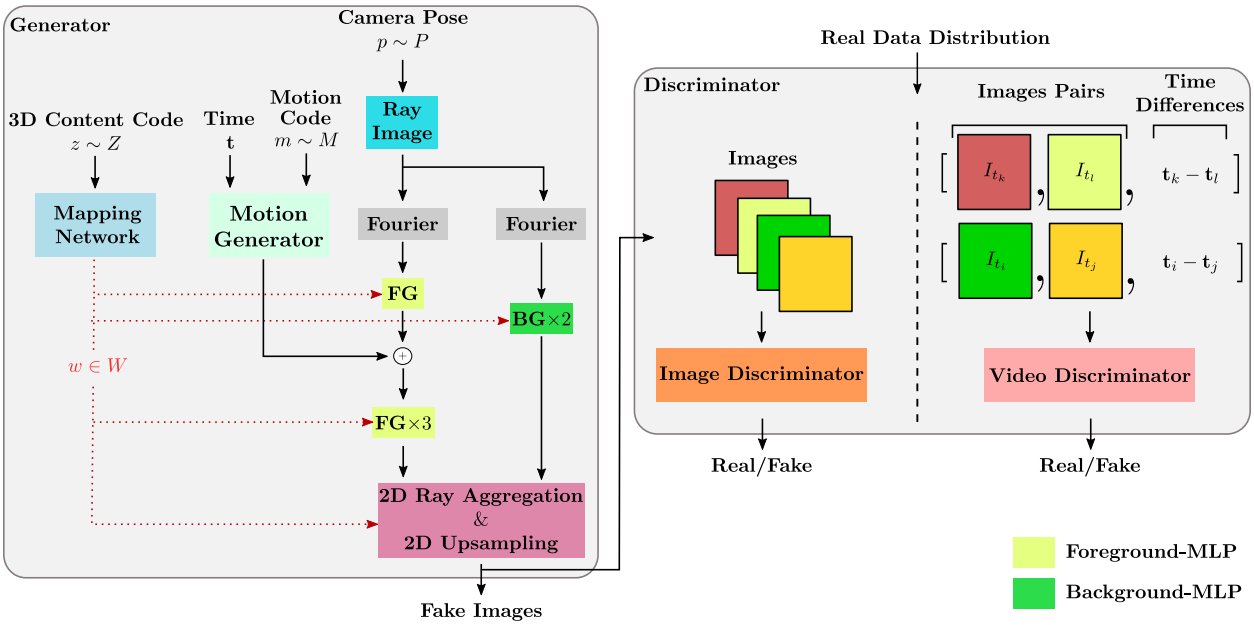

Figure 2: **Model architecture.** The generator (left) takes 3D content and motion codes and the query camera view to render an RGB image. Given a camera pose, we construct a ray image, from which we sample xyz positions along the rays to be passed to the fourier feature layer. The content code is transformed into w and modulate the intermediate features of MLP layers. The motion code, along with the time query t, is processed with the motion generator and added to the generator branch. The video discriminator takes two random frames from a video and their time difference and outputs real/fake prediction. The image discriminator takes individual frames and outputs real/fake label.

where ψ^i is a fully connected layer. Leaky ReLU activation is applied between the layers. Note that our use of continuous time variable t allows rendering the video at an arbitrary frame rate, unlike autoregressive models that can only sample discrete time steps.

The resulting motion vector $\boldsymbol{n}(\boldsymbol{m}, t)$ is then used to control the output of the generator MLP $g_{\boldsymbol{w}}(\boldsymbol{x}, \boldsymbol{n})$ along with the modulation parameters $\boldsymbol{w}(\boldsymbol{z})$ computed from the 3D content vector $\boldsymbol{z}$:

$$g_{\boldsymbol{w}}(\boldsymbol{x}, \boldsymbol{n}) = \phi^k_{\boldsymbol{w}} \circ \phi^{k-1}_{\boldsymbol{w}} \circ \ldots \phi^2_{\boldsymbol{w}} \circ (\boldsymbol{n}(\boldsymbol{m}, t) + \phi^1_{\boldsymbol{w}} \circ \gamma(\boldsymbol{x})), \tag{2}$$

where $\phi^i_{\boldsymbol{w}}$ is a fully connected layer (with Leaky ReLU activations between the layers) whose weights are modulated by the style vector $\boldsymbol{w}$, following the style-based generation techniques of Gu et al. (2022) and Karras et al. (2020). The style vector is produced with a mapping MLP network ζ: $\boldsymbol{w} = \zeta(\boldsymbol{z})$, where $\boldsymbol{z}$ is sampled from the surface of unit sphere. $\gamma(\boldsymbol{x})$ is a positional encoding vector of a spatial query $\boldsymbol{x}$ that induces sharper appearances, following Mildenhall et al. (2020). We omit positional encoding on the time step t, as the empirical results did not improve (see Sec. 4.3). Note that the additive conditioning allows us to easily disable the influence of motion by setting $\boldsymbol{n}$ to be 0, which promotes training on both image and video data when necessary.

The density value at $\boldsymbol{x}$ is then computed by passing the feature $g_{\boldsymbol{w}}(\boldsymbol{x}, \boldsymbol{n})$ to a two layer MLP φ_σ:

$$\sigma(\boldsymbol{x}, t; \boldsymbol{z}, \boldsymbol{m}) = \varphi_\sigma(g_{\boldsymbol{w}(\boldsymbol{z})}(\boldsymbol{x}, \boldsymbol{n}(t, \boldsymbol{m}))) \tag{3}$$

Image Rendering To render an image from a virtual camera with pose $\boldsymbol{p}$, we compute an appearance feature for a ray $\boldsymbol{r}(l) = \boldsymbol{o} + l\boldsymbol{d}$ going through each pixel that emanates from the camera focus $\boldsymbol{o}$ towards the direction $\boldsymbol{d}$. Specifically, we approximate the volume rendering process with discrete point sampling along the

rays and process the aggregated ray features with an MLP, denoted $h_{\boldsymbol{w}}$, conditioned on the view direction $\boldsymbol{d}$:

$$\boldsymbol{f}(\boldsymbol{r}, t, \boldsymbol{d}; \boldsymbol{z}, \boldsymbol{m}) = h_{\boldsymbol{w}} \circ \left[\gamma(\boldsymbol{d}), \int_{l_i}^{l_f} T(l)\sigma(\boldsymbol{r}(l), t) g_{\boldsymbol{w}}(\boldsymbol{r}(l), \boldsymbol{n}) dl \right], \tag{4}$$

where $T(l) = \exp(-\int_{l_i}^{l} \sigma(\boldsymbol{r}(s), t) ds)$. The volume rendering of Eq. 4 involves millions of MLP queries and thus becomes quickly intractable with high resolution images. To reduce the computational overhead, we adopt the 2D upsampling CNN of StyleNeRF (Gu et al., 2022) to convert the low-resolution volume rendering results of $\boldsymbol{f}$ into a high-resolution RGB image $\mathcal{I}$:

$$\mathcal{I}_{\boldsymbol{p}}(t; \boldsymbol{z}, \boldsymbol{m}) = \mathrm{CNN}_{\mathrm{up}}(\boldsymbol{f}(\mathcal{R}_{\boldsymbol{p}}, t, \boldsymbol{d}; \boldsymbol{z}, \boldsymbol{m})), \tag{5}$$

where $\mathcal{R}_{\boldsymbol{p}}$ denotes an image composed of rays from camera pose $\boldsymbol{p}$, with slight abuse of notation.

Background and Foreground Networks Videos consist of static background and moving foreground, hence it would be ineffective to model both foreground and background using one network and motion code. We follow the inverse sphere parameterization of Zhang et al. (2020) to model the background with a second MLP network of g_w that is modulated only with the content vector.

3.2 Training

We train our 4D generator via adversarial loss, leveraging time-aware discriminators from the video GAN literature. The key idea of our training is to encourage the generator to render realistic video frames for all sampled viewpoints and time steps, by scoring their realism with the discriminators.

Time-Aware Discriminator Unlike autoregressive generators, our continuous 4D generator can render frames at an arbitrary time step without knowing the 'past' frames. This feature allows using efficient time-aware discriminators (Skorokhodov et al., 2022a; Yu et al., 2022) that only look at sparsely sampled frames as opposed to the entire video that would require expensive 3D convolutions to process. We adopt the 2D CNN discriminator D_{time} of DIGAN (Yu et al., 2022) to score the realism of the generated motion from two sampled frames. Specifically, we render a pair of frames $\mathcal{I}_{\boldsymbol{p}}(t_1; \boldsymbol{z}, \boldsymbol{m})$ and $\mathcal{I}_{\boldsymbol{p}}(t_2; \boldsymbol{z}, \boldsymbol{m})$ from the same 3D scene and camera pose. The input to the discriminator D_{time} is a concatenation of the two RGB frames along with the time difference between the frames expanded to the image resolution $\mathcal{I}_{\mathrm{repeat}}(t_2 - t_1)$:

$$D_{\mathrm{time}} : [\mathcal{I}_{\boldsymbol{p}}(t_1; \boldsymbol{z}, \boldsymbol{m}), \mathcal{I}_{\boldsymbol{p}}(t_2; \boldsymbol{z}, \boldsymbol{m}), \mathcal{I}_{\mathrm{repeat}}(t_2 - t_1)] \to \mathbb{R}, \quad \text{where } t_2 > t_1. \tag{6}$$

For real videos, we similarly pass a random pair of frames along with their time difference to D_{time}.

Single Image Discriminator In theory, the time-aware discriminator D_{time} should be able to simultaneously measure the realism of the motion of the input pair and that of the individual frames. However, we empirically observe that training a separate discriminator that specializes on single frame discrimination improves quality (see Sec. 4.3). We therefore adopt another discriminator D_{image} that scores realism of individual images: $D_{\mathrm{time}} : \mathcal{I}_{\boldsymbol{p}}(t; \boldsymbol{z}, \boldsymbol{m}) \to \mathbb{R}$. Following 3D GAN approaches (Gu et al., 2022; Or-El et al., 2022a; Chan et al., 2022), we use StyleGAN2 (Karras et al., 2020) discriminator architecture without modifications.

Loss Functions Our training objectives include the adversarial losses from the two discriminators along with the R1 (Mescheder et al., 2018) and NeRF-Path regularizations (Gu et al., 2022):

$$\mathcal{L}(D, G) = \mathcal{L}_{\mathrm{adv}}(D_{\mathrm{time}}, G) + \mathcal{L}_{\mathrm{adv}}(D_{\mathrm{image}}, G) + \lambda_1 \mathcal{L}_{\mathrm{R1}}\left(D_{\mathrm{time}}, D_{\mathrm{image}}\right) + \lambda_2 \mathcal{L}_{\text{NeRF-path}}\left(G\right), \tag{7}$$

where G denotes the entire image generator machinery in Eq. 5, and λ's are balancing parameters. To compute the adversarial losses we use the non-saturating objective. Note that our networks are trained end-to-end without the progressive growing strategy. More details are provided in the supplementary.

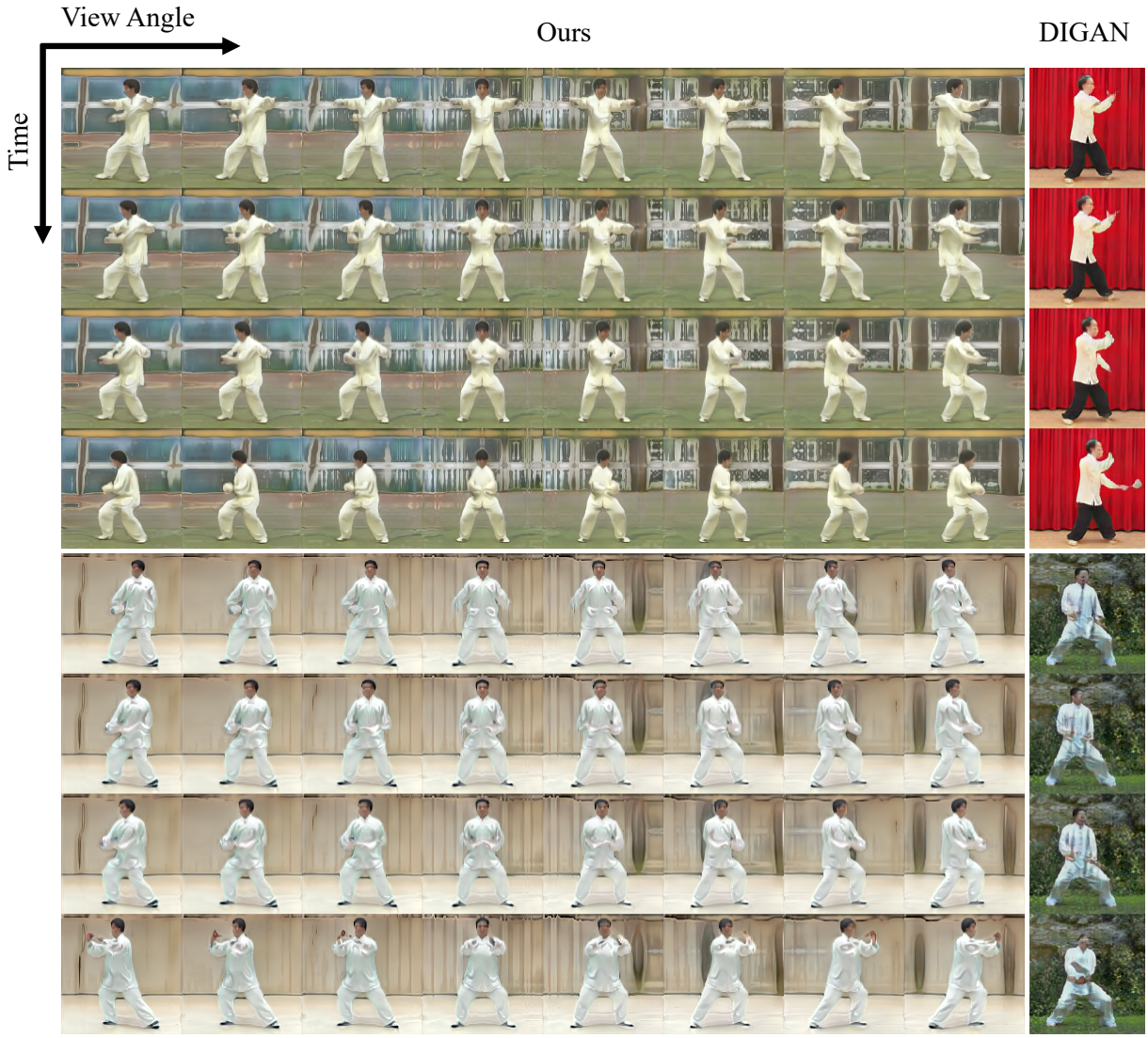

Figure 3: **Qualitative results on TaiChi Dataset.** We visualize the spatio-temporal renderings of two scenes sampled from our 4D GAN, where the horizontal axis indicates change of view angles while the vertical axis indicates progress of time. The rightmost column shows two videos sampled from DIGAN (Yu et al., 2022) that can only be viewed from a fixed camera angle. Our method extends DIGAN in the spatial dimension while producing frames of comparable quality.

Training on Image and Video Datasets The 3D structure of our generator emerges without 3D supervision by enforcing adversarial loss from sampled camera views. Thus, it is crucial that our video dataset features a diverse set of view angles. However, we notice that many of the popular video GAN datasets feature narrow range of viewpoints (e.g., FaceForensics). To address this issue, we seek to leverage existing 2D image datasets that are typically greater in quantity and diversity.

We explore two options: (1) We pre-train our generator model G on an image dataset and fine-tune it on a video dataset. During pre-training, we ignore the temporal components and sample the 3D content only. Such training can be done seamlessly by simply setting the motion vector $\boldsymbol{n}(\boldsymbol{m}, t)$ of Eq. 2 to be zero. After pre-training, we unfreeze the temporal components and minimize the whole objective (Eq. 7) on a video dataset. (2) We train our generator on image and video datasets simultaneously. We refer the reader to Sec. E in our supplementary for more details.

4 Experiments

We conduct experiments to demonstrate the effectiveness of our approach in generating 3D-aware videos, focusing on the new visual effects it enables and the quality of generated imagery. Moreover, we conduct extensive ablation studies on the design components of our model.

4.1 Experimental Setup

Datasets We evaluate our approach on three publicly available, unstructured video datasets: the FaceForensics (Rössler et al., 2019), the MEAD (Wang et al., 2020a), and the TaiChi (Siarohin et al., 2019) dataset. FaceForensics contains 704 training videos of human faces sourced from YouTube. While this dataset containing in-the-wild videos makes it a great testbed for synthesis tasks, many of its videos are captured from frontal views with limited view diversity. On the other hand, the MEAD dataset contains shorter videos capturing faces from discrete 7 angles, from which we randomly subsample 10,000. Note that we ignore the identity correspondences of the videos and treat them as independent unstructured videos. The TaiChi dataset contains 2942 in-the-wild videos of highly diverse TaiChi performances sourced from the internet. Following DIGAN (Yu et al., 2022), we use every fourth frame to make the motion more dynamic. Finally, we provide additional experiments on the SkyTimelapse dataset (Xiong et al., 2018) in the supplementary.

Metrics Following existing video GAN methods, we use Frechet Video Distance (FVD) (Unterthiner et al., 2018) as our main metric for measuring realism of the generated motion sequences. In particular, we use the FVD protocol of Skorokhodov et al. (2022a) that alleviates the inconsistency issues of the original FVD implementation (Unterthiner et al., 2018). Moreover, we consider the following additional metrics: Average Content Distance (ACD) (Tulyakov et al., 2018), which measures temporal consistency of a video and CPBD (Narvekar & Karam, 2011), which measures the sharpness of an image. We also measure the face identity consistency (ID) across viewpoints by computing ArcFace (Deng et al., 2019) cosine similarity, following Chan et al. (2022) (computed only for 3D-aware methods). To provide a common metric to 2D and 3D-aware image generation methods that cannot synthesize videos, we use the Frechet Image Distance (FID) (Heusel et al., 2017). The individual frames for FID scores are randomly selected from all generated video frames. Note that the checkpoints used for the results of StyleGAN2, VideoGPT, MoCoGAN, and MoCoGAN-HD on FaceForensics were not released by the StyleGAN-V authors. Hence, we are not able to compute the ACD and CPBD metrics for these methods. We can not compute ACD for StyleNeRF, as it is a static method.

Type	Method	FVD (↓)	FID (↓)	ACD (↓)	CPBD (↑)	ID (↑)
2D Image	StyleGAN2	-	8.4	-	-	-
2D Video	VideoGPT	185.9	22.7	-	-	-
	MoCoGAN	124.7	24.0	-	-	-
	MoCoGAN-HD	111.8	7.1	-	-	-
	DIGAN	62.5	19.1	1.09	0.171	-
	StyleGAN-V	47.4	9.4	1.11	0.155	-
3D Static	StyleNeRF	-	15.3	-	0.181	0.812
3D Video	Ours	68.7	13.7	0.965	0.196	0.861

Table 1: **Quantitative Results on FaceForensics.** We report metrics for all methods at 256^2 pixel resolution. For FVD and FID of 2D image and video methods, we use the numbers reported in StyleGAN-V. We omit ACD and CPBD metrics for methods whose checkpoints are not publicly available. View-consistent identity (ID) is only computed for 3D-aware methods.

4.2 Main Results

In this section, we provide experimental evidence that our model is able to learn a distribution of 3D-aware videos that permit spatio-temporal control. In particular, in Figures 1, 3, 4, and 5, we visualize snapshots of sampled 4D scenes rendered from various time steps and camera angles across various datasets. Moreover, we conduct both qualitative and quantitative comparisons against the strongest video GAN baselines on

Figure 4: **Qualitative results on FaceForensics Dataset.** The first 6 columns visualize two 4D fields sampled from of our model trained on FaceForensics rendered from various spatio-temporal snapshots. Note the high quality visuals and motions across diverse viewpoints. The last three columns show the video samples of MoCoGAN-HD (Tian et al., 2021), DIGAN (Yu et al., 2022), and StyleGAN-V (Skorokhodov et al., 2022a), in that order.

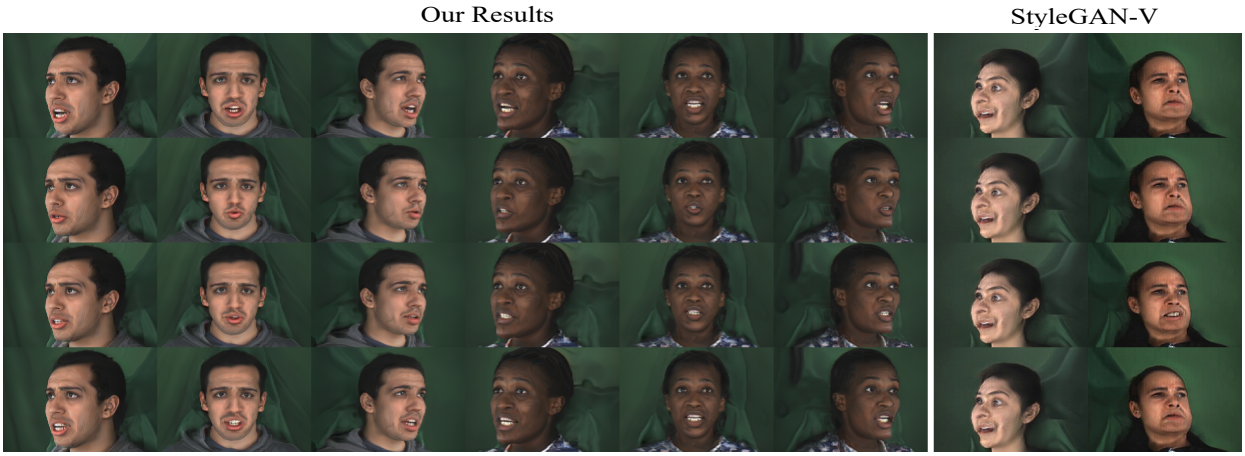

Figure 5: **Results on the MEAD Dataset.** The first six columns show the spatio-temporal renderings of our 4D GAN, where the vertical axis indicates the progress in time. The two right-most columns show generated videos from StyleGAN-V (Skorokhodov et al., 2022a). Zoom-in to inspect details.

Method	FVD (↓)	ACD(↓)	CBPD(↑)
StyleGAN-V	109.3	0.136	0.509
Ours	55.4	0.060	0.469

Table 2: **Results on MEAD Dataset.**

Method	FVD(↓)	ACD(↓)	CBPD(↑)
DIGAN	151.7	0.537	0.672
Ours	158.3	0.552	0.632

Table 3: **Results on TaiChi Dataset.**

the all datasets. For each method, we use video length of 16 frames and use image resolution of 256^2 for FaceForensics and MEAD, and 128^2 for TaiChi dataset.

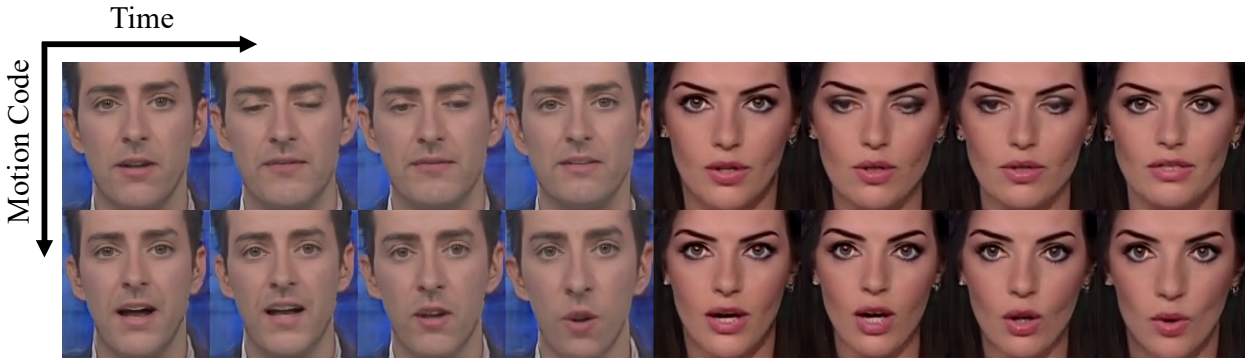

Figure 6: **Motion and content decomposition.** The two rows show respective video sequences that share the same 3D content latent vector applied to different motion vectors. Here we show motions with four time steps for two different identities. Note the difference in motions while the identities of the person appear unchanged.

FaceForensics is one of the most widely used dataset in the video generation literature, and thus we measure the scores across a wide range of methods, including: 2D image (Karras et al., 2020) and video generation works (Yu et al., 2022; Skorokhodov et al., 2022a; Tulyakov et al., 2018; Tian et al., 2021; Yan et al., 2021), and a 3D generation method (Gu et al., 2022). The qualitative results in Fig. 4 show our model's ability to generate 4D visual effects, displaying the spatiotemporal renderings and the physically-based zoom-in effects. The quantitative results in Table 1 confirm that our generated imagery are of competitive quality against the strongest 2D and 3D baselines. Note that the metrics for the 2D image and video methods are copied from the reports of Skorokhodov et al. (2022a). We train StyleNeRF (Gu et al., 2022) on FaceForensics from scratch.

Method	FVD
Ours	158.3
w/o Background NeRF	203.7
w/ Static and Dynamic Separation	207.6
w/o Motion Generator	166.3
w/ Positional Time Encoding	175.4
w/o Image Discriminator	234.3

Table 4: **Ablation studies on TaiChi Dataset.**

Fig. 3 presents sampled scenes from our TaiChi-trained model with varying time and viewpoints, showcasing our 4D learning on a challenging setup with complex motions and backgrounds. When compared with one of the latest 2D video generation methods, DIGAN (Yu et al., 2022), our method synthesizes frames with competitive visual fidelity while introducing another degree of freedom. Note that we used the provided pre-trained model to recompute their score following the FVD protocol used by Skorokhodov et al. (2022a).

The MEAD dataset contains videos of faces taken from a wide range of viewpoints. We hypothesize that, for the 2D-based methods, a significant amount of expressive power will be wasted for redundantly modeling the diverse views in 2D, while the 3D approach would only need to learn a shared representation across views. The results, shown in Fig. 5, indeed suggest higher visual quality for our 4D approach, while allowing rendering from 7 different viewpoints. Our quantitative analysis in Tab. 4.1 supports our observations – our FVD is noticeably better than that of StyleGAN-V (Skorokhodov et al., 2022a). Note that the pre-trained model provided by Skorokhodov et al. (2022a) was trained only on frontal views, so we included non-frontal faces and re-trained. We omit comparing against other video methods on MEAD that are already compared in Skorokhodov et al. (2022a). Finally, we demonstrate that the use of two independently sampled latent vectors to modulate motion and content makes these two components separable. Fig. 6 showcases such decomposition by applying two different motion codes to the same content code. We refer the reader to our supplementary for additional results and videos to fully appreciate 4D renderings.

4.3 Ablation

We conduct ablation studies to gauge how the design choices of our algorithm affect the quality of 4D scene generations. We use TaiChi dataset, as its diverse scenes and motions help us identify the effects of each component. Table 4 summarizes the numerical results, for further details see Sec. F in the appendix.

Background NeRF We analyze the importance of the background NeRF based on the inverse sphere parametrization of NeRF++ (Zhang et al., 2020) used for capturing unbounbed scenes. We observe that using the additional background NeRF is critical for the model to disentangle the dynamic and static parts in the video, leading to less motion artifacts and more static backgrounds.

Static and Dynamic Separation We forgo the use of the inverse sphere background NeRF of Zhang et al. (2020) and instead decompose the scene with static and dynamic NeRFs, following Gao et al. (2021a). We observe that this setup hurts the generation quality, suggesting that the background parameterization of NeRF++ (Zhang et al., 2020) is critical for training a 4D generative model.

Motion Generator Here we omit the use of motion vector and its mapping network, and instead pass the time value directly to the foreground NeRF. The use of motion network improves the output quality and induces useful decomposition of motion and 3D content.

Positional Time Encoding Our current motion mapping MLP is conditioned on the raw time value. We observe that applying positional encoding on the Fourier features leads to repetitive and unnatural motions even when tuning the frequencies of the Fourier features. We leave applying more complicated positional encoding of time as future work.

Image Discriminator Moreover, we demonstrate that adding a discriminator that specializes on single-frame discrimination improves quality compared to using only a time-aware discriminator.

5 Limitations & Discussions

Our approach models the entire dynamic foreground with a single latent vector, which limits the expressive capacity of our generative model. Learning with more number of independent latent vectors, as tried in (Niemeyer & Geiger, 2021a; Hudson & Zitnick, 2021), could promote handling of multi-objected scenes with more complex motions. Furthermore, our training scheme assumes static camera and dynamic scenes, so it cannot handle videos taken from a moving camera. Modeling plausible camera paths is an interesting but less explored problem. Being the first 4D GAN approach, our method can only generate short video sequences (16 time steps, following Yan et al. (2021), Tian et al. (2021), and Yu et al. (2022)). We leave applying recent progress (Skorokhodov et al., 2022a) in generating longer videos as future work. Finally, we model the scene motion as a purely statistical phenomenon, without explicitly considering physics, causality, semantics, and entity-to-entity interactions. These topics remain important challenges, which we continue to explore.

6 Conclusions

In this work, we introduced the first 4D generative model that synthesizes realistic 3D-aware videos, supervised only from a set of unstructured 2D videos. Our proposed model combines the benefits of video and 3D generative models to enable new visual effects that involve spatio-temporal renderings, while at the same time maintaining the visual fidelity. The resulting latent space of rich, decomposable motion and 3D content sets up the foundation to exciting new avenue of research towards interactive content generation and editing that requires knowledge of underlying 3D structures and motion priors.

7 Ethics Statement

7.1 Potential Misuse

We condemn the use of neural rendering technologies towards generating realistic fake content to harm specific entities or to spread misinformation, and we support future research on preventing such unintended applications.

While our method can plausibly synthesize moving human heads and bodies, the technology in the current form does not pose significant threat in enabling the spread of misinformation. We highlight that our 4D GAN is an unconditional generative model. This means that the current method can only generate *random* identities and motions and thus cannot be used for targeting specific individuals for misinformation (e.g., creating a video of a particular politician). Similarly, being an unconditional model, we cannot control a person with an user-desired motion, e.g., following a text prompt.

However, we acknowledge the possible development of future 4D GAN research that trains conditional models that are able to synthesize videos of given individuals. We note that there are various existing technologies to spot neural network-generated imagery with surprising level of accuracy (Cozzolino et al., 2021; Yang et al., 2021; de Lima et al., 2020; Güera & Delp, 2018; Kumar et al., 2020; Wang et al., 2020b). Among them, Wang et al. (2020b) suggests that the key of training such a detector is a realistic generator, which can provide ample amount of training data for fake content detection. In this context, we believe that our unconditional 4D GAN can be effectively used to fight against the potential misuse of AI-generated videos by generating such training data, because it can sample realistic videos from diverse identities, motions, and camera viewpoints.

7.2 Privacy Issues

We note readers that our experiments only used publicly available datasets that contain videos that belong to public domains. Moreover, we emphasize that all of our figures and videos across the main paper and the supplementary materials contain *synthetic* content. Therefore, *all* of the human faces or bodies shown in this work do not involve any privacy concerns.

7.3 Diversity and Potential Discrimination

We have shown in our main paper and the supplementary that our trained model on the FaceForensics and MEAD datasets are able to generate diverse human faces across gender and ethnicity. We acknowledge that our generation results might not cover diversity of some human traits, e.g., weights, disabilities, etc. Note the TaiChi dataset primarily contains videos of certain ethnicity due to the skewed popularity of the activity.

8 Reproducibility Statement

We provide detailed implementation-related information of our algorithms in Sec. C of the supplementary document. We will release the source code for training and testing our algorithms upon acceptance. For experiments. we describe our datasets (along with how we processed them) and metrics in Sec. 4.1 and provide in-depth details of our experiments in Sec. D of the supplementary document to further improve the reproducibility.

References

Dinesh Acharya, Zhiwu Huang, Danda Pani Paudel, and Luc Van Gool. Towards high resolution video generation with progressive growing of sliced wasserstein gans. *arXiv.org*, abs/1810.02419, 2018.

Abhishek Aich, Akash Gupta, Rameswar Panda, Rakib Hyder, M. Salman Asif, and Amit K. Roy-Chowdhury. Non-adversarial video synthesis with learned priors. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Matan Atzmon, Niv Haim, Lior Yariv, Ofer Israelov, Haggai Maron, and Yaron Lipman. Controlling neural level sets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Jonathan T. Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P. Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2021.

Alexander W. Bergman, Petr Kellnhofer, and Gordon Wetzstein. Fast training of neural lumigraph representations using meta learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2019.

Tim Brooks, Janne Hellsten, Miika Aittala, Ting-Chun Wang, Timo Aila, Jaakko Lehtinen, Ming-Yu Liu, Alexei A Efros, and Tero Karras. Generating long videos of dynamic scenes. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Rohan Chabra, Jan Eric Lenssen, Eddy Ilg, Tanner Schmidt, Julian Straub, Steven Lovegrove, and Richard A. Newcombe. Deep local shapes: Learning local sdf priors for detailed 3d reconstruction. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Proc. of the European Conf. on Computer Vision (ECCV)*, 2020.

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 *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021.

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 *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Hao Chen, Bo He, Hanyu Wang, Yixuan Ren, Ser Nam Lim, and Abhinav Shrivastava. Nerv: Neural representations for videos. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021a.

Jianchuan Chen, Ying Zhang, Di Kang, Xuefei Zhe, Linchao Bao, Xu Jia, and Huchuan Lu. Animatable neural radiance fields from monocular rgb videos. *arXiv.org*, abs/2012.05903, 2021b.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll. Implicit functions in feature space for 3d shape reconstruction and completion. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Yunjey Choi, Min-Je Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Aidan Clark, Jeff Donahue, and Karen Simonyan. Adversarial video generation on complex datasets. *arXiv.org*, abs/1907.06571, 2019.

Davide Cozzolino, Andreas Rössler, Justus Thies, Matthias Nießner, and Luisa Verdoliva. Id-reveal: Identity-aware deepfake video detection. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Oscar de Lima, Sean Franklin, Shreshtha Basu, Blake Karwoski, and Annet George. Deepfake detection using spatiotemporal convolutional networks. *arXiv preprint arXiv:2006.14749*, 2020.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. Gram: Generative radiance manifolds for 3d-aware image generation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Terrance DeVries, Miguel Ángel Bautista, Nitish Srivastava, Graham W. Taylor, and Joshua M. Susskind. Unconstrained scene generation with locally conditioned radiance fields. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2021.

Gereon Fox, Ayush Tewari, Mohamed Elgharib, and Christian Theobalt. Stylevideogan: A temporal generative model using a pretrained stylegan. In *Proc. of the British Machine Vision Conf. (BMVC)*, 2021.

Matheus Gadelha, Subhransu Maji, and Rui Wang. 3d shape induction from 2d views of multiple objects. In *Proc. of the International Conf. on 3D Vision (3DV)*, 2017.

Guy Gafni, Justus Thies, Michael Zollhofer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *CVPR*, pp. 8649–8658, 2021.

Chen Gao, Yi-Chang Shih, Wei-Sheng Lai, Chia-Kai Liang, and Jia-Bin Huang. Portrait neural radiance fields from a single image. *arXiv.org*, abs/2012.05903, 2020.

Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic view synthesis from dynamic monocular video. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2021a.

Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic view synthesis from dynamic monocular video. In *ICCV*, pp. 5712–5721, 2021b.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

Cade Gordon and Natalie Parde. Latent neural differential equations for video generation. In *Advances in Neural Information Processing Systems (NeurIPS) Workshops*, 2020.

Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *Proc. of the International Conf. on Machine learning (ICML)*, 2020.

Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2022.

David Güera and Edward J Delp. Deepfake video detection using recurrent neural networks. In *IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, 2018.

Zekun Hao, Arun Mallya, Serge J. Belongie, and Ming-Yu Liu. Gancraft: Unsupervised 3d neural rendering of minecraft worlds. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2021.

Paul Henderson and Vittorio Ferrari. Learning single-image 3d reconstruction by generative modelling of shape, pose and shading. *International Journal of Computer Vision (IJCV)*, 2019.

Paul Henderson and Christoph H. Lampert. Unsupervised object-centric video generation and decomposition in 3d. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Philipp Henzler, Niloy J Mitra, , and Tobias Ritschel. Escaping plato's cave: 3d shape from adversarial rendering. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2019.

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 (NIPS)*, 2017.

Aleksander Holynski, Brian L Curless, Steven M Seitz, and Richard Szeliski. Animating pictures with eulerian motion fields. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Drew A Hudson and Larry Zitnick. Generative adversarial transformers. In *Proc. of the International Conf. on Machine learning (ICML)*, 2021.

Sangeek Hyun, Jihwan Kim, and Jae-Pil Heo. Self-supervised video gans: Learning for appearance consistency and motion coherency. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Chiyu Jiang, Avneesh Sud, Ameesh Makadia, Jingwei Huang, Matthias Nießner, and Thomas Funkhouser. Local implicit grid representations for 3d scenes. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Emmanuel Kahembwe and Subramanian Ramamoorthy. Lower dimensional kernels for video discriminators. *arXiv.org*, abs/1912.08860, 2019.

Nal Kalchbrenner, Aäron van den Oord, Karen Simonyan, Ivo Danihelka, Oriol Vinyals, Alex Graves, and Koray Kavukcuoglu. Video pixel networks. In *Proc. of the International Conf. on Machine learning (ICML)*, 2017.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2018.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020.

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

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2015.

Akash Kumar, Arnav Bhavsar, and Rajesh Verma. Detecting deepfakes with metric learning. In *IEEE International Workshop on Biometrics and Forensics (IWBF)*, 2020.

Wonkwang Lee, Donggyun Kim, Seunghoon Hong, and Honglak Lee. High-fidelity synthesis with disentangled representation. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2020.

Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, and Zhaoyang Lv. Neural 3d video synthesis from multi-view video. In *CVPR*, 2022.

Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Yiyi Liao, Katja Schwarz, Lars Mescheder, and Andreas Geiger. Towards unsupervised learning of generative models for 3d controllable image synthesis. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Chen-Hsuan Lin, Chaoyang Wang, and Simon Lucey. Sdf-srn: Learning signed distance 3d object reconstruction from static images. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Connor Z. Lin, David B. Lindell, Eric Chan, and Gordon Wetzstein. 3d gan inversion for controllable portrait image animation. *arXiv.org*, abs/2203.13441, 2022.

David B Lindell, Dave Van Veen, Jeong Joon Park, and Gordon Wetzstein. Bacon: Band-limited coordinate networks for multiscale scene representation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Huan Ling, Karsten Kreis, Daiqing Li, Seung Wook Kim, Antonio Torralba, and Sanja Fidler. Editgan: High-precision semantic image editing. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. In *ACM Trans. on Graphics*, 2021a.

Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. *ACM Transactions on Graphics (TOG)*, 40(6):1–16, 2021b.

Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. In *ACM Trans. on Graphics*, 2019a.

Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *ACM Trans. Graph.*, 2019b.

Sebastian Lunz, Yingzhen Li, Andrew W. Fitzgibbon, and Nate Kushman. Inverse graphics gan: Learning to generate 3d shapes from unstructured 2d data. *arXiv.org*, abs/2002.12674, 2020.

Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Quan Meng, Anpei Chen, Haimin Luo, Minye Wu, Hao Su, Lan Xu, Xuming He, and Jingyi Yu. Gnerf: Gan-based neural radiance field without posed camera. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2021.

Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *Proc. of the International Conf. on Machine learning (ICML)*, 2018.

Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Mateusz Michalkiewicz, Jhony K Pontes, Dominic Jack, Mahsa Baktashmotlagh, and Anders Eriksson. Implicit surface representations as layers in neural networks. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2019.

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 *Proc. of the European Conf. on Computer Vision (ECCV)*, 2020.

Andrés Muñoz, Mohammadreza Zolfaghari, Max Argus, and Thomas Brox. Temporal shift gan for large scale video generation. In *Proc. of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021.

Niranjan D Narvekar and Lina J Karam. A no-reference image blur metric based on the cumulative probability of blur detection (cpbd). *IEEE Transactions on Image Processing*, 2011.

Thu Nguyen-Phuoc, Chuan Li, Stephen Balaban, and Yong-Liang Yang. Rendernet: A deep convolutional network for differentiable rendering from 3d shapes. In *Advances in Neural Information Processing Systems (NIPS)*, 2018.

Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. Hologan: Unsupervised learning of 3d representations from natural images. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2019.

Thu Nguyen-Phuoc, Christian Richardt, Long Mai, Yong-Liang Yang, and Niloy Mitra. Blockgan: Learning 3d object-aware scene representations from unlabelled images. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021a.

Michael Niemeyer and Andreas Geiger. Campari: Camera-aware decomposed generative neural radiance fields. In *Proc. of the International Conf. on 3D Vision (3DV)*, 2021b.

Michael Niemeyer, Lars M. Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Neural articulated radiance field. In *ICCV*, pp. 5762–5772, 2021.

Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2021.

Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. Stylesdf: High-resolution 3d-consistent image and geometry generation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022a.

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

Xingang Pan, Xudong Xu, Chen Change Loy, Christian Theobalt, and Bo Dai. A shading-guided generative implicit model for shape-accurate 3d-aware image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Eunbyung Park, Jimei Yang, Ersin Yumer, Duygu Ceylan, and Alexander C. Berg. Transformation-grounded image generation network for novel 3d view synthesis. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B. Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2021a.

Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *ICCV*, pp. 5865–5874, 2021b.

Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B. Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. Hypernerf: a higher-dimensional representation for topologically varying neural radiance fields. *ACM Trans. on Graphics*, 2021c.

Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *SIGGRAPH Asia*, 2021d.

Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *CVPR*, pp. 9054–9063, 2021.

Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2020.

Martin Piala and Ronald Clark. Terminerf: Ray termination prediction for efficient neural rendering. In *Proc. of the International Conf. on 3D Vision (3DV)*, 2021.

Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021a.

Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *CVPR*, pp. 10318–10327, 2021b.

Ruslan Rakhimov, Denis Volkhonskiy, Alexey Artemov, Denis Zorin, and Evgeny Burnaev. Latent video transformer. *arXiv.org*, abs/2006.10704, 2020.

Scott E. Reed, Kihyuk Sohn, Yuting Zhang, and Honglak Lee. Learning to disentangle factors of variation with manifold interaction. In *Proc. of the International Conf. on Machine learning (ICML)*, 2014.

Yurui Ren, Ge Li, Yuanqi Chen, Thomas H Li, and Shan Liu. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2021.

Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2019.

Masaki Saito, Eiichi Matsumoto, and Shunta Saito. Temporal generative adversarial nets with singular value clipping. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2017.

Masaki Saito, Shunta Saito, Masanori Koyama, and Sosuke Kobayashi. Train sparsely, generate densely: Memory-efficient unsupervised training of high-resolution temporal GAN. *International Journal of Computer Vision (IJCV)*, 2020.

Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2019.

Mehdi S. M. Sajjadi, Henning Meyer, Etienne Pot, Urs Bergmann, Klaus Greff, Noha Radwan, Suhani Vora, Mario Lucic, Daniel Duckworth, Alexey Dosovitskiy, Jakob Uszkoreit, Thomas A. Funkhouser, and Andrea Tagliasacchi. Scene representation transformer: Geometry-free novel view synthesis through set-latent scene representations. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Katja Schwarz, Axel Sauer, Michael Niemeyer, Yiyi Liao, and Andreas Geiger. Voxgraf: Fast 3d-aware image synthesis with sparse voxel grids. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Alon Shoshan, Nadav Bhonker, Igor Kviatkovsky, and Gerard Medioni. Gan-control: Explicitly controllable gans. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2021.

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

Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Vincent Sitzmann, Julien N. P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Ivan Skorokhodov, Savva Ignatyev, and Mohamed Elhoseiny. Adversarial generation of continuous images. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022a.

Ivan Skorokhodov, Sergey Tulyakov, Yiqun Wang, and Peter Wonka. Epigraf: Rethinking training of 3d gans. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022b.

Pratul P. Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T. Barron. Nerv: Neural reflectance and visibility fields for relighting and view synthesis. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Ayush Tewari, Mohamed Elgharib, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhöfer, and Christian Theobalt. Pie: Portrait image embedding for semantic control. In *ACM Trans. on Graphics*, 2020a.

Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. Stylerig: Rigging stylegan for 3d control over portrait images. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020b.

Ayush Tewari, Justus Thies, Ben Mildenhall, Pratul P. Srinivasan, Edgar Tretschk, Yifan Wang, Christoph Lassner, Vincent Sitzmann, Ricardo Martin-Brualla, Stephen Lombardi, Tomas Simon, Christian Theobalt, Matthias Nießner, Jonathan T. Barron, Gordon Wetzstein, Michael Zollhöfer, and Vladislav Golyanik. Advances in neural rendering. *arXiv.org*, abs/2111.05849, 2021.

Yu Tian, Jian Ren, Menglei Chai, Kyle Olszewski, Xi Peng, Dimitris N. Metaxas, and Sergey Tulyakov. A good image generator is what you need for high-resolution video synthesis. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2021.

Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2021a.

Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *ICCV*, pp. 12959–12970, 2021b.

Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv.org*, abs/1812.01717, 2018.

Jacob Walker, Ali Razavi, and Aäron van den Oord. Predicting video with VQVAE. *arXiv.org*, abs/2103.01950, 2021.

Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2020a.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020b.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Dirk Weissenborn, Oscar Täckström, and Jakob Uszkoreit. Scaling autoregressive video models. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2020.

Yue Wu, Yu Deng, Jiaolong Yang, Fangyun Wei, Qifeng Chen, and Xin Tong. Anifacegan: Animatable 3d-aware face image generation for video avatars. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim. Space-time neural irradiance fields for free-viewpoint video. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Wei Xiong, Wenhan Luo, Lin Ma, Wei Liu, and Jiebo Luo. Learning to generate time-lapse videos using multi-stage dynamic generative adversarial networks. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Xudong Xu, Xingang Pan, Dahua Lin, and Bo Dai. Generative occupancy fields for 3d surface-aware image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Yinghao Xu, Sida Peng, Ceyuan Yang, Yujun Shen, and Bolei Zhou. 3d-aware image synthesis via learning structural and textural representations. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using VQ-VAE and transformers. *arXiv.org*, abs/2104.10157, 2021.

Jiachen Yang, Aiyun Li, Shuai Xiao, Wen Lu, and Xinbo Gao. Mtd-net: learning to detect deepfakes images by multi-scale texture difference. *IEEE Transactions on Information Forensics and Security*, 2021.

Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Ronen Basri, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Sihyun Yu, Jihoon Tack, Sangwoo Mo, Hyunsu Kim, Junho Kim, Jung-Woo Ha, and Jinwoo Shin. Generating videos with dynamics-aware implicit generative adversarial networks. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2022.

Wentao Yuan, Zhaoyang Lv, Tanner Schmidt, and Steven Lovegrove. Star: Self-supervised tracking and reconstruction of rigid objects in motion with neural rendering. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021a.

Wentao Yuan, Zhaoyang Lv, Tanner Schmidt, and Steven Lovegrove. Star: Self-supervised tracking and reconstruction of rigid objects in motion with neural rendering. In *CVPR*, pp. 13144–13152, 2021b.

Vladyslav Yushchenko, Nikita Araslanov, and Stefan Roth. Markov decision process for video generation. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV) Workshops*, 2019.

Jiakai Zhang, Xinhang Liu, Xinyi Ye, Fuqiang Zhao, Yanshun Zhang, Minye Wu, Yingliang Zhang, Lan Xu, and Jingyi Yu. Editable free-viewpoint video using a layered neural representation. *ACM Trans. on Graphics*, 2021.

Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv.org*, abs/2010.07492, 2020.

Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient gan training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Peng Zhou, Lingxi Xie, Bingbing Ni, and Qi Tian. Cips-3d: A 3d-aware generator of gans based on conditionally-independent pixel synthesis. *arXiv.org*, abs/2110.09788, 2021.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2017.

A Video Results

We urge readers to view our video results by opening `supp.html` on an internet browser. We provide video results on the FaceForensics (Rössler et al., 2019), MEAD (Wang et al., 2020a), TaiChi (Siarohin et al., 2019), and SkyTimelapse (Xiong et al., 2018). In particular, we show the generated videos for all datasets from various camera viewpoints in order to showcase the ability of our model to learn a distribution of 3D-aware videos, which we highly encourage readers to view.

Specifically, for the case of FaceForensics, we show generated videos with a forward-facing camera (see Ours with Forward-facing Camera), with a camera that rotates along the yaw-axis (see Ours with Rotating Camera) and with a forward-facing camera that moves away from the depicted individual, thus creating a zoom-out effect, (see Ours with Forward-facing Camera and Zoom Effect). Moreover, we also provide generated videos using a variant of our model that uses a pre-trained generator on FFHQ (Karras et al., 2019), as discussed in the main submission. This variant of our model allows a wider range of viewpoint control than our original model. Particularly, for this variant of our model, we also show examples of generated videos, while we rotate the camera along both yaw and pitch axes, where the fine-tuned model generates realistically looking videos. In addition, we show "Motion and Content Decomposition" examples, where we show the ability of our approach to control shape and motion separately; i.e., we can create videos illustrating the same human performing different motions, and vice versa. We also provide example videos of prior work including MoCoGAN-HD (Tian et al., 2021), DIGAN (Yu et al., 2022) and StyleGAN-V (Skorokhodov et al., 2022a). We note that our generated videos are of comparable quality against those of the state-of-the-art video generation methods (Yu et al., 2022; Skorokhodov et al., 2022a), while at the same time permitting control on the camera viewpoint, e.g., zoom-out to reveal new content or rotate the camera around, which is not possible for the latest 2D video methods.

Similarly, we also showcase examples of our generated videos on the MEAD (Wang et al., 2020a) dataset using a similar setup. We observe that in comparison to StyleGAN-V (Skorokhodov et al., 2022a) our generations have significantly fewer visual artifacts (the faces of Skorokhodov et al. (2022a) appear uncanny), while at the same time our generated videos from different camera viewpoints (see Ours with Different Camera Positions) are consistently plausible.

We also show examples of generated videos on the TaiChi (Siarohin et al., 2019) dataset using a similar setup. In addition, we also consider two more setups, where we rotate the camera along the yaw-axis while having a static human (see Ours with Rotating Camera and Static Motion) and a moving human (see Ours with Rotating Camera and Dynamic Motion). Also for these scenarios, the quality of our generated videos are comparable to that of DIGAN (Yu et al., 2022) that does not allow viewpoint control.

Figure 7: **Interpolation in the content latent space.** We visualize, for each row, examples of linearly interpolating between two sampled latents in the content space with fixed motion vectors and camera viewpoints. Note the smooth and plausible transition of face appearances.

Figure 8: **Depth visualizations (FaceForensics).** We show depth maps (second row) obtained by volume rendering the sampled 4D fields at a given time step. The first row shows the corresponding RGB rendering of the same 4D fields. Here the depth is defined as the expected ray termination distance.

Finally, we show samples for the SkyTimelapse (Xiong et al., 2018) dataset. The dataset is in contrast with the other three datasets, as its videos often contain multiple objects or entities. We provide videos rendered from a fixed camera, and a rotating camera along the yaw-axis with and without scene dynamics (Ours with Rotating Camera along First Axis and Static Motion, and Ours with Rotating Camera along First Axis and Dynamic Motion). Similarly, we rotate the camera along the pitch-axis with and without scene motions (Ours with Rotating Camera along Second Axis and Static Motion, and Ours with Rotating Camera along Second Axis and Dynamic Motion). Note that we only model rotation of a camera located at the origin. A careful modeling of camera distribution for such a large-scale scene dataset is out of scope of this work, and thus we omit an in-depth analysis.

B Visualizing Latent Interpolations and Depth Maps

Our 4D GAN with decomposed content and motion latent spaces allow interpolation of content with fixed motion and vice versa. Moreover, our sampled neural fields can be used to obtain depth maps given the 3D nature of our representation.

As shown in Fig. 7, linearly interpolating between two sampled content vectors maps to smooth, plausible interpolation of content appearance in the 4D fields. Similarly, we fix the content vector and apply interpolated motion vectors. Such visualization is best viewed as videos, so we refer readers to the supplementary website `supp.html` (see Ours with Motion Interpolation). The video results show that the interpolation in the motion latent space leads to smooth transition of motions. Finally, in Fig. 8 and Fig. 9, we visualize example depth maps obtained via volume rendering our sampled neural fields.

Figure 9: **Depth visualizations (TaiChi).** We show depth maps (second row) obtained by volume rendering the sampled 4D fields at a given time step. The first row shows the corresponding RGB rendering of the same 4D fields. Here the depth is defined as the expected ray termination distance.

C Implementation Details

C.1 Architecture and Training Details

The 3D content code, motion code and style vector dimensions are all set to 512. Our motion generator (see Fig. 2) is implemented as an MLP with three fully connected (FC) layers and Leaky ReLU activations. The time step is repeated across the channel dimension and multiplied with the output of the first fully connected layer of the motion generator. We set the motion code and hidden dimension of the motion generator to 512, while the output dimension is 128. The output of the motion generator is then added to the output of the first FG NeRF Block, which also is a 128-dimensional representation. Besides the virtual camera setup and the motion generator dimensions, we adopt all hyperparameters determined in StyleNeRF (Gu et al., 2022). We set the motion generator hyperparameters empirically following previous GAN works. We believe that hyperparameter grid search and increasing the model dimensions could further boost the quality of our results, this, however, requires more computational resources. We analyze the virtual camera setup in the subsequent section.

Our foreground and background NeRF are modeled as MLPs (with Leaky ReLU activations) with 8 and 4 FC layers that each contain 128 and 64 hidden units, respectively. We use 10 frequency bands to map the positional input of the foreground background NeRF to the fourier features (Mildenhall et al., 2020). We do not apply positional encoding to the time input. We follow the implementation of StyleNeRF (Gu et al., 2022) for the the 2D ray aggregation and upsampling block (see Fig. 2) and the volume rendering process. Both the image and video discriminator follow the architecture of StyleGAN2 (Karras et al., 2020) with hidden dimensions of 512, and the input channels being 3 and 7, respectively. We apply the Differentiable Augmentation technique (Zhao et al., 2020) with all augmentations except CutOut, to prevent the discriminators from overfitting to the relatively small video datasets. In contrast to StyleNeRF (Gu et al., 2022), we do not use progressive-growing training (Karras et al., 2018) but directly train on the final image resolution, as we did not observe any change in visual quality.

For both the generator and discriminator, we use the Adam optimizer (Kingma & Ba, 2015) with a learning rate of 0.0025, $\beta_1 = 0$, $\beta_2 = 0.99$ and $\epsilon = 10^{-8}$. We follow the setup of StyleGAN2 (Karras et al., 2020) to use 8 fully connected layer content mapping network and apply $100\times$ lower learning rate compared to that of the main generator layers. For our objective function (Eq. 7), we set $\lambda_1 = 0.5$ and $\lambda_2 = 0.2$. We use 16 samples for the NeRF path regularization (Gu et al., 2022).

C.2 Virtual Camera Setup

For the three main datasets we empirically set the virtual camera on the surface of unit sphere and parameterize the camera viewpoint distribution with pitch and yaw angles. The standard deviation for pitch sampling is 0.15 for all three datasets. For yaw sampling the standard deviation is 0.3, 0.3, and 0.8 for FaceForensics, MEAD, TaiChi. The field-of-view of the camera is set to be 18 degrees.

For the SkyTimelapse dataset we do not sample on a sphere, but place the camera at the origin and make the camera look outwards. We uniformly sample a point on a hemisphere and set the camera to look towards the direction. The field-of-view is set to be 80 degrees. Note that this setup only models rotation of the camera. We leave a more complicated camera sampling method as future work.

C.3 Motion and Content Disentanglement

The motion and content are mainly disentangled because of our two-frame discrimination in combination with the different sampled latent codes. While the content code is fixed across the two frames, the intermediate motion code computed from the sampled motion code and variable timesteps are not constant across the two frames. Hence, the FG network receives two different motion features for the same content during training. This inductive bias allows the model to disentangle content from motion. Note that we mainly follow the 2D video generator MoCoGAN (Tulyakov et al., 2018), which disentangles motion and content in a similar fashion and sets a default approach for motion and content decomposition in video generation for most of the follow-up works. In early experiments, we observed slightly more content inconsistencies when injecting the content at deeper layers, hence we inject the motion directly after the first layer. We also choose the output of the first layer and not the input of the first layer as this allows us to use the addition operation by first mapping the number of channels to a common channel number for motion generator output and FG MLP. Otherwise we have to use, e.g., concatenation which does not allow the easy pre-training technique by disabling motion in the sum as discussed in Sec. 3.2.

D Experiment Details

D.1 Evaluations

We used the code and evaluation protocol of StyleGAN-V (Skorokhodov et al., 2022a) for computing FVD (Unterthiner et al., 2018) and FID (Heusel et al., 2017). The FVD protocol requires 2048 16-frame videos, while the FID score uses 50K images. For MEAD dataset (Wang et al., 2020a), we re-trained StyleGAN-V (Skorokhodov et al., 2022a) using their official code as the authors only provided results on the front view videos of MEAD at 1024^2 resolution. We randomly choose 10,000 videos across all viewpoints, including non-frontal views and follow the identical training setup provided by Skorokhodov et al. (2022a) to process 25,000K images with batch size 64. For TaiChi dataset, we use the officially provided checkpoint of DIGAN (Yu et al., 2022) to evaluate their model with the new FVD protocol (Skorokhodov et al., 2022a). For FaceForensics dataset, we use the reported numbers provided by the StyleGAN-V (Skorokhodov et al., 2022a) authors for all models in Table 1, except for StyleNeRF (Gu et al., 2022) and our model, which we train from scratch. We provide sampled 3D models of StyleNeRF trained on FaceForensics, rendered from a horizontally moving camera, as shown in Fig. 10.

We train our model and StyleNeRF using 4 NVIDIA V100 GPUs. For our approach we train for a maximum of 3,000K images, which takes two days at 256^2 resolution. For StyleNeRF we abort the training after 15,000K images due to the FID diverging after 11,400K images. It is generally true that 2D models like StyleGAN-V (Skorokhodov et al., 2022a) have a significantly faster throughput, as there is no costly volume rendering. However, we observe that our 4D model converges at much lower iterations, already converging after 2,500K processed images, while some of the 2D video models such as VideoGPT (Yan et al., 2021) does not converge even at 25,000K processed images. We hypothesize that the extraordinary fast convergence of our model is due to the explicit disentanglement of 3D content, camera viewpoints, and motions.

D.2 Statistical Reproducibility

For computing the FVD, we follow the protocol of StyleGAN-V (Skorokhodov et al., 2022a), which aims to reduce the score variations significantly compared to the original protocol (Unterthiner et al., 2018). Nevertheless, in Table 5 we report the relative standard deviations after evaluating our three main dataset results for 10 rounds with the FVD protocol (Skorokhodov et al., 2022a). We observe that the standard deviation is rather small across all three datasets, as reported in the extensive analysis of Skorokhodov et al. (2022a).

Figure 10: **StyleNeRF results.** We show qualitative results of StyleNeRF model trained on FaceForensics, rendered from five different cameras (columns) for three identities (rows).

Dataset	FVD Relative Standard Deviation
FaceForensics	2.62%
TaiChi	1.65%
MEAD	2.17%

Table 5: Relative standard deviation for the FVD metric on FaceForensics, TaiChi, and MEAD datasets, computed as percentage of standard deviation with respect to the mean.

Method	CD (↓)
StyleNeRF	1.27
Ours	1.03

Table 6: Chamfer distance evaluation on FaceForensics between StyleNeRF and our approach, following the protocol of Or-El et al. (2022b).

D.3 3D Consistency Evaluation

In Table 6 we further report an evaluation of the underlying 3D model using the chamfer distance (CD) metric. Specifically, we follow the protocol of Or-El et al. (2022b) and evaluate chamfer distance between the front pose and a randomly sampled side pose on 1000 generated FaceForensics samples. We compare our results to StyleNeRF, as it is the only 3D-aware baseline. We observe a lower chamfer distance, which aligns with our previous results that we can extend a 3D-aware image generator in the temporal domain to generate 3D videos without sacrificing any visual quality or geometric consistency.

E Training on Image and Video Datasets

E.1 Training Setup

In this section we discuss the use of an auxiliary 2D image dataset for training our 4D GANs, as described in Sec. 3.2 in the main paper. As motivated in the main paper, the lack of diversity of the provided videos could diminish the 3D accuracy of 4D GANs. In fact, we observe that the model trained only on the FaceForensics dataset (Rössler et al., 2019) exhibits narrow range of camera angles that generates view-consistent renderings.

As described in the main paper, we explored two options to leverage an image dataset to complement the video dataset: (i) pre-training on the image dataset, and (ii) simultaneously training both on image and video

Figure 11: **Comparing the range of pitch control.** We show that our 4D model trained solely on the FaceForensics (Rössler et al., 2019) dataset (top row) has a narrow range of pitch where high quality renderings can be produced, likely due to the lack of diverse training views. Note that the upper and bottom parts of the head becomes unnaturally large at both extremes. Fine-tuning a model pre-trained on FFHQ (Karras et al., 2019) image dataset (bottom row), which features more diverse view angles, results in a wider range of views that generate plausible images.

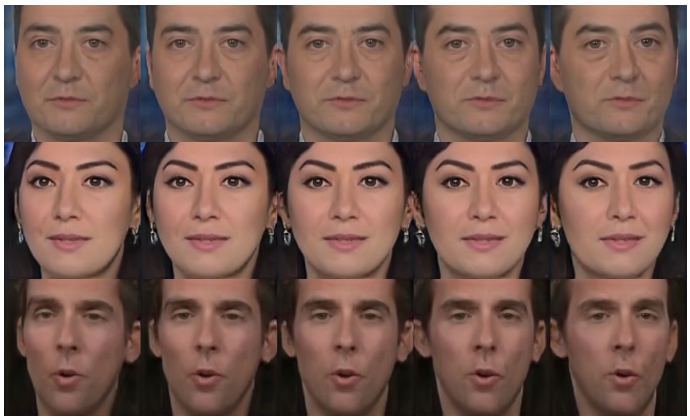

Figure 12: **Qualitative results for pre-training on FFHQ before training on FaceForensics.** We show qualitative results of our model pre-trained on FFHQ and then trained on FaceForensics, rendered from five different cameras (columns) for three identities (rows). We observe high view-consistency and high quality results.

datasets. For both options, we follow the training setup of StyleNeRF (Gu et al., 2022) on computing and backpropagating the image losses.

In early experiments we also explored pre-training and simultaneous training for MEAD to better interpolate between the discrete poses in the data distribution. However, combining MEAD with FFHQ does not lead to any improvement in neither quality nor 3D behavior. We hypothesize that training on the image dataset mainly leads to improvements when the camera poses of the video dataset are already diverse but narrow as in FaceForensics. MEAD provides a wide but extremely sparse range of camera poses, which can not be interpolated by the image dataset distribution. We did not leverage image datasets for TaiChi as the viewpoint distribution is already diverse. For a fair comparison to previous works, we only used our vanilla approach for both qualitative and quantitative comparisons in our main paper.

E.2 Qualitative Results

Figure 13: **FFHQ and FaceForensics simultaneous training results.** We show qualitative results of our model simultaneously trained on FFHQ and FaceForensics, rendered from five different cameras (columns) for three identities (rows). We observe lower view-consistency compared to our model only pre-trained on FFHQ.

Method	FVD (↓)	FID (↓)	ACD (↓)	CPBD (↑)	ID (↑)
Pre-trained	127.2	17.2	0.819	0.2084	0.982
Simultaneously	62.1	18.2	0.945	0.2077	0.893
Vanilla	68.7	13.7	0.965	0.196	0.861

Table 7: **Quantitative Results on FaceForensics in combination with FFHQ.** We report metrics for pre-training on FFHQ before training on FaceForensics, simultaneously training on FFHQ and FaceForensics, and only training on FaceForensics (vanilla).

The model pre-trained on the FFHQ image dataset and fine-tuned on the video dataset generates high quality, 3D-aware renderings, as can be seen on the supplementary videos. We generally observe sharper and higher quality renderings compared to our default model. Moreover, we notice that the range of allowed view angles increases, which is most obvious for the pitch (or elevation) angles, as described in Fig. 11. We note, however, that the colors seem less vivid compared to the pure-video model.

We further provide qualitative results for pre-training and simultaneous training in Fig. 12 and Fig. 13, respectively. For the case of simultaneous image and video training, the resulting model fails to output view-consistent videos. We hypothesize that training the generators to fit two different data distributions does not lead to consistent 4D models. However, only pre-training on FFHQ leads to view-consistent renderings while maintaining high quality.

E.3 Quantitative Results

In Table 7 we provide a quantitative comparison between i) pre-training on FFHQ before training on FaceForensics, ii) simultaneously training on FFHQ and FaceForensics, and iii) only training on FaceForensics as a vanilla approach. Generally, it is difficult to interpret FVD and FID when using two training datasets as these scores compare real and fake data distributions. Due to the usage of the FFHQ dataset the latent space of the model is manipulated with a dataset that is not used as part of the evaluation process. We still provide these two scores for completeness. For our pre-trained model, we observe that the FVD gap is significantly higher than the FID gap in comparison to our vanilla approach. We hypothesize that due to the model converging to a static dataset in the first training phase, namely FFHQ, the model has difficulties in learning motion in the second training phase with FaceForensics. However, we observe a significantly improved ID, which confirms the better multi-view consistency in our videos. Simultaneous training results

Method	FVD (↓)
w/ Positional Time Encoding	175.4
w/ Concatenation	164.7
w/ Multiplication	158.3

Table 8: **Ablation of time incorporation in motion generator on TaiChi Dataset.** We choose simple multiplication for our default model due to its low FVD score in comparison to positional time encoding and and concatenation.

Method	FVD (↓)
w/o Image Discriminator	234.3
w/ Deterioration of Video Discriminator	190.9
w/ Separate Image Discriminator	158.3

Table 9: **Ablation of image discriminator on TaiChi Dataset.**

in similar scores as our vanilla approach, however the multi-view consistency is significantly worse than using pre-training, which aligns with the lower ID score.

F Ablations

We provide further details for the ablations conducted in Sec. 4.3.

F.1 Motion Vector

Generally, there are different ways to incorporate time into the motion generator. We use a simple multiplication due to our empirical results on TaiChi showing the best FVD score of 158.3 for this setup, as shown in Table 8. As explained in Sec. 4.3, using positional encoding leads to a worse FVD with 175.4. Furthermore, we experimented with concatenation of time, which also results in a slightly worse FVD with 164.7. We leave more complex incorporation of time, like acyclic positional encoding (Skorokhodov et al., 2022a), for future work.

F.2 Image Discriminator

Theoretically, our video discriminator should suffice for the model to learn from videos. However, we observe that using a separate image discriminator significantly boosts the quality of our results, as shown in Table 9. Removing the image discriminator leads to a FVD score of 234.3. Another option is to use a deterioration of our video discriminator as an image discriminator. For this, we repeat every image along the channel dimension to form a 6 channel tensor and set the time difference to zero, which we further concatenate to obtain a 7 channel tensor for our video discriminator. This leads to a FVD score of 190.9, which is substantially better than just removing the image discriminator, however worse than using a separate image discriminator (158.3). Consequently, we observe that using two discriminators with disentangled weights significantly boosts the quality.

G Related Works with Complete References

In this section, we discuss, in more detail, relevant prior works that was omitted in our main paper due to space constraints.

Neural Implicit Representations Neural Implicit Representation (NIR) (Mescheder et al., 2019; Park et al., 2019; Chen & Zhang, 2019; Michalkiewicz et al., 2019) have demonstrated impressive results on various

tasks due to their continuous, efficient, and differentiable representation. Among the most widely utilized coordinate-based representations are Neural Radiance Fields (NeRFs) (Mildenhall et al., 2020) that combine an implicit neural network with volumetric rendering to enforce 3D consistency, while performing the novel view synthesis task. Implicit-based representations have been proven beneficial for other tasks including but not limited to 3D reconstruction of objects (Mescheder et al., 2019; Park et al., 2019; Chen & Zhang, 2019; Michalkiewicz et al., 2019; Oechsle et al., 2021; Gropp et al., 2020; Saito et al., 2019; Atzmon et al., 2019; Sitzmann et al., 2019) and scenes (Jiang et al., 2020; Chibane et al., 2020; Peng et al., 2020; Chabra et al., 2020; Sitzmann et al., 2020), novel-view synthesis of static (Barron et al., 2021; Bergman et al., 2021; Gao et al., 2020; Jiang et al., 2020; Liu et al., 2020; Lindell et al., 2022; Srinivasan et al., 2021; Piala & Clark, 2021; Martin-Brualla et al., 2021; Oechsle et al., 2021; Sajjadi et al., 2022) and dynamic (Lombardi et al., 2019a; Li et al., 2021; Pumarola et al., 2021a; Park et al., 2021a; Xian et al., 2021; Yuan et al., 2021a; Park et al., 2021c; Tretschk et al., 2021a; Tewari et al., 2021) environments, inverse graphics (Niemeyer et al., 2020; Yariv et al., 2020; Lin et al., 2020) as well as for representing videos (Chen et al., 2021a; Yu et al., 2022).

3D-Aware Image Generations Many recent models investigate how 3D representations can be incorporated in generative settings for improving the image quality (Park et al., 2017; Nguyen-Phuoc et al., 2018) and increasing the controllability over various aspects of the image formation process (Gadelha et al., 2017; Chan et al., 2021; Henderson & Ferrari, 2019; Henzler et al., 2019; Henderson & Lampert, 2020; Liao et al., 2020; Lunz et al., 2020; Nguyen-Phuoc et al., 2019; 2020; Schwarz et al., 2020; Gu et al., 2022; DeVries et al., 2021; Hao et al., 2021; Meng et al., 2021; Niemeyer & Geiger, 2021b; Zhou et al., 2021). Towards this goal, Henzler et al. (2019) proposed a GAN-based architecture that combines voxel-based representations with differentiable rendering. However, due to the low voxel resolution, their generated images suffered from various artifacts. Concurrently, HoloGAN (Nguyen-Phuoc et al., 2019) demonstrated that adding inductive biases about the 3D structure of the world allows control over the pose of the generated objects. Nguyen-Phuoc et al. (2020) extend this by also considering the compositional structure of the world into objects. While both Nguyen-Phuoc et al. (2019; 2020) demonstrated impressive results, their performance was less consistent in higher resolutions. Another line of work Liao et al. (2020) propose to combine a 3D generator with a differentiable renderer and a 2D image generator to enable 3D controllability. Despite their promising results in synthetically generated environments, their model struggles to generalize in real-world scenarios. More recent approaches propose generative models using MLP-based radiance fields. These methods use volume rendering to obtain images from the 3D fields to model single (Schwarz et al., 2020; Chan et al., 2021) or multiple (Niemeyer & Geiger, 2021a) objects. Similarly, Zhou et al. (2021), Chan et al. (2021), and DeVries et al. (2021) explored the idea of combining NeRF with GANs for designing 3D-aware image generators. Xu et al. (2021) propose a transition from cumulative rendering to rendering with only the surface points. ShadeGAN (Pan et al., 2021) introduces a multi-lighting constraint to improve the 3D representation. More recently, StyleSDF (Or-El et al., 2022a) and StyleNeRF (Gu et al., 2022) proposed to combine an MLP-based volume renderer with a style-based generator (Karras et al., 2020) to produce high-resolution 3D-aware images. Moreover, VolumeGAN (Xu et al., 2022) characterizes the underlying 3D structure with a feature volume. Deng et al. (2022) explored learning a generative radiance field on 2D manifolds and Chan et al. (2022) introduced a 3D-aware architecture that exploits both implicit and explicit representations. To improve view-consistency, EpiGRAF (Skorokhodov et al., 2022b) proposes patch-based training to discard the 2D upsampling network, while VoxGRAF (Schwarz et al., 2022) uses sparse voxel grids for efficient rendering without a superresolution module.

GAN-based Video Synthesis Inspired by the success of GANs and adversarial training on photorealistic image generation, researchers shifted their attention to various video synthesis tasks (Saito et al., 2017; Tulyakov et al., 2018; Acharya et al., 2018; Clark et al., 2019; Yushchenko et al., 2019; Kahembwe & Ramamoorthy, 2019; Aich et al., 2020; Saito et al., 2020; Gordon & Parde, 2020; Holynski et al., 2021; Muñoz et al., 2021; Tian et al., 2021; Fox et al., 2021; Yan et al., 2021; Hyun et al., 2021; Yu et al., 2022; Skorokhodov et al., 2022a). Several recent works pose the video synthesis as an autoregressive video prediction task and seek to generate discrete future frames conditioned on the previous using either recurrent (Kalchbrenner et al., 2017; Walker et al., 2021) or attention-based (Rakhimov et al., 2020; Weissenborn et al., 2020; Yan et al., 2021) models. Other works on video generation (Saito et al., 2017; Tulyakov et al., 2018; Aich et al., 2020; Saito et al., 2020) tried to disentangle the motion from the image generation during the video synthesis

process. While this paradigm has been widely adopted, these approaches typically struggle to generate realistic videos. To facilitate generating high-quality frames, Tian et al. (2021) and Fox et al. (2021) employed a pre-trained image generator of Karras et al. (2020). Recently, LongVideoGAN (Brooks et al., 2022) has investigated synthesizing longer videos of more complex datasets. Closely related to our method are the recent and concurrent work of DIGAN (Yu et al., 2022) and StlyeGAN-V (Skorokhodov et al., 2022a) that generate videos at continuous time step, without conditioning on previous frames. DIGAN (Yu et al., 2022) employs an NIR-based image generator (Skorokhodov et al., 2021) for learning continuous video synthesis and introduces two discriminators: the first discriminates the realism of each frame and the second operates on image pairs and seeks to determine the realism of the motion. Similarly, StyleGAN-V (Skorokhodov et al., 2022a) employs a style-based GAN (Karras et al., 2020) and a single discriminator that operates on sparsely sampled frames.

Dynamic View Synthesis Given a video of a dynamic scene, dynamic view synthesis methods create novel views at arbitrary viewpoints and time steps. Neural Volumes (Lombardi et al., 2019b) uses an encoder-decoder network with ray marching to render dynamic objects with a multi-view capture system. STaR (Yuan et al., 2021b) trains on multi-view videos to reconstruct rigid object motion by complementing a static NeRF with a dynamic NeRF. Similarly, Gao et al. (2021b) introduce a dynamic network, which predicts scene flow to create a warped radiance field from monocular videos. Inspired by level set methods, HyperNeRF (Park et al., 2021d) models each frame as a nonplanar slice through a hyperdimensional NeRF. Another line of works (Gafni et al., 2021; Liu et al., 2021b; Noguchi et al., 2021; Peng et al., 2021) model dynamic humans from videos based on radiance fields. Other approaches (Tretschk et al., 2021b; Pumarola et al., 2021b; Park et al., 2021b) encode the scene into a canonical space to then deform the canonical radiance field. Recently, DyNeRF (Li et al., 2022) uses multi-view videos and proposes a novel hierarchical training scheme in combination with ray importance sampling. In contrast to our method, the aforementioned works require videos with different viewpoints to train a single network per scene, hence are not able to generate novel scenes. Instead, we learn a generative model from unstructured single-view video collections.

H Limitations & Discussions (continued)

As mentioned in the main paper, we inherit the features of existing 3D-aware GANs that typically work the best when there is a single target object at the center of the scenes, e.g., human faces. A possible way of modeling multi-object scenes is to add in compositionality to scene distribution modeling (as in Niemeyer & Geiger (2021a) and Hudson & Zitnick (2021)). Another factor that limits the modeling of larger and more complex scene is sampling of virtual cameras. Training of GANs involves approximating the input data distribution with the generators. For the case of 3D GANs, image generation requires sampling virtual viewpoints from a plausible viewpoint distribution, which could be designed manually for single-object scenes with outside-inward looking cameras. However, for larger scenes, obtaining or designing the plausible viewpoint distribution becomes challenging, due to lack of correspondences across scenes and high structural variability. An interesting future direction would be to model the camera viewpoint distribution using neural networks.

While our 4D GAN compares competitively in terms of training time until convergence, we note that the volume rendering process or our approach leaves huge memory footprints. In fact, training of our model consumes “2.4 GB / batch size," while the 2D video method of StyleGAN-V (Skorokhodov et al., 2022a) only consumes “0.8 GB / batch size." This relatively high memory consumption prevents us from using larger models with the highest generation qualities and limits the resolution of output renderings. We leave exploration of more efficient implicit representations, such as tri-plane representations (Chan et al., 2022), as future work to improve the efficiency and quality 4D GANs.