PREDICTING 3D STRUCTURE BY LATENT POSTERIOR SAMPLING

Azmi Haider
Department of Computer Science
University of Haifa
Haifa, Israel
ahaide03@campus.haifa.ac.il

Dan Rosenbaum
Department of Computational Science
University of Haifa
Haifa, Israel
danro@cs.haifa.ac.il

ABSTRACT

The remarkable achievements of both generative models of 2D images and neural field representations for 3D scenes present a compelling opportunity to integrate the strengths of both approaches. In this work, we propose a methodology that combines a NeRF-based representation of 3D scenes with probabilistic modeling and reasoning using diffusion models. We view 3D reconstruction as a perception problem with inherent uncertainty that can thereby benefit from probabilistic inference methods. The core idea is to represent the 3D scene as a stochastic latent variable for which we can learn a prior and use it to perform posterior inference given a set of observations. We formulate posterior sampling using the score-based inference method of diffusion models in conjunction with a likelihood term computed from a reconstruction model that includes volumetric rendering. We train the model using a two-stage process: first we train the reconstruction model while auto-decoding the latent representations for a dataset of 3D scenes, and then we train the prior over the latents using a diffusion model. By using the model to generate samples from the posterior we demonstrate that various 3D reconstruction tasks can be performed, differing by the type of observation used as inputs. We showcase reconstruction from single-view, multi-view, noisy images, sparse pixels, and sparse depth data. These observations vary in the amount of information they provide for the scene and we show that our method can model the varying levels of inherent uncertainty associated with each task. Our experiments illustrate that this approach yields a comprehensive method capable of accurately predicting 3D structure from diverse types of observations.

1 INTRODUCTION

3D prediction using neural networks (Mildenhall et al., 2020; Sitzmann et al., 2019; Park et al., 2019) has garnered significant attention, tackling two main challenges: **3D reconstruction** (predicting 3D representations from limited observations) and **3D generation** (sampling new 3D scenes using generative models conditioned on signals like text or images). While 3D generation employs probabilistic generative models, 3D reconstruction is in most cases an ill-posed problem that requires incorporating prior knowledge and could therefore benefit from probabilistic inference methods.

In this work, we propose a probabilistic framework for 3D reconstruction. By combining a generative **prior** over latent 3D representations with a **likelihood** term from a reconstruction model, our approach predicts the full **posterior** distribution of a scene's 3D structure given sparse or noisy observations. This framework leverages a volumetric renderer based on a shared conditional neural field (CNF), trained in two stages:

1. **Auto-decoding** optimizes the shared CNF and latent representations for scenes in the training set.
2. **A diffusion model** captures the prior distribution over the latent representations.

Our method uses a tri-plane latent structure (Chan et al., 2021; Chen et al., 2022) for efficient representation, balancing global and local 3D information. Diffusion-based posterior sampling, guided by

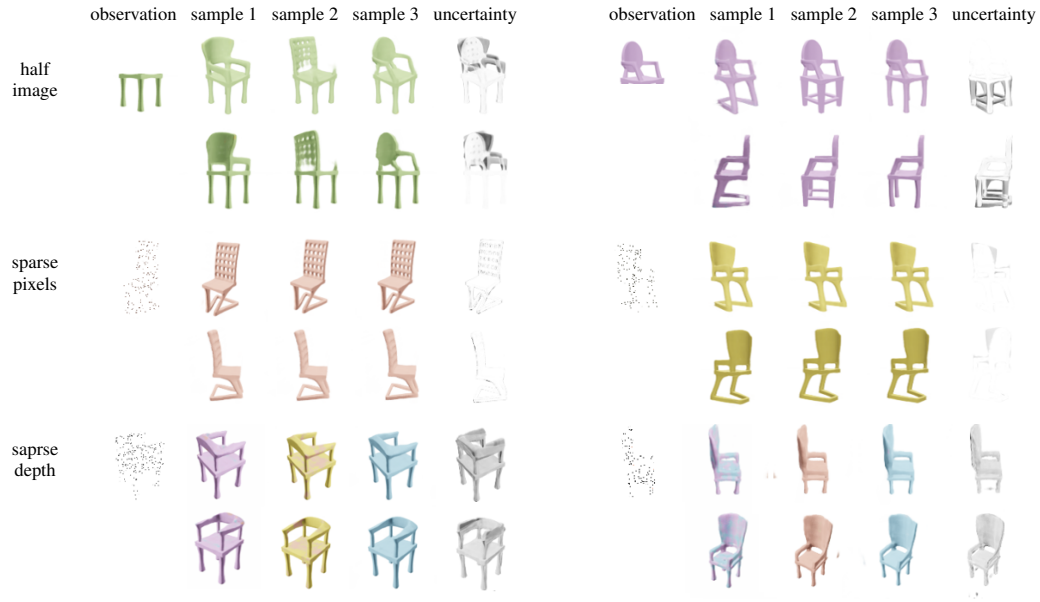

Figure 1: Examples of various 3D prediction tasks performed by generating posterior samples with our method. For each task we show the observation, three samples of the scene and an uncertainty map computed from the variance of 10 samples. Two different views are shown in subsequent rows. **Top:** reconstruction from half of an image. The variance is high in the hidden half of the scene. **Middle:** reconstruction from only a few pixels (5% of a single image). **Bottom:** reconstruction from a few depth values (5% of a full depth image from a single direction). Samples and uncertainty map suggest sparse depth is enough to reconstruct the 3D shape and uncertainty remains only about color.

reconstruction gradients, enables probabilistic reasoning and uncertainty quantification. In contrast to previous work suggesting amortizing posterior inference (Kosiorek et al., 2021a) in a variational autencoder setting, which experimentally demonstrated only limited results. By guiding the Langevin sampling with the gradient of the reconstruction model we combine the strength of two recently successful methods (1) iterative sampling with diffusion models and (2) gradient based optimization for translating observations to 3D representations.

We validate our method on tasks like single-view reconstruction, reconstruction from sparse pixels or depth, and noisy observations, demonstrating improved coverage of ground truth structures and generating uncertainty maps for unobserved regions.

Our contributions are as follows:

1. A probabilistic framework for 3D reconstruction, leveraging a diffusion prior and NeRF-based decoder.
2. A two-stage training approach: auto-decoding latent 3D representations and training a diffusion model as a prior.
3. Demonstrations on diverse 3D reconstruction tasks, showcasing robustness to sparse and noisy observations.
4. Enhanced reconstruction quality and uncertainty quantification through posterior sampling.

All code, models, and data will be released upon publication.

2 Related Work

Latent variable models over 3D scenes Early approaches like GQN (Eslami et al., 2018) used variational autoencoders for probabilistic reasoning in simple 3D scenes but lacked specialized 3D

geometry. Later, NeRF-based models (Kosiorek et al., 2021b) integrated rendering pipelines but did not fully leverage NeRF's capacity for complex scenes. Our work advances this by replacing amortized inference with high-capacity diffusion models and Langevin posterior sampling. Other efforts (Shen et al., 2022; Sünderhauf et al., 2022; Goli et al., 2023) modeled uncertainty but lacked data-driven priors like ours.

Generating 3D with 2D Generative models Given limited 3D ground truth data, several works (Poole et al., 2022; Watson et al., 2022; Liu et al., 2023) use pretrained 2D diffusion models to infer 3D representations. In a similar approach, Liu et al. (2024) use a 2D prior to compute 3D uncertainty maps. While these approaches achieve impressive generative visual results, they do not explicitly reason about the 3D structure of the scene, which leads to less consistency in generation (see experiment in Sec. C in the appendix) and prevents them from performing the full range of 3D probabilistic reasoning tasks, e.g. reconstruction from depth information.

Generative models of observed 3D representations Despite data scarcity, some works (Shue et al., 2022; Erkoç et al., 2023) train diffusion models directly on 3D datasets using representations like tri-planes or neural fields. Our approach uniquely avoids reliance on 3D ground truth, using only 2D datasets to generate 3D scenes.

Generative models of latent 3D representations Inspired by (Dupont et al., 2022), we enhance conditional neural fields (CNFs) with a compressed tri-plane representation and diffusion-based posterior sampling. Similar models (Bautista et al., 2022; Yang et al., 2023) focus on generative tasks rather than reconstruction, while (Chen et al., 2023) unify training stages but lack our latent compression. Concurrent work (Le et al., 2024) explores full posterior inference but with different objectives.

The concurrent work of Le et al. (2024) shares a similar motivation to ours. It focuses on specific type of noisy observations using a 3D modeling of the corruption field, and extensively demonstrate the advantages of the full posterior distribution over the maximum only (MAP inference). In Zhang et al. (2024) a 3D generative model is trained based on a Gaussian splatting representation which could also be combined with posterior sampling in future research.

3 BACKGROUND

3.1 AUTO-DECODING 3D REPRESENTATIONS

Recent advances in 3D scene representation have leveraged deep neural networks. NeRF (Mildenhall et al., 2020) introduced a method to reconstruct 3D scenes by training a neural network on multi-view images, requiring separate models for each scene. Subsequent work, such as PixelNeRF (Yu et al., 2021) and IBRNet (Wang et al., 2021), developed generalizable models that integrate prior knowledge of 3D scenes, reducing the number of views needed. These models often rely on conditional neural fields (CNFs), where a shared neural field is conditioned on scene-specific representations.

Recent studies (Dupont et al., 2022; Bautista et al., 2022; Chen et al., 2023; Yang et al., 2023) proposed the use of CNFs to train representations of scenes that can later be used in downstream tasks. Such models use an *auto-decoding* approach (Bojanowski et al., 2019; Park et al., 2019), where the representations are optimized for each scene concurrently with the training of the shared CNF.

Tri-plane representations (Chan et al., 2021; Chen et al., 2022) have proven effective, maintaining spatial structure and balancing global and local information. These representations condition the CNF by interpolating queried 3D positions across orthogonal planes, as shown in Fig. 2.

3.2 POSTERIOR SAMPLING WITH DIFFUSION MODELS

Diffusion models, such as Denoising Diffusion Probabilistic Models (DDPM)(Ho et al., 2020), generate high-quality samples by reversing a forward diffusion process that progressively adds noise. Many different variants (Sohl-Dickstein et al., 2015; Song et al., 2020) use U-Net architectures (Ronneberger et al., 2015) to predict noise and iteratively refine samples. Given a noisy input x_t, the

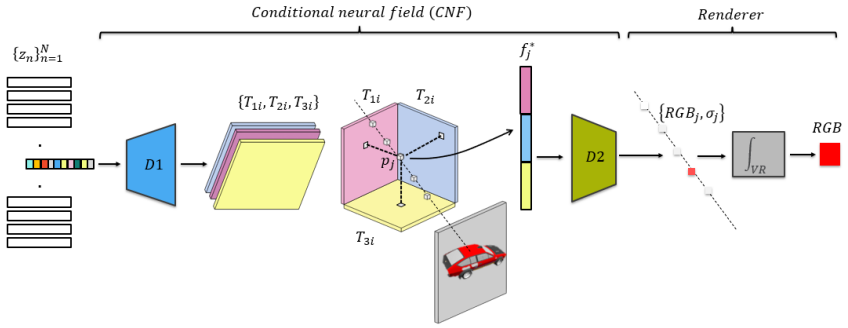

Figure 2: The reconstruction model mapping latent representations to images of a 3D scene. The latent decoder $D1$ maps the latent vector z_i corresponding to scene i into three multichannel planes (tri-planes) $\{T_{1i}, T_{2i}, T_{3i}\}$. Given an image and a camera position from which the image was taken, a ray is projected onto the scene from each pixel of the image, and multiple 3D points are sampled along the ray. Each 3D point p_j, is projected onto the multichannel tri-planes where each plane produces a feature vector f_j using bilinear interpolation. The three feature vectors are concatenated to form one feature vector f_j^*, and the decoder $D2$ is used to produce RGB and σ values for each 3D point along the ray. Volumetric rendering is then used to generate a single RGB value to be compared to the ground truth value of the pixel in the image.

denoised sample x_{t-1} and the clean estimate $\hat{x}_0$ are computed as:

$$\begin{aligned} x_{t-1} &\sim \mathcal{N}\left(\frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)\right),\ \tilde{\beta}_t I\right), \\ \hat{x}_0(x_t, t) &= \frac{1}{\sqrt{\bar{\alpha}_t}}\left(x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t, t)\right). \end{aligned} \tag{1}$$

Where, $\bar{\alpha}_t = \prod_{s=0}^{t} \alpha_s$ and $\tilde{\beta}_t$ is the noise variance. Diffusion models are widely used as priors for image restoration tasks (e.g., denoising, inpainting) (Choi et al., 2021; Chung et al., 2023; Kawar et al., 2022), with posterior sampling defined as:

$$\nabla_{x_t} \log p_t(x_t|y) = \nabla_{x_t} \log p_t(x_t) + \nabla_{x_t} \log p_t(y|x_t) \tag{2}$$

This approach combines the prior score $\nabla_{x_t} \log p_t(x_t)$ with the likelihood gradient $\nabla_{x_t} \log p_t(y|x_t)$, often referred to as guidance. While exact likelihoods typically depend on clean images x_0, approximations have been proposed, as detailed in Sec. 4.3.

4 METHOD

In this section we describe our method both at training time and at inference time. Training is based on two stages: (1) training the reconstruction model (RM) while optimizing the latent representation of the training scenes (auto-decoding), and (2) training a diffusion model over the latents as a prior. At inference time we use the trained prior and the reconstruction model to perform posterior sampling of the latents. For all implementation details please refer to Sec. B in the appendix.

4.1 TRAINING THE REPRESENTATION AND RECONSTRUCTION

The reconstruction model is a CNF followed by a volumetric renderer. Conditioned on a scene representation, the CNF predicts the values of 3D positions within the scene that are subsequently used by the volumetric renderer. The CNF is trained while concurrently auto-decoding the representation of each scene. The role of the reconstruction model is to form a mapping from the representation vectors to the values of the observations, i.e. image pixels, and also serve as the model through which the representation is optimized, effectively mapping the 3D scene observations back into the representation.

	functa, as reported in (Dupont et al., 2022)	ours
latent dim	1024	1024
PSNR train	24.4	27.67
PSNR test	23.1	26.9

Table 1: Reconstruction from latent.

Figure 3: Novel view reconstruction for held out 3D scenes. Each pair shows the ground truth image (left) and the reconstructed image (right). Top row: SRN cars. Bottom row: Objaverse chairs.

The model is depicted in Fig. 2. The latent vector $z_i \in \mathbb{R}^d$, corresponding to the i-th scene, is first reshaped into a 2D map of shape $r \times r \times c$. The latent decoder $D1$ decodes z_i into a 3D tensor $T \in \mathbb{R}^{R \times R \times 3C}$ using a series of ResNet blocks. T is reshaped to form a tri-plane representation $T_{1i}, T_{2i}, T_{3i} \in \mathbb{R}^{R \times R \times C}$. The tri-planes structure is used for reconstruction as follows: given an image of a scene, rays are projected from each pixel into the 3D scene, and multiple 3D points are sampled along each ray. Each 3D point is projected onto the tri-planes, and using bi-linear interpolation, each plane produces a single corresponding feature vector $f \in \mathbb{R}^C$. The three feature vectors are concatenated to form f^*. The decoder $D2$, an MLP, transforms f^* into RGB and σ values for the corresponding 3D position. This process is repeated for all 3D points along the ray and volumetric rendering is applied on the ray's points to generate a single RGB value for the pixel from which the ray was projected into the scene.

The reconstruction model (RM) and the latents are trained using the auto-decoding approach as following: at each training iteration, a minibatch of scenes $\mathcal{B}$ is randomly selected along with the corresponding latent vectors, where for each scene a random set of images, and random set of pixels within the images are used. The minibatch is used to apply a forward pass of the reconstruction model on the latents, and backpropagate the loss between the model's output and ground-truth pixel values to all network weights and latent values.

$$\mathcal{L}_{rec} = \sum_{i \in \mathcal{B}} \sum_{x \in \mathcal{X}_i} \|x - RM_\phi(z_i)\|^2 \tag{3}$$

where $\mathcal{B}$ is a random minibatch of scenes, and $\mathcal{X}_i$ is a random set of pixels from a random set of images from each scene i. The network weights are updated using $\partial \mathcal{L}_{rec}/\partial \phi$, and the latents are updated using $\partial \mathcal{L}_{rec}/\partial z_i$. In this way the latent representation for each scene is optimized while the network weights converge to their final values. For all experiments in the paper we use a latent dimension of 1024, which forms a highly compressed representation of the scenes. For more implementation details, see Sec. B in the appendix.

Fig. 3 shows examples of reconstruction for a few selected scenes using two models that were trained on the SRN Cars (Sitzmann et al., 2019) and Objaverse-lvis chair category (Deitke et al., 2022). See Sec. A in the appendix for details about the datasets.

After training the reconstruction models, 125 images of held-out test scenes are used to optimize the scene latents while freezing the reconstruction model's weights, and the latents are then used to reconstruct novel views of the scenes. The results show that the latent representation captures the 3D scenes with high fidelity. In Tab. 1 we compare the reconstruction accuracy of our compressed representation to Dupont et al. (2022). Our results are favorable, and we argue that this is due to the spatial structure of the tri-plane representation.

4.2 Training the prior

The goal of the second stage is to obtain a prior over the latent representation. This is achieved by training a generative model based on a Denoising Diffusion Probabilistic Model (DDPM) (Ho et al., 2020) on the latent data obtained in the first stage. As is standard in diffusion models, the model is based on a U-net architecture (Ronneberger et al., 2015) that is trained to denoise the latent representations $\{z_i\}_{n=1}^N \in \mathbb{R}^d$. To comply with the U-net architecture, the latents are reshaped to be $\{z_i\}_{n=1}^N \in \mathbb{R}^{r \times r \times c}$. The training loss is computed by:

$$\mathcal{L}_{gen} = \mathbb{E}_{z \in \{z\}, \epsilon \in \mathcal{N}(0,1), t \in U[0,T]} \left\| \epsilon_\theta \left(\sqrt{\bar{\alpha}_t} z + \sqrt{1 - \bar{\alpha}_t} \epsilon, t \right) - \epsilon \right\|^2 \tag{4}$$

Figure 4: Samples from the trained diffusion model. Each row corresponds to a different sample of the latent representation, corresponding to a different 3D scene, and each column shows a different view reconstructed from the same scene. Left: SRN cars. Right: Objaverse-lvis chairs.

Our implementation is based on Graikos et al. (2022). Fig 4 shows examples of random samples generated from the learned prior. Each image is generated by first sampling a latent from the prior, corresponding to a sampled scene (rows), and then using the reconstruction model to render images of the scene from different views (columns). The resulting samples show both coherence and diversity.

4.3 Sampling from the posterior

As described in Sec. 3.2, different methods have been proposed to sample from posterior distributions given a trained diffusion model as a prior. These methods consist of adding a likelihood term to each step in the iterative process of sampling from the prior. Here, the likelihood term comes from applying the reconstruction model (RM) on the estimated latent, and computing a squared loss compared to the given observation y , which corresponds to a Gaussian log-likelihood.

$$\log p(y|z) = -s\|y - RM_\phi(z)\|^2 + const. = -\mathcal{L}_{rec} + const. \tag{5}$$

where s is a scaling factor corresponding to the assumed variance of the reconstruction.

The method is depicted in Fig. 5, and described in Alg. 1. In more detail, at each step t the output of the U-net $\epsilon_\theta(z_t, t)$ is used to compute the one-step denoised latent z_{t-1} and a fully denoised estimate $\hat{z_0}$ (Eq. 1). The clean estimate is fed to the reconstruction model (RM) which outputs a prediction of the input views. A gradient of the log-likelihood with repsect to z_t can be computed by back-propagating the reconstruction error (Eq. 5) between the predicted images and the observed ground-truth images. However, this requires back-propagating through the U-net at each step. In order to accelerate inference, we approximate this gradient by computing $\tilde{z}_0(z_{t-1}) = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(z_{t-1} - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(z_t, t)\right)$, and the gradient with respect to z_{t-1}. When using many sampling steps we empirically observe that the difference between z_t and z_{t-1} is negligible and this approximation can be used to efficiently compute the posterior score:

$$\nabla_{z_t} \log p_t(z_t \mid y) \approx \nabla_{z_t} \log p_t(z_t) + \nabla_{z_{t-1}} \log p\big(y \mid \tilde{z}_0(z_{t-1}, t)\big), \tag{6}$$

Repeating this process for $t = T...1$ forms an approximated Langevin sampling process from the posterior distribution.

As the reconstruction loss is calculated with no regards to pixel order or quantity, this approach allows training a single prior model, and then use it to generate posterior samples for various types of conditioning signals. Examples include conditioning on many images, few images, or even a few random pixels per scene. Moreover, the desired inference task does not even need to be known at training time, as long as a corresponding reconstruction term can be formulated and differentiated at inference time.

5 Experiments

For all experiments we use the same model and the same configuration.

Algorithm 1: Posterior Sampling

Input: images y, scale s

Initialize $z_T \sim N(0, 1)$

for $t = T$ **to** 1 **do**

$\epsilon \leftarrow \text{U-net}_\theta(z_t, t)$

$z_{t-1} \sim \mathcal{N}\left(\frac{1}{\sqrt{\alpha_t}}\left(z_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon\right), \tilde{\beta}\right)$

$\tilde{z}_0(z_{t-1}) = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(z_{t-1} - \sqrt{1-\bar{\alpha}_t}\epsilon\right)$

$\mathcal{L}_{rec} = \|y - RM_\phi\left(\tilde{z}_0(z_{t-1})\right)\|^2$

$z_{t-1} \leftarrow z_{t-1} - s\partial\mathcal{L}_{rec}/\partial z_{t-1}$

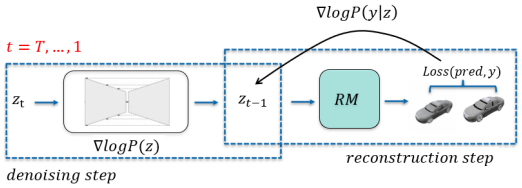

Figure 5: **Left:** The posterior sampling algorithm. **Right:** Illustration of a single step in the iterative process. Conditioned on the previous estimate z_t, the U-net predicts the noise, which is used to compute both z_{t-1} and $\tilde{z}_0$. The latter is fed to the reconstruction model to predict an image from the given view which is compared to the ground truth image y. The error is backpropagated through the frozen networks to compute a gradient which is then added to z_{t-1}.

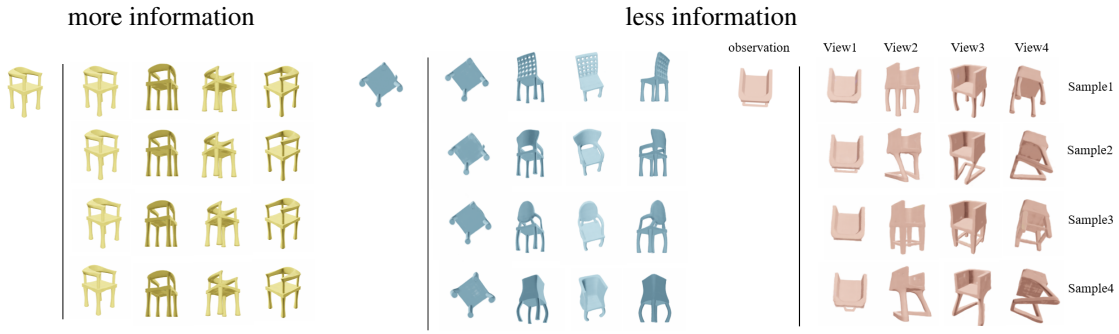

Figure 6: Posterior samples given a single view for Objaverse chairs. Each row corresponds to a different sample of the scene, and each column shows a different view. The observation in the example on the left carries high information about the scene, resulting in very similar samples. The observations in the middle and right scenes are less informative, and therefore result in more diverse samples, where the chairs are completed with different possible configurations of legs, armrests and backrests. These example demonstrate a coherent merging of observed data and prior information.

GENERATING CONDITIONAL SAMPLES

We show results of generating posterior samples given one observed image per scene. In Fig. 6, three examples from Objaverse chairs are shown. In the scene shown on the left, the given image contains enough information to predict any view of the scene with certainty. This results in multiple samples (rows) that are almost identical. In the other examples the observed image is less informative and does not provide enough information about the scene from all angles. Therefore, samples from the posterior exhibit more diversity in the way they complete the missing information. More concretely, the chairs observed from uninformative views are predicted to have different possible leg, armrest and backrest configurations. Note that while the samples are different, the generated latent is a 3D representation, so each sample can be used to predict a coherent set of images from different views.

In Fig. 1, we demonstrate the ability of the method to perform more diverse probabilistic reasoning tasks. We show prediction from half-image inputs, from a sparse set of pixels of one image (5%), and from a sparse set of depth map pixels (5% of a depth map from a single view). For each scene we show three samples, showing two different views for each, and an uncertainty map. The uncertainty is computed by generating 10 samples of the scene, rendering corresponding 10 images for each view and computing the variance in the rendered images. Using our method for partial RGB observations (half-image or sparse pixels) is trivial to implement since the reconstruction model operates per pixel and can be used to predict any subset of pixels in the scene. In the case of depth data, we implement

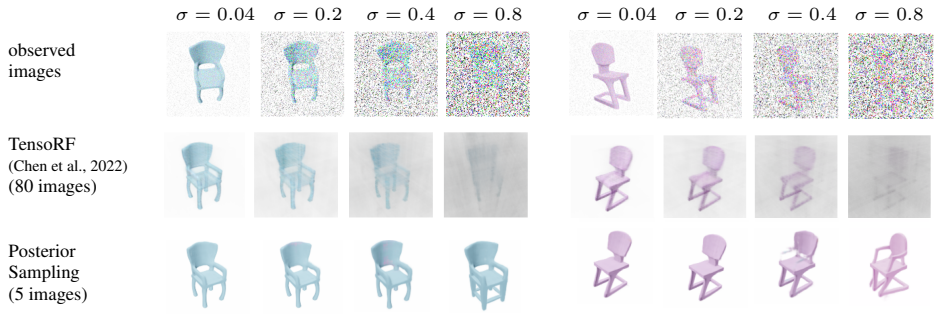

Figure 7: 3D reconstruction from noisy images. Reconstruction from 80 images without a prior (TensoRF) quickly deteriorates as noise increases. Using our prior to perform posterior sampling results in a much more robust method, significantly outperforming TensoRF even when using an order of magnitude less images (5 images).

a different reconstruction loss comparing the predicted σ values to ground truth values without using the RGB prediction and the renderer in Fig. 2. Given a depth pixel value, the ground truth value of σ is set to 1 for the 3D point on the ray sampled at the given depth value, and 0 for all the other 3D points. We emphasize that this reconstruction model is formulated at inference time and is not used at training. The results show the different plausible predictions of the scene and the resulting uncertainty. For the first case we see that the uncertainty is high for the hidden half of the scene as expected. For the other two cases, samples generated from sparse pixel observations demonstrate a high degree of similarity, suggesting, perhaps surprisingly, that even just 5% of the pixels from a single view is sufficient for accurate 3D scene prediction. In case of the sparse depth data, the only uncertain aspect is the object color.

In Fig. 7, we evaluate the robustness of our method in 3D reconstruction from noisy images by comparing it to TensoRF (Chen et al., 2022) without a prior. By generating samples from multiple scene images under increasing noise levels, we demonstrate that posterior sampling significantly improves resilience to noise. While TensoRF, trained on 80 images, experiences a sharp performance drop as noise increases, our method maintains stable performance even with only 5 input images.

6 Conclusion

In this work we introduced a methodology that combines the strengths of NeRF-based 3D reconstruction together with the probabilistic reasoning of diffusion models. Our method views 3D reconstruction as an ill-posed perception problem that requires reconciling the observed information with prior knowledge. We showed that (1) 3D scenes can be efficiently represented by compact latent vectors, using a reconstruction model that consists of a tri-plane representation, which preserves spatial structure within the 3D model; and (2) this representation is amenable to training a strong diffusion-model based prior that can later be used to solve various inference tasks. We highlight the importance of predicting the full posterior distribution rather than optimizing for an average sample with higher PSNR (see Sec. C). Averaging tends to produce oversmoothed results that may score well numerically but fail to capture the full variability of plausible 3D structures. By emphasizing diverse posterior samples, our approach better represents the inherent uncertainty in 3D scene synthesis, leading to more robust and generalizable models and solving various 3D reconstruction tasks.

Limitations and future work: A main challenge that remains in 3D reconstruction is scaling to more complex and more diverse data towards developing methods that can reliably predict real 3D scenes from different levels of observations. Another challenge is the slow sampling time with diffusion models. While our results are demonstrated on small scale data, we believe that the compressed representation and the principled way of handling uncertainty that we propose, combined with recent developments in accelerating diffusion model sampling, are key for scaling up these models to larger and more complex datasets.

ACKNOWLEDGMENTS

We gratefully acknowledge the PhD funding provided by the data science research center (DSRC) at the university of Haifa, Israel.

REFERENCES

Miguel Angel Bautista, Pengsheng Guo, Samira Abnar, Walter Talbott, Alexander Toshev, Zhuoyuan Chen, Laurent Dinh, Shuangfei Zhai, Hanlin Goh, Daniel Ulbricht, Afshin Dehghan, and Josh Susskind. Gaudi: A neural architect for immersive 3d scene generation, 2022.

Piotr Bojanowski, Armand Joulin, David Lopez-Paz, and Arthur Szlam. Optimizing the Latent Space of Generative Networks, May 2019. URL http://arxiv.org/abs/1707.05776.

Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient Geometry-aware 3D Generative Adversarial Networks. *arXiv preprint arXiv:2112.07945*, 2021.

Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *European Conference on Computer Vision (ECCV)*, 2022.

Hansheng Chen, Jiatao Gu, Anpei Chen, Wei Tian, Zhuowen Tu, Lingjie Liu, and Hao Su. Single-stage diffusion nerf: A unified approach to 3d generation and reconstruction, 2023. URL https://arxiv.org/abs/2304.06714.

Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938*, 2021.

Hyungjin Chung, Jeongsol Kim, Michael Thompson Mccann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *ICLR*, 2023. URL https://openreview.net/forum?id=OnD9zGAGT0k.

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. *arXiv preprint arXiv:2212.08051*, 2022.

Emilien Dupont, Hyunjik Kim, S. M. Ali Eslami, Danilo J. Rezende, and Dan Rosenbaum. From data to functa: Your data point is a function and you should treat it like one. *CoRR*, abs/2201.12204, 2022. URL https://arxiv.org/abs/2201.12204.

Ziya Erkoç, Fangchang Ma, Qi Shan, Matthias Nießner, and Angela Dai. Hyperdiffusion: Generating implicit neural fields with weight-space diffusion, 2023. URL https://arxiv.org/abs/2303.17015.

S. Eslami, Danilo Jimenez Rezende, Frederic Besse, Fabio Viola, Ari Morcos, Marta Garnelo, Avraham Ruderman, Andrei Rusu, Ivo Danihelka, Karol Gregor, David Reichert, Lars Buesing, Theophane Weber, Oriol Vinyals, Dan Rosenbaum, Neil Rabinowitz, Helen King, Chloe Hillier, Matt Botvinick, and Demis Hassabis. Neural scene representation and rendering. *Science*, 360:1204–1210, 06 2018. doi: 10.1126/science.aar6170.

Lily Goli, Cody Reading, Silvia Sellán, Alec Jacobson, and Andrea Tagliasacchi. Bayes' rays: Uncertainty quantification for neural radiance fields, 2023. URL https://arxiv.org/abs/2309.03185.

Alexandros Graikos, Nikolay Malkin, Nebojsa Jojic, and Dimitris Samaras. Diffusion models as plug-and-play priors. In *Thirty-Sixth Conference on Neural Information Processing Systems*, 2022. URL https://arxiv.org/pdf/2206.09012.pdf.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *NeurIPS*, 35:23593–23606, 2022.

Adam R. Kosiorek, Heiko Strathmann, Daniel Zoran, Pol Moreno, Rosalia Schneider, Soňa Mokrá, and Danilo J. Rezende. Nerf-vae: A geometry aware 3d scene generative model, 2021a.

Adam R. Kosiorek, Heiko Strathmann, Daniel Zoran, Pol Moreno, Rosalia Schneider, Soňa Mokrá, and Danilo J. Rezende. Nerf-vae: A geometry aware 3d scene generative model, 2021b. URL https://arxiv.org/abs/2104.00587.

Tuan Anh Le, Pavel Sountsov, Matthew D. Hoffman, Ben Lee, Brian Patton, and Rif A. Saurous. Robust inverse graphics via probabilistic inference, 2024. URL https://arxiv.org/abs/2402.01915.

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

Xinhang Liu, Jiaben Chen, Shiu hong Kao, Yu-Wing Tai, and Chi-Keung Tang. Deceptive-nerf/3dgs: Diffusion-generated pseudo-observations for high-quality sparse-view reconstruction, 2024. URL https://arxiv.org/abs/2305.15171.

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

Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation, 2019. URL https://arxiv.org/abs/1901.05103.

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

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234–241, 2015.

Jianxiong Shen, Antonio Agudo, Francesc Moreno-Noguer, and Adria Ruiz. Conditional-flow nerf: Accurate 3d modelling with reliable uncertainty quantification, 2022. URL https://arxiv.org/abs/2203.10192.

J. Ryan Shue, Eric Ryan Chan, Ryan Po, Zachary Ankner, Jiajun Wu, and Gordon Wetzstein. 3d neural field generation using triplane diffusion, 2022. URL https://arxiv.org/abs/2211.16677.

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

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Int. conf. machine learning*, pp. 2256–2265, 2015.

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.

Niko Sünderhauf, Jad Abou-Chakra, and Dimity Miller. Density-aware nerf ensembles: Quantifying predictive uncertainty in neural radiance fields, 2022. URL https://arxiv.org/abs/2209.08718.

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, 2021.

Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models, 2022. URL https://arxiv.org/abs/2210.04628.

Guandao Yang, Abhijit Kundu, Leonidas J. Guibas, Jonathan T. Barron, and Ben Poole. Learning a diffusion prior for nerfs, 2023. URL https://arxiv.org/abs/2304.14473.

Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images, 2021.

Bowen Zhang, Yiji Cheng, Jiaolong Yang, Chunyu Wang, Feng Zhao, Yansong Tang, Dong Chen, and Baining Guo. Gaussiancube: A structured and explicit radiance representation for 3d generative modeling, 2024. URL https://arxiv.org/abs/2403.19655.

A DATA

We use two datasets in our experiments. The first dataset is SRN Cars (Sitzmann et al., 2019), which comprises 3,200 scenes with 250 images each. We randomly divide the images in each scene evenly between training images and test images, and we use 3,000 scenes for training, holding out 200 scenes for testing. The second dataset we use is the Objaverse-lvis chair category (Deitke et al., 2022), which comprises 439 instances with 100 images generated for each scene. While this dataset is smaller, it is more diverse in terms of shapes. We use 80% of the images in each scene for training, and hold out 8 scenes for testing and visualizations. For both datasets we use image resolution of 128×128.

B IMPLEMENTATION DETAILS

In this section we describe the implementation details of our model and experiments. All code, models and data will be made available upon publication.

TRAINING THE REPRESENTATION AND RECONSTRUCTION MODEL

The weights of the reconstruction model ϕ are randomly initialized, while the latent representations z_i are initialized to zero. The size of the data set corresponds to the number of latent vectors, each latent representing a single scene $\{z_i\}_{i=1}^{N}$ (N scenes = N latents).

During training, the images of each scene optimize only its respective latent, while the entire model, including decoders, is jointly trained.

The latent representation z_i dimensions are $d = 1024$, $r = 16, c = 4$. D1 is constructed using a series of six ResNet blocks where at each block the number of channels is the following: [4, 32, 64, 96, 128, 192]. Blocks are followed by a self-attention layer and alternating upsampling. The resulting 3D tensor T is divided into two tensors responsible for generating RGB and density, $T_{RGB} \in \mathbb{R}^{R \times R \times 3C_{RGB}}$, $T_\sigma \in \mathbb{R}^{R \times R \times 3C_\sigma}$, respectively. T_{RGB} is reshaped to form a tri-plane representation $T_{1i}, T_{2i}, T_{3i} \in \mathbb{R}^{R \times R \times C_{RGB}}$. Similarly, T_σ forms a tri-plane representation $T'_{1i}, T'_{2i}, T'_{3i} \in \mathbb{R}^{R \times R \times C_\sigma}$. Dimensions are $R = 128, C_{RGB} = 48, C_\sigma = 16$.
For each scene i, we randomly select 4096 rays from pixels in the training images. Along each ray, we sample 220 3D points and project them onto the tri-planes of both the RGB and density planes separately.
For each (RGB and density), this projection extracts three feature vectors from the three planes for further processing. Three vectors are concatenated into a single feature vector $f^*_{RGB} \in \mathbb{R}^{3C_{RGB}}$ for RGB and $f^*_\sigma \in \mathbb{R}^{3C_\sigma}$ for density. While the density feature vector f^*_σ produces density for 3D points by simply summing its elements, the RGB feature vector f^*_{RGB} is passed through D2 to produce a single RGB value. D2 is an MLP of 7 layers. Once all 3D points along the ray have RGB and density values, volumetric rendering, a parameterless process, produces a single RGB value to be compared with the pixel's color.

We train the model with a minibatch $\mathcal{B}$ size of 2 scenes, and with an Adam optimizer using three different learning rates: 1e-3 for the latents, 1e-4 for the D1 parameters and $1e{-}3$ for D2 parameters.

Our model is based on the code published in Chen et al. (2022).

At test time, a new latent (initialized to zeros) is coupled with the new scene and optimized using the learned/frozen decoders.

TRAINING THE PRIOR

As in Sec.4.1, latent representation z_i dimensions are $d = 1024$, $r = 16, c = 4$. The diffusion model used is implemented by Graikos et al. (2022) with the following parameters: The noise scheduler is a linear schedule with parameters $T = 1000, \beta_0 = 1e^{-4}, \beta_T = 2e^{-2}$. The U-net parameters are $model\_channels = 64$, $num\_resnet\_blocks = 2$, $channel\_mult = (1, 2, 3, 4)$, $attention\_resolutions = [8, 4]$, $num\_heads = 4$. We train the model with a minibatch $\mathcal{B}$ size of 32 scenes, and with an Adam optimizer with learning rate equal to 1e-3.

The reconstruction model and the diffusion model were trained on an NVIDIA GeForce RTX 4090 for Approximately one day each.

C EXPERIMENTS

GENERATING 3D WITH 2D GENERATIVE MODELS

As mentioned in Sec. 2, 2D generative models for 3D generation approaches do not fully capture the underlying 3D structure of a scene. To evaluate the impact of explicit 3D structure reasoning, we trained a Neural Radiance Field (NeRF) on images generated by both standard 2D diffusion models and our proposed model, which incorporates an inherent understanding of 3D structure. Tab. 2 results indicate that NeRF trained on images from our model produces more consistent 3D reconstructions, highlighting the importance of explicit 3D reasoning in generative models for robust 3D scene synthesis.

Method	**PSNR↑**	**SSIM↑**
3Dim (SRN cars)	28.53	0.96
Ours (SRN cars)	34.7	0.98
zero123 (Objaverse chairs)	26.8	0.925
Ours (Objaverse chairs)	43.4	0.99

Table 2: 3D Consistency Comparison: To evaluate 3D consistency, we compare our model with 3Dim (Watson et al., 2022) and Zero-1-to-3 (Liu et al., 2023), trained on SRN Cars and Objaverse Chairs, respectively. Given an input image of a scene, each model generates multiple novel views, which are then used to train a TensoRF (NeRF) model. Since higher 3D consistency in the generated images facilitates NeRF training, models producing more consistent views enable NeRF to achieve a higher PSNR. Our results demonstrate that NeRF trained on images from our model attains the highest PSNR, highlighting the benefits of our model's built-in 3D structural understanding for improved 3D scene synthesis.

POSTERIOR SAMPLING

Posterior sampling involves two types of computations: 1) denoising, using the diffusion as a prior to generate a plausible latent, and 2) reconstruction, using the reconstruction model to align the latent with the observed views.

For all experiments we use the same model using the same inference process. We generate posterior samples using 1000 iterations as described in Alg. 1 with the same scale factor s =5e-3 for all experiments. The only exception is the experiment with noisy data in Fig 7, where the scale factor for most extreme noise level $\sigma = 0.8$ was decreased to a value of s =3e-3, corresponding to the high noise variance in the observation.

FULL POSTERIOR VS HIGHER PSNR

In Fig. 8, we present two examples (rows) of conditional posterior sampling, where the leftmost image serves as the observed image. The reconstructions are displayed alongside their corresponding PSNR values. We generate 20 samples from the posterior distribution and showcase three individual samples under "single samples," highlighting the variability in reconstruction quality—some being closer to the ground truth than others. Additionally, we display the averaged reconstructions using 5, 10, and 20 latent samples. While averaging improves numerical PSNR scores, it often leads to oversmoothed results that fail to capture the full diversity of plausible 3D structures.

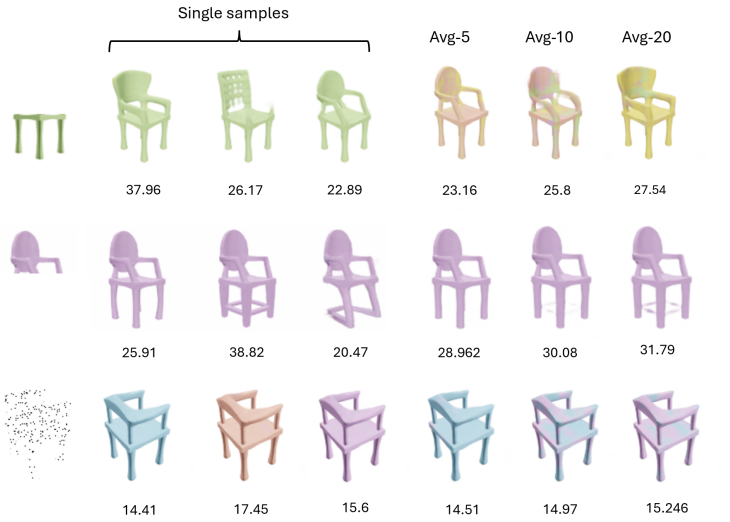

Figure 8: Averaging latent samples results in higher PSNR scores but fails to capture the full posterior distribution. In the first row, the bottom half of an image is used as guidance, while in the second row, the top half is used. Since most generated samples closely match the ground truth, averaging leads to a higher PSNR but collapses the distribution into a single reconstruction, limiting the diversity of plausible outcomes. In the third row, guidance is provided using only a few pixels from a depth image. The generated samples vary in color, as color information is not available from the depth input. Averaging across multiple samples produces an intermediate color that achieves a higher PSNR but fails to reflect the full posterior, which contains diverse color variations.