Generative Novel View Synthesis with 3D-Aware Diffusion Models

Eric R. Chan [*†1,2], Koki Nagano[*2], Matthew A. Chan[*2], Alexander W. Bergman[*1], Jeong Joon Park[*1], Axel Levy[1], Miika Aittala[2], Shalini De Mello[2], Tero Karras[2], and Gordon Wetzstein[1]

[1]Stanford University [2]NVIDIA

Abstract

We present a diffusion-based model for 3D-aware generative novel view synthesis from as few as a single input image. Our model samples from the distribution of possible renderings consistent with the input and, even in the presence of ambiguity, is capable of rendering diverse and plausible novel views. To achieve this, our method makes use of existing 2D diffusion backbones but, crucially, incorporates geometry priors in the form of a 3D feature volume. This latent feature field captures the distribution over possible scene representations and improves our method's ability to generate view-consistent novel renderings. In addition to generating novel views, our method has the ability to autoregressively synthesize 3D-consistent sequences. We demonstrate state-of-the-art results on synthetic renderings and room-scale scenes; we also show compelling results for challenging, real-world objects.

1. Introduction

In this work, we address multiple open problems in novel view synthesis (NVS): to design an NVS framework that (1) operates from as little as a single image and is capable of (2) generating long-range of sequences far from the input views as well as (3) handling both individual objects and complex scenes (see Fig. 1). While existing few-shot NVS approaches, trained on a category of objects with a regression objective, can generate geometrically consistent renderings, i.e., sequences whose frames share a coherent scene structure, they are ineffective in handling extrapolation and unbounded scenes (see Fig. 2). Dealing with long-range extrapolation (2) requires using a generative prior to deal with the innate ambiguity that comes with completing portions of the scenes that were unobserved in the input. In this work, we propose a diffusion-based few-shot NVS framework that can generate plausible and competitively geometrically consistent renderings, pushing the boundaries of NVS towards a solution that can operate in a wide range of challenging real-world data.

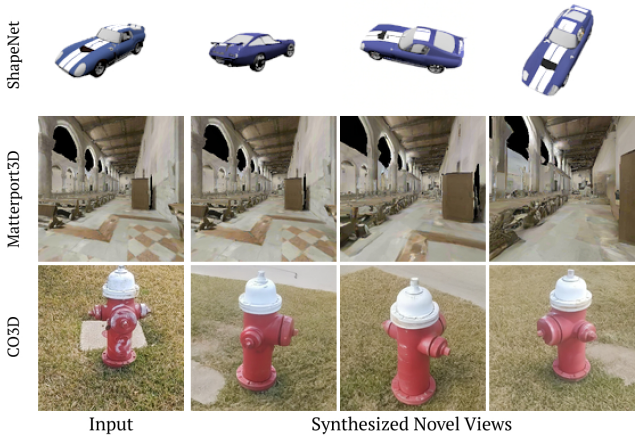

Figure 1. Our 3D-aware diffusion model synthesizes realistic novel views from as little as a single input image. These results are generated with the ShapeNet [12], Matterport3D [11], and Common Objects in 3D [58] datasets.

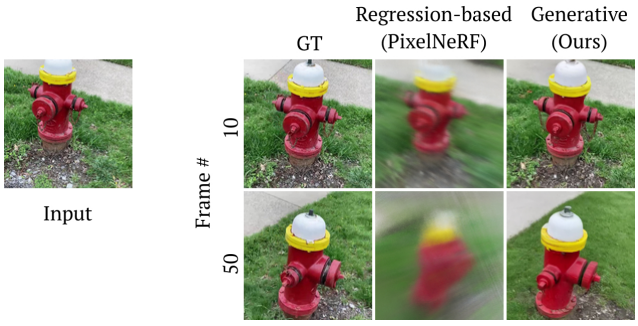

Figure 2. While regression-based models are capable of effective view synthesis near input views (top row), they blur across ambiguity when extrapolating. Generative approaches can continue to sample plausible renderings far from input views (second row, third column).

Previous approaches to few-shot novel view synthesis can broadly be grouped into two categories. Geometry-prior-based methods [61, 60, 48, 43, 49, 3, 104] have drawn from work on scene representations and neural rendering [87]. While they achieve impressive results on interpolating near

*Equal contribution.
†Work was done during an internship at NVIDIA.

input views, most methods are trained purely with regression objectives and struggle in dealing with ambiguity or longer-range extrapolations. When challenged with the task of novel view synthesis from sparse inputs, they can only tackle mildly ambiguous cases, *i.e.*, cases where the conditional distribution of novel renderings is well approximated by the mean estimator of this distribution — obtained by minimizing a pixel-wise L1 or L2 loss [106, 78, 104]. However, in highly ambiguous cases, for example when parts of the scene are occluded in all the given views, the conditional distribution of novel renderings becomes multi-modal and the mean estimator produces blurry novel views (see Fig. 2). Because of these limitations, regression-based approaches are limited to short-range view interpolation of object-centric scenes and struggle in long range extrapolation of unconstrained scenes.

By contrast, generative approaches solve the novel view synthesis problem by sampling random plausible samples from a conditional distribution modeled by a generative prior. Existing generative models for view synthesis [64, 97, 59, 44] autoregressively extrapolate one or a few input images with few or no geometry priors. For this reason, most of these methods struggle with generating geometrically consistent sequences — renderings are only approximately consistent between frames and lack a coherent rigid scene structure. In this work, we present an NVS method that bridges the gap between geometry-based and generative view synthesis approaches for both geometrically consistent and generative rendering.

Our method leverages recent developments in diffusion models. Specifically, conditional diffusion models [67, 65, 57, 63, 66] can be directly applied to the task of NVS. Conditioned on input images, these models can sample from the conditional distribution of output renderings. As a generative model, they naturally handle ambiguity and lend themselves to continued autoregressive extrapolation of plausible outputs. However, as we show in Sec. 4 (Tab. 1), an image diffusion framework alone struggles to synthesize 3D-consistent views.

Geometry priors remain valuable for ensuring view consistency when operating on complex scenes, and pixel-aligned features [68, 104, 92] have been shown to be successful for conditioning scene representations on images. We incorporate these ideas into the architecture of our diffusion-based NVS model with the inclusion of a latent 3D feature field and neural feature rendering [51]. Unlike previous view synthesis works that include neural fields, however, our latent feature field captures a distribution of scene representations rather than the representation of a specific scene. A rendering from this latent field is distilled into the rendering of a particular scene realization through diffusion sampling at inference. This novel formulation is able to both handle ambiguity resulting from long-range extrapolation and generate geometrically consistent sequences.

In summary, contributions of our work include:

- A novel view synthesis method that extends 2D diffusion models to be 3D-aware by conditioning them on 3D neural features extracted from input image(s).
- A demonstration that our 3D feature-conditioned diffusion model can generate realistic novel views given as little as a single input image on a wide variety of datasets, including object level, room level, and complex real-world scenes.
- A showcase that with our proposed method and sampling strategy, our method can generate long trajectories of realistic, multi-view consistent novel views without suffering from the blurring of regression models or the drift of pure generative models.

2. Related work

Focusing on novel view synthesis (NVS) from as little as a single image, our work touches on several areas at the intersection of 3D reconstruction, NVS, and generative models.

Geometry-based novel view synthesis. A large body of prior works for NVS recovers the 3D structure of a scene by estimating the input images' camera parameters [80, 70] and running multi-view stereo (MVS) [1, 23]. The recovered explicit geometry proxies enable NVS but fail to synthesize photorealistic and complete novel views especially for occluded regions. Some recent methods [60, 61] combine 3D geometry from an MVS pipeline with deep learning–based NVS, but the overall quality may suffer if the MVS pipeline fails. Other explicit geometric representations, such as depth maps [22, 91], multi-plane images [21, 110], or voxels [77, 46] are also used by many recent NVS approaches, as surveyed by Tewari et al. [87].

Regression-based novel view synthesis. Many deep learning–based approaches to NVS are supervised to predict training views with regression. These works often employ 3D representations for scenes and differentiable neural rendering [78, 48]. While many methods are optimized on a per-scene basis with dense input views [48], few-shot NVS approaches are designed to generalize across a class of 3D scenes, which enable them to make predictions from one or a few input images at inference. Among few-shot NVS methods, some rely on test-time optimization [78, 32] or meta learning [75, 85], while others lift input observations via encoders [91, 52, 110, 104, 88, 13, 92] and predict novel views in a feed-forward fashion. A recent trend has some NVS methods forgoing geometry priors for light fields [76] or transformers [69, 39], but these geometry-free methods are otherwise trained similarly to other regression-based NVS algorithms.

Generative models for novel view synthesis. A separate line of work studies methods for long-range view extrapolation. Because venturing far beyond the observed views requires generating parts of the scene, these methods are typically grounded in generative models. A common thread amongst these methods is that they often contain only weak geometry priors, e.g., sparse feature point clouds [97, 62, 37], or lack geometry priors altogether [64, 59]. As image-translation-based generative models, they are capable of

conditioning on their own previous generations to autoregressively synthesize long camera trajectories, sometimes infinitely [44, 41]. Because the focus is on extrapolating at large scales, these methods ordinarily achieve only approximate view consistency at longer ranges.

3D GANs. 3D GANs [50, 73, 9, 8, 25, 53, 101, 109, 79, 103, 108, 16, 98, 4, 102] combine an adversarial [24] training strategy with implicit neural scene representations to learn generative models for 3D objects. While typically tasked with unconditional synthesis of 3D objects, a trained 3D GAN contains a strong prior for 3D shapes and can be inverted for NVS of detailed scenes [9, 8]. 3D GANs have been extensively developed to achieve compositionality [51], higher rendering resolution [8, 25, 79], video generation [2], and scalability to larger scenes [17]. GANs, however, are notoriously difficult to train, and their 3D inversions from an input image are often brittle without additional 3D priors [100] or an accurate camera input [36]. Moreover, most 3D GANs assume canonical camera poses and limit their optimal operating ranges to single objects.

2D diffusion models. 2D diffusion models [29, 81, 83, 34] have transformed image synthesis. Favorable properties such as mode coverage and a stable training objective have enabled them to outperform [18] previous generative models [24] on unconditional generation. Diffusion models have also been shown to be excellent at modeling conditional distributions of images, where the conditioning information may be a class label [84, 18], text [57, 63, 66] or another image [30, 67, 65, 10].

Recent 3D diffusion works. Recently, DreamFusion [55] and 3DiM [96] apply 2D image diffusion models to build 3D generative models. DreamFusion performs text-guided 3D generation by optimizing a NeRF from scratch. 3DiM performs novel view synthesis conditioned on input images and poses (similar to [59]) and does not employ any explicit geometry priors; it aggregates multiple observations at inference using a unique stochastic conditioning scheme. By contrast, the geometry priors present in our approach enable 3D consistency with a much lighter-weight model (90M for ours vs 471M or 1.3B for 3DiM [96]), and because our model naturally handles multiple input views, we have the flexibility to choose efficient sampling schemes at inference. While code for 3DiM is unavailable, we compare to a similar geometry-free variant in Sec. 4 (Tab. 1) and to stochastic view conditioning in the supplement.

A number of concurrent works have been proposed which utilize 2D diffusion models for novel view synthesis and 3D generation [105]. Many such works build on DreamFusion, either providing improvements to the score distillation loss formulation [95, 27, 35] or underlying 3D representation [42, 15, 5, 38, 33, 7, 89, 56]. More similar to our method are those which use pose-conditioned image diffusion models to synthesize novel views [45, 90, 86, 99, 26, 111]. These methods either use specialized architectures to inject the pose conditioning [45, 90], use warped images as conditioning [99], or use a partially optimized 3D representation to serve as a geometric prior for conditioning [86]. Our method strikes a balance between these approaches: no scene-specific training is needed, but we still capture the strong geometric prior of feature re-projection in a 3D space.

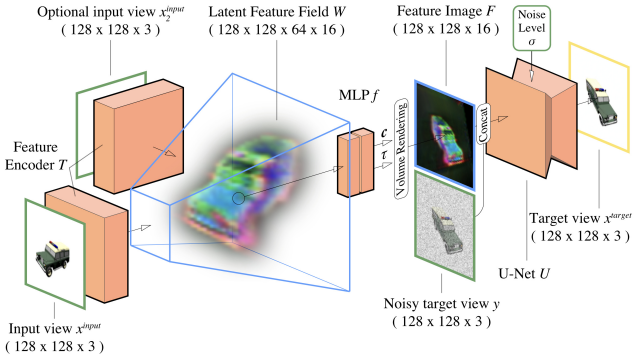

Figure 3. Illustration of our framework D. The pipeline receives as input one or more input views $\boldsymbol{x}$ and the camera parameters associated with input and target views. We extract features from each input view $\boldsymbol{x}$ using T and unproject them into a feature volume $\boldsymbol{W}$. These volumes are aggregated using a mean-pooling operation, decoded by a small MLP f, and a feature image F is created by projecting into the target view $\boldsymbol{x}^{\text{target}}$ using volume rendering. The U-Net denoiser U then takes in the resulting feature image F as well as a noisy image of the target view $\boldsymbol{y}$ and noise level σ, and produces a denoised image of the target view $\boldsymbol{x}^{\text{target}}$.

A final line of related work [74, 93] adapts 2D diffusion models to 3D object generation by training diffusion models on datasets of optimized tri-planes [8, 54]. They have achieved impressive results on object synthesis but require two-stage training and are limited to object generation. While our work also adapts 2D diffusion architectures for 3D generation, it can be applied to larger scenes and is trained end-to-end.

3. Method

Here we describe the architecture of our NVS model for both single and multiple-view conditioning, and we explain our training and inference methods.

In novel view synthesis, we are given a set of input images $\boldsymbol{x}^{\text{inputs}}$ and camera parameters $\mathbf{P}^{\text{inputs}}$ with associated pose and intrinsics and are tasked with making a prediction for a query view given a set of query camera parameters.

Our goal is to sample novel views from the corresponding conditional distribution:

$$p(\boldsymbol{x}^{\text{target}} | \boldsymbol{x}^{\text{inputs}}, \mathbf{P}^{\text{inputs}}, \mathbf{P}^{\text{target}}). \tag{1}$$

3.1. 3D-aware diffusion model architecture

Diffusion models rely on a denoiser trained to predict $\mathbb{E}_{p(\boldsymbol{x}|\boldsymbol{y})}[\boldsymbol{x}]$ given $\boldsymbol{y}$, a noisy version of $\boldsymbol{x}$ with noise standard deviation σ. An image is generated by drawing $\boldsymbol{y}_0 \sim \mathcal{N}(0, \sigma_{\text{max}}^2 \mathbf{I})$ and iteratively denoising it according to a sequence of noise levels $\sigma_0 = \sigma_{\text{max}} > \ldots > \sigma_N = 0$.

In our work, we directly repurpose 2D diffusion models to model the distribution in Eq. 1. The intuition is that generative novel view synthesis is identical to any other conditional image generation task — all we need to do is condition a 2D image diffusion model on the input image and the relative camera pose. However, while there are many ways of applying this conditioning, some may be more effective than others (see Tab. 1 and ablation studies of different options in Sec. 4.4). By incorporating geometry priors in the form of a 3D feature field and neural rendering, we give our architecture a strong inductive bias towards geometrical consistency.

Fig. 3 summarizes the design of our conditional-desnoiser-based pipeline D that takes as inputs a noisy target view $\boldsymbol{y}$, conditioning information $(\boldsymbol{x}^{\text{inputs}}, \mathbf{P}^{\text{inputs}}, \mathbf{P}^{\text{target}})$ and a noise level σ. Our strategy builds upon pixel-aligned implicit functions [68, 104] and neural rendering. Following Fig. 3, given a single input image $\boldsymbol{x}$ taken from an input view camera $\mathbf{P}$, we use an image-to-image translation network T to predict a feature image with $c \times d$ channels and reshape it into a feature volume $\boldsymbol{W}$ that spans the source camera frustum. d then corresponds to the depth dimension of the volume and c to the number of channels in each cell of the volume (typically, $c = 16$ and $d = 64$). Given a query camera $\mathbf{P}^{\text{target}}$, we cast rays in 3D space. Continuing on Fig. 3, for any point $\mathbf{r}$ along a ray, we sample the volume $\boldsymbol{W}$ with trilinear interpolation and decode the obtained feature $w = \boldsymbol{W}(\mathbf{r})$ with a small multi-layer perceptron (MLP) f to obtain a density τ and a feature vector $\mathbf{c}$

$$(\tau, \mathbf{c}) = f(w). \tag{2}$$

By projecting this feature field into the target view using volume rendering [47, 48], we obtain a feature image F in Fig. 3:

$$F(\boldsymbol{x}, \mathbf{P}, \mathbf{P}^{\text{target}}) = \text{RENDER}(f \circ T(\boldsymbol{x}), \mathbf{P}, \mathbf{P}^{\text{target}}). \tag{3}$$

In practice, we employ the image segmentation architecture *DeepLabV3+* [14, 31] for T, and implement f as a two-layer ReLU MLP with 64 channels. We perform volume rendering over features in the same way as *NeRF* [48]. We use input/output image resolution 128^2 in all experiments.

The feature image F is concatenated to the noisy image $\boldsymbol{y}$ and passed as input to a denoiser network U to produce the final target view $\boldsymbol{x}^{\text{target}}$ (see Fig. 3). We use *DDPM++* [84, 34] for U, where

$$D(\boldsymbol{y}; \boldsymbol{x}^{\text{inputs}}, \mathbf{P}^{\text{inputs}}, \mathbf{P}^{\text{target}}, \sigma) = U(\boldsymbol{y}, F; \sigma) \tag{4}$$

Fig. 3 and Eq. 4 summarize the design of D. The total number of trainable parameters in D is 90M.

3.2. Incorporating multiple views

The previous section describes our approach to conditioning on a single input view. However, additional information in the form of multiple input views reduces uncertainty and enables our model to sample renderings from a narrower distribution. When multiple conditioning views are available, we process each input image independently into a separate feature volume.

Eq. 2 can be generalized to n conditioning views by averaging the features $w_j = \boldsymbol{W}_j(\mathbf{r})$ obtained for each input image $\boldsymbol{x}_j$, as in [104]:

$$(\tau, \mathbf{c}) = f\left(\frac{1}{n}\sum_{j=1}^{n} w_j\right). \tag{5}$$

To leverage this strategy during inference, we train our model by conditioning with multiple (variable) input images. Conditioning using multiple input images helps to ensure smooth, loop-consistent video synthesis. While conditioning on only the previous frame is sufficient for view consistency in a small view change, it does not guarantee loop closure. In practice, we find that conditioning on a subset of previous views helps to enforce correct loop closure while maintaining reasonable view to view consistency.

3.3. Training

At each iteration during training, we sample a batch of target images, input images, and their associated camera poses, where the targets and inputs are constrained to be from the same scene. Our model is trained end-to-end from scratch to minimize the following objective

$$L := \mathbb{E}_{(\boldsymbol{x}^{\text{target}}, \boldsymbol{x}^{\text{inputs}}, \mathbf{P}^{\text{target}}, \mathbf{P}^{\text{inputs}}) \sim p_{\text{data}}} \mathbb{E}_{\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I})} \left[\| D(\boldsymbol{x}^{\text{target}} + \varepsilon; \boldsymbol{x}^{\text{inputs}}, \mathbf{P}^{\text{inputs}}, \mathbf{P}^{\text{target}}, \sigma) - \boldsymbol{x}^{\text{target}} \|_2^2 \right], \tag{6}$$

where σ is sampled during training according to the strategy proposed by *EDM* [34]. The number of conditioning views for a query is drawn uniformly from $\{1, 2, 3\}$ at every iteration. During training, we apply non-leaking augmentation [34] to U and augment input images with small amounts of random noise. Please see the supplement for hyperparameters and additional training details.

3.4. Generating novel views at inference

Sampling a novel view with our method is identical to sampling an image with a conditional diffusion model. The specific update rule for the denoised image is determined by the choice of sampler. In our experiments, we use a deterministic 2^{nd} order sampling strategy [34], with 25 or fewer denoising steps. Other sampling strategies [84, 82] can be dropped in if other properties (*e.g.*, stochastic sampling) are desired.

In order to improve efficiency at inference, we decouple Γ and U. Rather than running both Γ and U at every step during sampling, we first render the feature image F as a preprocessing step and reuse it for each iteration of the sampling loop – while U must run every step during inference, Γ is run only once.

Alternative "one-step" inference. An alternative variant of our model to generating an image with iterative denoising is to produce the image with a single step of denoising. Intuitively, the one-step prediction of a model trained with Eq. 6

Figure 4. Qualitative comparison on ShapeNet [12] with one input view. Unlike regression-based approaches, our method produces sharp realizations. With one-step inference, our approach behaves like a mean estimator of the novel view, similarly to PixelNeRF.

should behave identically to the prediction of a model trained to minimize pixel-wise MSE. Thus, this alternative inference mode is representative of regression-based methods. A model trained as described is capable of both generative sampling and deterministic one-step inference—no architecture or training modifications are required.

3.5. Autoregressive generation

In order to generate consistent sequences, we take an autoregressive approach to synthesizing sequential frames. Instead of independently generating each frame conditioned only on the input images, which would lead to large deviations between frames, we generate each frame conditioned on the inputs as well as a subset of previously generated frames. While there are many possible ways of selecting conditioning views, a reasonable setting that we use in our experiments is to condition on the input image(s), the most recently generated image, and five additional images drawn at random from the set of previously generated frames.

We found this default conditioning setting to be a good starting point that balances short range, frame-to-frame consistency, long-range consistency across the scene, and compute cost, but other variants may be preferred to emphasize specific qualities.

While one might expect errors and artifacts to accumulate throughout long autoregressive sequences, in practice we find that our model effectively suppresses such errors, making it suitable for extended sequence generation. Please see the supplement for alternative autoregressive schemes.

4. Experiments

We evaluate the performance of our generative NVS method on ShapeNet [12] "cars" and Matterport3D [11], two starkly different datasets. ShapeNet is representative of synthetic, object-centric datasets that have long been dominated by regression-based approaches to NVS (e.g., [104, 76]). Meanwhile, long-range NVS on Matterport3D is prototypical of unbounded scene exploration, where generative models with weak geometry priors [97, 64, 59] have seen more success. Finally, we stress-test our method on the challenging

		FID↓	LPIPS↓	DISTS↓	PSNR↑	SSIM↑
	PixelNeRF [104]	65.83	0.146	0.203	**23.2**	0.90
	ViewFormer [39]	20.82	0.146	0.161	19.0	0.83
	EG3D-PTI [8]	27.23	0.150	0.310	19.0	0.85
	3DiM (autoregressive) [96]†	8.99			21.01	0.57
Ours	Explicit	8.09	0.129	0.158	19.1	0.86
Ours	Geom.-Free	16.68	0.342	0.329	13.1	0.74
Ours	One-Step	42.07	0.150	0.178	**23.2**	**0.91**
Ours	Full (autoregressive)	11.08	0.120	0.146	20.6	0.89
Ours	Full	**6.47**	**0.104**	**0.145**	20.7	0.89

Table 1. Quantitative comparison of single-view novel view synthesis on ShapeNet cars [12, 78]. † As reported by [96].

Common Objects in 3D (CO3D) [58], an unconstrained real-world dataset — to our knowledge, our work is the first to attempt single-shot NVS on this dataset while including its complex backgrounds. Our method improves upon the state-of-the-art for all tasks. For additional results, please refer to the videos contained in the supplement.

Baselines and implementation details. For ShapeNet and CO3D, we compare our method to PixelNeRF [104], a state-of-the-art NeRF-based method for NVS, and ViewFormer [39], a transformer-based, geometry-free approach to NVS. For ShapeNet, we additionally provide a comparison with EG3D-PTI [8], which is based on a state-of-the-art 3D GAN for object-scale scenes, and a numerical comparison with 3DiM [96], a recent geometry-free diffusion method for NVS. For Matterport3D, we compare our method against the state-of-the-art on this dataset: Look Outside The Room [59], a transformer-based, geometry-free NVS method designed for room-scale scenes, and to additional SOTA methods, including SynSin [97] and GeoGPT [64] in Tab. 2.

Metrics. We evaluate the task of novel view synthesis along three axes: ability to (1) recreate the image quality and diversity of the ground truth dataset, (2) generate novel views consistent with the ground truth, and (3) generate sequences that are geometrically consistent. For (1), we use distribution-comparison metrics, FID [28] and KID [6], which are commonly used to evaluate generative models for image synthesis. For (2), we use perceptual metrics LPIPS [107] and DISTS [20], which measure structural and texture similarity between the synthesized novel view and ground-truth novel view. For completeness, we include PSNR and SSIM, although the drawbacks of these metrics are well-studied: these raw pixel metrics have been shown to be poor evaluators of generative models as they favor conservative, blurry estimates that lack detail [67, 65]. For (3), we provide COLMAP [71, 72] reconstructions of generated video sequences, a standard evaluation for 3D consistency in 3D GANs [73, 9, 8]. Dense, well-defined point clouds are indicative of geometrically consistent frames. We calculate Chamfer distances between reconstructions of the ground-truth images and reconstructions of generated sequences to quantitatively evaluate geometrical consistency.

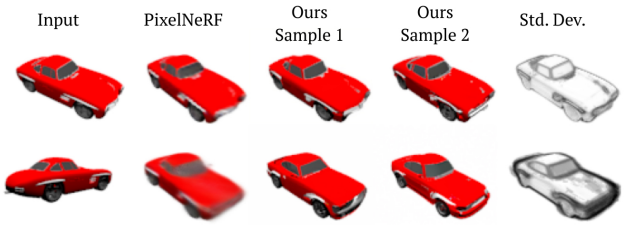

Figure 5. Generating new views from more (bottom) or less (top) ambiguous conditioning information. PixelNeRF [104] is constrained to output deterministic novel views and renders an average of all plausible renderings that are consistent with the input view. In comparison, our method samples the conditional distribution, leading to sharp but different realizations. In the last column, we show the per-pixel standard deviation of the novel view and show that unseen areas are more ambiguous, *i.e.*, vary more from one sample to the other. Pixel-wise standard deviation is computed over 50 samples. Dark pixels indicate higher ambiguity.

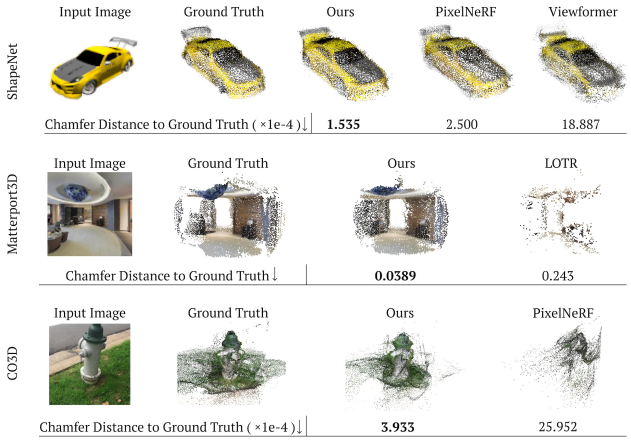

Figure 6. COLMAP reconstructions from video sequences produced by our method are dense, well-defined, and highly similar to reconstructions of the ground-truth images, demonstrating a high degree of geometric consistency, as measured by Chamfer distance. The three rows show results on ShapeNet, Matterport3D, and CO3D, respectively.

4.1. ShapeNet

We standardize our training and evaluation on the single-class, single-view NVS benchmark described in [104, 78, 39]. The ShapeNet training set contains 2,458 cars, each with 50 renderings randomly distributed on the surface of a sphere. For evaluation, we use the provided test set with 704 cars, each with 250 rendered images and poses on an Archimedean spiral. All evaluations are conducted with a single input image. For our model, we evaluate both independently generated frames and frames generated with autoregressive conditioning. In addition to our model and the baselines, we provide additional comparisons to several ablative variants of our approach, which are discussed in more detail in Sec. 4.4.

Fig. 4 provides a qualitative comparison against baselines for single-view novel view synthesis on ShapeNet. In contrast to PixelNeRF, which predicts a blurry mean of the conditional distribution, our method (Ours Full) generates sharp realizations. While ViewFormer also produces sharp images due to training with a perceptual loss, its renderings fail to transfer some small details, such as headlight shape, from the input.

In Tab. 1, we report the quality of novel renderings produced by our method and baselines, as measured by FID [28], LPIPS [107], DISTS [19], PSNR, and SSIM [94]. As a generative model, our method creates sharp, diverse outputs, which closely match the image distribution; it thus scores more favorably in FID than regression baselines [104, 39], which tend to produce less finely detailed renderings. Our method outperforms baselines in LPIPS and DISTS, which indicates that our method produces novel views that achieve greater structural and textural similarity to the ground truth novel views. We would not expect a generative model to outperform a regression model in PSNR and SSIM, and indeed, renderings from PixelNeRF achieve higher scores in these pixel-wise metrics than realizations from our model. However, we note that the one-step denoised prediction of our model (described in Sec. 3.4) is able to match PixelNeRF's state-of-the-art PSNR and SSIM. While our method with autoregressive conditioning does not surpass 3DiM [96], it achieves competetive scores with a lighter weight model (90M vs 471M params) and fewer diffusion steps (25 vs 512).

In Fig. 5, we demonstrate that for a given observation, our model is capable of producing multiple plausible realizations. When conditioning information is reliable, such as when the query view is close to the input view, ambiguity is low and samples are drawn from a narrow conditional distribution. For more ambiguous inputs, such as when the model is tasked with recreating regions that were occluded in the input image, our model produces plausible realizations with more variation. In contrast, regression-based methods such as PixelNeRF deterministically predict the mean of the conditional distribution and are therefore unable to create high quality realizations when the target view is far from conditioning information and the conditional distribution is large.

Fig. 6 shows that our method can also achieve high geometrical consistency when combined with autoregressive generation as validated by dense point cloud reconstruction and the Chamfer distance to the ground truth.

4.2. Matterport3D

Beyond ShapeNet, we seek to show the effectiveness of our method on the Matterport3D (MP3D) dataset that features building-scale, real-world scans. We use the provided code of [59] to sample trajectories of embodied agents and generate 6,000 videos for training and 200 videos for testing, using the provided 61/18 training and test splits. We train our model by sampling random pairs of input and target images from the same video sequence, where 50% of input views are drawn from within ten frames of the target view and the rest are sampled randomly from the video sequence. The rest of the training procedure is equivalent to the one we use with ShapeNet.

For evaluation, we randomly select an input frame in the

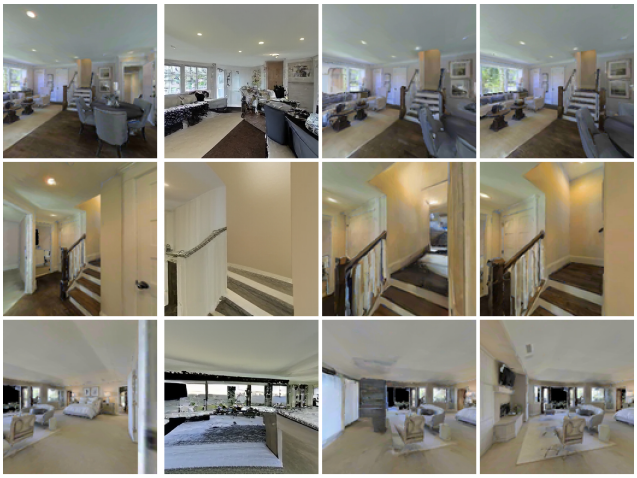

Figure 7. Qualitative comparison on Matterport3D [11] for NVS. Given a single input image (1st col.), we autoregressively run our method and LOTR [59] for 10 frames to synthesize novel view images (2nd and 3rd columns). Ground truth images for the corresponding query camera poses are shown in the fourth column. Best viewed zoomed-in.

	KID↓	LPIPS↓	DISTS↓	PSNR↑	SSIM↑
LOTR [59] (10 f.)	0.050	0.33	0.27	16.57	0.49
Ours (10 f.)	**0.002**	**0.14**	**0.14**	**20.80**	**0.71**
SynSin-6X* [97]	0.072	0.48	0.34	14.89	0.41
GeoGPT* [64]	0.039	0.33	0.27	16.47	0.49
LOTR [59]	0.027	0.25	0.22	18.00	0.55
Ours	**0.002**	**0.09**	**0.11**	**22.79**	**0.79**

Table 2. Quantitative comparison of single-view novel view synthesis on Matterport3D [11]. Here, we use KID since it provides an unbiased estimate when the number of images is small. "10 f." indicates novel view synthesis for 10 frames from the input image (used 5 frames for the bottom rows). *For SynSin and GeoGPT, we obtained the rendered images from the authors of LOTR.

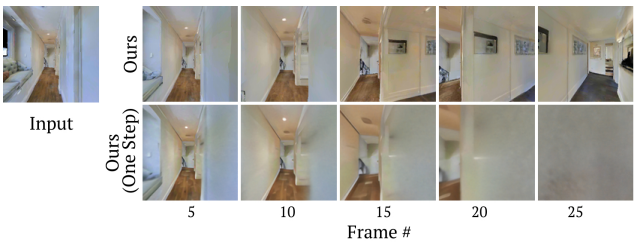

Figure 8. Regression-based models, such as the one-step variant of our approach, struggle to model ambiguity and therefore fail to create plausible renderings far from the input. Generative sampling enables plausible synthesis in ambiguity. When combined with autoregressive generation, we are able to explore areas that were completely occluded in the input.

test video set (one input frame for each test video), and run ten steps of autoregressive synthesis, following the test camera trajectory; we calculate metrics using all ten synthesized frames. Beyond 10 frames, input and the target frusta rarely

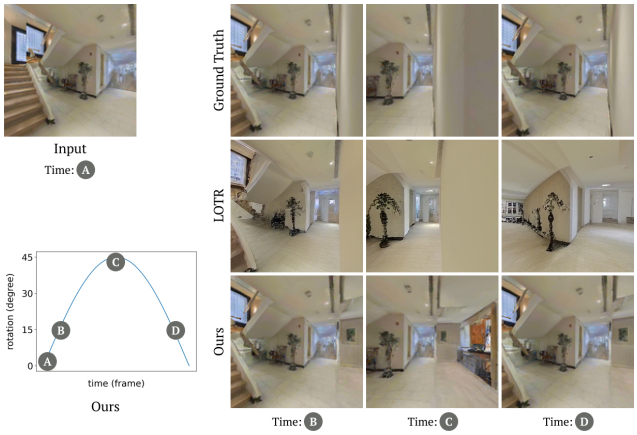

Figure 9. Loop closure test on Matterport3D [11]. We run our method and LOTR [59] on a small cyclic rotation angle trajectory ($0°\rightarrow15°\rightarrow45°\rightarrow15°$). Without 3D representations, transformer-based methods, such as LOTR, rely on interpreting raw camera parameters, resulting in weak spatial awareness. Our 3D feature representation more effectively aggregates past observations and provides better loop closure. Best viewed zoomed-in.

overlap, making comparisons against ground truth frames less meaningful. We compare against Look Outside the Room (LOTR) [59], the current state-of-the-art (SOTA) for single-view NVS on Matterport3D that outperforms prior NVS works (i.e., [97, 62, 64, 40]). We additionally compare against SynSin [97] and GeoGPT [64], using the 5-frame renderings provided by the authors of LOTR. Note that, since the trajectories of embodied agents are randomly sampled, the trajectories used for these two baselines are different from those used for our method and LOTR. This comparison measures performance on 200 random trajectories, which is statistically meaningful and the results align with the trends reported in LOTR. For all baselines, we downsample the outputs to our output resolution, i.e., 128^2, and compute the aforementioned metrics against the ground truth images. To measure the realism of the outputs, we choose KID [6], as it is known to be less biased than FID when the number of test images is small (we use 2000 images).

The results, summarized in Tab. 2, show that our approach generates novel view predictions that outperform baselines in terms of quality and consistency with the input view. Fig. 7 supports the trends observed in the metrics—our NVS is noticeably more accurate and realistic than the current SOTA.

In Fig. 9, we compare against LOTR on a cyclic trajectory. Our method produces better loop closure, indicating higher geometric consistency and showing the effectiveness of incorporating 3D priors. Fig. 6 additionally validates the consistency of our results with superior reconstructed point clouds and Chamfer distances.

4.3. Common Objects in 3D (CO3D)

We challenge our method with real-world scenes from the Common Objects in 3D (CO3D) [58] dataset with complete

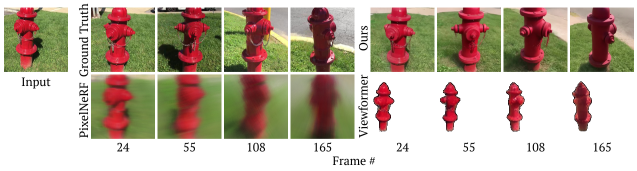

Figure 10. While PixelNeRF produces severe artifacts when the rendering view is far away from the input and ViewFormer requires masks for training on this dataset, our method generates compelling sequences from single-views on challenging, real-world objects of the CO3D dataset [58].

		KID↓	LPIPS↓	DISTS↓	PSNR↑	SSIM↑
	PixelNeRF [104]	0.210	0.705	0.487	16.26	0.271
Ours	One-Step	0.106	0.641	0.492	**16.78**	**0.331**
Ours	Full	**0.012**	**0.369**	**0.446**	15.48	0.266

Table 3. Quantitative comparison of single-view novel view synthesis on CO3D [58].

backgrounds. To our knowledge, no prior method has attempted single-shot NVS on CO3D without object masks. We train our method on the hydrant category of the CO3D dataset, which contains 726 RGB videos of real-world fire hydrants. Most videos contain a walkaround trajectory looking in at the hydrant spanning between 60 and 360 degrees, and most videos consist of about 200 frames. We use a 95:5 train/test split to train our model. CO3D is a highly unconstrained and extraordinarily difficult benchmark: scene scale, camera intrinsics, complex backgrounds, and lighting conditions are highly variable between (and sometimes within) scenes.

Fig. 10 compares predictions from our method against baselines on CO3D. Our method produces plausible and sharp foregrounds and backgrounds that do not deteriorate in quality with increasing distance from the source pose. While we include a qualitative comparison against ViewFormer for reference, we exclude it from numerical comparisons because of its reliance on object masks. Fig. 6 demonstrates the degree of geometric consistency that is attainable by our approach. Tab. 3 additionally provides a quantitative comparison against PixelNeRF. On complex scenes rife with ambiguity, the generative nature of our approach enables synthesis of plausible realizations.

4.4. Ablation Studies

Choice of intermediate representations. Tab. 1 (bottom) compares several choices of intermediate representations within our method. While we have described a specific approach to the task of generative novel view synthesis using diffusion, there is ample freedom to choose how D interprets information from input views. In fact, the simplest approach forgoes any geometry priors and instead directly conditions the model on an input view by concatenation. In our experiments, this *geometry-free* approach struggled compared to variants that incorporated geometry priors. However, greater model capacity and effective use of cross-attention [96] may be key to making this approach work. We additionally compare against an "Explicit" intermediate representation similar to our described approach but without the MLP decoder; while slightly faster, this representation generally produced worse results. We compare to the *one-step* inference mode of our method on ShapeNet in Fig. 4 and Tab. 1, on MP3D in Fig. 8, and on CO3D in Tab. 3. Like regression-based methods, it obtains excellent PSNR and SSIM scores but lacks the ability to generate plausible results far from the input. On Matterport3D, Fig. 8 illustrates the motivation of using a generative prior for long-range synthesis. While the quality of regression-based predictions rapidly degrades with increasing ambiguity, a generative model can create a plausible rendering even in regions with little or no conditioning information, such as behind an occlusion.

Figure 11. *Without* autoregressive conditioning (top), our method generates plausible, albeit geometrically incoherent, novel views conditioned on the input image. *With* autoregressive conditioning (bottom), our method generates plausible sequences that achieve greater geometric consistency between frames.

Effect of autoregressive generation. Although autoregressive conditioning slightly trades off image quality (Tab. 1), Fig. 11 demonstrates the necessity of autoregressive conditioning for generating geometrically consistent multi-view images. Without autoregressive conditioning, independently sampled frames are each plausible, but lack coherence—when conditioning information is ambiguous, e.g., when the model is predicting novel views far from the input view, it samples from a wide conditional distribution and accordingly, subsequent frames exhibit significant variance. Autoregressive conditioning effectively conditions the network not only on the source image, but also on previously generated frames that closely overlap with the current view, helping narrow this conditional distribution.

Additional studies. Additional ablations, including experiments that evaluate out-of-distribution extrapolation, classifier-free guidance, effect of number of input views, stochastic conditioning, and effect of distance to input views, can be found in the supplement.

5. Discussion

Conclusion. We proposed a generative novel view synthesis approach from a single image using geometry-based priors and diffusion models. Our hybrid method combines the benefit of explicit 3D representations with the generative power of diffusion models for generating realistic and

3D-aware novel views, demonstrating the state-of-the-art performance in both object-scale and room-scale scenes. We also demonstrate the compelling results on a challenging real-world dataset of CO3D with background — a challenge never attempted. While our results are not perfect, we believe we presented a significant step towards a practical NVS solution that can operate on a wide range of real-world data.

Limitations and future work. While our method effectively combines explicit geometry priors with 2D diffusion models, the output resolution is currently limited to 128^2 and the diffusion-based sampling is not fast enough for interactive visualization. Since our model can leverage existing 2D diffusion architectures for U, it can directly benefit from future advances in the underlying 2D diffusion models. While our method achieves reasonable geometrical consistency, it can still exhibit minor inconsistencies and drift in challenging real-world datasets, which should be addressed by future work. While our method can operate for novel view synthesis from a single view during inference, training the method requires multi-view supervision with accurate camera poses. In this work, we implemented our method using a 3D feature volume representation. Possible future work includes investigating other types of intermediate 3D representations.

Ethical considerations. Diffusion models could be extended to generate DeepFakes. These pose a societal threat, and we do not condone using our work to generate fake images or videos with the intent of spreading misinformation.

Acknowledgements

We thank David Luebke, Samuli Laine, Tsung-Yi Lin, and Jaakko Lehtinen for feedback on drafts and early discussions. We thank Jonáš Kulhánek and Xuanchi Ren for thoughtful communications and for providing results and data for comparisons. We thank Trevor Chan for help with figures. Koki Nagano and Eric Chan were partially supported by DARPA's Semantic Forensics (SemaFor) contract (HR0011-20-3-0005). JJ park was supported by ARL grant W911NF-21-2-0104. This project was in part supported by Samsung, the Stanford Institute for Human-Centered AI (HAI), and a PECASE from the ARO. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. Distribution Statement "A" (Approved for Public Release, Distribution Unlimited).

References

[1] Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M Seitz, and Richard Szeliski. Building Rome in a day. *Communications of the ACM*, 54(10):105–112, 2011. 2

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

[3] 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 *IEEE International Conference on Computer Vision (ICCV)*, pages 5855–5864, 2021. 1

[4] Alexander W. Bergman, Petr Kellnhofer, Wang Yifan, Eric R. Chan, David B. Lindell, and Gordon Wetzstein. Generative neural articulated radiance fields. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 3

[5] Alexander W. Bergman, Wang Yifan, and Gordon Wetzstein. Articulated 3d head avatar generation using text-to-image diffusion models. In *arXiv preprint arXiv:2307.04859*, 2023. 3

[6] Mikołaj Bińkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations (ICLR)*, 2018. 5, 7

[7] Yukang Cao, Yan-Pei Cao, Kai Han, Ying Shan, and Kwan-Yee K. Wong. Dreamavatar: Text-and-shape guided 3d human avatar generation via diffusion models, 2023. 3

[8] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3D generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16123–16133, 2022. 3, 5

[9] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5799–5809, 2021. 3, 5

[10] Trevor Chan, Nada Kamona, Brian-Tinh Vu, Felix Wehrli, and Chamith Rajapakse. Deep learning super-resolution of mr images of the distal tibia improves image quality and assessment of bone microstructure. 3

[11] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. 1, 5, 7

[12] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. ShapeNet: An information-rich 3D model repository. *arXiv preprint arXiv:1512.03012*, 2015. 1, 5

[13] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. MVSNeRF: Fast generalizable radiance field reconstruction from multi-view stereo. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 2

[14] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 4

[15] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of*

the IEEE/CVF International Conference on Computer Vision (ICCV), October 2023. 3

[16] Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. GRAM: Generative radiance manifolds for 3D-aware image generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[17] Terrance DeVries, Miguel Angel Bautista, Nitish Srivastava, Graham W Taylor, and Joshua M Susskind. Unconstrained scene generation with locally conditioned radiance fields. In *IEEE International Conference on Computer Vision (ICCV)*, pages 14304–14313, 2021. 3

[18] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 3

[19] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 6

[20] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 5

[21] John Flynn, Michael Broxton, Paul Debevec, Matthew DuVall, Graham Fyffe, Ryan Overbeck, Noah Snavely, and Richard Tucker. DeepView: View synthesis with learned gradient descent. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[22] John Flynn, Ivan Neulander, James Philbin, and Noah Snavely. Deep stereo: Learning to predict new views from the world's imagery. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[23] Michael Goesele, Noah Snavely, Brian Curless, Hugues Hoppe, and Steven M Seitz. Multi-view stereo for community photo collections. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1–8. IEEE, 2007. 2

[24] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 3

[25] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. StyleNeRF: A style-based 3D-aware generator for high-resolution image synthesis. *arXiv preprint arXiv:2110.08985*, 2021. 3

[26] Jiatao Gu, Alex Trevithick, Kai-En Lin, Josh Susskind, Christian Theobalt, Lingjie Liu, and Ravi Ramamoorthi. Nerfdiff: Single-image view synthesis with nerf-guided distillation from 3d-aware diffusion. In *International Conference on Machine Learning*, 2023. 3

[27] Amir Hertz, Kfir Aberman, and Daniel Cohen-Or. Delta denoising score. 2023. 3

[28] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Günter Klambauer, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a Nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 5, 6

[29] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 3

[30] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *J. Mach. Learn. Res.*, 23:47–1, 2022. 3

[31] Pavel Iakubovskii. Segmentation models pytorch. https://github.com/qubvel/segmentation_models.pytorch, 2019. 4

[32] Wonbong Jang and Lourdes Agapito. CodeNeRF: Disentangled neural radiance fields for object categories. In *IEEE International Conference on Computer Vision (ICCV)*, pages 12949–12958, 2021. 2

[33] Ruixiang Jiang, Can Wang, Jingbo Zhang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Avatarcraft: Transforming text into neural human avatars with parameterized shape and pose control, 2023. 3

[34] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 3, 4

[35] Subin Kim, Kyungmin Lee, June Suk Choi, Jongheon Jeong, Kihyuk Sohn, and Jinwoo Shin. Collaborative score distillation for consistent visual synthesis. *arXiv preprint arXiv:2307.04787*, 2023. 3

[36] Jaehoon Ko, Kyusun Cho, Daewon Choi, Kwangrok Ryoo, and Seungryong Kim. 3d gan inversion with pose optimization. *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2023. 3

[37] Jing Yu Koh, Honglak Lee, Yinfei Yang, Jason Baldridge, and Peter Anderson. Pathdreamer: A world model for indoor navigation. In *IEEE International Conference on Computer Vision (ICCV)*, pages 14738–14748, 2021. 2

[38] Nikos Kolotouros, Thiemo Alldieck, Andrei Zanfir, Eduard Gabriel Bazavan, Mihai Fieraru, and Cristian Sminchisescu. Dreamhuman: Animatable 3d avatars from text. 2023. 3

[39] Jonáš Kulhánek, Erik Derner, Torsten Sattler, and Robert Babuška. ViewFormer: NeRF-free neural rendering from few images using transformers. In *European Conference on Computer Vision (ECCV)*, 2022. 2, 5, 6

[40] Zihang Lai, Sifei Liu, Alexei A Efros, and Xiaolong Wang. Video autoencoder: self-supervised disentanglement of static 3D structure and motion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9730–9740, 2021. 7

[41] Zhengqi Li, Qianqian Wang, Noah Snavely, and Angjoo Kanazawa. InfiniteNature-Zero: Learning perpetual view generation of natural scenes from single images. In *European Conference on Computer Vision (ECCV)*, pages 515–534. Springer, 2022. 3

[42] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[43] David B Lindell, Dave Van Veen, Jeong Joon Park, and Gordon Wetzstein. BACON: Band-limited coordinate networks for multiscale scene representation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16252–16262, 2022. 1

[44] Andrew Liu, Richard Tucker, Varun Jampani, Ameesh Makadia, Noah Snavely, and Angjoo Kanazawa. Infinite nature: Perpetual view generation of natural scenes from

a single image. In *IEEE International Conference on Computer Vision (ICCV)*, pages 14458–14467, 2021. 2, 3

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

[46] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *ACM Trans. Graph.*, 38(4):65:1–65:14, July 2019. 2

[47] Nelson Max. Optical models for direct volume rendering. *IEEE Transactions on Visualization and Computer Graphics*, 1(2):99–108, 1995. 4

[48] 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. *Communications of the ACM*, 65(1):99–106, 2021. 1, 2, 4

[49] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *arXiv preprint arXiv:2201.05989*, 2022. 1

[50] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. Hologan: Unsupervised learning of 3d representations from natural images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7588–7597, 2019. 3

[51] Michael Niemeyer and Andreas Geiger. GIRAFFE: Representing scenes as compositional generative neural feature fields. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11453–11464, 2021. 2, 3

[52] Simon Niklaus, Long Mai, Jimei Yang, and Feng Liu. 3D Ken Burns effect from a single image. *ACM Transactions on Graphics (ToG)*, 38(6):1–15, 2019. 2

[53] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13503–13513, 2022. 3

[54] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 523–540. Springer, 2020. 3

[55] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. DreamFusion: Text-to-3D using 2D diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 3

[56] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, et al. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. *arXiv preprint arXiv:2306.17843*, 2023. 3

[57] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 2022. 2, 3

[58] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3D: Large-scale learning and evaluation of real-life 3D category reconstruction. In *IEEE International Conference on Computer Vision (ICCV)*, pages 10901–10911, 2021. 1, 5, 7, 8

[59] Xuanchi Ren and Xiaolong Wang. Look outside the room: Synthesizing a consistent long-term 3D scene video from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3563–3573, 2022. 2, 3, 5, 6, 7

[60] Gernot Riegler and Vladlen Koltun. Free view synthesis. In *European Conference on Computer Vision (ECCV)*, 2020. 1, 2

[61] Gernot Riegler and Vladlen Koltun. Stable view synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2

[62] Chris Rockwell, David F Fouhey, and Justin Johnson. PixelSynth: Generating a 3D-consistent experience from a single image. In *IEEE International Conference on Computer Vision (ICCV)*, pages 14104–14113, 2021. 2, 7

[63] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 2, 3

[64] R. Rombach, P. Esser, and B. Ommer. Geometry-free view synthesis: Transformers and no 3D priors. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 2, 5, 7

[65] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM Transactions on Graphics (SIGGRAPH)*, pages 1–10, 2022. 2, 3, 5

[66] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 2, 3

[67] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2, 3, 5

[68] 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 *IEEE International Conference on Computer Vision (ICCV)*, pages 2304–2314, 2019. 2, 4

[69] Mehdi SM Sajjadi, Henning Meyer, Etienne Pot, Urs Bergmann, Klaus Greff, Noha Radwan, Suhani Vora, Mario Lučić, Daniel Duckworth, Alexey Dosovitskiy, et al. Scene representation transformer: Geometry-free novel view synthesis through set-latent scene representations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[70] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4104–4113, 2016. 2

[71] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5

[72] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 5

[73] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. GRAF: Generative radiance fields for 3D-aware image synthesis. *Advances in Neural Information Processing Systems*, 33:20154–20166, 2020. 3, 5

[74] J Ryan Shue, Eric Ryan Chan, Ryan Po, Zachary Ankner, Jiajun Wu, and Gordon Wetzstein. 3d neural field generation using triplane diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20875–20886, 2023. 3

[75] Vincent Sitzmann, Eric Chan, Richard Tucker, Noah Snavely, and Gordon Wetzstein. MetaSDF: Meta-learning signed distance functions. *Advances in Neural Information Processing Systems*, 33:10136–10147, 2020. 2

[76] Vincent Sitzmann, Semon Rezchikov, William T. Freeman, Joshua B. Tenenbaum, and Fredo Durand. Light field networks: Neural scene representations with single-evaluation rendering. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2, 5

[77] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhöfer. DeepVoxels: Learning persistent 3D feature embeddings. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[78] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3D-structure-aware neural scene representations. *Advances in Neural Information Processing Systems*, 32, 2019. 2, 5, 6

[79] Ivan Skorokhodov, Sergey Tulyakov, Yiqun Wang, and Peter Wonka. EpiGRAF: Rethinking training of 3D GANs. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 3

[80] Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3D. In *ACM Transactions on Graphics (SIGGRAPH)*, pages 835–846. 2006. 2

[81] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning (ICML)*, pages 2256–2265. PMLR, 2015. 3

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

[83] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[84] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 3, 4

[85] Matthew Tancik, Ben Mildenhall, Terrance Wang, Divi Schmidt, Pratul P Srinivasan, Jonathan T Barron, and Ren Ng. Learned initializations for optimizing coordinate-based neural representations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2846–2855, 2021. 2

[86] Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior, 2023. 3

[87] Ayush Tewari, Justus Thies, Ben Mildenhall, Pratul Srinivasan, Edgar Tretschk, W Yifan, Christoph Lassner, Vincent Sitzmann, Ricardo Martin-Brualla, Stephen Lombardi, et al. Advances in neural rendering. In *Computer Graphics Forum*, volume 41, pages 703–735. Wiley Online Library, 2022. 1, 2

[88] Alex Trevithick and Bo Yang. GRF: Learning a general radiance field for 3D scene representation and rendering. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 2

[89] Christina Tsalicoglou, Fabian Manhardt, Alessio Tonioni, Michael Niemeyer, and Federico Tombari. Textmesh: Generation of realistic 3d meshes from text prompts, 2023. 3

[90] Hung-Yu Tseng, Qinbo Li, Changil Kim, Suhib Alsisan, Jia-Bin Huang, and Johannes Kopf. Consistent view synthesis with pose-guided diffusion models. In *CVPR*, 2023. 3

[91] Richard Tucker and Noah Snavely. Single-view view synthesis with multiplane images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[92] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. IBRNet: Learning multi-view image-based rendering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[93] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, et al. Rodin: A generative model for sculpting 3d digital avatars using diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4563–4573, 2023. 3

[94] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

[95] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *arXiv preprint arXiv:2305.16213*, 2023. 3

[96] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. *arXiv preprint arXiv:2210.04628*, 2022. 3, 5, 6, 8

[97] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. SynSin: End-to-end view synthesis from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7467–7477, 2020. 2, 5, 7

[98] Jianfeng Xiang, Jiaolong Yang, Yu Deng, and Xin Tong. GRAM-HD: 3D-consistent image generation at high resolution with generative radiance manifolds, 2022. 3

[99] Jianfeng Xiang, Jiaolong Yang, Binbin Huang, and Xin Tong. 3d-aware image generation using 2d diffusion models, 2023. 3

[100] Jiaxin Xie, Hao Ouyang, Jingtan Piao, Chenyang Lei, and Qifeng Chen. High-fidelity 3d gan inversion by pseudo-multi-view optimization. *arXiv preprint arXiv:2211.15662*, 2022. 3

[101] Yinghao Xu, Sida Peng, Ceyuan Yang, Yujun Shen, and Bolei Zhou. 3D-aware image synthesis via learning structural and textural representations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18430–18439, 2022. 3

[102] Yinghao Xu, Wang Yifan, Alexander W. Bergman, Menglei Chai, Bolei Zhou, and Gordon Wetzstein. Efficient 3d

articulated human generation with layered surface volumes. In *arXiv preprint arXiv:2307.05462*, 2023. 3

[103] Yang Xue, Yuheng Li, Krishna Kumar Singh, and Yong Jae Lee. GIRAFFE HD: A high-resolution 3D-aware generative model. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[104] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4578–4587, 2021. 1, 2, 4, 5, 6, 8

[105] Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, and In So Kweon. Text-to-image diffusion models in generative ai: A survey. *arXiv preprint arXiv:2303.07909*, 2023. 3

[106] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European Conference on Computer Vision (ECCV)*, pages 649–666. Springer, 2016. 2

[107] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 5, 6

[108] Xuanmeng Zhang, Zhedong Zheng, Daiheng Gao, Bang Zhang, Pan Pan, and Yi Yang. Multi-view consistent generative adversarial networks for 3D-aware image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[109] Peng Zhou, Lingxi Xie, Bingbing Ni, and Qi Tian. CIPS-3D: A 3D-aware generator of GANs based on conditionally-independent pixel synthesis. *arXiv preprint arXiv:2110.09788*, 2021. 3

[110] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. In *ACM Transactions on Graphics (SIGGRAPH)*, 2018. 2

[111] Zhizhuo Zhou and Shubham Tulsiani. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In *CVPR*, 2023. 3