CARFF: Conditional Auto-encoded Radiance Field for 3D Scene Forecasting

Jiezhi "Stephen" Yang[1]*, Khushi Desai[2]*, Charles Packer[3], Harshil Bhatia[4], Nicholas Rhinehart[3], Rowan McAllister[5], and Joseph E. Gonzalez[3]

[1] Harvard University MA 02138, USA
[2] Columbia University NY 10025, USA
[3] UC Berkeley CA 94720, USA
[4] Avataar.ai KA 560103, India
[5] Toyota Research Institute CA 94022, USA

Abstract. We propose CARFF: Conditional Auto-encoded Radiance Field for 3D Scene Forecasting, a method for predicting future 3D scenes given past observations. Our method maps 2D ego-centric images to a distribution over plausible 3D latent scene configurations and predicts the evolution of hypothesized scenes through time. Our latents condition a global Neural Radiance Field (NeRF) to represent a 3D scene model, enabling explainable predictions and straightforward downstream planning. This approach models the world as a POMDP and considers complex scenarios of uncertainty in environmental states and dynamics. Specifically, we employ a two-stage training of Pose-Conditional-VAE and NeRF to learn 3D representations, and auto-regressively predict latent scene representations utilizing a mixture density network. We demonstrate the utility of our method in scenarios using the CARLA driving simulator, where CARFF enables efficient trajectory and contingency planning in complex multi-agent autonomous driving scenarios involving occlusions. Video and code are available at www.carff.website.

1 Introduction

Humans often imagine what they cannot see given partial visual context. Consider a scenario where reasoning about the unobserved is critical to safe decision-making: for example, a driver navigating a blind intersection. An expert driver will plan according to what they believe may or may not exist in occluded regions of their vision. The driver's belief – defined as the understanding of the world modeled with consideration for inherent environment uncertainties – is informed by their partial observations (i.e., the presence of other vehicles on the road), as well as their prior knowledge (e.g., past experience navigating this intersection).

When reasoning about the unobserved, humans form complex beliefs about the existence, position, shapes, colors, and textures of occluded scene portions

* Core contributors

(e.g., an oncoming car). Autonomous systems with high-dimensional sensor data, like video or LiDAR, traditionally reduce this data to low-dimensional state information (e.g., position and velocity of tracked objects) for prediction and planning.

In addition to tracking fully observed objects, this object-centric framework handles partially observed settings by considering potentially dangerous unobserved objects. These systems often plan for worst-case scenarios, such as a "ghost car" at the edge of the visible field of view [45].

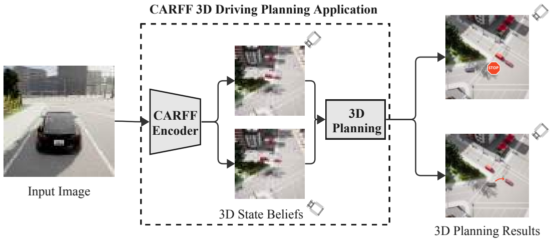


Fig. 1: CARFF 3D planning application for driving. An input image containing a partially observable view of an intersection is processed by CARFF's encoder to establish 3D environment state beliefs, i.e. the predicted possible state of the world: whether or not there could be another vehicle approaching the intersection. These beliefs are used to forecast the future in 3D for planning, generating one among two possible actions for the vehicle to merge into the other lane.

Recent advances in neural rendering, particularly Neural Radiance Fields (NeRF), have significantly improved 3D scene representation learning. NeRF enables novel view synthesis, thus simplifying the process of viewing behind occlusions. NeRF decouples the dependancy of scene representation from traditional object detection and tracking, allowing for the capture of vital visual information that might be missed by detectors, yet is crucial for safe decision-making. NeRF's implicit density representation of explicit geometry also facilitates its direct application in motion planning without the need for rendering. NeRF's ability to represent both visual and geometric information makes them a more general and intuitive 3D representation for autonomous systems.

Despite NeRF's advantages, achieving probabilistic predictions in 3D based on reasoning from occluded views is challenging. For example, discriminative models that yield categorical predictions are unable to capture the underlying 3D structure, impeding their ability to model uncertainty. While prior work on 3D representation captures view-invariant structures, their application is primarily confined to simple scenarios [20]. We present CARFF, which to our knowledge, is the first forecasting approach in scenarios with partial observations that uniquely facilitates stochastic predictions in a partially observable Markov decision process (POMDP) within a 3D representation, effectively integrating visual perception and geometry. Specifically, we make the following contributions:

1. We propose a novel architecture *PC-VAE*: Pose-Conditioned Variational Autoencoder. The encoder maps potentially partially observable ego-centric images to pose-invariant latent scene representations, which hold state beliefs of the POMDP with implicit probability distributions (see Sec. 3.1).
2. We develop the two-stage training pipeline that uniquely enables complex scene modeling with a probabilistic objective. This involves separately training the PC-VAE and a latent conditioned neural radiance field that functions as a 3D decoder, enabling interpretable predictions (see Sec. 3.1).
3. We design a mixture density model to predict the evolution of 3D scenes over time stochastically and regressively in the encoder belief space (see Sec. 3.2). This allows for an effective sampling based-controller to output actions in the POMDP.

We demonstrate how CARFF can be used to enable contingency planning in complex driving scenarios that require reasoning into visual occlusions on CARLA simulated datasets inspired by autonomous driving planning tasks [29, 56, 57, 58]. A potential application of CARFF is illustrated in Fig. 1.

2 Related work

2.1 NeRF and 3D representations

Neural radiance fields. Neural Radiance Fields (NeRF) [2, 27, 44] for 3D representations generate high-resolution, photorealistic scenes. Instant Neural Graphics Primitive (Instant-NGP) [28] speeds up training and rendering time by introducing a multi-resolution hash encoding. Other works like Plenoxels [12] and DirectVoxGo (DVGO) [42] also provide similar speedups. Recent advancements in volumetric representations such as 3D Gaussian Splatting [18] enhance rendering efficiency while maintaining compatibility with traditional NeRF applications [11]. We utilize Instant-NGP for its accessibility, although our approach is adaptable to alternative rendering methods. NeRFs have also been extended for several tasks such as modeling large-scale unbounded scenes [2, 43, 50], scene from sparse views [7, 39, 49] and multiple scenes [20, 51]. For an in-depth survey on neural representation learning and its applications we refer the reader to [46].

Generalizable novel view synthesis models, like pixelNeRF and pixelSplat [5, 55], learn a scene prior to render novel views from sparse existing ones. In contrast, CARFF is based on a VAE, encoding a probabilistic objective and decoding to future 3D scenes. Dynamic NeRF models scenes with moving or deforming objects, within which a widely used approach is to construct a canonical space and predict a deformation field [22, 33, 34, 36]. The canonical space is usually a static scene, and the model learns an implicitly represented flow field [33, 36]. A recent line of work also models dynamic scenes via different representations and decomposition [3, 41]. These approaches tend to perform better for spatially bounded and predictable scenes with relatively small variations [3, 23, 33, 55]. Moreover, these methods only solve for changes in the environment but are limited in incorporating stochasticity in the environment.

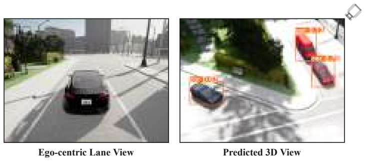


Fig. 2: Novel view planning application. CARFF allows reasoning behind occluded views from the ego car as simple as moving the camera to see the sampled belief predictions, allowing simple downstream planning using, for example, density probing or 2D segmentation models from arbitrary angles.

Multi-scene NeRF: Our approach builds on multi-scene NeRF approaches [20, 48, 51, 52] that learn a global latent scene representation, which conditions the NeRF, allowing a single NeRF to effectively represent various scenes. A similar method, NeRF-VAE, was introduced by Kosiorek *et al.* [20] to create a geometrically consistent 3D generative model with generalization to out-of-distribution cameras. However, NeRF-VAE [20] is prone to mode collapse when faced with complex visual information (see Sec. 4.2).

2.2 Scene Forecasting

Planning in 2D space: Planning in large, continuous state-action spaces is challenging due to exponentially large search spaces [32], leading to various approximation methods for tractability [26, 35]. Model-free [13, 31, 47] and model-based [4] reinforcement learning frameworks, along with other learning-based methods [6, 29], have emerged as viable approaches. Additionally, methods forecast for downstream control [16], learn behavior models for contingency planning [38], or predict the existence and intentions of unobserved agents [30]. While these methods operate in 2D, we reason under partial observations and account for these factors in 3D.

NeRF in robotics: Recent works have applied NeRFs in robotics for localization [54], navigation [1, 25], dynamics modeling [10, 22],and robotic grasping [15, 19]. Adamkiewicz *et al.*[1] propose quadcopter motion planning in NeRF models by sampling the learned density function, useful for forecasting and planning. Driess *et al.*[10] employ a graph neural network to learn dynamics in a multi-object NeRF scene. Li *et al.* [21] focus on pushing tasks and address grasping and planning with NeRF and a separate latent dynamics model. Prior approaches work in simple, static scenes [1] or uses deterministic dynamics models [21]. CARFF addresses complex, realistic environments with both state and dynamics uncertainty, considering potential object existence and unknown movements.

3 Method

Recent advancements in 3D scene representation allow for modeling environments in a contextually rich and interactive 3D space. This offers analytical benefits, such as spatial analysis with soft occupancy grids and object detection through novel view synthesis. Given these advantages, our primary objective is to develop a model for probabilistic 3D scene forecasting in dynamic environments. However, direct integration of 3D scene representation via NeRF and probabilistic models like VAE often involves non-convex and inter-dependent optimization, which causes unstable training. For instance, NeRF's optimization may rely on the VAE's latent space being structured to provide informative gradients. To navigate these complexities, our method bifurcates the training process into two stages (see Fig. 3). First, we train the PC-VAE to learn view-invariant scene representations. Next, we replace the decoder with a NeRF to learn a 3D scene from the latent representations. The latent scene representations capture the environmental states and dynamics over possible underlying scenes, while NeRF synthesizes novel views within the belief space, giving us the ability to see the unobserved (see Fig. 2 and Sec. 3.1). During prediction, uncertainties can be modeled by sampling latents auto-regressively from a predicted Gaussian mixture, allowing for effective decision-making. To this extent, we approach scene forecasting as a POMDP over latent distributions, which enables us to capture multi-modal beliefs for planning amidst perceptual uncertainty (see Sec. 3.2).

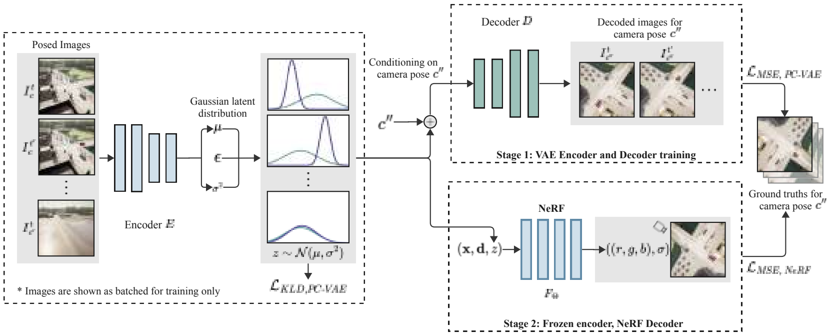

Posed Images
I_c^t
$I_c^{t'}$
$I_{c'}^t$
Encoder E
Gaussian latent distribution
μ
ϵ
σ^2
$z \sim \mathcal{N}(\mu, \sigma^2)$
$\mathcal{L}_{KLD,PC\text{-}VAE}$
* Images are shown as batched for training only
Conditioning on camera pose c''
c''
Decoder D
Decoded images for camera pose c''
$I_{c''}^t$
$I_{c''}^{t'}$
Stage 1: VAE Encoder and Decoder training
$\mathcal{L}_{MSE,\ PC\text{-}VAE}$
Ground truths for camera pose c''
NeRF
$(\mathbf{x}, \mathbf{d}, z)$
$((r, g, b), \sigma)$
F_Θ
Stage 2: Frozen encoder, NeRF Decoder
$\mathcal{L}_{MSE,\ NeRF}$

Fig. 3: Visualizing CARFF's two stage training process. **Left:** The convolutional VIT-based encoder encodes each image I at timestamps t, t' and camera poses c, c' into Gaussian latent distributions. Assuming two timestamps and an overparameterized latent, one Gaussian distribution will have a smaller σ^2, and different μ across timestamps. **Upper Right:** The pose-conditional decoder stochastically decodes the sampled latent z using the camera pose c'' into images $I_{c''}^t$ and $I_{c''}^{t'}$. The decoded reconstruction and ground truth images are used for the loss $\mathcal{L}_{\text{MSE, PC-VAE}}$. **Lower Right:** A NeRF is trained by conditioning on the latent variables sampled from the optimized Gaussian parameters. These parameters characterize the distinct timestamp distributions derived from the PC-VAE. An MSE loss is calculated for NeRF as $\mathcal{L}_{\text{MSE, NeRF}}$.

3.1 Pose-Conditional VAE (PC-VAE) and NeRF

Architecture: We assume that the model follows a Markovian process, and thus each belief state only depends on the previous. Given a scene S_t at timestamp t, we have an ego-centric observation image I_c^t captured from camera pose c. The objective is to formulate a 3D representation of the image that holds implicit probability distributions of the possible states, where we can perform a forecasting step that evolves the scene forward. Here, the POMDP can be seen as an MDP in belief space [17]. To achieve this, we utilize a radiance field conditioned on latent variable z sampled from the posterior distribution $q_\phi(z|I_c^t)$. Now, to learn the posterior, we utilize PC-VAE. We construct an encoder using convolutional layers and a pre-trained ViT on ImageNet [8]. The encoder learns a mapping from the image space to a Gaussian distributed latent space $q_\phi(z|I_c^t) = \mathcal{N}(\mu, \sigma^2)$ parametrized by mean μ and variance σ^2. The decoder, $p(I|z,c)$, conditioned on camera pose c, maps the latent $z \sim \mathcal{N}(\mu, \sigma^2)$ into the image space I. This helps the encoder to generate latents that are invariant to the camera pose c.

To enable 3D scene modeling, we employ Instant-NGP [28], which incorporates a hash grid and an occupancy grid to enhance computation efficiency. Additionally, a smaller multilayer perceptron (MLP), $F_\theta(z)$ can be utilized to model the density and appearance, given by:

$$F_\theta(z) : (\mathbf{x}, \mathbf{d}, z) \rightarrow ((r, g, b), \sigma) \tag{1}$$

Here, $\mathbf{x} \in \mathbb{R}^3$ and $\mathbf{d} \in (\theta, \phi)$ represent the location vector and the viewing direction respectively. The MLP is conditioned on the sampled scene latents $z \sim q_\phi(z|I_c^t)$ (see Appendix B).

Training methodology: The architecture alone does not enable us to model complex scenarios, as seen through a similar example in NeRF-VAE [20]. A crucial contribution of our work is our two-stage training framework which stabilizes the training. First, we optimize the convolutional ViT based encoder and pose-conditional convolutional decoder in the pixel space for reconstruction. This enables our method to deal with more complex and realistic scenes as the encoding is learned in a semantically rich 2D space. By conditioning the decoder on camera poses, we achieve disentanglement between camera view angles and scene context, making the representation view-invariant and the encoder 3D-aware. Once rich latent representations are learned, we replace the decoder with a latent-conditioned NeRF over the latent space of the frozen encoder. The NeRF reconstructs encoder beliefs in 3D for novel view synthesis.

Loss: PC-VAE is trained using standard VAE loss, with mean square error (MSE) and a Kullback–Leibler (KL) divergence given by evidence lower bound:

$$\begin{aligned}\mathcal{L}_{\mathit{PC\text{-}VAE}} &= \mathcal{L}_{\mathit{MSE,\ PC\text{-}VAE}} + \mathcal{L}_{\mathit{KLD,\ PC\text{-}VAE}} = \\ &||p(I|z,c'') - I_{c''}^t||^2 + \mathbb{E}_{q(z|I_c^t)}[\log p(I|z)] - w_{KL} D_{KL}(q_\phi(z|I_c^t) \,||\, p(I|z))\end{aligned} \tag{2}$$

where w_{KL} denotes the KL divergence loss weight and $z \sim q_\phi(z|I_c^t)$. To make our representation 3D-aware, our posterior is encoded using camera c while the decoder is conditioned on a randomly sampled pose c''.

KL divergence regularizes the latent space to balance conditioned reconstruction and stochasticity under occlusion. An elevated KL divergence loss weight w_{KL} pushes the latents closer to a standard normal distribution, $\mathcal{N}(0,1)$, thereby ensuring probabilistic sampling in scenarios under partial observation. However, excessive regularization causes the latents to be less separable, leading to mode collapse. To mitigate this, we adopt delayed linear KL divergence loss weight scheduling to strike a balanced w_{KL}.

Next, we learn a NeRF decoder on the posterior of the VAE to model scenes. At any timestamp t we use a standard photometric loss for training the NeRF, given by the following equation:

$$\mathcal{L}_{MSE,\ NeRF} = \|I_c^t - \mathit{render}(F_\theta(\cdot|q_\phi(z|I_c^t)))\|^2 \tag{3}$$

We use a standard rendering algorithm as proposed by Müller *et al.* [28]. Next, we build a forecasting module over the learned latent space of our pose-conditional encoder.

3.2 Scene Forecasting

Formulation: The current formulation allows us to model scenes with different configurations across timestamps. In order to forecast future configurations of a scene given an ego-centric view, we need to predict future latent distributions. We formulate the forecasting as a POMDP over the posterior distribution $q_\phi(z|I_c^t)$ in the PC-VAE's latent space.

During inference, we observe stochastic behaviors under occlusion, which motivates us to learn a mixture of several Gaussian distributions that potentially denote different scene possibilities. Therefore, we model the POMDP using a Mixture Density Network (*MDN*), with multi-headed MLPs, that predicts a mixture of K Gaussians. At any timestamp t the distribution is given as:

$$q'_\phi(z_t|I_c^{t-1}) = MDN(q_\phi(z_{t-1}|I_c^{t-1})) \tag{4}$$

The model is conditioned on the posterior distribution $q_\phi(z_{t-1})$ to learn a predicted posterior distribution $q'_\phi(z_t|I_c^{t-1})$ at each timestamp. The predicted posterior distribution is given by the mixture of Gaussian:

$$q'_\phi(z_t) = \sum_{i=1}^{K} \pi_i\, \mathcal{N}(\mu_i, \sigma_i^2) \tag{5}$$

here, π_i, μ_i, and σ_i^2 denote the mixture weight, mean, and variance of the i^{th} Gaussian distribution within the posterior distribution. Here, K is the total number of Gaussians. For brevity we remove their conditioning on the posterior $q_\phi(z_{t-1})$ and sampled latent z_{t-1}. We sample z_t from the mixture of Gaussians $q'_\phi(z_t)$, where z_t likely falls within one of the Gaussian modes. The configuration corresponding to the mode is reflected in the 3D scene rendered by NeRF.

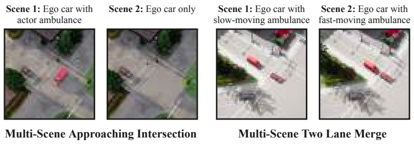

Scene 1: Ego car with actor ambulance
Scene 2: Ego car only
Scene 1: Ego car with slow-moving ambulance
Scene 2: Ego car with fast-moving ambulance
Multi-Scene Approaching Intersection
Multi-Scene Two Lane Merge

Fig. 4: Multi-scene CARLA datasets. Varying car configurations and scenes for the Multi-Scene Two Lane Merge dataset (**left**) and the Multi-Scene Approaching Intersection dataset (**right**).

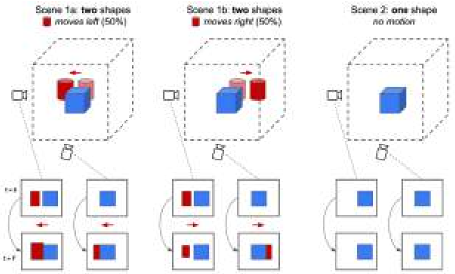


Fig. 5: Blender dataset. Blender dataset with a blue cube and a potential red cylinder exhibiting probabilistic temporal movement. The possible occlusions from different camera angles demonstrate how movement needs to be modeled probabilistically.

Loss: To optimize the MDN, we minimize a negative log-likelihood function, given by:

$$\mathcal{L}_{MDN} = -\sum_{j=1}^{N} log \left(\sum_{i=1}^{K} \pi_i \mathcal{N}(y_j; \mu_i, \sigma_i^2) \right) \tag{6}$$

where $y_i \sim q_\phi(z_t)$ is sampled from the distribution of latent z_t, learned by the encoder, and N denotes the total number of samples.

Inference: We consider an unseen ego-centric image and retrieve its posterior $q_\phi(z_t)$ through the encoder. Next, we predict the possible future posterior distribution $q'_\phi(z_{t+1})$. From the predicted posterior, we sample a scene latent and perform localization. We achieve this via (a) density probing the NeRF or (b) segmenting the rendered novel views using off-the-shelf methods such as YOLO [37] (see Fig. 2). These allow us to retrieve a corresponding Gaussian distribution $q_\phi(z_{t+1})$ in encoder latent space. This is auto-regressively fed back into the MDN to predict the next timestamp. See Fig. 6 for an overview of the pipeline.

4 Results

Decision-making under perceptual uncertainty is a pervasive challenge faced in robotics and autonomous driving, as the real environment is mostly likely partially observable, making it a POMDP. In a partially observable driving scenario, accurate inference regarding the presence of potentially obscured agents is pivotal. We evaluate the effectiveness of CARFF on common driving situations with partial observability and added complexity. We implemented several scenarios in the CARLA driving simulator [9] (see Fig. 4). A single NVIDIA RTX 3090 GPU is used to train PC-VAE, NeRF, and the MDN. All models, trained sequentially, tend to converge within a combined time frame of 24 hours. A detailed experimental setup can be found in Appendix B. We show that, given partially observable 2D inputs, CARFF performs well in predicting latent distributions that represent complete 3D scenes. Using these predictions we design a CARFF-based controller for performing downstream planning tasks.

4.1 Data Generation

We conduct experiments on (a) synthetic blender dataset for principle experiments to test the probabilistic modeling capacities in isolation of the vision encoder (it is visually as simple as possible, but requires the full predictive model proposed in CARFF) and (b) CARLA-based driving datasets for more complex driving scenarios [9]. To deliver convincing results, we model these driving scenarios off of related works [29, 56, 57, 58] that concern planning for driving under difficult situations. We generate the datasets in 3D by programming an ego object and varying actor objects in different configurations.

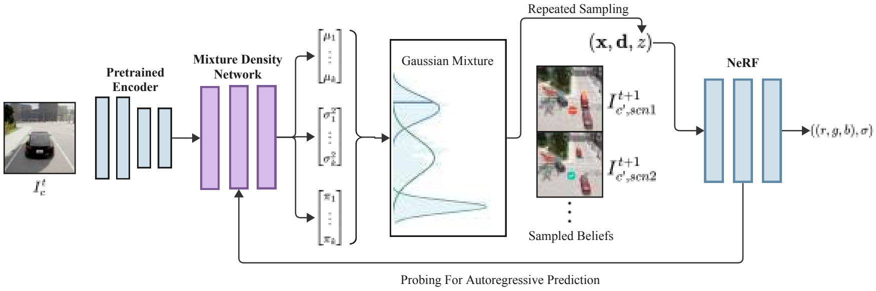


Fig. 6: Auto-regressive inference in scene prediction. The input image at timestamp t, I_c^t, is encoded using the pre-trained encoder from PC-VAE. The corresponding latent distribution is fed into the Mixture Density Network, which predicts a mixture of Gaussians. Each of the K Gaussians is a latent distribution that may correspond to different beliefs at the next timestamp. The mixture of Gaussians is sampled repeatedly for the predicted latent beliefs, visualized as $I_{c',scni}^{t+1}$, representing potentially the ith possible outcome. This is used to condition the NeRF to generate 3D views of the scene. To accomplish autoregressive predictions, we probe the NeRF for the location of the car and feed this information back to the pre-trained encoder to predict the scene at the next timestamp.

Blender synthetic dataset: This comprises of a stationary blue cube (ego) accompanied by a red cylinder (actor) that may or may not be present (see Fig. 5). If the actor is present, it exhibits lateral movement as depicted in Fig. 5. This simplistic setting provides an interpretable framework to evaluate our model.

CARLA dataset: Each dataset is simulated for N timestamps and uses $C = 100$ predefined camera poses to capture images of the environment under full observation, partial observation, and no visibility. These datasets are modeled after common driving scenarios involving state uncertainty that have been proposed in related works such as Active Visual Planning [29].

a) Single-Scene Approaching Intersection: The ego vehicle is positioned at a T-intersection. An actor vehicle traverses the crossing along an evenly spaced, predefined trajectory. We simulate this for $N = 10$ timestamps. We mainly use this dataset to predict the evolution of timestamps under full observation.

b) Multi-Scene Approaching Intersection: We extend the previous scenario to a more complicated setting with state uncertainty, by making the existence of the actor vehicle probabilistic. A similar intersection crossing is simulated for $N = 3$ timestamps for both possibilities. The ego vehicle's view of the actor may be occluded as it approaches the T-intersection over the N timestamps. The ego vehicle either moves forward or halts at the junction (see Fig. 4).

c) Multi-Scene Multi-actor Two Lane Merge: To add more environment dynamics uncertainty, we consider a multi-actor setting at an intersection of two merging lanes. We simulate the scenario at an intersection with partial occlusions, with the second approaching actor having variable speed. Here the ego vehicle can either merge into the left lane before the second actor or after all the actors pass, (see Fig. 4). Each branch is simulated for $N = 3$ timestamps.

4.2 CARFF Evaluation

Method	3D	Complex Scenarios	State Uncertainty	Dynamics Uncertainty	Prediction	Planning	Code Released
CARFF	✓	✓	✓	✓	✓	✓	✓
[20]	✓		✓				
[21]	✓	✓			✓	✓	
[1]	✓	✓				✓	✓
[29]		✓	✓	✓	✓	✓	

Table 1: Qualitative comparison of CARFF to related works. CARFF accomplishes all highlighted objectives as opposed to NeRF-VAE [20], NeRF for Visuomotor Control [21], Vision-only NeRF Navigation [1], and AVP [29]. We compare whether methods reason in a 3D environment and perform novel view synthesis; work on complex scenarios; predict probabilistically under state and dynamics uncertainty; forecast into the future; and use model predictions for decision-making.

A desirable behavior from our model is that it should predict a complete set of possible scenes consistent with the given ego-centric image, which could be partially observable. This is crucial for autonomous driving in unpredictable environments as it ensures strategic decision-making based on potential hazards. To achieve this we require a rich PC-VAE latent space, high-quality novel view synthesis, and auto-regressive probabilistic predictions of latents at future timestamps. We evaluate CARFF on a simple synthetic blender-based dataset and each CARLA-based dataset. Additionally, we extend our model application to a hand-manipulation dataset in Appendix A.

Comparisons with related work: We attempt to compare CARFF to existing approaches. NeRF-VAE has the most comparable objective, but during our experiments, it collapse to black using CARLA datasets. We make further qualitative comparisons to other most similar methods in Tab. 1, but none aligns with ours enough to make any possible quantitative comparisons.

Evaluation on blender dataset: In Fig. 5, for both Scene 1a and 1b, our model correctly forecasts the lateral movement of the cylinder to be in either position approximately 50% of the time, considering a left viewing angle. In Scene 2, with the absence of the red cylinder in the input camera angle, the model predicts the potential existence of the red cylinder approximately 50% of the time, and predicts lateral movements with roughly equal probability. This validates PC-VAE's ability to predict and infer occlusions in the latent space, aligning with human intuitions. These intuitions, shown in the Blender dataset's simple scenes, can transfer to driving scenarios in our CARLA datasets.

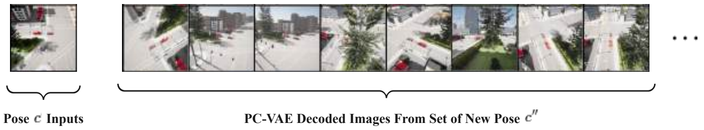


Fig. 7: PC-VAE reconstructions. The encoder input, I_c^t, among the other ground truth images I_c viewed from camera pose c at different timestamps, is reconstructed across a new set of poses c'' respecting timestamp t, generating $I_{c''}^t$. A complete grid is in Appendix D.

PC-VAE performance and ablations: We evaluate the performance of PC-VAE on CARLA datasets with multiple encoder architectures. We show that PC-VAE effectively reconstructs complex environments involving variable scenes, actor configurations, and environmental noise given potentially partially observable inputs (see Fig. 9). We calculated an average Peak Signal-to-Noise Ratio (PSNR) over the training data, as well as novel view encoder inputs. To evaluate the quality of the latent space generated by the encoder, we utilize t-SNE [24] plots to visualize the distribution of latent samples for each image in a given dataset (see Appendix D). We introduce a Support Vector Machine (SVM) [14] based metric to measure the visualized clustering quantitatively, where a higher value indicates better clustering based on timestamps. Most latent scene samples are separable by timestamps, which indicates that the latents are view-invariant. Samples that are misclassified or lie on the boundary usually represent partially or fully occluded regions. This is desirable for forecasting, as it enables us to

Ground Truth Prediction Pair	Avg. PSNR (Scene 1)	Avg. PSNR (Scene 2)
Single-Scene Intersection		
Matching Pairs	**29.06**	N.A
Un-matching P.	24.01	N.A
Multi-Scene Intersection		
Matching Pairs	**28.00**	**28.26**
Un-matching P.	23.27	24.56
Multi-Scene Two Lane Merge		
Matching Pairs	**28.14**	**28.17**
Un-matching P.	22.74	23.32

Table 2: Averaged PSNR for fully observable 3D predictions. CARFF correctly predicts scene evolution across all timestamps for each dataset. The average PSNR is high for predictions $\hat{I_{t_i}}$ and matching ground truths, I_{t_i}. PSNR values for incorrect correspondences, $\hat{I_{t_i}}, I_{t_j}$, is a result of matching surroundings. See complete table in Appendix D.

Multi-Scene Intersection		
Controller Type	Actor	No Actor
Underconfident	30/30	0/30
Overconfident	0/30	30/30
CARFF (n=2)	17/30	30/30
CARFF (n=10)	**30/30**	**30/30**
CARFF (n=35)	30/30	19/30
Multi-Scene Two Lane Merge		
Controller Type	Fast	Slow
Underconfident	30/30	0/30
Overconfident	0/30	30/30
CARFF (n=2)	21/30	30/30
CARFF (n=10)	**30/30**	**30/30**
CARFF (n=35)	30/30	22/30

Table 3: Planning in 3D with controllers with varying sampling numbers n. CARFF-based controllers outperform baselines in success rate over 30 trials. For $n = 10$, the CARFF-based controller consistently chooses the optimal action in potential collision scenarios.To maintain consistency, we use one single image input across 30 trials.

model probabilistic behavior over these samples. In this process, balancing KL divergence weight scheduling maintains the quality of the PC-VAE's latent space and reconstructions (see Appendix B). Additionally, we substantiate the benefits of our PC-VAE encoder architecture through our ablations (see Appendix D.3).

3D novel view synthesis: Given an unseen ego-centric view with potentially partial observations, our method maintains all possible current state beliefs in 3D, and faithfully reconstructs novel views from arbitrary camera angles for each belief. Fig. 2 illustrates one of the possible 3D beliefs that CARFF holds. This demonstrates our method's ability to generate 3D beliefs that could be used for novel view synthesis in a view-consistent manner. Our model's ability to achieve accurate and complete 3D environmental understanding is important for applications like prediction-based planning.

Inference under full and partial observations: Under full observation, we use MDN to predict the subsequent car positions in all three datasets. PSNR values are calculated based on bird-eye view NeRF renderings and ground truth bird-eye view images of the scene across different timestamps. In Tab. 2 we report the PSNR values for rendered images over the predicted posterior with the ground truth images at each timestamp. We also evaluate the efficacy of our prediction model using the accuracy curve given in Fig. 8. This represents CARFF's ability to generate stable beliefs, without producing incorrect predictions, based on actor(s) localization results. For each number of samples between $n = 0$ to $n = 50$, we choose a random subset of 3 fully observable ego images and take an average of the accuracies. In scenarios with partial observable ego-centric images where several plausible scenarios exist, we utilize recall instead of accuracy using a similar setup. This lets us evaluate the encoder's ability to avoid false negative predictions of potential danger.

Fig. 8 shows that our model achieves high accuracy and recall in both datasets, demonstrating the ability to model state uncertainty (Approaching Intersection) and dynamic uncertainty (Two Lane Merge). The results indicate CARFF's resilience against randomness in resampling, and completeness in probabilistic modeling of the belief space. Given these observations, we now build a reliable controller to plan and navigate through complex scenarios.

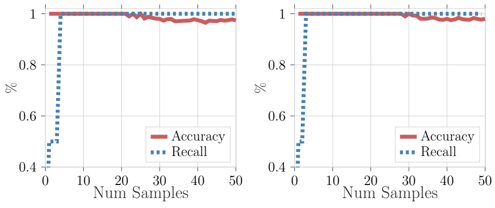
4.3 Planning

In all our experiments, the ego vehicle must make decisions to advance under certain observability. The scenarios are designed such that the ego views contain partial occlusion and the state of the actor(s) is uncertain in some scenarios.

In order to facilitate decision-making using CARFF, we design a controller that takes ego-centric input images and outputs an action. Decisions are made incorporating sample consistency from the mixture density network. For instance, the controller infers occlusion and promotes the ego car to pause when scenes alternate between actor presence and absence in the samples. We use the two multi-scene datasets to assess the performance of the CARFF-based controller as they contain actors with potentially unknown behaviors.

To design an effective controller, we need to find a balance between accuracy and recall (see Fig. 8). A lowered accuracy from excessive sampling means unwanted randomness in the predicted state. However, taking insufficient samples would generate low recall i.e., not recovering all plausible states. This would lead to incorrect predictions as we would be unable to account for the plausible uncertainty present in the environment. To achieve optimal balance, we designed an open-loop planning controller using a sampling strategy that generates $n = 2, 10, 35$ samples. The hyperparameter n is tuned per scene for peak performance but is expected to remain relatively stable across scenes. We demonstrate that $n = 10$ performs well consistently in varying CARLA scenarios (Fig. 8) and do not anticipate this being very different for other experiments.

For sampling values that lie on the borders of the accuracy and recall margin, for example, $n = 2$ and 35, we see that the CARFF-based controller obtains lower success rates, whereas $n = 10$ produces the best result. For actor exists and fast-actor scenes in Tab. 3, we consider occluded ego-centric inputs to test the controller's ability to avoid collisions. For no-actor and slow-actor scenes, we consider state observability and test the controllers' ability to recognize the optimal action to advance. Across the two datasets, the overconfident controller



(a) Intersection **(b)** Lane Merge

Fig. 8: Multi-Scene dataset accuracy and recall curves from predicted beliefs. We test our framework across $n = 1$ and $n = 50$ samples from MDN's predicted latent distributions from ego-centric image input. Across the number of samples n, we achieve an ideal margin of belief state coverage generated under partial observation (recall), and the proportion of correct beliefs sampled under full observation (accuracy). As we significantly increase the number of samples, the accuracy starts to decrease due to randomness in latent distribution resampling.

will inevitably experience collisions in case of a truck approaching, since it does not cautiously account for occlusions. On the other hand, an overly cautious approach results in stasis, inhibiting the controller's ability to advance in the scene. This nuanced decision-making using CARFF-based controller is especially crucial in driving scenarios, as it enhances safety and efficiency by adapting to complex and unpredictable road environments, thereby fostering a more reliable and human-like response in autonomous vehicles.

5 Discussion

Limitations: Like other NeRF-based methods, CARFF currently relies on posed images of specific scenes such as road intersections, limiting its direct applicability to unseen environments. However, we anticipate enhanced generalizability with the increasing deployment of cameras around populated areas, such as traffic cameras at intersections. Additionally, handling very complex dynamics with an extremely large number of actors still poses a challenge for our method, requiring per-scene optimization to balance comprehensive dynamics modeling against accuracy. Potentially stronger models in the near future may offer a promising avenue for further enhancements in this regard.

Conclusion: We present CARFF, a novel method for probabilistic 3D scene forecasting from partial observations. By employing a Pose-Conditional VAE, a NeRF conditioned on the learned posterior, and a mixture density network that forecasts future scenes, we effectively model, predict, and plan in complex environments with state and dynamics uncertainty in a POMDP. We further demonstrate the capabilities of our method in simulated autonomous driving scenarios. Overall, CARFF offers an intuitive framework to perceiving, forecasting, and acting under uncertainty that could prove invaluable for vision-based algorithms in unstructured environments.

A Datasets

A.1 CARLA Datasets

A complete figure of the actor and ego configurations across scenes and the progression of timestamps for the Single-Scene Approaching Intersection, Multi-Scene Approaching Intersection, and the Multi-Scene Two Lane Merge is visualized in Fig. 14.

A.2 Hand-manipulation Dataset

We generate an additional hand-manipulation dataset involving a robotic manipulator engaged in probabilistic reaching tasks utilizing VUER fig[53]. The experimental setup consists of two target objects placed on a table in a room: a Rubik's cube and a green tennis ball. The robotic hand's configurations during the reaching and grasping phases for both objects are illustrated in Fig. 14.

B Implementation Details

B.1 Pose-Conditional VAE

PC-VAE Hyperparameters	
Latent Size	8
LR	0.004
KLD Weight Start	0.000001
KLD Weight End	0.00001 − 0.00004*
KLD Increment Start	50 epochs
KLD Increment End	80 epochs

Table 4: PC-VAE experimental setup and hyperparameters. The main hyperparameters in PC-VAE training on the three datasets are latent size, LR, and KLD weight. For KLD scheduling, the KLD increment start refers to the number of epochs at which the KLD weight begins to increase from the initial KLD weight. KLD increment end is the number of epochs at which the KLD weight stops increasing at the maximum KLD weight. The asterisk (*) marks the hyperparameter that is dataset-dependent.

Architecture: We implement PC-VAE on top of a standard PyTorch VAE framework. The encoder with convolutional layers is replaced with a single convolutional layer and a Vision Transformer (ViT) Large 16 [8] pre-trained on ImageNet [40]. We modify fully connected layers to project ViT output of size 1000 to mean and variances with size of the latent dimension, 8. During training, the data loader returns the pose of the camera angle represented by an integer value. This value is one-hot encoded and concatenated to the re-parameterized encoder outputs, before being passed to the decoder. The decoder input size is increased to add the number of poses to accommodate the additional pose information.

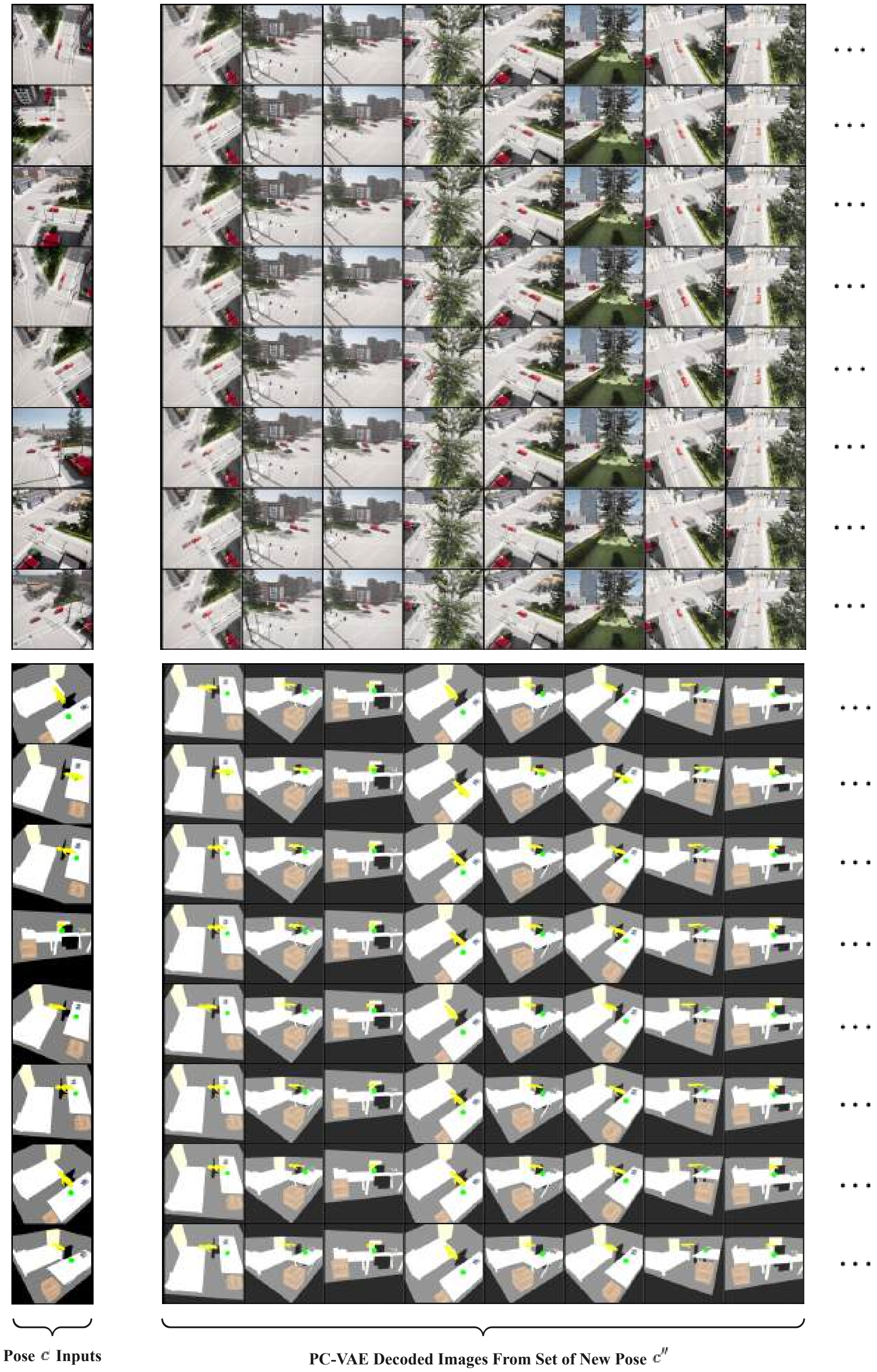


Fig. 9: PC-VAE encoder inputs, ground truth timestamps, and reconstructions for a CARLA dataset and Hand-manipulation dataset. The encoder input, I_c^t, among the other ground truth images I_c viewed from camera pose c at different timestamps, is reconstructed across a new set of poses c'' respecting timestamp t, generating $I_{c''}^t$. This is a full grid of the reconstructions.

Optimization: We utilize a single RTX 3090 graphics card for all our experiments. The PC-VAE model takes approximately 22 hours to converge using this GPU. During this phase, we tune various hyperparameters including the latent size, learning rate and KL divergence loss weight to establish optimal training tailored to our model (see Tab. 4). In order to optimize for the varied actor configurations and scenarios generated within the CARLA [9] simulator, we slightly adjust hyperparameters differently for each dataset.

The learning rate (LR) and KL divergence (KLD) weight are adjusted to find an appropriate balance between the effective reconstruction of pose conditioning in the latent space, and the regularization of latents. Regularization pushes the latents toward Gaussian distributions and keeps the non-expressive latents in an over-parameterized latent space to be standard normal. This stabilizes the sampling process and ensures stochastic behavior of latent samples in case of occlusion. To achieve this balance, we use a linear KLD weight scheduler, where the weight is initialized at a low value for KLD increment start epoch (see Tab. 4). This allows the model to initially focus on achieving highly accurate conditioned reconstructions. The KLD weight is then steadily increased until KLD increment end epoch is reached, ensuring probabilistic behavior under partial observability.

B.2 Mixture Density Network

With the POMDP, the mixture density network (MDN) takes in the mean and variances of the latent distributions of the current belief state $q_\phi(z_{t-1}|I_c^{t-1})$ and outputs the next belief state's estimated posterior distribution. To better model the uncertainty of the predicted belief state distribution, the output is a mixture of Gaussian $q'_\phi(z_t|I_c^{t-1})$ modeled through a multi-headed MLP.

Architecture: The shared backbone simply contains 2 fully connected layers and rectified linear units (ReLU) activation with hidden layer size of 512. Additional heads with 2 fully connected layers are used to generate μ_i and σ_i^2. The mixture weight, π_i, is generated from a 3 layer MLP network. We limit the number of Gaussians, $K = 2$.

Optimization: We train our network for $30,000$ epochs using the batch size of 128 and an initial LR of 0.005, and apply LR decay to optimize training. This takes approximately 30 minutes to train utilizing the GPU. During training, the dataloader outputs the means and variances at the current timestamp and indexed view, and the means and variances for the next timestamp, at a randomly sampled neighboring view. This allows the MDN to learn how occluded views advance into all the possible configurations from potentially unoccluded neighboring views, as a mixture of Gaussian.

At each iteration, the negative log-likelihood loss is computed for 1000 samples drawn from the predicted mixture of distributions $q'_\phi(z_t|I_c^{t-1})$ with respect to the ground truth distribution $q_\phi(z_t|I_c^t)$. While the MDN is training, additional Gaussian noise, given by $\epsilon \sim \mathcal{N}(0, \sigma^2)$, is added to the means and variances of the current timestamp $t-1$, where $\sigma \in [0.001, 0.01]$. The Gaussian noise and

LR decay help prevent overfitting and reduce model sensitivity to environmental artifacts like moving trees, moving water, etc.

B.3 NeRF

Architecture: We implement our NeRF as a decoder to our belief state to recover 3D observations utilizing an existing PyTorch implementation of Instant-NGP [28]. We concatenate the latents to the inputs of two parts of the Instant-NGP architecture: the volume density network, $\sigma(\mathbf{x})$, for the density values, and the color network, $C(\mathbf{r})$, for conditional RGB generation. While the overall architecture is kept constant, the input dimensions of each network are modified to allow additional latent concatenation.

Optimization: Empirically, we observe that it is essential to train the NeRF such that it learns the distribution of scenes within the PC-VAE latent space. Using only pre-defined learned samples to train may run the risk of relying on non-representative samples. On the other hand, direct re-sampling during each training iteration in Instant-NGP may lead to delayed training progress, due to NeRF's sensitive optimization. In our optimization procedure, we use an LR of 0.002 along with an LR decay and start with pre-defined latent samples. Then we slowly introduce the re-sampled latents. We believe that this strategy progressively diminishes the influence of a single sample, while maintaining efficient training. Based on our observations, this strategy contributes towards Instant-NGP's ability to rapidly assimilate fundamental conditioning and environmental reconstruction, while simultaneously pushing the learning process to be less skewed towards a single latent sample.

C GUI Interface

For ease of interaction with our inference pipeline, our NeRF loads a pre-trained MDN checkpoint, and we build a graphical user interface (GUI) using DearPyGUi for visualization purposes. We implement three features in the GUI: (a) predict, (b) probe and predict, and (c) toggle.

Predict: We implement the function to perform prediction directly from a given image path in the GUI. We use the distribution $q_\phi(z_{t-1}|I_c^{t-1})$ from PC-VAE encoder, corresponding to the input image I_c^{t-1}, to predict the latent distribution for the next timestamp belief state $q'_\phi(z_t|I_c^{t-1})$. This process is done on the fly through the MDN. A sample from the predicted distribution is then generated and used to condition the NeRF. This advances the entire scene to the next timestamp.

Probe and predict: The sampled latent from the predicted distribution does not correspond to a singular distribution and hence we can not directly predict the next timestamp. To make our model auto-regressive in nature, we perform

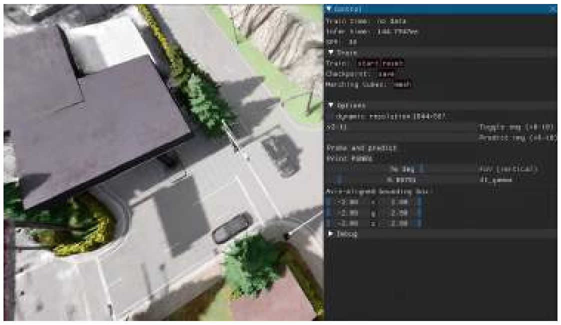


Fig. 10: NeRF graphical user interface. The GUI allows us to toggle and predict with an input image path. The probe and predict function probes the current location of the car and predicts the next. The screenshot is sharpened for visual clarity in the paper.

density probing. We probe the density of the NeRF at the possible location coordinates of the car to obtain the current timestamp and scene. This is then used to know the actual state sampled from the belief state probability distributions. The new distribution enables auto-regressive predictions using the predict function described above.

Toggle: The NeRF generates a scene corresponding to the provided input image path using learned latents from PC-VAE. This function tests the NeRF decoder's functionality with a given belief state. When the input image is a fully observable view (corresponding to a unknown belief state), the NeRF renders clear actor and ego configurations respecting the input. This allows us to visualize the scene at different timestamps and in different configurations.

D CARFF Evaluation

D.1 Pose-Conditional VAE

Reconstruction Quality: To analyze the reconstruction performance of the model during training, we periodically plot grids of reconstructed images. These grids consist of (a) randomly selected encoder inputs drawn from the dataset, (b) the corresponding ground truth images for those inputs at each timestamp at the same camera pose, and (c) reconstructed outputs at randomly sampled poses respecting the input scene and timestamp. An example reconstruction grid is provided in Fig. 9. The grid enables visual assessment of whether the model is capable of accurately reconstructing reasonable images using the encoder inputs, conditioned on the poses. This evaluation provides us with visual evidence of improvement in reconstruction quality. We also quantitatively analyze the progressive improvement of reconstruction through the average PSNR calculated over the training data (see Fig. 12).

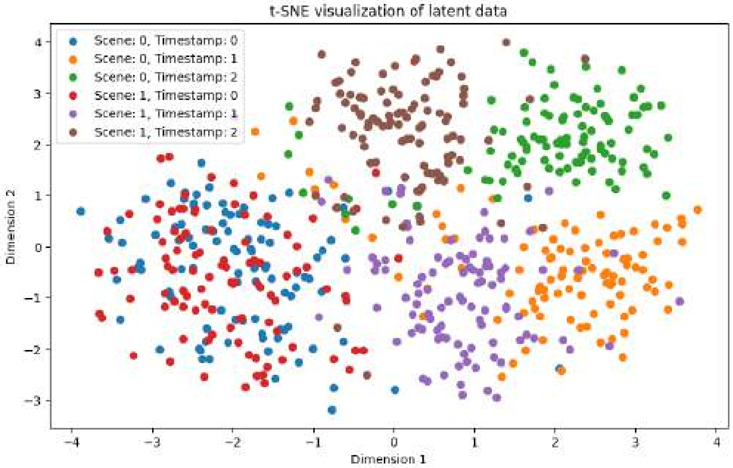


Fig. 11: Latent sample distribution clustering. The distributions of latent samples for the Multi-Scene Two Lane Merge dataset are separable through t-SNE clustering. In the figure, the clusters for *Scene 0, Timestamp 0* and *Scene 1, Timestamp 0* overlap in distribution because they represent the same initial state of the environment under dynamics uncertainty.

The PC-VAE outputs in Fig. 9 only provides visual confirmation to assess the quality of the latents learned by PC-VAE. Utilization of the 3D decoder later in our method allows us to produce more high resolution visualizations of the scene that can be used for further downstream tasks.

Latent Space Analysis To assess the quality of the latents generated by PC-VAE, we initially use t-SNE plots to visualize the latent distributions as clusters. Fig. 11 shows that the distributions of the latent samples for the Multi-Scene Two Lane Merge dataset are separable. While t-SNE is good at retaining nearest-neighbor information by preserving local structures, it performs weakly in preserving global structures. Therefore, t-SNE may be insufficient in capturing the differences in distributions for all our datasets.

Instead, we pivot to Support Vector Machine to perform a quantitative evaluation of the separability of the latents. We utilize a Radial Basis Function (RBF) kernel with the standard regularization parameter ($C = 1$). We perform 10-fold validation on the latents to calculate the accuracy as a metric for clustering. See Tab. 5 for the results.

Beyond separability, we analyze the recall and accuracy of the learned latents directly from PC-VAE under partial and full observations. This achieves very high accuracy even under a large number of samples while retraining decent recall, enabling downstream MDN training. (See Fig. 13)

For the additional Hand-manipulation Object Reaching dataset, we used a similar setup as detailed in Fig 8 with $n = 1$ to $n = 50$ samples from the

Architectures	Train PSNR	SVM Accuracy	NV PSNR
Multi-Scene Approaching Intersection			
PC-VAE	**26.47**	**89.17**	**26.37**
PC-VAE w/o CL	26.20	83.83	26.16
Vanilla PC-VAE	25.97	29.33	25.93
PC-VAE w/o Freezing	24.82	29.83	24.78
PC-VAE w/ MobileNet	19.37	29.50	19.43
Vanilla VAE	26.04	14.67	9.84
Multi-Scene Two Lane Merge			
PC-VAE	**25.50**	**88.33**	**25.84**
PC-VAE w/o CL	24.38	29.67	24.02
Vanilla PC-VAE	24.75	29.67	24.96
PC-VAE w/o Freezing	23.97	28.33	24.04
PC-VAE w/ MobileNet	17.70	75.00	17.65
Vanilla VAE	25.11	28.17	8.49

Table 5: PC-VAE metrics and ablations across Multi-Scene datasets. CARFF's PC-VAE outperforms other encoder architectures across the Multi-Scene datasets in reconstruction and pose-conditioning.

MDN's predicted latent distributions from an potentially partially observable image input. Similar to the CARLA dataset results, we achieve an ideal margin of belief state coverage generated under partial observation (recall), and the proportion of correct beliefs sampled under full observation (accuracy).

D.2 Fully Observable Predictions

One of the tasks of the MDN is to forecast the future scene configurations under full observation. We quantitatively evaluate our model's ability to forecast future scenes by comparing bird's-eye views rendered from the NeRF with chosen ground truth images of the scene for the various timestamps (see Tab. 6). The values are calculated and displayed for all four datasets. In Tab. 6, images are marked as either toggled ($\tilde{I}_{t_i}$) or predicted ($\hat{I}_{t_i}$). Toggled images in the table cannot be predicted deterministically due to it being the first timestamp in the dataset, or the state of the previous timestamps across scenes being the same in case of dynamics uncertainty. Due to the same reason, in the Multi-Scene Two Lane Merge and the Hand-manipulation Object Reaching Datasets, there are additional bolded PSNR values for the pairs $(I_{t_1}, \tilde{I}_{t_4})$ and $(I_{t_4}, \tilde{I}_{t_1})$.

We demonstrate that the toggled or predicted images that correspond to the correct ground truth show a PSNR value around 29, indicating high fidelity 3D reconstruction and clear visual decodings as the output of CARFF.

Single-Scene Approaching Intersection

Result	I_{t_1}	I_{t_2}	I_{t_3}	I_{t_4}	I_{t_5}	I_{t_6}	I_{t_7}	I_{t_8}	I_{t_9}	$I_{t_{10}}$
$\tilde{I}_{t_1}$	**29.01**	−5.97	−6.08	−6.52	−6.44	−6.03	−6.31	−6.36	−6.26	−6.28
$\hat{I}_{t_2}$	−5.42	**27.51**	−3.07	−4.67	−4.58	−4.17	−4.43	−4.51	−4.39	−4.39
$\hat{I}_{t_3}$	−6.06	−2.81	**28.12**	−4.47	−4.68	−4.19	−4.05	−4.61	−4.47	−4.52
$\hat{I}_{t_4}$	−7.01	−5.37	−5.03	**29.40**	−4.99	−5.08	−5.03	−5.41	−5.28	−5.32
$\hat{I}_{t_5}$	−6.87	−5.2	−4.93	−5.00	**29.44**	−4.53	−4.46	−5.19	−5.05	−5.09
$\hat{I}_{t_6}$	−6.29	−4.55	−4.27	−4.8	−4.24	**29.02**	−4.02	−4.53	−4.38	−4.44
$\hat{I}_{t_7}$	−6.76	−5.05	−4.76	−5.31	−5.14	−4.36	**29.50**	−4.50	−4.86	−4.93
$\hat{I}_{t_8}$	−6.73	−5.02	−4.74	−5.25	−5.10	−4.64	−4.76	**29.46**	−4.41	−4.86
$\hat{I}_{t_9}$	−6.75	−5.00	−4.70	−5.23	−5.07	−4.64	−4.85	−4.52	**29.55**	−4.42
$\hat{I}_{t_{10}}$	−6.79	−5.06	−4.75	−5.30	−5.15	−4.69	−4.93	−5.01	−4.34	**29.55**

Multi-Scene Approaching Intersection

Result	I_{t_1}	I_{t_2}	I_{t_3}	I_{t_4}	I_{t_5}	I_{t_6}
$\tilde{I}_{t_1}$	**28.10**	−5.24	−5.50	−1.67	−3.29	−3.92
$\hat{I}_{t_2}$	−5.23	**28.02**	−6.11	−4.70	−3.21	−4.84
$\hat{I}_{t_3}$	−5.43	−6.03	**27.97**	−4.85	−4.53	−2.93
$\tilde{I}_{t_4}$	−1.71	−4.73	−5.00	**28.26**	−2.25	−3.08
$\tilde{I}_{t_5}$	−3.68	−3.24	−4.91	−2.76	**28.21**	−2.99
$\hat{I}_{t_6}$	−4.02	−4.91	−3.27	−3.13	−2.61	**28.26**

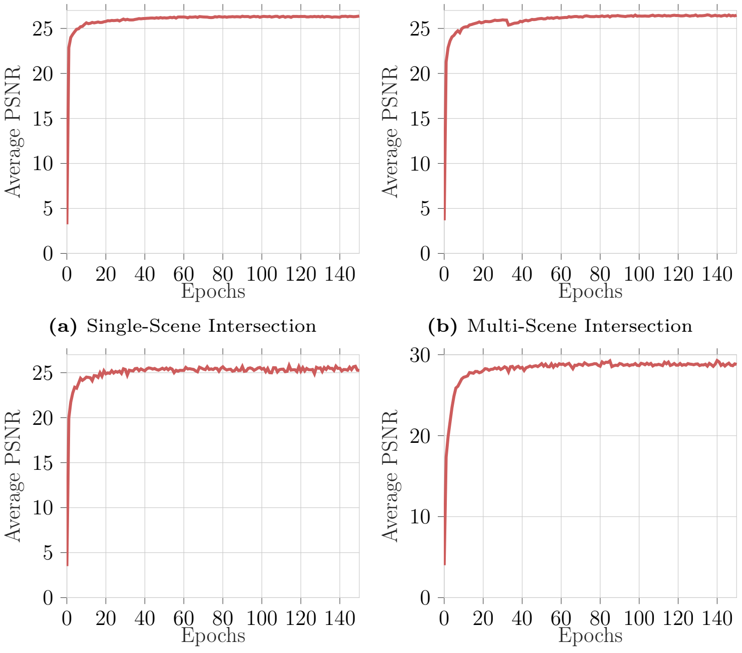
Multi-Scene Two Lane Merge

Result	I_{t_1}	I_{t_2}	I_{t_3}	I_{t_4}	I_{t_5}	I_{t_6}
$\tilde{I}_{t_1}$	**28.27**	−5.31	−6.41	**28.23**	−4.77	−5.42
$\tilde{I}_{t_2}$	−5.22	**28.23**	−5.17	−5.27	−2.91	−4.01
$\hat{I}_{t_3}$	−6.32	−5.09	**28.14**	−6.33	−5.01	−4.28
$\tilde{I}_{t_4}$	**28.27**	−5.27	−6.37	**28.23**	−4.72	−5.37
$\tilde{I}_{t_5}$	−4.64	−2.73	−5.01	−4.71	**28.08**	−5.29
$\hat{I}_{t_6}$	−5.32	−4.02	−4.32	−5.33	−5.34	**28.17**

Hand-manipulation Object Reaching

Result	I_{t_1}	I_{t_2}	I_{t_3}	I_{t_4}	I_{t_5}	I_{t_6}
$\tilde{I}_{t_1}$	**30.71**	−10.74	−10.63	**30.71**	−9.81	−11.13
$\tilde{I}_{t_2}$	−10.07	**30.32**	−9.25	−10.07	−10.88	−10.17
$\hat{I}_{t_3}$	−9.98	−9.11	**30.37**	−9.98	−10.78	−10.02
$\tilde{I}_{t_4}$	**30.66**	−10.69	−10.58	**30.66**	−9.77	−11.09
$\tilde{I}_{t_5}$	−8.19	−10.21	−10.1	−8.19	**29.49**	−9.99
$\hat{I}_{t_6}$	−10.96	−10.6	−10.49	−10.96	−11.08	**30.67**

Table 6: Complete PSNR values for fully observable predictions for all CARLA datasets and the Hand-manipulation dataset. The table contains PSNR values between the ground truth images and either a toggled image (marked as $\tilde{I}_{t_i}$), or a predicted image from the NeRF decoder (marked as $\hat{I}_{t_i}$). Toggled or predicted images that correspond to the correct ground truth are bolded and have a extremely high PSNR value, indicating high fidelity results. The PSNR values for incorrect correspondances are replaced with the difference between the incorrect PSNR and the bolded PSNR associated with a correct correspondance.



(a) Single-Scene Intersection

(b) Multi-Scene Intersection

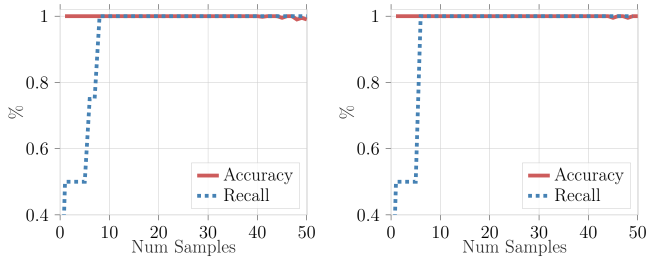
(c) Multi-Scene Two Lane Merge

(d) Hand-manipulation

Fig. 12: Average train PSNR plot for all CARLA datasets and the hand-manipulation dataset. The plot shows the increase in average training PSNR of all images for each dataset, over the period of the training process.

D.3 Architecture Ablations

The results presented in Tab. 7 substantiate the benefits of our PC-VAE encoder architecture compared to other formulations. Specifically, a non-conditional VAE fails in SVM accuracy as it only reconstructs images and does not capture the underlying 3D structures. Vanilla PC-VAE and PC-VAE without freezing weights require careful fine-tuning of several hyper-parameters and don't generalize well to drastic camera movements. Our experiments show that our proposed model is capable of sustaining stochastic characteristics via latent representations in the presence of occlusion, while simultaneously ensuring precise reconstructions.



(a) Approaching Intersection **(b)** Two Lane Merge

Fig. 13: Multi-Scene dataset accuracy and recall curves from learned latents. We test our framework across $n = 1$ and $n = 50$ samples from PC-VAE's latent distributions from ego-centric image input. Across the number of samples n, we achieve an ideal margin of belief state coverage generated under partial observation (recall), and the proportion of correct beliefs sampled under full observation (accuracy) for the MDN to learn. As we significantly increase the number of samples, the accuracy starts to decrease due to randomness in latent distribution resampling.

Encoder Architectures	Train PSNR	SVM Acc.	NV PSNR
PC-VAE	**26.30**	**75.20**	**25.24**
PC-VAE w/o CL	26.24	70.60	24.80
Vanilla PC-VAE	26.02	25.70	24.65
PC-VAE w/o Freezing	24.57	5.80	24.60
PC-VAE w/ MobileNet	17.14	19.70	17.16
Vanilla VAE	24.15	10.60	11.43

Table 7: PC-VAE ablations. CARFF's PC-VAE encoder outperforms other architectures in image reconstruction and pose-conditioning. We evaluate on: PC-VAE without Conv. Layer, PC-VAE with a vanilla encoder, PC-VAE without freezing ViT weights, PC-VAE replacing ViT with MobileNet, and non pose-conditional VAE.

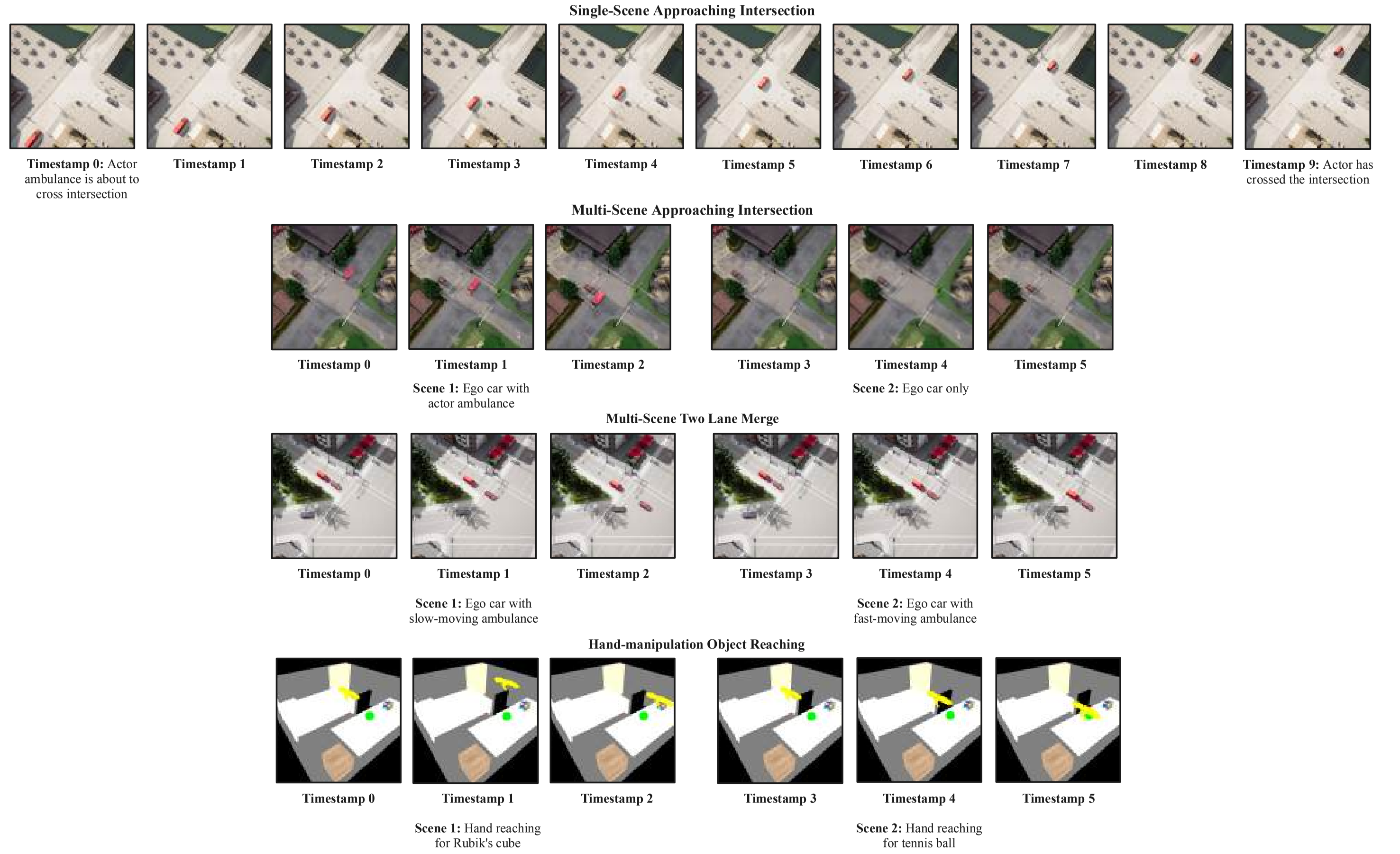

Single-Scene Approaching Intersection
Timestamp 0: Actor ambulance is about to cross intersection
Timestamp 1
Timestamp 2
Timestamp 3
Timestamp 4
Timestamp 5
Timestamp 6
Timestamp 7
Timestamp 8
Timestamp 9: Actor has crossed the intersection
Multi-Scene Approaching Intersection
Timestamp 0
Timestamp 1
Timestamp 2
Timestamp 3
Timestamp 4
Timestamp 5
Scene 1: Ego car with actor ambulance
Scene 2: Ego car only
Multi-Scene Two Lane Merge
Timestamp 0
Timestamp 1
Timestamp 2
Timestamp 3
Timestamp 4
Timestamp 5
Scene 1: Ego car with slow-moving ambulance
Scene 2: Ego car with fast-moving ambulance
Hand-manipulation Object Reaching
Timestamp 0
Timestamp 1
Timestamp 2
Timestamp 3
Timestamp 4
Timestamp 5
Scene 1: Hand reaching for Rubik's cube
Scene 2: Hand reaching for tennis ball

Fig. 14: Single-Scene Approaching Intersection, Multi-Scene Approaching Intersection, Multi-Scene Two Lane Merge, and Hand-manipulation Object Reaching Datasets. The actor and ego car configurations and the hand configurations for the timestamps and scenes of the three CARLA datasets, and the hand-manipulation dataset are visualized at a single camera pose. The colors of the cars for the Multi-Scene Approaching Intersection have been slightly modified for greater contrast and visual clarity in the paper.

Bibliography

[1] Adamkiewicz, M., Chen, T., Caccavale, A., Gardner, R., Culbertson, P., Bohg, J., Schwager, M.: Vision-only robot navigation in a neural radiance world. IEEE Robotics and Automation Letters **7**(2), 4606–4613 (2022)

[2] Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In: Int. Conf. Comput. Vis. pp. 5855–5864 (2021)

[3] Cao, A., Johnson, J.: Hexplane: A fast representation for dynamic scenes. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 130–141 (2023)

[4] Cao, J., Wang, X., Darrell, T., Yu, F.: Instance-aware predictive navigation in multi-agent environments. In: IEEE Int. Conf. on Robotics and Automation. pp. 5096–5102. IEEE (2021)

[5] Charatan, D., Li, S., Tagliasacchi, A., Sitzmann, V.: pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction (2023)

[6] Codevilla, F., Santana, E., López, A.M., Gaidon, A.: Exploring the limitations of behavior cloning for autonomous driving. In: Int. Conf. Comput. Vis. pp. 9329–9338 (2019)

[7] Deng, K., Liu, A., Zhu, J.Y., Ramanan, D.: Depth-supervised nerf: Fewer views and faster training for free. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 12882–12891 (2022)

[8] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: Int. Conf. Learn. Represent. (2021)

[9] Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V.: CARLA: An open urban driving simulator. In: Conf. on Robol Learning. pp. 1–16 (2017)

[10] Driess, D., Huang, Z., Li, Y., Tedrake, R., Toussaint, M.: Learning multi-object dynamics with compositional neural radiance fields. arXiv preprint arXiv:2202.11855 (2022)

[11] Fei, B., Xu, J., Zhang, R., Zhou, Q., Yang, W., He, Y.: 3d gaussian as a new vision era: A survey (2024)

[12] Fridovich-Keil, S., Yu, A., Tancik, M., Chen, Q., Recht, B., Kanazawa, A.: Plenoxels: Radiance fields without neural networks. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5501–5510 (2022)

[13] Hausknecht, M., Stone, P.: Deep recurrent q-learning for partially observable mdps. In: AAAI (2015)

[14] Hearst, M.A., Dumais, S.T., Osuna, E., Platt, J., Scholkopf, B.: Support vector machines. IEEE Intelligent Systems and their applications **13**(4), 18–28 (1998)

[15] Ichnowski, J., Avigal, Y., Kerr, J., Goldberg, K.: Dex-nerf: Using a neural radiance field to grasp transparent objects. arXiv preprint arXiv:2110.14217 (2021)

[16] Ivanovic, B., Elhafsi, A., Rosman, G., Gaidon, A., Pavone, M.: Mats: An interpretable trajectory forecasting representation for planning and control. In: Conf. on Robol Learning (2021)

[17] Kaelbling, L.P., Littman, M.L., Cassandra, A.R.: Planning and acting in partially observable stochastic domains. Artificial Intelligence **101**(1), 99–134 (1998). https://doi.org/https://doi.org/10.1016/S0004-3702(98)00023-X, https://www.sciencedirect.com/science/article/pii/S000437029800023X

[18] Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (July 2023), https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/

[19] Kerr, J., Fu, L., Huang, H., Avigal, Y., Tancik, M., Ichnowski, J., Kanazawa, A., Goldberg, K.: Evo-nerf: Evolving nerf for sequential robot grasping of transparent objects. In: Conf. on Robol Learning (2022)

[20] Kosiorek, A.R., Strathmann, H., Zoran, D., Moreno, P., Schneider, R., Mokrá, S., Rezende, D.J.: NeRF-VAE: A geometry aware 3d scene generative model. In: ICML. pp. 5742–5752 (2021)

[21] Li, Y., Li, S., Sitzmann, V., Agrawal, P., Torralba, A.: 3d neural scene representations for visuomotor control. In: Conf. on Robol Learning. pp. 112–123 (2022)

[22] Liu, J.W., Cao, Y.P., Mao, W., Zhang, W., Zhang, D.J., Keppo, J., Shan, Y., Qie, X., Shou, M.Z.: Devrf: Fast deformable voxel radiance fields for dynamic scenes. In: Adv. Neural Inform. Process. Syst. vol. 35, pp. 36762–36775 (2022)

[23] Luiten, J., Kopanas, G., Leibe, B., Ramanan, D.: Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis (2023)

[24] van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. Journal of Machine Learning Research **9**, 2579–2605 (2008), http://www.jmlr.org/papers/v9/vandermaaten08a.html

[25] Marza, P., Matignon, L., Simonin, O., Wolf, C.: Multi-object navigation with dynamically learned neural implicit representations. In: Int. Conf. Comput. Vis. pp. 11004–11015 (2023)

[26] McAllister, R., Rasmussen, C.E.: Data-efficient reinforcement learning in continuous state-action gaussian-pomdps. In: Adv. Neural Inform. Process. Syst. (2017)

[27] Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: Eur. Conf. Comput. Vis. (2020)

[28] Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Trans. Graph. **41**(4), 1–15 (2022)

[29] Packer, C., Rhinehart, N., McAllister, R.T., Wright, M.A., Wang, X., He, J., Levine, S., Gonzalez, J.E.: Is anyone there? learning a planner contingent on perceptual uncertainty. In: Conf. on Robol Learning (2022)

[30] Packer, C., Rhinehart, N., McAllister, R.T., Wright, M.A., Wang, X., He, J., Levine, S., Gonzalez, J.E.: Is anyone there? learning a planner contingent on perceptual uncertainty. In: Liu, K., Kulic, D., Ichnowski, J. (eds.) Conf. on Robol Learning. pp. 1607–1617 (2023)

[31] Pan, X., You, Y., Wang, Z., Lu, C.: Virtual to real reinforcement learning for autonomous driving. arXiv preprint arXiv:1704.03952 (2017)

[32] Papadimitriou, C.H., Tsitsiklis, J.N.: The complexity of markov decision processes. Mathematics of operations research **12**(3), 441–450 (1987)

[33] Park, K., Sinha, U., Barron, J.T., Bouaziz, S., Goldman, D.B., Seitz, S.M., Martin-Brualla, R.: Nerfies: Deformable neural radiance fields. In: Int. Conf. Comput. Vis. pp. 5865–5874 (2021)

[34] Park, K., Sinha, U., Hedman, P., Barron, J.T., Bouaziz, S., Goldman, D.B., Martin-Brualla, R., Seitz, S.M.: Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. ACM Trans. Graph. (2021)

[35] Pineau, J., Gordon, G., Thrun, S., et al.: Point-based value iteration: An anytime algorithm for pomdps. In: IJCAI. vol. 3, pp. 1025–1032 (2003)

[36] Pumarola, A., Corona, E., Pons-Moll, G., Moreno-Noguer, F.: D-NeRF: Neural Radiance Fields for Dynamic Scenes. In: IEEE Conf. Comput. Vis. Pattern Recog. (2020)

[37] Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: IEEE Conf. Comput. Vis. Pattern Recog. (2016)

[38] Rhinehart, N., He, J., Packer, C., Wright, M.A., McAllister, R., Gonzalez, J.E., Levine, S.: Contingencies from observations: Tractable contingency planning with learned behavior models. In: IEEE Int. Conf. on Robotics and Automation. pp. 13663–13669 (2021)

[39] Roessle, B., Barron, J.T., Mildenhall, B., Srinivasan, P.P., Nießner, M.: Dense depth priors for neural radiance fields from sparse input views. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 12892–12901 (2022)

[40] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: Imagenet large scale visual recognition challenge (2015)

[41] Shao, R., Zheng, Z., Tu, H., Liu, B., Zhang, H., Liu, Y.: Tensor4d: Efficient neural 4d decomposition for high-fidelity dynamic reconstruction and rendering. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 16632–16642 (2023)

[42] Sun, C., Sun, M., Chen, H.T.: Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5459–5469 (2022)

[43] Tancik, M., Casser, V., Yan, X., Pradhan, S., Mildenhall, B., Srinivasan, P.P., Barron, J.T., Kretzschmar, H.: Block-nerf: Scalable large scene neural view synthesis. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 8248–8258 (2022)

[44] Tancik, M., Srinivasan, P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J., Ng, R.: Fourier features let

networks learn high frequency functions in low dimensional domains. In: Adv. Neural Inform. Process. Syst. pp. 7537–7547 (2020)

[45] Tas, O.S., Stiller, C.: Limited visibility and uncertainty aware motion planning for automated driving. In: IEEE Intelligent Vehicles Symposium (IV) (jun 2018)

[46] Tewari, A., Thies, J., Mildenhall, B., Srinivasan, P., Tretschk, E., Yifan, W., Lassner, C., Sitzmann, V., Martin-Brualla, R., Lombardi, S., Simon, T., Theobalt, C., Nießner, M., Barron, J.T., Wetzstein, G., Zollhöfer, M., Golyanik, V.: Advances in Neural Rendering. Comput. Graph. Forum (2022)

[47] Toromanoff, M., Wirbel, E., Moutarde, F.: End-to-end model-free reinforcement learning for urban driving using implicit affordances. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 7153–7162 (2020)

[48] Tretschk, E., Golyanik, V., Zollhoefer, M., Bozic, A., Lassner, C., Theobalt, C.: Scenerflow: Time-consistent reconstruction of general dynamic scenes. In: International Conference on 3D Vision (3DV) (2023)

[49] Truong, P., Rakotosaona, M.J., Manhardt, F., Tombari, F.: Sparf: Neural radiance fields from sparse and noisy poses. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 4190–4200 (2023)

[50] Turki, H., Ramanan, D., Satyanarayanan, M.: Mega-nerf: Scalable construction of large-scale nerfs for virtual fly-throughs. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 12922–12931 (2022)

[51] Wang, Q., Wang, Z., Genova, K., Srinivasan, P.P., Zhou, H., Barron, J.T., Martin-Brualla, R., Snavely, N., Funkhouser, T.: Ibrnet: Learning multi-view image-based rendering. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 4690–4699 (2021)

[52] Xu, Q., Xu, Z., Philip, J., Bi, S., Shu, Z., Sunkavalli, K., Neumann, U.: Point-nerf: Point-based neural radiance fields. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5438–5448 (2022)

[53] Yang, G.: VUER: A 3D visualization and data collection environment for robot learning (2024), https://github.com/vuer-ai/vuer

[54] Yen-Chen, L., Florence, P., Barron, J.T., Rodriguez, A., Isola, P., Lin, T.Y.: inerf: Inverting neural radiance fields for pose estimation. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 1323–1330. IEEE (2021)

[55] Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelnerf: Neural radiance fields from one or few images (2021)

[56] Yu, M.Y., Vasudevan, R., Johnson-Roberson, M.: Occlusion-aware risk assessment for autonomous driving in urban environments. IEEE Robotics and Automation Letters **4**(2), 2235–2241 (Apr 2019). https://doi.org/10.1109/lra.2019.2900453, http://dx.doi.org/10.1109/LRA.2019.2900453

[57] Zhang, C., Steinhauser, F., Hinz, G., Knoll, A.: Improved occlusion scenario coverage with a pomdp-based behavior planner for autonomous urban driving. In: 2021 IEEE International Intelligent Transportation Systems Conference (ITSC). pp. 593–600 (2021). https://doi.org/10.1109/ITSC48978.2021.9564424

[58] Zhang, C., Steinhauser, F., Hinz, G., Knoll, A.: Occlusion-aware planning for autonomous driving with vehicle-to-everything communication. IEEE Transactions on Intelligent Vehicles **9**(1), 1229–1242 (2024). `https://doi.org/10.1109/TIV.2023.3308098`