LASER: LATENT SET REPRESENTATIONS FOR 3D GENERATIVE MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural Radiance Field (NeRF) provides unparalleled fidelity of novel view synthesis—rendering a 3D scene from an arbitrary viewpoint. NeRF requires training on a large number of views that fully cover a scene, which limits its applicability. While these issues can be addressed by learning a prior over scenes in various forms, previous approaches have been either applied to overly simple scenes or struggling to render unobserved parts. We introduce Latent Set Representations for NeRF-VAE (LASER-NV)—a generative model which achieves high modelling capacity, and which is based on a set-valued latent representation modelled by normalizing flows. Similarly to previous amortized approaches, LASER-NV learns structure from multiple scenes and is capable of fast, feed-forward inference from few views. To encourage higher rendering fidelity and consistency with observed views, LASER-NV further incorporates a geometry-informed attention mechanism over the observed views. LASER-NV further produces *diverse and plausible* completions of occluded parts of a scene while remaining consistent with observations. LASER-NV shows state-of-the-art novel-view synthesis quality when evaluated on ShapeNet and on a novel simulated City dataset, which features high uncertainty in the unobserved regions of the scene.

1 INTRODUCTION

Probabilistic scene modelling aims to learn stochastic models for the structure of 3D scenes, which are typically only partially observed (Eslami et al., 2018; Kosiorek et al., 2021; Burgess et al., 2019). Such models need to reason about unobserved parts of a scene in way that is consistent with the observations and the data distribution. Scenes are usually represented as latent variables, which are ideally compact and concise, yet expressive enough to describe complex data.

Such 3D scenes can be thought of as projections of light rays onto an image plane. Neural Radiance Field (NeRF, (Mildenhall et al., 2020)) exploits this structure explicitly. It represents a scene as a radiance field (a.k.a. a scene function), which maps points in space (with the corresponding camera viewing direction) to color and mass density values. We can use volumetric rendering to project these radiance fields onto any camera plane, thus obtaining an image. Unlike directly predicting images with a CNN, this rendering process respects 3D geometry principles. NeRF represents scenes as parameters of an MLP, and is trained to minimize the reconstruction error of observations from a single scene—resulting in unprecedented quality of novel view synthesis.

For generative modelling, perhaps the most valuable property of NeRF is the notion of 3D geometry embedded in the rendering process, which does not need to be learned, and which promises strong generalisation to camera poses outside the training distribution. However, since NeRF's scene representations are high dimensional MLP parameters, they are not easily amenable to generative modelling (Dupont et al., 2022). NeRF-VAE (Kosiorek et al., 2021) embeds NeRF in a generative model by conditioning the scene function on a latent vector that is inferred from a set of 'context views'. It then uses NeRF's rendering mechanism to generate outputs. While NeRF-VAE admits efficient inference of a compact latent representation, its outputs lack visual fidelity. This is not surprising, given its simple latent structure, and the inability to directly incorporate observed features. In addition, NeRF-VAE does not produce varied samples of unobserved parts of a scene.

A number of recent deterministic methods (Yu et al., 2021; Trevithick & Yang, 2021; Wang et al., 2021) uses local image features to directly condition radiance fields in 3D. This greatly improves

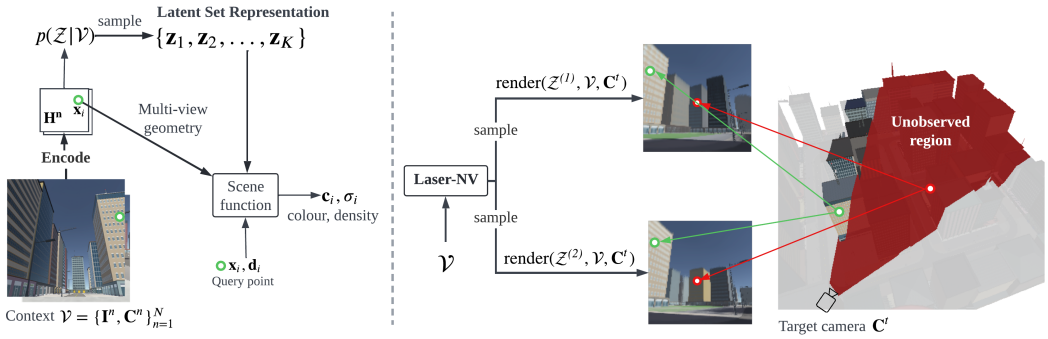


Figure 1: **Left:** LASER-NV infers a set-valued latent $\mathcal{Z}$ from the context $\mathcal{V}$ that consists of N image and camera pairs $(\mathbf{I}^n, \mathbf{C}^n)$. On querying the scene function at a point $\mathbf{x}_i$ with direction $\mathbf{d}_i$, the latents are combined with local features $\mathbf{H}^n$ that are back-projected from the context views—producing color and density. **Right**: Rendering a novel viewpoint may include observed (green, an example query point on the left) and unobserved parts (red) of the scene. Conditioned on the context $\mathcal{V}$, LASER-NV allows sampling multiple scene completions that are consistent with the context views (green arrows) while providing varied explanations for the unobserved parts of the scene (red arrows). We show two such samples for the same target camera $\mathbf{C}^t$. Also see the `gif` in supp. material showing a fly-through for four prior samples conditioned on the same views.

reconstruction quality of observed parts of the scene. However, these methods are still unable to produce plausible multimodal predictions for the unobserved parts of a scene.

In this work, we address NERF-VAE's shortcomings by proposing Latent Set Representations for NERF-VAE (LASER-NV). To increase modelling capacity, LASER-NV uses an arbitrarily-sized set of latent variables (instead of just one vector) modelled with normalizing flows. To further enable producing samples which are consistent with observed parts, we make the generative model conditional on a set of context views (as opposed to conditioning only the approximate posterior). LASER-NV offers superior visual quality with the ability to synthesise multiple varied novel views compatible with observations. Figure 1 shows LASER-NV's key components and abilities. We include a `gif` in supp. material showing fly-throughs of additional prior samples, see Section 4.5 for details.

Our contributions are as follows:

- We introduce a novel set-valued latent representation modelled by purpose-built permutation-invariant normalizing flows conditioned on context views. We show that increasing the number of latent set elements improves modelling performance, providing a simple way to trade off computation for quality without adding new model parameters. We also verify that increasing latent dimensionality in NERF-VAE offers no such benefits.
- We develop a novel attention mechanism to condition the scene function on the set-valued latent as well as additional local features computed from context views. We show that including local features further improves visual quality.
- We evaluate LASER-NV on the category-agnostic ShapeNet dataset, and on a novel "City" dataset that contains a large simulated urban area and poses significant challenges as a benchmark for novel view synthesis due to high uncertainty in the unobserved parts of the scene. Our model overcomes some of the main limitations of NERF-VAE and also outperforms deterministic NERF models on novel view synthesis in the face of uncertainty.

2 BACKGROUND: NERF & NERF-VAE

Neural Radiance Field (NERF, (Mildenhall et al., 2020)) represents a 3D scene as a *scene function* $F(\boldsymbol{x}, \boldsymbol{d})$ of a point coordinate $\boldsymbol{x} \in \mathbb{R}^3$ and direction $\boldsymbol{d} \in \mathbb{R}^3$, and outputs the colour $\boldsymbol{c} \in \mathbb{R}^3$ and volume density $\sigma \geq 0$. This function is parameterized using a neural network (typically a fully connected MLP). To obtain a pixel colour, the scene function is evaluated at many points along the corresponding ray in the volumetric rendering process.

This volumetric rendering integral is in practice approximated using numerical integration; see (Mildenhall et al., 2020; Blinn, 1982) for details. We use $\hat{\boldsymbol{I}} = \texttt{render}(F(\cdot), \boldsymbol{C})$ to denote the image rendering process that outputs the image $\hat{\boldsymbol{I}} \in \mathbb{R}^{H \times W \times 3}$ for the rays of a camera $\boldsymbol{C}$ and a scene function F. The camera $\boldsymbol{C} = (\boldsymbol{K}, \boldsymbol{R}, \boldsymbol{t})$ is specified by its intrinsic parameters $\boldsymbol{K} \in \mathbb{R}^{3\times 3}$ and extrinsic parameters given by a rotation $\boldsymbol{R} \in \mathbb{R}^{3\times 3}$ and translation $\boldsymbol{t} \in \mathbb{R}^3$. In its original formulation, the parameters of NeRF are optimized by minimizing an error function (typically the mean-squared error) between rendered images and the ground-truth images across many views of a single scene.

While NeRF learns high-fidelity scene representations when many views are available, it does not learn a prior over scenes, does not provide compact scene representations, and does not admit efficient inference (i.e. fast estimation of the scene parameters from a few input views). To address these issues, NeRF-VAE (Kosiorek et al., 2021) embeds NeRF as a decoder in a variational auto-encoder (VAE, (Kingma & Welling, 2014; Rezende et al., 2014)). Each scene is represented by a latent vector $\boldsymbol{z}$ (instead of the set of parameters of an MLP as in NeRF) which can be either sampled from a Gaussian prior $p(\boldsymbol{z})$ or from the approximate posterior $q(\boldsymbol{z} \mid \boldsymbol{C})$ inferred from a set of context views and associated cameras $\boldsymbol{C}$. The latent $\boldsymbol{z}$ is used to condition an MLP that parameterizes the scene function, $F_\theta(\cdot, \boldsymbol{z}) : (\boldsymbol{x}_i, \boldsymbol{d}_i) \mapsto \boldsymbol{c}_i, \boldsymbol{\sigma}_i$. Here, parameters θ are shared between scenes, while $\boldsymbol{z}$ is scene specific. NeRF-VAE uses a Gaussian likelihood $p(\boldsymbol{I} \mid \boldsymbol{z}, \boldsymbol{C})$ with the image rendered from the sampled scene function as mean, and a fixed standard deviation.

3 Latent Set Representations for NeRF-VAE

LASER-NV is a conditional generative model, conditioned on a set of input views $\mathcal{V} := \{\boldsymbol{I}^n, \boldsymbol{C}^n\}_{n=1}^N$. Given these views the model infers a latent representation that can be used for novel view synthesis.

In order to represent large-scale and complex scenes we use **La**tent **Se**t **R**epresentations (LASERs), a set of latents of the form $\mathcal{Z} := \{\boldsymbol{z}_k\}_{k=1}^K$, where K is a hyperparameter. In the NeRF-VAE, the scene function MLP is directly conditioned on the inferred latent. In contrast, LASER-NV's scene function is conditioned on the latent set $\mathcal{Z}$ and features $\mathcal{H} := \{\boldsymbol{H}^n\}_{n=1}^N$ extracted from N input views, and integrates information from both. This results in the form $F_\theta(\cdot, \mathcal{Z}, \mathcal{H}) : (\boldsymbol{x_i}, \boldsymbol{d_i}) \mapsto \boldsymbol{c}_i, \boldsymbol{\sigma}_i$, which we detail further in Section 3.3.

We obtain the input features $\mathcal{H}$ by separately encoding the input views using a convolutional neural net that maintains spatial structure. These feature maps are subsequently used in the conditional prior, the posterior, and the resulting scene function, all of which we describe next. Further architectural details are provided in Appendix A.

3.1 Conditional Prior

Given a set of context views $\mathcal{V} := \{\boldsymbol{I}^n, \boldsymbol{C}^n\}_{n=1}^N$ consisting of images $\boldsymbol{I}^n$ and corresponding cameras $\boldsymbol{C}^n$, LASER-NV defines a conditional prior $p(\mathcal{Z}|\mathcal{V})$ over a latent set of size K, parameterized by a permutation-invariant normalizing flow— a distribution model that allows for both fast and exact sampling and efficient density evaluation (Rezende & Mohamed, 2015; Papamakarios et al., 2021).

Flows are defined by one or more compositions of invertible mappings of random variables. Our flow design closely follows that of Wirnsberger et al. (2020; 2021), who parameterize transformations in each layer using transformers, preserving permutation invariance. Specifically, we propose a novel design to condition the flow distribution on the features $\mathcal{H}$ extracted from the context views. To achieve that, we use split coupling layers (Dinh et al., 2017), where for each layer, we first apply self-attention to the set of latents, followed by cross-attention to attend the context features. The output of the second transformer defines the parameters for an invertible affine transformation of the random variables. Implementation details of our normalizing flow architecture is given in Appendix A.1.

3.2 Conditional Posterior

Since LASER-NV is trained as a VAE, we need to define an approximate posterior $q(\mathcal{Z}|\mathcal{U} \cup \mathcal{V})$, where $\mathcal{U} = \{\boldsymbol{I}^t, \boldsymbol{C}^t\}_{t=1}^M$ are target views that the model is trained to reconstruct, denoted as posterior

context. The posterior and conditional prior are both distributions over the latent space, where the posterior is conditioned on the union of prior and posterior context. We model the posterior distribution using the same architecture as the conditional prior, but with a separate set of learned parameters.

3.3 Decoder & Scene Function

When evaluating the scene function, i. e. computing colour and density for a point $\mathbf{x}_i, \mathbf{d}_i$, LASER-NV has access to both the latent set $\mathcal{Z}$ as well as the features $\mathcal{H}$ extracted from the context views $\mathcal{V}$. We now describe how to integrate those two, which consists of querying the latent set for each point, and using 3D projection to extract relevant features.

Querying the Latent Set While the latent set representation $\mathcal{Z}$ contains information regarding the entire scene, when querying for a specific point $\boldsymbol{x}_i$, we would like to extract only relevant information from $\mathcal{Z}$. We propose using a transformer cross-attention model to compute a single feature vector from $\mathcal{Z}$. In the first layer, queries are computed from the positional encoding of $\boldsymbol{x}_i$, and keys and values are computed from the elements of $\mathcal{Z}$. In subsequent layers, queries are computed from previous layers' output. We denote the output of the latent querying as latent features $\tilde{\boldsymbol{z}}_i$, each corresponding to a query point $\boldsymbol{x}_i$.

Local Features When synthesizing a novel view given a few input views, a number of existing NeRF-based methods uses 3D projection to determine which features from the context views are useful for explaining a particular 3D point (Yu et al., 2021; Trevithick & Yang, 2021; Wang et al., 2021). LASER-NV incorporates this idea, and in particular follows the design of pixelNeRF (Yu et al., 2021) as it is shown to achieve strong results in a wide range of domains.

We use the encoded input views $\{\boldsymbol{H}^n\}_{n=1}^N$ described above, along with their corresponding camera matrices $\{\boldsymbol{C}^n\}_{n=1}^N$. Given a query 3D point $\boldsymbol{x}_i$ and direction $\boldsymbol{d}_i$, we obtain the point $\boldsymbol{p}_i^n \in \mathbb{R}^2$ in the image space of the nth context view using the known intrinsic parameters. The corresponding local features are then bilinearly interpolated $\boldsymbol{h}_i^n = \boldsymbol{H}^n[\boldsymbol{p}_i^n]$. We refer to $\boldsymbol{h}_i$ as *local features* since they are spatially localized based on the projected coordinates, as opposed to being a global aggregation of all context features. We finally obtain $\hat{\boldsymbol{h}}_i^n$ for each context view by separately processing the features $\boldsymbol{h}_i^n$ using a residual MLP that has as an additional conditioning on the image-space projected points and directions.

Integrating Latents and Local Features In contrast to pixelNeRF, which does not have latents, LASER-NV integrates the both local and the latent features. It does so with a transformer:

$$\boldsymbol{f}_i = M_\theta(\tilde{\boldsymbol{z}}_i; \bigcup_{n=1}^{N} \{\boldsymbol{h}_i^n : \text{is_visible}(\mathbf{x}_i, \mathbf{C}^n)\} \cup \{\tilde{\boldsymbol{z}}_i\}) \tag{1}$$

The attention queries are computed from latent features $\tilde{\boldsymbol{z}}_i$ (see above), and attend to the combined set of local features and to $\tilde{\boldsymbol{z}}_i$. Note that, for a point $\mathbf{x}_i$, we use only those context views from which this point is visible. Using a transformer to process encoded features from multiple views with epipolar constraints has been proposed in prior work on Multi-view Stereo, see e.g. (He et al., 2020; Xi et al., 2022; Wang et al., 2022; Ding et al., 2022).

The final step in evaluating the scene function is to use $\boldsymbol{f}_i$ to condition a residual MLP similar to NeRF-VAE resulting in LASER-NV's scene function $F_\theta(\cdot, \mathcal{Z}, \mathcal{H}) : (\mathbf{x}_i, \mathbf{d}_i) \mapsto \boldsymbol{c}_i, \sigma_i$ which is used to volume-render image outputs.

3.4 Loss

Similarly to NeRF-VAE, our objective function is given by the evidence lower-bound:

$$\mathcal{L}(\mathcal{U}, \mathcal{V}) = \mathbb{E}_{\mathcal{Z}\sim q}\left[\sum_{t=1}^{M} \log p\big(\boldsymbol{I}^t \mid \boldsymbol{C}^t, \mathcal{Z}, \mathcal{V}\big)\right] - \mathrm{KL}(q(\mathcal{Z} \mid \mathcal{U} \cup \mathcal{V}) \,||\, p(\mathcal{Z} \mid \mathcal{V})) \tag{2}$$

where $\boldsymbol{I}^t$ and $\boldsymbol{C}^t$ are the images and cameras from the posterior context views $\mathcal{U}$. We use a Gaussian likelihood term for $p(\boldsymbol{I}^t \mid \boldsymbol{C}^t, \mathcal{Z}, \mathcal{V})$ with a fixed variance.

4 EXPERIMENTS

4.1 DATASETS

We first evaluate LASER-NV on a novel dataset based on a procedurally generated outdoors urban environment. This dataset contains posed images of a virtual city with objects of multiple scales, such as house blocks, roads, buildings, street lights, etc (but no transient objects such as cars or people). As such, it is an extremely challenging scene representation benchmark, especially when representations are inferred from few views only—resulting in partial observability and high uncertainty. In this situation, when evaluated on unobserved parts of a scene, we want a model to produce predictions that are varied, self-consistent, and aligned with the data distribution. Each scene consists of a neighborhood block, and a randomly chosen designated point of interest (road intersection or park) is selected within. Images are then rendered by placing the camera at random nearby positions and viewpoints from the point of interest. The dataset has 100K training scenes, and 3200 test scenes. We provide further details and dataset samples in Appendix B.

To test LASER-NV's ability to produce near-deterministic scene completions in a low-uncertainty setting, we compare it against a number of previous methods on the ShapeNet NMR dataset (Kato et al., 2018), which consists of several views of single objects from 13 different ShapeNet categories. Since one view is enough to largely define ShapeNet objects (unlike scenes in the City), novel-view synthesis in this dataset requires deterministic prediction rather than uncertainty modelling.

4.2 MODELS & EVALUATION

On the City dataset, we compare LASER-NV with Multi-view Conditional NeRF (MVC-NeRF) and NeRF-VAE; see Figure 8 in Appendix A for model diagrams. MVC-NeRF (Figure 8c) is a variant of LASER-NV (Figure 8a) which does not use LASER-NV's latent set representations but instead only relies on deterministically rendering new views using local features. It therefore closely resembles the design of pixelNERF but uses the same architecture for the image encoder and scene function F_θ as LASER-NV to ensure they are comparable. On ShapeNet, we further show quantitative results of a number of non-NeRF novel view synthesis methods: Scene Representation Networks (SRN, (Sitzmann et al., 2019)), Differentiable Volumetric Rendering (DVR, (Niemeyer et al., 2020)), and Scene Representation Transformers (SRT, (Sajjadi et al., 2021)).

During training, LASER-NV uses a random subset of 4 of the 23 target views as posterior context views[1]. MVC-NeRF and NeRF-VAE are trained to directly predict the 23 views given a single context view. At test time, LASER-NV samples the latents from its conditional prior whereas NeRF-VAE uses its posterior (given that its prior is unconditional). Importantly, all models in this evaluation use one input view and predict 23 target views.

At evaluation time, we feed each model with a set of context views, and evaluate *reconstructions* of those same views, and *predictions* of novel views of the same scene. We evaluate reconstruction ability via likelihood estimates, and standard image similarity metrics PSNR, and SSIM (Wang et al., 2004). For novel view predictions we report the Fréchet Inception Distance (FID, (Heusel et al., 2017)) between the marginal distribution of samples and the evaluation dataset[2]. This metric highlights the incapability of deterministic models to generate varied and plausible samples in face of uncertainty. Note that since NeRF-VAE is unconditional, we use the learned approximate posterior as the conditional distribution. In the City, we thus train a separate model (one to use for reconstruction metrics that has 8 posterior context views and another one with 2 views for predictions). For ShapeNet results all models use one context view.

4.3 ABLATIONS

We further ablate LASER-NV to isolate contributions of individual model components. To quantify the effects of conditioning the prior we compare LASER-NV to a conditional version of NeRF-VAE.

[1] While using the 23 views as posterior context would be the standard set up (Bayer et al., 2021), we find that using 4 randomly-sampled views is much cheaper computationally and yet sufficient for training.

[2] Image similarity metrics such as PSNR and SSIM are not useful when there exists a large multimodal space of possible predictions as is the case in the City.

Note that the original NeRF-VAE used an unconditional prior (Kosiorek et al., 2021). We also compare LASER-NV to a version that does not condition the scene function with local context features, cf. Equation (1) and Figure 8b in Appendix A, denoted by LASER-NV (NO GEOM.).

LASER-NV, LASER-NV (NO GEOM.), NeRF-VAE variants, and MVC-NeRF share architectures for the image encoder and the scene function F_θ.

4.4 TRAINING AND IMPLEMENTATION DETAILS

We use the hierarchical sampling technique of (Mildenhall et al., 2020) when rendering a pixel colour, which involves learning two separate sets of scene function parameters as well as optimizing a corresponding colour likelihood term for each step.

For experiments with the City, we leverage ground-truth depth maps using the method proposed by Stelzner et al. (2021) in order to train all the models with fewer scene function evaluations per ray. Not using ground-truth depth significantly increases training time while only marginally lowering reconstruction PSNR, see Appendix C for a discussion and Fig. 14 in Appendix B for visualisations. Importantly, when evaluating our models we revert to volumetric rendering.

All generative models are trained by optimizing the evidence lower bound; where we anneal the KL term in Eq. (2). All experiments use the Adam (Kingma & Ba, 2014) optimizer. We provide full details of model architectures in Appendix A and training procedures in Appendix D.

4.5 RESULTS

Results on the City dataset are shown in Table 1. In terms of reconstructing observed input views, both LASER-NV and MVC-NeRF, due to the use of explicit geometrical knowledge, have the ability to almost directly output the received input, leading to excellent reconstructions. NeRF-VAE, however, needs to reconstruct input views purely from its latent representation, resulting in low fidelity reconstructions. For prediction of unseen views, LASER-NV outperforms both NeRF-VAE and MVC-NeRF in terms of pixel error (higher PSNR, SSIM). At the same time, its output distribution is closer to the actual evaluation data (lower FID), in contrast to MVC-NeRF which outputs a single prediction, and NeRF-VAE which produces low quality outputs that do not vary much. We show representative example reconstructions and predictions in Fig. 2. We also include a `gif`[3] in supp. material showing fly-throughs of additional prior samples.

	Reconstruction			Prediction
	Log-Likelihood ↑	PSNR↑	SSIM↑	FID↓
MVC-NeRF	-	28.63	0.910	91.92
Cond. NeRF-VAE	3.51	23.94	0.709	50.98
NeRF-VAE	3.50	23.92	0.708	52.12
LASER-NV (NO GEOM.)	**3.82**	27.13	0.870	24.19
LASER-NV	**3.83**	**29.61**	**0.923**	**22.54**

Table 1: Results in City. The color likelihood is computed via importance sampling (10 samples); note that MVC-NeRF is deterministic and thus we do not compare it on this metric. **Reconstruction**: We report performance of reconstruction of the 2 context views. **Prediction**: We evaluate the models on previously unobserved views using the FID score. MVC-NeRF provides a single deterministic prediction whereas for the other models we average over 10 independent samples.

Results on ShapeNet are shown in Table 2. Both MVC-NeRF and LASER-NV outperform existing published methods when predicting novel views of these single-object scenes. We further observe low quality predictions of NeRF-VAE in example predictions in Fig. 3. Finally, Fig. 4 shows a hand-selected example where the input view has ambiguity. As in the City experiment, LASER-NV produces plausibly varied predictions of the object in contrast to the other methods.

[3]To create this visualisation we trained the model on a modified City dataset with included birds-eye views. This showcases LASER-NV's ability to represent large environments and features at multiple scale levels.

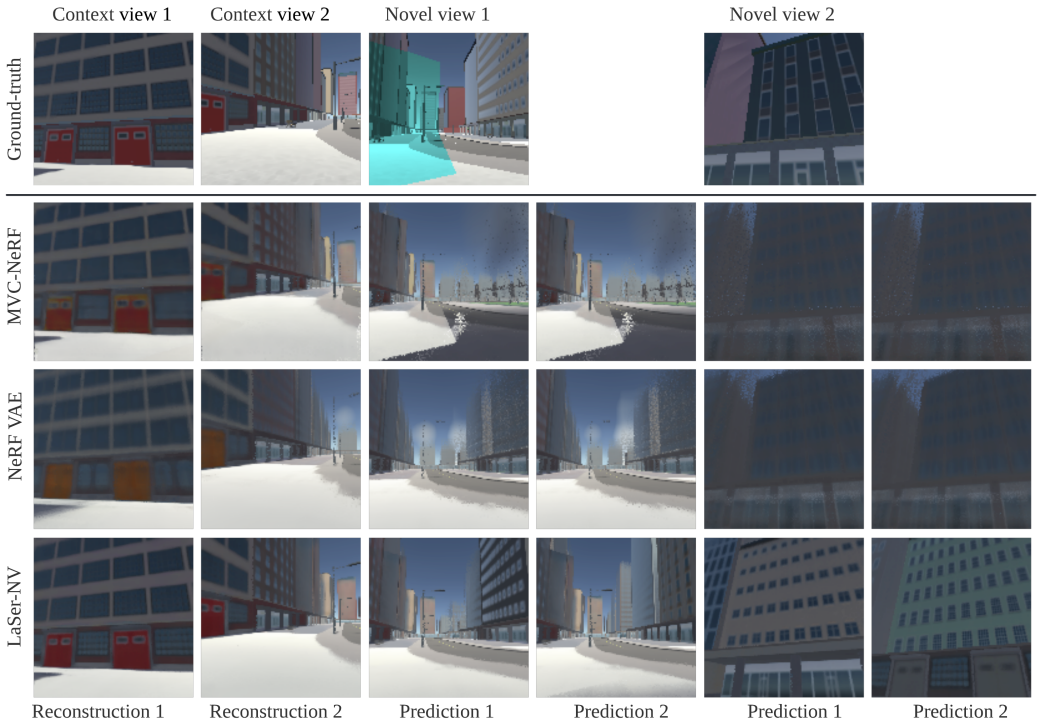


Figure 2: Representative reconstructions of observed views and predictions of partially unobserved novel views. The models are conditioned on two context views (top left). The first two columns show reconstructions of those views. Note how NeRF-VAE struggles to reconstruct both views accurately, while the other models do well. We show predictions of two novel views (top right), where the first view is partially observed in the context (highlighted in cyan color) and the second view is fully unobserved. LASER-NV's predictions are consistent with the observed part of the context, diverse, and plausibly so. NeRF-VAE's predictions lack quality and diversity. MVC-NeRF's only produces a single prediction, which is not very plausible.

	DVR*	SRN*	SRT*	pixelNeRF*	NeRF-VAE	MVC-NeRF	LASER-NV
PSNR ↑	22.70	23.28	27.87	26.80	25.36	**28.03**	27.92
SSIM ↑	0.860	0.849	0.912	0.910	0.875	**0.925**	**0.923**

Table 2: Category-agnostic ShapeNet evaluation by reconstructing 23 views from one input view for all models. We further report numbers from the respective papers for DVR* (Niemeyer et al., 2020), SRN* (Sitzmann et al., 2019), SRT* (Sajjadi et al., 2021), and pixelNeRF* (Yu et al., 2021).

Figure 3: Representative samples of category-agnostic ShapeNet for different ablations of LASER-NV. MVC-NeRF and LASER-NV both result in similar high-fidelity predictions, while the NeRF-VAE's outputs are significantly more blurry.

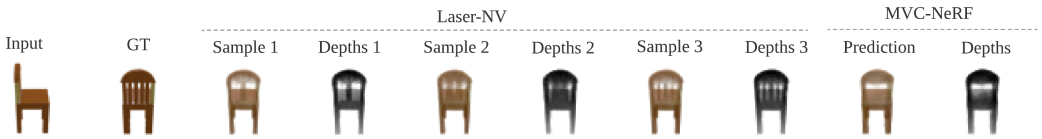
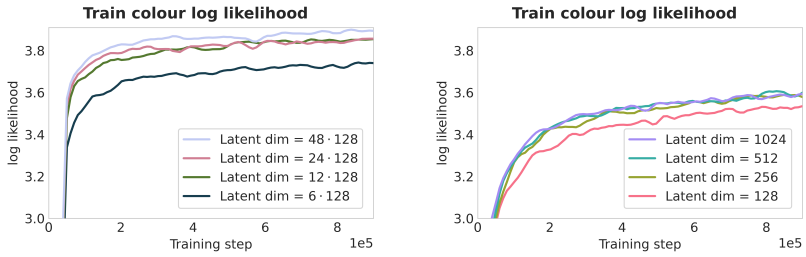


Figure 4: When predicting novel view from an ambiguous input, LASER-NV generates plausible variations, whereas MVC-NERF results in a single, more blurry prediction.



(a) LASER-NV (b) Conditional NERF-VAE

Figure 5: Training reconstruction performance as a function of the capacity of latent representations in City. We compare LASER-NV with an increasing number of latent vectors K (each with 128 dimensions) vs a conditional version of NERF-VAE with increasing dimensionality of its latent vector.

4.6 TRAIN- AND TEST-TIME SCALING & DATA EFFICIENCY

Latent set representations allow trading computation for increased model capacity. We train LASER-NV on the City dataset with increasing latent set sizes K, and compare it to a conditional NERF-VAE also trained with increasing latent dimensions[4]. Figure 5 shows reconstruction log-likelihoods as a function of latent capacity. While LASER-NV consistently improves its reconstructions with larger latent sets, cond. NERF-VAE quickly saturates at 256 latent dimensions, at worse reconstructions compared to LASER-NV. Further, in Appendix C, we provide an analysis of test-time scaling—we show that increasing the latent set size at test time does not result in clear change of performance. Finally, in Figure 11 we show that LASER-NV performs better at test time than NERF-VAE and MVC-NERF for different training dataset sizes.

5 RELATED WORK

5.1 NEURAL SCENE REPRESENTATIONS

Inferring neural scene representations using a deep generative model of scenes is an active research area, and we mention some of the most closely related approaches below.

NERF-VAE learns a generative model of 3D scenes which is only shown to work in relatively simple scenes. When applied to more complicated data (see Fig. 2), it produces blurry reconstructions and novel views. We posit that this is caused by insufficient capacity of its latent representation $\boldsymbol{z}$. Recall that $\boldsymbol{z}$ in NERF-VAE is just a single d-dimensional vector, and the posterior is an independent Gaussian, whose statistics are predicted from the context views.

Generative Query Network (Eslami et al., 2018) similarly learns a distribution over scenes, but its decoder is a convolutional neural network. Subsequent work (Rosenbaum et al., 2018) improves over its representational capacity by using an attentive mechanism when conditioning the decoder from context in an analogous way to how LASER-NV improves over NERF-VAE. However, by implementing a NeRF in our decoder we gain the benefits of more robust generalization capabilities to novel viewpoints compared to the convolutional counterparts (Kosiorek et al., 2021). Simone (Kabra et al., 2021) is a convolutional generative model that performs high quality novel view

[4]Note that this increases the number of NERF-VAE's parameters.

synthesis and while also performing unsupervised segmentation of input videos. The authors in (Dupont et al., 2022) propose directly modelling the underlying radiance fields as functions, which they refer to as 'functa'. More specifically, 'functa' learns a distribution over modulations of the scene function parameters. From this perspective, both our method and theirs learn a prior over scene functions with a model that has a large representational capacity. Bautista et al. (2022) propose a two-stage method to learn a generative model of 3D scenes. As LASER-NV, their model can conditionally sample radiance fields. However, their first stage involves optimizing the embeddings using NeRF on datasets with a large number of views per scene. In contrast, LASER-NV is trained on datasets with sparse views.

5.2 DETERMINISTIC NOVEL VIEW SYNTHESIS

NeRF relies on optimizing the parameters of an MLP scene function, separately for every scene. A number of methods in the literature propose amortizing the mapping of input views to the scene function using multi-view geometry, see e.g. IBRNet (Wang et al., 2021), General Radiance Fields (GRF) (Trevithick & Yang, 2021) and pixelNeRF (Yu et al., 2021). These methods extract features of the context views by using projective geometry to select which features to attend to. These methods are shown to generalize well and efficiently learn to carry out high-fidelity predictions. Each of these models have different strengths, for instance pixelNeRF is designed to work well in a solely view-centered coordinate system. MVSNerf (Chen et al., 2021) proposes an alternative way to leverage multi-view geometry by lifting context features to a 3D representation with strong novel-view synthesis performance and generalization.

Scene Representation Transformers (SRT) (Sajjadi et al., 2021) use global features obtained from context views using a generalization of the Vision Transformer (Dosovitskiy et al., 2020), which bears similarities to LASER-NV. SRT bypasses the use of 3D geometry entirely, yet shows remarkable performance on real world data.

5.3 NERF FOR LARGE-SCALE SCENES

A number of methods have recently been developed with the aim of reconstructing large-scale scenes using NeRF. These works incorporate architectures and optimization techniques that enable scaling up to whole neighborhoods or even city-wide scenes. They work by for example optimizing multiple separate scene functions for separate blocks of a scene, which are then dynamically used when rendering (Tancik et al., 2022; Turki et al., 2021). Another example works by designing a scene function that represents the scene at multiple levels of scale or detail (Martel et al., 2021; Xiangli et al., 2021). Similarly, Mip-NeRF (Barron et al., 2021) allows prefiltering the scene function representation in order to better capture coarse and fine details. These methods are orthogonal to our contributions since they do not learn a conditional NeRF model from a few input views, nor address the uncertainty in predictions due to partial observation of the scenes.

6 CONCLUSION

We propose LASER-NV, a conditional generative model of neural radiance fields capable of efficient inference of large and complex scenes under partial observability conditions. While recent advances in neural scene representation and neural rendering lead to unprecedented novel view synthesis capabilities, producing sharp and multi-modal completions of unobserved parts of the scene remains a challenging problem. We experimentally show that LASER-NV can model scenes of different scale and uncertainty structure, and isolate the usefulness of each contribution through ablation studies.

While NeRF serves as a strong inductive bias for learning 3D structure from images, LASER-NV also inherits some of its drawbacks. Volumetric rendering with neural radiance fields, especially for large-scale scenes, is computationally costly and can limit the model from e.g. real-time rendering. It remains to be seen whether fast NeRF implementations (Müller et al., 2022) can be of help here. Furthermore, accurate ground-truth camera information is required for learning and novel view synthesis; (Moreau et al., 2022) recently made progress on localization methods using NeRF. Finally, an interesting direction for future research is incorporating object-centric structure and dynamics into LASER-NV, which might improve learned representations for downstream reasoning tasks (Sajjadi et al., 2022; Ding et al., 2021).

REFERENCES

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 *International Conference on Computer Vision*. IEEE, 2021.

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.

Justin Bayer, Maximilian Soelch, Atanas Mirchev, Baris Kayalibay, and Patrick van der Smagt. Mind the gap when conditioning amortised inference in sequential latent-variable models. In *International Conference on Representation Learning*, volume abs/2101.07046, 2021.

Christopher M. Bender, Kevin O'Connor, Yang Li, Juan Jose Garcia, Junier Oliva, and Manzil Zaheer. Exchangeable generative models with flow scans. In *AAAI Conference on Artificial Intelligence*. AAAI Press, 2020.

James F. Blinn. Light reflection functions for simulation of clouds and dusty surfaces. In *SIGGRAPH*, 1982.

Christopher P Burgess, Loic Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matt Botvinick, and Alexander Lerchner. MONET: Unsupervised scene decomposition and representation. *arXiv/1901.11390*, 2019.

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 *International Conference on Computer Vision*, volume abs/2103.15595, 2021.

David Ding, Felix Hill, Adam Santoro, Malcolm Reynolds, and Matt Botvinick. Attention over learned object embeddings enables complex visual reasoning. *Advances in Neural Information Processing Systems*, 34, 2021.

Yikang Ding, Wentao Yuan, Qingtian Zhu, Haotian Zhang, Xiangyue Liu, Yuanjiang Wang, and Xiao Liu. Transmvsnet: Global context-aware multi-view stereo network with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. In *International Conference on Representation Learning*, volume abs/1605.08803, 2017.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

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

SM Ali Eslami, Danilo Jimenez Rezende, Frederic Besse, Fabio Viola, Ari S Morcos, Marta Garnelo, Avraham Ruderman, Andrei A Rusu, Ivo Danihelka, Karol Gregor, et al. Neural scene representation and rendering. *Science*, (6394):1204–1210, 2018. 360.

Yihui He, Rui Yan, Katerina Fragkiadaki, and Shoou-I Yu. Epipolar transformers. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 2020.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Rishabh Kabra, Daniel Zoran, Goker Erdogan, Loic Matthey, Antonia Creswell, Matt Botvinick, Alexander Lerchner, and Chris Burgess. Simone: View-invariant, temporally-abstracted object representations via unsupervised video decomposition. *Advances in Neural Information Processing Systems*, 34, 2021.

Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, 2018.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv/1412.6980*, 2014.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

Adam R. Kosiorek, Heiko Strathmann, Daniel Zoran, Pol Moreno, Rosália G. Schneider, Sovna Mokr'a, and Danilo Jimenez Rezende. Nerf-vae: A geometry aware 3d scene generative model. In *International Conference on Machine Learning*, volume abs/2104.00587, 2021.

Julien N. P. Martel, David B. Lindell, Connor Z. Lin, Eric Chan, Marco Monteiro, and Gordon Wetzstein. Acorn: Adaptive coordinate networks for neural scene representation. *ACM Trans. Graph.*, 40:58:1–58:13, 2021.

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 *European Conference on Computer Vision*, 2020.

Arthur Moreau, Nathan Piasco, Dzmitry Tsishkou, Bogdan Stanciulescu, and Arnaud de La Fortelle. Lens: Localization enhanced by nerf synthesis. In *Conference on Robot Learning*. PMLR, 2022.

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, July 2022.

Michael Niemeyer, Lars M. Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 2020.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

Razvan Pascanu, Tomás Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, volume 28 of *JMLR Workshop and Conference Proceedings*. JMLR.org, 2013.

Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International Conference on Machine Learning*, 2015.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, 2014.

Dan Rosenbaum, Frederic Besse, Fabio Viola, Danilo J. Rezende, and S. M. Ali Eslami. Learning models for visual 3d localization with implicit mapping. *CoRR*, abs/1807.03149, 2018.

Mehdi S. M. Sajjadi, Henning Meyer, Etienne Pot, Urs M. Bergmann, Klaus Greff, Noha Radwan, Suhani Vora, Mario Lucic, Daniel Duckworth, Alexey Dosovitskiy, Jakob Uszkoreit, Thomas A. Funkhouser, and Andrea Tagliasacchi. Scene representation transformer: Geometry-free novel view synthesis through set-latent scene representations. abs/2111.13152, 2021.

Mehdi S. M. Sajjadi, Daniel Duckworth, Aravindh Mahendran, Sjoerd van Steenkiste, Filip Paveti'c, Mario Luvci'c, Leonidas J. Guibas, Klaus Greff, and Thomas Kipf. Object scene representation transformer. volume abs/2206.06922, 2022.

Vincent Sitzmann, M. Zollhöfer, and G. Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *Advances in Neural Information Processing Systems*, 2019. arXiv/1906.01618.

Karl Stelzner, Kristian Kersting, and Adam R. Kosiorek. Decomposing 3d scenes into objects via unsupervised volume segmentation. In *arXiv*, volume abs/2104.01148, 2021.

Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P. Srinivasan, Jonathan T. Barron, and Henrik Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. *ArXiv*, abs/2202.05263, 2022.

Alex Trevithick and B. Yang. GRF: Learning a general radiance field for 3d scene representation and rendering. In *International Conference on Learning Representations*, 2021. arXiv/2010.04595.

Haithem Turki, Deva Ramanan, and Mahadev Satyanarayanan. Mega-nerf: Scalable construction of large-scale nerfs for virtual fly-throughs. *CoRR*, abs/2112.10703, 2021.

Arash Vahdat and Jan Kautz. NVAE: A deep hierarchical variational autoencoder. *arXiv/2007.03898*, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 30.

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

Xiaofeng Wang, Zheng Zhu, Fangbo Qin, Yun Ye, Guan Huang, Xu Chi, Yijia He, and Xingang Wang. Mvster: Epipolar transformer for efficient multi-view stereo. *arXiv preprint arXiv:2204.07346*, 2022.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, 2004.

Peter Wirnsberger, Andrew J. Ballard, George Papamakarios, Stuart Abercrombie, Sébastien Racanière, Alexander Pritzel, Danilo Jimenez Rezende, and Charles Blundell. Targeted free energy estimation via learned mappings. *The Journal of Chemical Physics*, 2020.

Peter Wirnsberger, George Papamakarios, Borja Ibarz, Sébastien Racanière, Andrew J Ballard, Alexander Pritzel, and Charles Blundell. Normalizing flows for atomic solids. *arXiv preprint arXiv:2111.08696*, 2021.

Junhua Xi, Yifei Shi, Yijie Wang, Yulan Guo, and Kai Xu. Raymvsnet: Learning ray-based 1d implicit fields for accurate multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Yuanbo Xiangli, Linning Xu, Xingang Pan, Nanxuan Zhao, Anyi Rao, Christian Theobalt, Bo Dai, and Dahua Lin. Citynerf: Building nerf at city scale. *ArXiv*, abs/2112.05504, 2021.

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 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 2021.