

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

¹Stanford University ²NVIDIA

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 challenge ourselves to 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 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

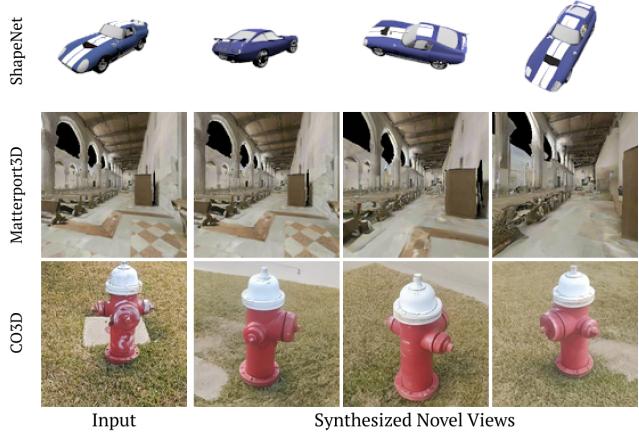


Figure 1. Our 3D-aware diffusion model synthesizes realistic novel views from as little as a single input image. These results are generated with the ShapeNet [10], Matterport3D [9], and Common Objects in 3D [50] datasets.

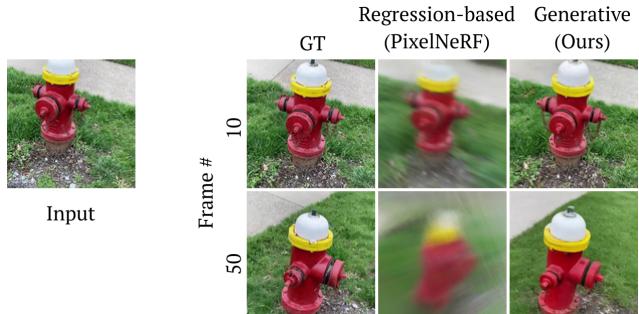


Figure 2. While regression-based models are capable of effective view synthesis near input views (top row), they blur across ambiguity when extrapolating. Generative approaches can continue to sample plausible renderings far from input views (second row, third column).

the boundaries of NVS towards a solution that can operate in a wide range of challenging real-world data.

Previous approaches to few-shot novel view synthesis can broadly be grouped into two categories. Geometry-prior-based methods [53, 52, 41, 37, 42, 3, 89] have drawn from work on scene representations and neural rendering [78].

^{*}Equal contribution.

[†]Work was done during an internship at NVIDIA.

While they achieve impressive results on interpolating near 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 [90, 70, 89]. 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.

In contrast, generative approaches rely on generative priors and solve the novel view synthesis problem by generating random plausible samples from this conditional distribution. Existing generative models for view synthesis [56, 84, 51, 38] 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 [59, 57, 49, 55, 58] 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 [60, 89, 81] 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 [44]. 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:

- We present 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).

- We demonstrate 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.
- We show 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.

We will make the code and pre-trained models available.

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 [72, 63] and running multi-view stereo (MVS) [1, 20]. The recovered explicit geometry proxies enable NVS but fail to synthesize photorealistic and complete novel views especially for occluded regions. Some recent methods [52, 53] 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 [19, 80], multi-plane images [18, 94], or voxels [69, 39] are also used by many recent NVS approaches, as surveyed by Tewari et al. [78].

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 [70, 41]. While many methods are optimized on a per-scene basis with dense input views [41], 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 [70, 29] or meta learning [67, 77], while others lift input observations via encoders [80, 45, 94, 89, 79, 11, 81] and predict novel views in a feed-forward fashion. A recent trend has some NVS methods forgoing geometry priors for light fields [68] or transformers [61, 34], 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 [84, 54, 33], or lack geometry priors altogether [56, 51]. As image-translation-based generative models, they are capable of conditioning on their own previous generations to autoregressively synthesize long camera trajectories, sometimes infinitely [38, 36]. Because the focus is on extrapolating at large scales, these methods ordinarily achieve only approximate view consistency at longer ranges.

3D GANs. 3D GANs [43, 66, 7, 6, 22, 46, 87, 93, 71, 88, 92, 13, 85, 4] combine an adversarial [21] 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 [7, 6]. 3D GANs have been extensively developed to achieve compositionality [44], higher rendering resolution [6, 22, 71], video generation [2], and scalability to larger scenes [14]. GANs, however, are notoriously difficult to train, and their 3D inversions from an input image are often brittle without additional 3D priors [86] or an accurate camera input [32]. Moreover, most 3D GANs assume canonical camera poses and limit their optimal operating ranges to single objects.

2D diffusion models. 2D diffusion models [25, 73, 75, 30] have transformed image synthesis. Favorable properties such as mode coverage and a stable training objective have enabled them to outperform [15] previous generative models [21] 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 [76, 15], text [49, 55, 58] or another image [26, 59, 57, 8].

Recent 3D diffusion works. Recently, DreamFusion [48] and 3DiM [83] 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 [51]) 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 [83]), 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.

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.

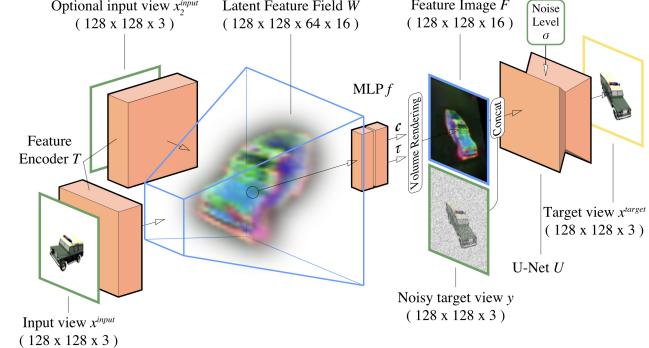


Figure 3. Illustration of our framework D . The pipeline receives as input one or more input views \mathbf{x} and the camera parameters associated with input and target views. We extract features from each input view \mathbf{x} using T and unproject them into a feature volume \mathbf{W} . These volumes 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 $\mathbf{x}^{\text{target}}$ 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 y and noise level σ , and produces a denoised image of the target view $\mathbf{x}^{\text{target}}$.

In novel view synthesis, we are given a set of input images $\mathbf{x}^{\text{inputs}}$ 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:

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

3.1. 3D-aware diffusion model architecture

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

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 \mathbf{y} , conditioning information ($\mathbf{x}^{\text{inputs}}, \mathbf{P}^{\text{inputs}}, \mathbf{P}^{\text{target}}$) and a noise level σ . Our strategy builds upon pixel-aligned implicit functions [60, 89] and neural rendering. Following Fig. 3, given a single input image \mathbf{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 \mathbf{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 \mathbf{W} with trilinear interpolation and decode the obtained feature $w = \mathbf{W}(\mathbf{r})$ with a small multi-layer perceptron (MLP) f to obtain a density τ and a feature vector \mathbf{c}

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

By projecting this feature field into the target view using volume rendering [40, 41], we obtain a feature image F in Fig. 3:

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

In practice, we employ the image segmentation architecture *DeepLabV3+* [12, 28] 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* [41]. We use input/output image resolution 128^2 in all experiments.

The feature image F is concatenated to the noisy image \mathbf{y} and passed as input to a denoiser network U to produce the final target view $\mathbf{x}^{\text{target}}$ (see Fig. 3). We use *DDPM++* [76, 30] for U , where

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

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 $w_j = \mathbf{W}_j(\mathbf{r})$ obtained for each input image \mathbf{x}_j , as in [89]:

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

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

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

where σ is sampled during training according to the strategy proposed by *EDM* [30]. 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 [30] 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 2nd order sampling strategy [30], with 25 or fewer denoising steps. Other sampling strategies [76, 74] 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 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.



Figure 4. Qualitative comparison on ShapeNet [10] 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.

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 [10] ‘‘cars’’ and Matterport3D [9], 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., [89, 68]). Meanwhile, long-range NVS on Matterport3D is prototypical of unbounded scene exploration, where generative models with weak geometry priors [84, 56, 51] have seen more success. Finally, we stress-test our method on the challenging Common Objects in 3D (CO3D) [50], 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 [89], a state-of-the-art NeRF-based method for NVS, and ViewFormer [34], a transformer-based, geometry-free approach to NVS. For ShapeNet, we additionally provide a comparison with EG3D-PTI [6], which is based on a state-of-the-art 3D GAN for object-scale scenes, and a numerical comparison with 3DiM [83], 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 [51], a transformer-based, geometry-free NVS method designed for room-scale scenes, and to additional SOTA methods, including SynSin [84] and GeoGPT [56] in Tab. 2.

	FID \downarrow	LPIPS \downarrow	DISTS \downarrow	PSNR \uparrow	SSIM \uparrow
PixelNeRF [89]	65.83	0.146	0.203	23.2	0.90
ViewFormer [34]	20.82	0.146	0.161	19.0	0.83
EG3D-PTI [6]	27.23	0.150	0.310	19.0	0.85
3DiM (autoregressive) [83] †	8.99			21.01	0.57
Ours					
Explicit	8.09	0.129	0.158	19.1	0.86
Geom.-Free	16.68	0.342	0.329	13.1	0.74
One-Step	42.07	0.150	0.178	23.2	0.91
Full (autoregressive)	11.08	0.120	0.146	20.6	0.89
Full	6.47	0.104	0.145	20.7	0.89

Table 1. Quantitative comparison of single-view novel view synthesis on ShapeNet cars [10, 70]. † As reported by [83].

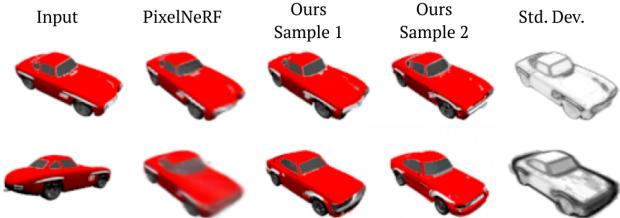


Figure 5. Generating new views from more (bottom) or less (top) ambiguous conditioning information. PixelNeRF [89] is constrained to output deterministic novel views and renders an average of all plausible renderings that are consistent with the input view. In comparison, our method samples the conditional 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.

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 [24] and KID [5], which are commonly used to evaluate generative models for image synthesis. For (2), we use perceptual metrics LPIPS [91] and DISTS [17], 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 [59, 57]. For (3), we provide COLMAP [64, 65] reconstructions of generated video sequences, a standard evaluation for 3D consistency in 3D GANs [66, 7, 6]. 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.

4.1. ShapeNet

We standardize our training and evaluation on the single-class, single-view NVS benchmark described in [89, 70, 34]. The ShapeNet training set contains 2,458 cars, each with 50

	Input Image	Ground Truth	Ours	PixelNeRF	Viewformer
ShapeNet					
		Chamfer Distance to Ground Truth ($\times 10^{-4}$) ↓	1.555	2.500	18.887
Matterport3D					
		Chamfer Distance to Ground Truth ↓	0.0389	0.243	
CO3D					
		Chamfer Distance to Ground Truth ($\times 10^{-4}$) ↓	3.953	25.952	

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.

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 [24], LPIPS [91], DISTs [16], PSNR, and SSIM [82]. 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 [89, 34], 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 [83],



Figure 7. Qualitative comparison on Matterport3D [9] for NVS. Given a single input image (1st col.), we autoregressively run our method and LOTR [51] 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.

it achieves competitive 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 [51] 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.

	KID \downarrow	LPIPS \downarrow	DISTS \downarrow	PSNR \uparrow	SSIM \uparrow
LOTR [51] (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* [84]	0.072	0.48	0.34	14.89	0.41
GeoGPT* [56]	0.039	0.33	0.27	16.47	0.49
LOTR [51]	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 [9]. 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.



Figure 9. Loop closure test on Matterport3D [9]. We run our method and LOTR [51] on a small cyclic rotation angle trajectory ($0^\circ \rightarrow 15^\circ \rightarrow 45^\circ \rightarrow 15^\circ$). 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.

For evaluation, we randomly select an input frame in the 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 overlap, making comparisons against ground truth frames

less meaningful. We compare against Look Outside the Room (LOTR) [51], the current state-of-the-art (SOTA) for single-view NVS on Matterport3D that outperforms prior NVS works (i.e., [84, 54, 56, 35]). We additionally compare against SynSin [84] and GeoGPT [56], 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 [5], 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) [50] dataset with complete 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.

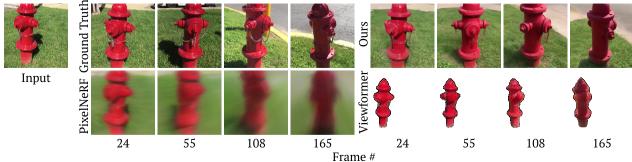


Figure 10. While PixelNeRF produces severe artifacts when the rendering view is far away from the input and ViewFormer requires masks for training on this dataset, our method generates compelling sequences from single-views on challenging, real-world objects of the CO3D dataset [50].

	KID \downarrow	LPIPS \downarrow	DISTS \downarrow	PSNR \uparrow	SSIM \uparrow
PixelNeRF [89]	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 [50].

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 [83] 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.

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



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.

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 Jónáš 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] Mikołaj Bińkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations (ICLR)*, 2018. [5](#), [7](#)
- [6] 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](#)
- [7] 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](#)
- [8] 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](#)
- [9] 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](#), [6](#), [7](#), [19](#)
- [10] 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](#), [16](#), [17](#), [19](#)
- [11] 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](#)
- [12] 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](#), [17](#)
- [13] 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](#)
- [14] 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](#)
- [15] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. [3](#), [20](#)
- [16] 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](#)
- [17] 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](#)
- [18] 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](#)
- [19] 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](#)
- [20] 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](#)

- [21] 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](#), [20](#)
- [22] 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](#)
- [23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. [17](#)
- [24] 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](#)
- [25] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. [3](#)
- [26] 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](#)
- [27] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. [14](#)
- [28] Pavel Iakubovskii. Segmentation models pytorch. https://github.com/qubvel/segmentation_models.pytorch, 2019. [4](#), [17](#)
- [29] 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](#)
- [30] 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](#), [17](#), [18](#)
- [31] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *Proc. NeurIPS*, 2021. [18](#)
- [32] 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](#)
- [33] Jing Yu Koh, Honglak Lee, Yinfai 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. [3](#)
- [34] 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](#), [17](#), [18](#)
- [35] 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](#)
- [36] 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](#)
- [37] 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](#)
- [38] 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](#)
- [39] 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](#)
- [40] Nelson Max. Optical models for direct volume rendering. *IEEE Transactions on Visualization and Computer Graphics*, 1(2):99–108, 1995. [4](#)
- [41] 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](#), [17](#)
- [42] 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](#)
- [43] 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](#)
- [44] 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](#), [17](#)
- [45] 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](#)
- [46] 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](#)
- [47] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *CVPR*, 2022. [18](#)
- [48] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. DreamFusion: Text-to-3D using 2D diffusion. *arXiv preprint arXiv:2209.14988*, 2022. [3](#)
- [49] 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](#)
- [50] 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, 16, 19
- [51] 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, 19
- [52] Gernot Riegler and Vladlen Koltun. Free view synthesis. In *European Conference on Computer Vision (ECCV)*, 2020. 1, 2
- [53] Gernot Riegler and Vladlen Koltun. Stable view synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2
- [54] 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. 3, 7, 19
- [55] 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
- [56] 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, 3, 5, 7, 19
- [57] 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
- [58] 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
- [59] 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
- [60] 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, 3
- [61] 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, 20
- [62] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9339–9347, 2019. 19
- [63] Johannes L Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4104–4113, 2016. 2
- [64] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5
- [65] 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
- [66] 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
- [67] 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
- [68] Vincent Sitzmann, Semon Rezhikov, 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, 20
- [69] 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
- [70] 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, 18
- [71] Ivan Skorokhodov, Sergey Tulyakov, Yiqun Wang, and Peter Wonka. EpiGRAF: Rethinking training of 3D GANs. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 3
- [72] 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
- [73] 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
- [74] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 4
- [75] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019. 3
- [76] 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, 17
- [77] 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
- [78] 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

- [79] 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
- [80] 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
- [81] 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
- [82] 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
- [83] 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, 15, 17
- [84] 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, 3, 5, 7, 19
- [85] Jianfeng Xiang, Jiaolong Yang, Yu Deng, and Xin Tong. GRAM-HD: 3D-consistent image generation at high resolution with generative radiance manifolds, 2022. 3
- [86] Jiaxin Xie, Hao Ouyang, Jingtian Piao, Chenyang Lei, and Qifeng Chen. High-fidelity 3d gan inversion by pseudo-multi-view optimization. *arXiv preprint arXiv:2211.15662*, 2022. 3
- [87] 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
- [88] 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
- [89] 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, 3, 4, 5, 6, 8, 17, 18, 19
- [90] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European Conference on Computer Vision (ECCV)*, pages 649–666. Springer, 2016. 2
- [91] 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, 13
- [92] 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
- [93] 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
- [94] 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, 19

Supplementary Material

In this supplement, we first provide additional experiments (Sec. A). We follow with details of our implementation (Sec. B), including further descriptions of the model architecture and training process, as well as hyperparameters. We then discuss experimental details (Sec. C). Lastly, we consider artifacts and limitations (Sec. D) that may be targets for future work. We encourage readers to view the accompanying supplemental videos, which contain additional visual results.

A. Additional experiments & ablations



Figure 12. New views generated from out-of-distribution poses. Extreme zooms and large translations may lead to unrealistic views.

A.1. Extrapolation to unseen camera poses.

In the ShapeNet dataset, cameras are located on a sphere, point towards the centers of the objects and have the same “up” direction during training. We investigate the results of our method when querying out-of-distribution poses at test time in Fig. 12. From a fixed pose, we generate a zoom, a one-dimensional translation of the camera, and a camera roll. Although novel views deteriorate with large deviations from the training pose distribution, the 3D prior present in our method can reasonably tolerate small extrapolations.

A.2. Percentile results based on LPIPS

Fig. 13 shows our synthesized results on ShapeNet ordered by the percentile of the LPIPS [91] score, with examples that scored best according to the metric at the top and examples that scored worst at the bottom. We compute predictions for the same input and output views across the entire test set. To reduce the effects of randomness, we evaluate 9 realizations for each input, and use only the median image/score when ordering our results. Our method produces consistently sharp outputs (even at the 10th percentile) and maintains overall textures and shapes from the input image.

A.3. Handling multiple input images

Fig. 14 shows our generated novel views when more than one image is given as the input conditioning information. When only 1 view is given from the back side of the car,

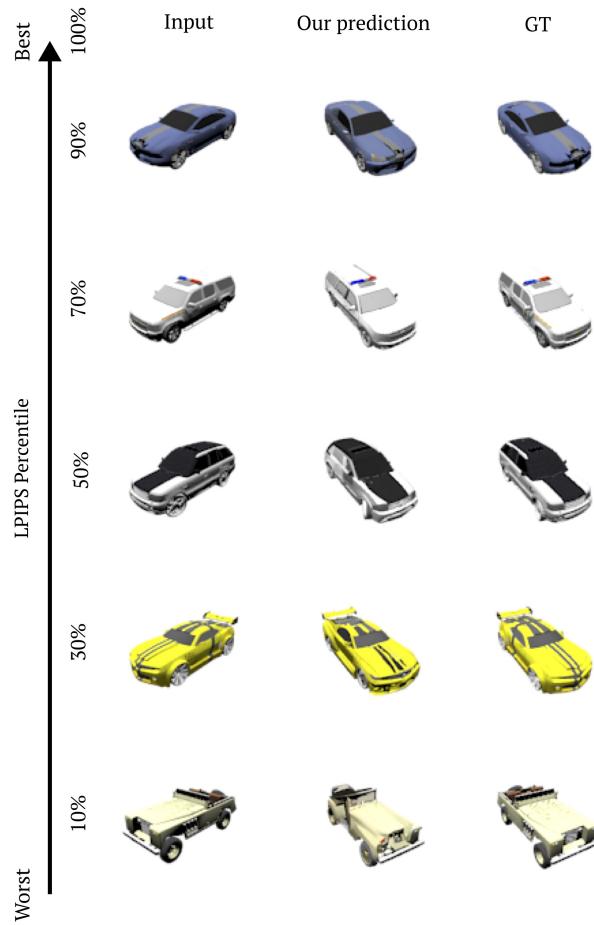


Figure 13. Our synthesized novel views sorted by the percentile of the LPIPS [91] score, with results that scored best according to LPIPS at the top.

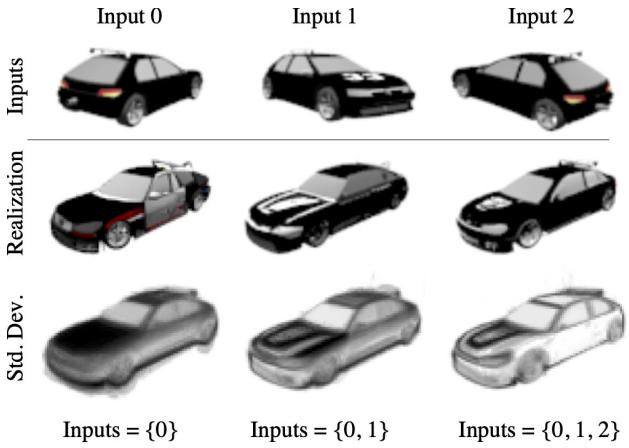


Figure 14. Effect of varying the number of input views. Increasing the number of input views reduces uncertainty, decreasing the pixel-wise standard deviation in novel renderings. Dark pixels in the third row represent higher standard deviation and indicate greater variation in the realizations.

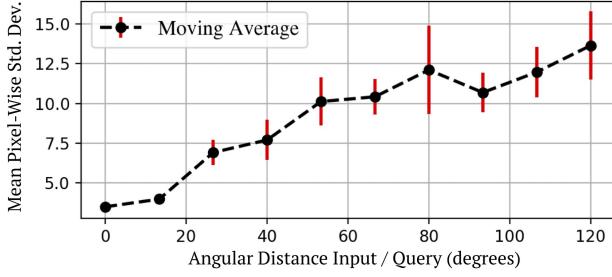


Figure 15. Average pixel variance of generated views vs. the distance between the query camera and the input camera. Input views close to the camera are valuable—the model can directly observe many of the details it must transfer to the output rendering. Input views distant to the camera are more ambiguous—the model is tasked with generating large parts of the rendering from scratch. As the conditioning information gets increasingly ambiguous, novel views get increasingly diverse. Pixel variance is calculated across 50 renderings per pose. Red bars indicate the empirical standard deviation of the moving average.

the model has the freedom to choose multiple plausible completions for the unseen front side of the car, leading to a high standard deviation (high uncertainty). Adding 2 or 3 views reduces uncertainty (low standard deviation), and the model generates a novel view that is compatible with multiple input views.

A.4. Effect of distance to input view

As Fig. 15 demonstrates, nearby views provide more valuable information than distant views, thus reducing variance in the output rendering. Consequently, by conditioning autoregressively on nearby views, we narrow the conditional distribution of possible outputs, improving geometric consistency compared to non-autoregressive conditioning.

A.5. Classifier-free guidance

Recently, [27] suggested classifier-free diffusion guidance technique to effectively trade off diversity and sample quality. At training, we implement classifier-free guidance by dropping out the feature image with 10% probability; in its place, we replace this conditioning image with a sample of Gaussian noise. At inference, we can linearly interpolate between unconditional and unconditional predictions of the denoised image in order to boost or decrease the effect of the conditioning information.

Fig. ?? shows the effect of classifier-free guidance [27] (CFG) when making predictions in isolation. In general, positive classifier-free guidance increases the effect of the conditioning information and improves sample quality. With guidance = 0, our model produces greater variation of generated views (note the different realizations of the passenger-side door). However, we would consider some of these realizations to be unlikely given the input. Increased CFG strength narrows the distribution of possible outputs, and while we would consider such a set of realizations to be

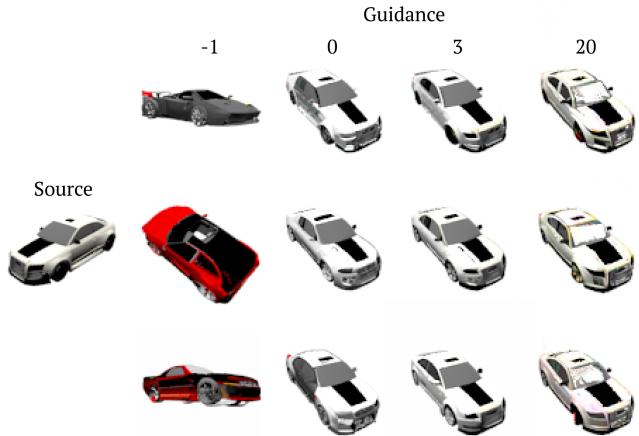


Figure 16. Independent (single-frame input) NVS with various classifier-free guidance (CFG) strengths. For each level of CFG, we show three realizations. With guidance = 0, we sample a “diverse” set of novel views, each plausible, but with variations (e.g. doors). Higher guidance strength reduces diversity but improves sample quality. Excessively high guidance begins to introduce saturation and visual artifacts. Negative guidance upweights the unconditional contribution; with guidance = −1, generation is unconditional.

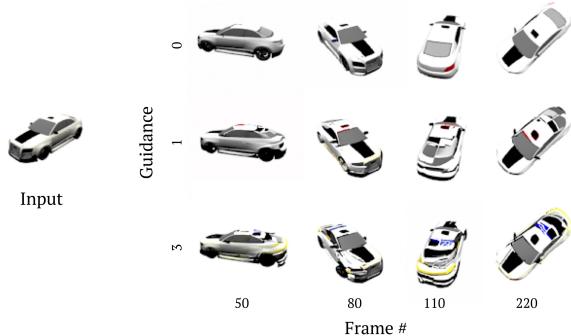


Figure 17. Autoregressive sequence generation with varying CFG strength. With low guidance, we can generate extended autoregressive sequences with little deterioration over time. Higher guidance tends to carry over errors from previous frames, which gradually degrades the quality of subsequent generations.

less diverse, each one is of high fidelity. Excessively high guidance strength begins to introduce artifacts and color saturation. Negative guidance upweights unconditional prediction; guidance = −1 produces unconditional samples without influence from the input image. In general, when making independent novel view predictions, we find moderate levels of CFG to be beneficial. However, as described in Sec. A.6, CFG has an adverse effect on the quality of autoregressively generated sequences. As a default, we refrain from using CFG in our experiments.

A.6. Extended autoregressive generation

Fig. 17 shows autoregressively generated sequences made with varying levels of classifier-free guidance. When making

long autoregressive sequences, the ability to suppress errors and return to the image manifold is an important attribute. Unchecked, gradual accumulation of errors could lead to progressive deterioration in image quality. Intuitively, unconditional samples do not suffer from error buildup, since unconditional ($\text{CFG} = -1$) samples make use of *no* information from previous frames. On the other end of the spectrum, highly conditioned ($\text{CFG} >> 0$) samples should be more likely to suffer from error accumulation because they *emphasize* information from previous frames. A happy medium between these two extremes allows the model to use information from previous frames while preventing undesired error accumulation. Empirically, we find that while small positive guidance can reduce frame-to-frame flicker, it enhances the model’s tendency to carry over visual errors from previous frames. We observe saturation buildup and artifact accumulation to be significant roadblocks to using CFG when synthesizing long video sequences. For these reasons, we default to using $\text{CFG} = 0$, which we found to enable autoregressive generation of long sequences without significant error accumulation. A solution that enables higher CFG weights for autoregressive generation may make a valuable contribution in the future.

A.7. Alternative autoregressive conditioning schemes

Baseline strategy When generating a sequence autoregressively, there are many possible strategies, each with a set of tradeoffs. To produce the visual results presented in our work, we used the following baseline strategy, with minor variations for different datasets. As described in the main paper, our baseline strategy is to condition our model on the input image(s), the most recently generated rendering, and five previously generated images, selected at random.

For Matterport3D, when generating long sequences, we select the five previously generated frames from a set of only the 20 most recently generated frames; we additionally condition on every 15th previously generated frame.

For CO3D, we use the two-pass conditioning method discussed below to improve temporal consistency.

Alternative strategies and tradeoffs As described in the main paper, our baseline autoregressive strategy can induce noticeable flickering. One way to reduce flickering is to condition on *only* the previous frame. Doing so almost completely eliminates frame-to-frame flicker. However, this strategy sacrifices long-term consistency and does little to prevent drift; new renderings might not be consistent with frames rendered at the start of the sequence. By contrast, to promote long-term consistency, one could avoid conditioning on previously-generated frames at all and instead condition on only the input image(s). Because drift is the result of error accumulation from conditioning on previous generations, this strategy eliminates potential for drift. However, it suffers from short-term inconsistency (i.e. frame-to-frame flicker). We found our baseline strategy, which conditions on the inputs, the most re-

cent rendering, and several previous renderings, to be a good compromise between long-term and short-term consistency. The number of previously generated images we condition upon affects the behavior. Because we equally weight the contribution of all images we condition upon, increasing the number of previous renderings (which are sampled uniformly from the generated sequence) reduces the relative contribution of the most recent rendering. Increasing the size of this “buffer” of previously-generated conditioning images thus improves long-term consistency at the cost of short-term consistency; reducing the size of the buffer has the opposite effect.

One way to suppress flickering is to generate frames in two passes, where in the second pass, we condition on the nearby frames from the first pass in a sliding window fashion. Empirically, conditioning on only the nearest 4 frames during the second pass results in videos with reduced flicker, at the expense of higher inference computation. However, unless otherwise noted, we render all videos shown with our baseline autoregressive strategy, i.e. *without* these alternative methods.

A.8. Stochastic Conditioning

To demonstrate the effectiveness of our autoregressive synthesis method, which aggregates conditioning feature volumes from autoregressively selected generated images, we compare to an adaptation of the stochastic conditioning method proposed in 3DiM [83]. We adapt the stochastic conditioning method to our architecture by replacing the feature volume aggregation from autoregressively selected generated images with a single feature volume generated from an image randomly sampled from all previously generated images. As done in 3DiM, the number of diffusion denoising steps is increased significantly and the randomly sampled image is varied at each individual step of denoising. Each generated final image is then added to the set of all previous images and can be used as conditioning in subsequent view generations. This alternative form of conditioning is also able to provide the model with information from many generated views, but they are processed independently with each step of denoising, rather than together after a feature volume aggregation.

In Fig. 20, we show 3D reconstruction results from sequences of images generated by our autoregressive synthesis method and with our adaptation of stochastic conditioning [83]. Here, we find that our autoregressive synthesis method performs slightly better than stochastic conditioning in terms of 3D consistency of generated frames as seen by the COLMAP 3D reconstruction and corresponding Chamfer distance. Additionally, we are able to generate novel views significantly faster – in practice, stochastic conditioning requires 256 denoising steps to generate each novel view while our method only requires 25, leading to a 10 \times improvement in speed.

A.9. Additional Common Objects in 3D results

We provide additional results for single-view novel view synthesis (NVS) with real-world objects for CO3D *Hydrants*

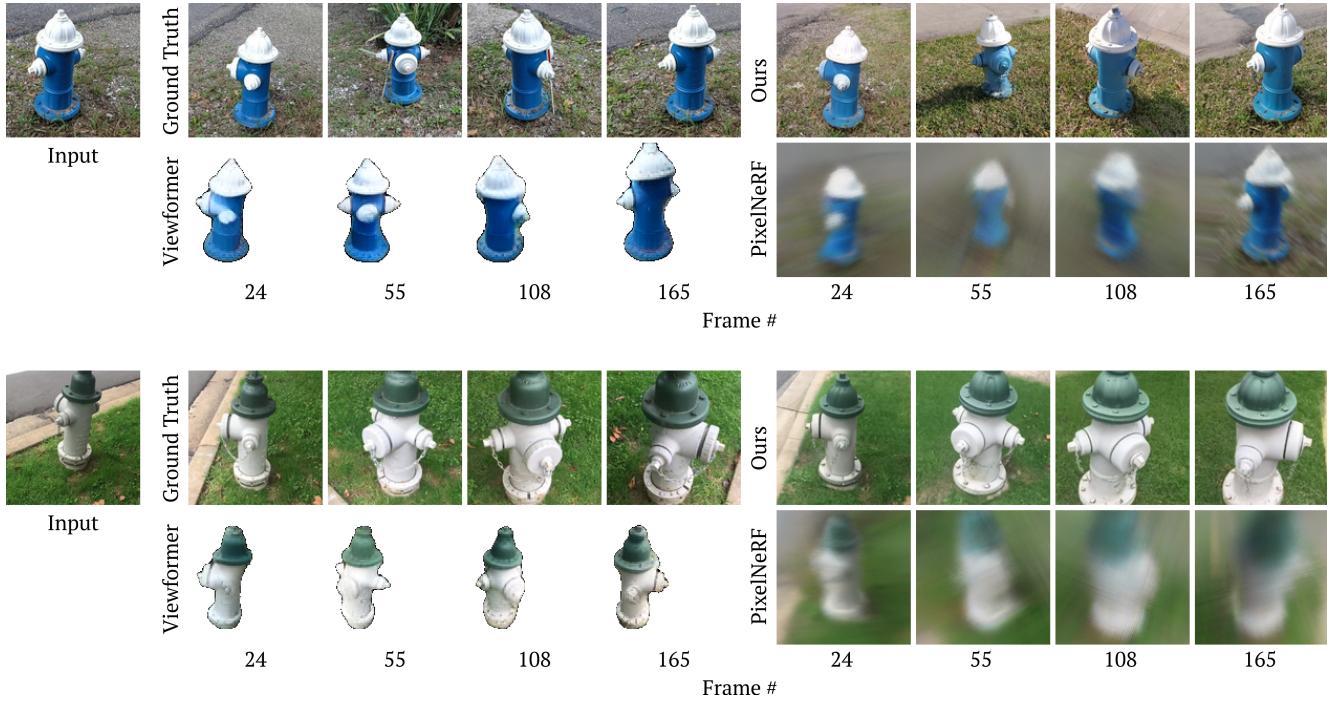


Figure 18. Qualitative comparison for single-view novel view synthesis on CO3D [50] Hydrants.

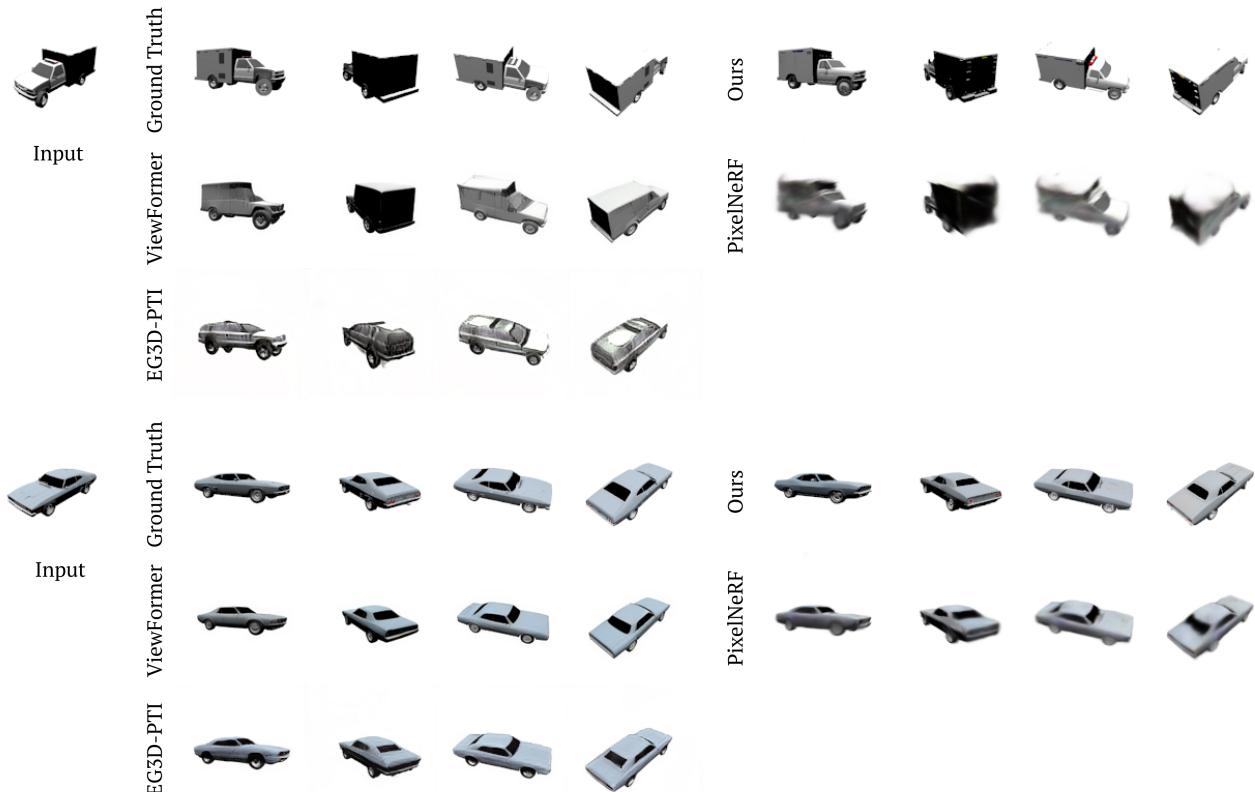


Figure 19. Additional qualitative comparisons against baselines on ShapeNet [10].

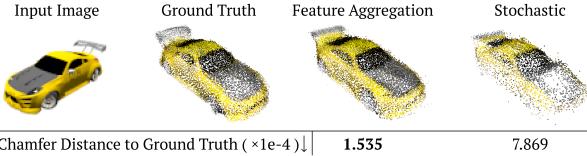


Figure 20. Our default autoregressive conditioning strategy, which aggregates information from multiple views within a feature volume, typically performs at least on par with stochastic view conditioning [83] in geometric consistency, but requires many fewer steps of diffusion to remain effective. Here, we compare COLMAP reconstructions of a sequenced produced by feature aggregation, using 25 steps of denoising, against a sequence produced by stochastic conditioning, using 256 steps of denoising.

in Fig. 18. We compare against ViewFormer [34], which has demonstrated success in few-shot NVS on CO3D, and PixelNeRF [89]. However, we note that ViewFormer is not a 1:1 comparison for two reasons: 1. ViewFormer operates with object masks, whereas our method operates with backgrounds. 2. ViewFormer train/test splits did not align with other methods. For this figure, and for comparison videos, we selected objects that were contained in our *test* split but were part of ViewFormer’s *train* split. Despite these disadvantages, our method demonstrates a compelling ability to plausibly complete complex scenes.

A.10. Additional ShapeNet results

Fig. 19 provides additional visual comparisons on the ShapeNet [10] dataset against baselines. In general, our method renders images with sharper details and higher perceived quality than PixelNeRF, while better transferring details from the input image than ViewFormer and EG3D. In this figure, renderings from our method are selected from autoregressively-generated sequences.

B. Implementation details

We implemented our 3D-aware diffusion models using the official source code of EDM [30], which is available at <https://github.com/NVlabs/edm>. Most of our training setup and hyperparameters follow [30]; the exceptions are detailed here.

Feature volume encoder, T . Our encoder backbone is based on DeepLabV3+ [12]. We use a Pytorch reimplementation [28] available at https://github.com/qubvel/segmentation_models.pytorch, and ResNet34 [23] as the encoder backbone. We found unmodified DeepLabV3+, to struggle because the output branch contains several unlearned, bilinear upsampling layers; this resolution bottleneck makes it difficult to effectively reconstruct fine details from the input. We replace these unlearned upsampling layers with learnable convolutional layers and skip connections from previous layers. We disable batchnorm and dropout throughout the feature volume encoder. The

feature volume encoder expects as input a $3 \times 128 \times 128$ image; it produces a $(16 \times 64) \times 128 \times 128$ feature image, which we reshape into a $16 \times 64 \times 128 \times 128$ volume.

Multiview aggregation. We aggregate information from multiple input views by predicting a feature volume W_i for each input image independently, projecting the query point into each feature volume, sampling a separate feature vector from each feature volume, and mean-pooling across the sampled feature vectors to produce a single aggregated feature. We experimented with two alternative aggregation strategies: 1. max-pooling, and 2. weighted average pooling, where the feature volumes have an additional channel that is interpreted as a weight by a softmax function. We found these alternative aggregation strategies to perform similarly to mean-pooling.

MLP, f . We use a two-layer ReLU MLP to aggregate features drawn from multiple input images. Our MLP has an input dimension of 16, two hidden layers of dimension 64, and an output dimension of 17, which is interpreted as a 1-channel density τ and a 16-channel feature c . We additionally skip the MLP’s input feature to the output feature.

Rendering. We render feature images from the model using neural volume rendering [41] of features [44], from the neural field parameterized by the set of feature volumes W and the MLP f . For computational efficiency, we render at half spatial resolution, i.e. 64×64 and use bilinear upsampling to produce a 128×128 feature image. We use 64 depth samples by default, scattered along each ray with stratified sampling. We do not use importance sampling.

UNet, U . The design of U is based on DDPM++ [76], using the implementation and preconditioning scheme of [30]. U accepts as input 19 total channels (a noisy RGB image, plus a 16-channels feature rendering) of spatial dimension 128^2 . It produces a 3-channel 128^2 denoised rendering. For experiments shown in the manuscript, our models contain five downsampling blocks with channel multipliers of [128, 128, 256, 256, 256]. As in [76], we utilize a residual skip connection from the input to U to each block in the encoder of U .

Training. We use a batch size of 96 for all training runs, split across 8 A100 GPUs, with a learning rate of 2×10^{-5} . During training, we sample the noise level σ according to the method proposed by [30] by drawing σ from the following distribution:

$$\log(\sigma) \sim \mathcal{N}(P_{\text{mean}}, P_{\text{std}}^2). \quad (7)$$

We use $P_{\text{mean}} = -1.0$, $P_{\text{std}} = 1.4$. During training, we randomly drop out the conditioning information with a probability 0.1 to enable classifier-free guidance. In place of the rendered feature image, we insert random noise.

Our dataset is composed of posed multi-view images, where for each training image, we are given the 4×4 camera pose matrix, the camera field of view, and a near/far plane. For all experiments, we specify a global near/far value for each dataset, where the values are chosen such that a camera frustum with the chosen near/far planes adequately covers the visible portion of the scene. For ShapeNet, near/far = (0.8, 1.8); for MP3D, near/far = (0., 12.5); for CO3D, near/far = (0.5, 40). We found our method to be fairly robust to the chosen values of near/far planes.

For ShapeNet, we train until the model has processed 140M images, which takes approximately 9 days on eight A100 GPUs. For MP3D, we train for 110M images, which takes approximately 7 days on eight A100 GPUs. For CO3D, we train for 170M images, which takes approximately eleven days on eight A100 GPUs.

Augmentation. During training, we introduce two forms of augmentation. First, with probability 0.5, we add Gaussian white noise to the input images. For input images in the range $[-1, 1]$, we sample the standard deviation of the added noise uniformly from $[0, 0.5]$. Second, we apply non-leaking augmentation [30] to U . With probability 0.1, we apply random flips, random integer translations (up to 16 pixels), and random 90° rotations, where the transformations are applied to the input noisy image, the input feature image, and the target denoised image. We condition U with a vector that informs it of the currently applied augmentations; we zero this vector at inference.

Inference. We use the deterministic second order sampler proposed in [30] at inference. As a default, we use $N = 25$ timesteps, with a noise schedule governed by $\sigma_{\max} = 80$, $\sigma_{\min} = 0.002$, and $\rho = 7$, where ρ is a constant that controls the spacing of noise levels. The noise level at a timestep i is given in Eq. 8:

$$\sigma_{i < N} = \left(\sigma_{\max}^{\frac{1}{\rho}} + \frac{i}{N-1} \left(\sigma_{\min}^{\frac{1}{\rho}} - \sigma_{\max}^{\frac{1}{\rho}} \right) \right). \quad (8)$$

Rendering an image from scratch with 25 denoising steps takes approximately 1.8 seconds per image at inference on an RTX 3090 GPU.

“Production” settings for CO3D. For rendering videos of CO3D, we use more computationally expensive “production” hyperparameters to obtain better image quality. Seeking better image quality and detail, we use 256 denoising steps instead of the default 25 denoising steps. Seeking better temporal consistency, we increase the number of samples per ray cast through the latent feature field, from 64 to 128; we also use the two-pass form of autoregressive conditioning described in Sec. A.7.

C. Experiment details

C.1. Evaluation details

FID Calculation. We compute FID by sampling 30,000 images randomly from both the ground truth testing dataset and corresponding generated frames. We use an inception network provided in the StyleGAN3 [31] repository for computing image features.

KID Calculation. We compute KID by sampling all images from both the ground truth testing dataset and corresponding generated frames. We use the implementation of *clean-fid* [47], available at <https://github.com/GaParmar/clean-fid>.

COLMAP Reconstructions. We compute COLMAP reconstructions using frames from rendered video sequences. We provide the ground-truth camera pose trajectory as input for all reconstructions. For ShapeNet evaluations, we additionally compute masks by thresholding images to remove white pixels. We leave all settings at their recommended default.

Chamfer Distance Calculation. For all datasets, we compute the bi-directional Chamfer distance between the reconstructed point cloud from synthesized images to the reconstructed point cloud from ground truth images. Additionally, for CO3D, we translate and scale the reconstructed point clouds to lie within the unit cube.

C.2. Baselines

PixelNeRF [89]. We compare to PixelNeRF for the ShapeNet and CO3D single-image novel view synthesis benchmark. For ShapeNet, we use the official implementation and pre-trained weights for single-category (car), single-image, ShapeNet novel view synthesis evaluation provided at: <https://github.com/sxyu/pixel-nerf>. We follow the protocol described in the original PixelNeRF paper and SRNs [70] for data pre-processing. We use the provided dataset and splits in the PixelNeRF repository for training and testing of both our method and PixelNeRF (this dataset is slightly different from that used in the SRNs paper due to a bug; see PixelNeRF supplementary information). We follow the same protocol for evaluation as we do for our method and SRNs: view 64 is used as input, and the remaining 249 views are synthesized conditioned on this. For CO3D, we train PixelNeRF from scratch using our train/test splits and using the recommended hyperparameters.

ViewFormer [34]. We compare to ViewFormer on the ShapeNet single-image novel view synthesis benchmark and qualitatively on single-image novel view synthesis for CO3D. We received the data and results for single-image novel view synthesis for the entire ShapeNet testing set

from the authors. We compute metrics using their provided ground truth data and synthesized results. The training and testing splits are the same as those used in our method and in PixelNeRF. They use the previously introduced protocol for single-image novel view synthesis evaluation: view 64 is used as input, and the remaining 249 views are synthesized conditioned on this. For CO3D, we instead condition on the first frame from each shown sequence, and generate a video based on this conditioning information. We use provided source code from the official repository at: <https://github.com/jkulhanek/viewformer>. We do not generate quantitative metrics, as ViewFormer operates on masked and center-cropped images. Additionally, the images, which we use for comparison are in the training set for ViewFormer, while for our method they are in the test set.

Look Outside the Room [51]. We compare against Look-outside-the-room (LOTR), the current state-of-the-art method on novel view synthesis on Matterport3D (MP3D)[9] and RealEstate10K [94] datasets. For LOTR, we obtained the pretrained weights for the MP3D dataset from their official codebase <https://github.com/xrenaa/Look-Outside-Room>. We match LOTR’s data preparation methodology, including identical train/test splits, and we use LOTR’s implementation for generating multi-view images from MP3D RGB-D scans. For testing their method, we prepare a common set of 200 input images from the test split with the trajectories and ground truth images for the next 10 frames for each input. Then, we run the LOTR method on the given input using the code from their Github repository, using 3 overlapping frame windows, as stated in their paper. We run LOTR on the next 10 frames, given the input frame, and measure the metrics against the ground truth.

Additional Baselines for MP3D To further evaluate our method’s effectiveness on the novel-view synthesis task on MP3D scenes, we compare against additional baselines of GeoGPT [56] and SynSin [84]. Note that these two baselines, along with another recent work of PixelSynth [54], have been already shown to underperform against LOTR [51]. Since GeoGPT does not provide pre-trained models or rendered images for MP3D, we inquired the authors of LOTR for the images they used for the benchmarks. The acquired NVS images of GeoGPT and SynSin are rendered by the exact same protocol as our experiments, except that they proceeded five frames from the initial input images for 200 sequences (thus we have 1,000 images in total). We note that the trajectories used for these acquired images are different from the trajectories we used for our experiments because the trajectories are generated randomly via the Habitat embodied agent simulation [62]. However, at 1000 trajectory samples, we believe our comparisons are statistically significant. The final numbers we computed show similar trends to those reported in the LOTR paper, further confirming the validity of the comparisons. Both qualitatively and quantitatively, we observe that

our novel-view renderings are significantly more desirable.

C.3. Dataset details

ShapeNet [10]. We extensively evaluate our method on the ShapeNet dataset. The full ShapeNet dataset contains different object categories, each with a synthetically generated posed images in pre-defined training, validation, and testing sets. In our work, we specifically evaluate with the “cars” category, and focus on single-image novel view synthesis. We use the version of the dataset provided in PixelNeRF [89] for consistency in training and evaluation, keeping all frames in the dataset at 128^2 resolution and doing no additional pre-processing. As described in the main paper, the training set contains 2,458 cars, each with 50 renderings randomly distributed on the surface of a sphere. The test split contains 704 cars, each with 250 rendered images and poses on an Archimedean spiral. During the training of our method, we use the defined training split, randomly sampling between one and three input frames with the objective of synthesizing a randomly selected target frame for a specific object instance. In evaluation, we use the defined testing split, use image number 64 as input, and synthesize the other 249 ground truth images. We note that since these images are synthetically generated at only 128^2 , they lack backgrounds and fine detail. However, the accuracy of poses in the constrained environment and consistent evaluation method between baselines allows for easily providing quantitative benchmarks for single-image novel view synthesis.

Matterport3D [9]. We showcase our algorithm on a highly complex, large-scale indoor dataset, Matterport3D (MP3D). MP3D contains RGB-D scans of real-world building interiors. Scenes are calibrated to metric scale, and thus there is no scale ambiguity. We preprocess MP3D scans into a dataset of posed multi-view images following the procedure detailed in LOTR [51] and SynSin [84]. Specifically, we generate the image sequences by simulating a navigation agent in the room scans, using the popular Habitat [62] API. We randomly select the start and end position within the MP3D scenes and simulate the navigation towards the goal via Habitat. The agent is only allowed to take limited actions, including going forward and rotating 15 degrees. During training, we randomly sample a target frame and then select 1 to 3 random source frames in the neighborhood of 20 frames for conditioning.

Common Objects in 3D [50]. We validate our method on a real-world dataset: Common Objects in 3D (CO3D). The CO3D dataset consists of several categories. We train on CO3D Hydrants, which contains 726 scenes. The average scene consists of around 200 frames of RGB video, object masks, poses, and semi-sparse depth. We note that the CO3D dataset is quite unconstrained: even across scenes within a category, aspect ratio, resolution, FOV, camera trajectory, object scale, and global orientation all vary. Additionally, we note that the dataset is noisy, with several examples of

miscategorized objects and numerous extremely short or low-quality videos. Such noise adds to the challenge of single-image NVS.

In preparing data, we first center-crop to the largest possible square, then resize to 128^2 using Lanczos resampling. We adjust the camera intrinsics to reflect this change. We also seek to normalize the canonical scale of scenes across the dataset. To do so, we examine the provided depths within each scene, and consider the depth values that fall within the object segmentation mask. For each image, we calculate the median value of the masked depth. Taking the mean of these median values across the scene gives us a rough approximation of the distance between the camera and object. We adjust the scale of the scene so that this camera-object distance is identical across every scene in the dataset.

To help resolve scale, which is highly variable across the dataset, and to provide information parity with PixelNeRF, which has access to a global reference frame, we provide our feature encoder, T , with the location of the global origin. In addition to each input RGB image, we concatenate a channel that contains a depth rendering of the three coordinate planes, as rendered from the input camera. We modify T to accept the four-channel input. We find this input augmentation to improve our model’s ability to localize objects.

D. Discussion

D.1. Alternative approaches

GAN-based generative novel view synthesis. We have presented a diffusion-based generative model for novel view synthesis, but in principle, it is possible to construct a similar framework around other types of generative models. Generative Adversarial Networks [21] (GANs), are a natural fit, and adversarial training could drop in to replace our diffusion objective with minor changes. While recent work [15] has demonstrated that diffusion models often outperform GANs in mode coverage and image quality, GANs have a major advantage in speed. Future work that aims for real-time synthesis may prefer a GAN-based 3D-aware NVS approach.

Transformer-based, geometry-free multi-view aggregation strategies. A promising alternative to explicit geometry priors, such as the type we have presented in this work, is to instead make use of powerful attention mechanisms for effectively combining multiple observations. Scene Representation Transformers [61] utilize a transformer-based approach to merge information from multiple views, which is effective for NVS on both simple and complex scenes. We explored an SRT-based variant of Γ , which would forego explicit geometry priors for a transformer and light field [68] based conditioning scheme. However, we had difficulty achieving sufficient convergence and in justifying the additional compute cost. Nevertheless, related approaches could be a promising area for future study.

D.2. Limitations

We believe our method to be a valuable step towards in-the-wild single-view novel view synthesis but we acknowledge several limitations. While we demonstrate our method to be competitively geometrically consistent, it is not inherently 3D or temporally consistent. Noticeable flicker and other artifacts are sometimes visible in rendered sequences.

While our model generally produces plausible renderings, it may not always perfectly transfer details from the input. On ShapeNet, this sometimes manifests as an inability to replicate the angle of car tires or the style of windows across the line of symmetry; on more complex datasets, the model sometimes struggles to transfer fine details. We use a relatively lightweight, ResNet-backed Deeplab feature encoder. A more powerful encoder, potentially one that makes use of attention to improve long-range information flow, may resolve these issues.