# Learning semantic traversability priors using diffusion models for uncertainty-aware global path planning

Ethan Fahnestock[1], Erick Fuentes[1], Philip R. Osteen[2], Siddharth Ancha[1], Nicholas Roy[1]

Website: https://difftrav.d1njdpg6jia6z4.amplifyapp.com

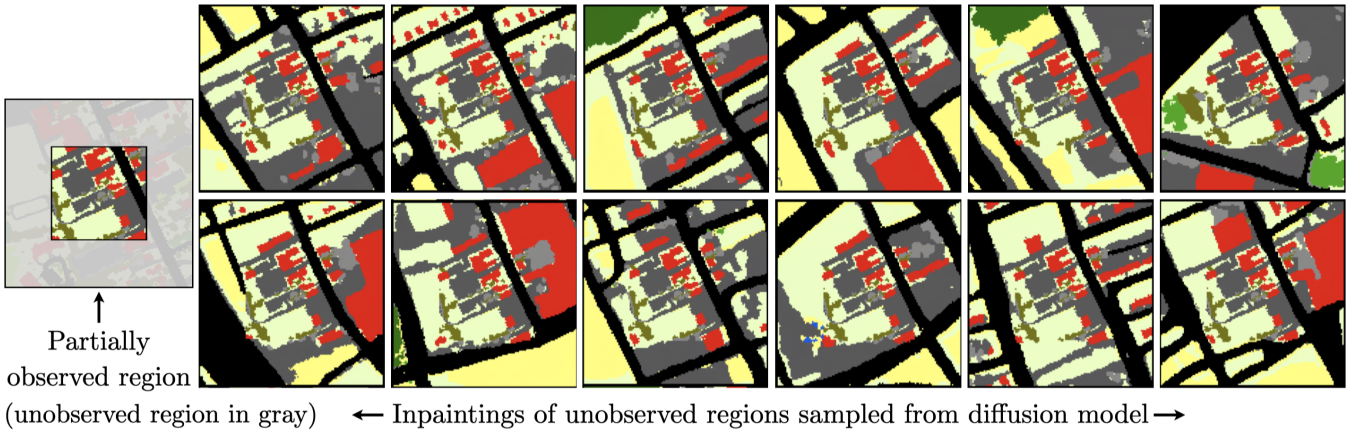


Fig. 1: As a robot navigates an unknown environment, it only observes its immediate surroundings due to sensor range limits. However, long-horizon global path planning could be significantly improved if the robot can estimate semantics of far-field regions outside its visibility. We propose learning priors over the semantic structure of navigation environments using a diffusion model that "inpaints" the semantics of unobserved regions conditioned on observed regions. By sampling a diverse set of high-fidelity far-field environment maps that are consistent with already-observed regions, our method can plan efficient, low-cost paths in an uncertainty-aware manner that improves navigation performance.

***Abstract*— Robots have limited sensor ranges, restricting what they can observe, complicating navigation through a-priori unknown environments. If environment structure is present, priors over this structure can extend the utility of local observations and improve navigation performance. In this work, we propose learning priors over the semantic structure of navigation environments using state-of-the-art generative diffusion models. We show that diffusion models can capture complex spatial dependencies in overhead semantic maps, and are able to infer the semantics of far-away unobserved regions *conditioned on* local semantics already observed by the robot. By sampling a diverse, multi-modal set of high-fidelity semantic maps that are consistent with observed regions, we are able to estimate far-field navigation costs in an uncertainty-aware manner. Our preliminary investigations suggest that diffusion-based uncertainty-aware navigation costs can enable a downstream global planner to find more efficient paths and improve navigation performance.**

## I. Introduction

As autonomous mobile robots navigate in a-priori unknown environments, they must plan paths to reach goal locations while optimizing for desired mission objectives such as time-to-goal, safety or energy efficiency, while relying solely on information obtained from on-board sensors with limited sensing range. Whereas *local* path planners operate within the sensing range primarily to avoid immediate obstacles, *global* planners reason about long-horizon paths that extend beyond the sensing range. Global reasoning is essential for efficient navigation. For example, deciding whether to stay on a road even if it is long and windy versus cutting through a dense forest could have a significant impact on navigation performance.

However, such decisions require estimating which directions the road likely extends in, and how expansive the forest might be. In other words, global planners need estimates of environment features that lie far beyond the robot's sensing range.

Estimating the semantics of such "far-field" invisible terrain is challenging due to the immense diversity and complexity of natural environments. For example, given a partially observed region in Fig. 1 (*left*), the unobserved region (in gray) might contain many different environment layouts Fig. 1 (*right*) that are consistent with the observed data. It is nearly impossible to make precise predictions about far-away terrain based on a local knowledge of environment semantics. This inherent unpredictability requires that the robot's perception system also produce estimates of *uncertainty* in far-field environment structure to enable principled uncertainty-aware global planning, which is challenging for

[1]Computer Science and Artifical Intelligence Laboratory (CSAIL), MIT, Cambridge, MA 02139. E-mail: {ekf, erick, sancha, nickroy}@csail.mit.edu.

[2]DEVCOM Army Research Laboratory, Adelphi, MD 20783, USA. E-mail: philip.r.osteen.civ@army.mil. Distribution Statement A. Approved for public release: distribution unlimited

most deep-learning methods.

In this work, we propose using *diffusion models* [20, 8, 21] to learn priors over structures of far-field navigation environments. We show that diffusion models are able to capture complex spatial dependencies in overhead semantic images such as the linearity of roads and co-location of terrains like marshes and water bodies. Then, we *condition* diffusion-based priors on local semantics observed by the robot to infer the structure of far-field regions that are not directly visible. Our diffusion model is able to sample a diverse set of high-fidelity far-field semantic maps that are consistent with already-observed regions ( Fig. 1). This diverse yet feasible sampling enables us to estimate probability distributions over far-field navigation costs in an uncertainty aware manner.

We contribute two diffusion-based approaches for far-field semantics prediction: (1) training a conventional *unconditional* diffusion model that formulates far-field prediction given local semantics as a test-time inference problem (Sec. III-C), and (2) training a *conditional* diffusion model to take observed semantics as additional input and directly predict far-field semantics (Sec. III-D). Our preliminary investigation indicates that uncertainty-aware far-field predictions could enable a downstream global planner to plan more efficient paths and improve navigation performance.

## II. Problem Formulation

We consider the problem of 2D navigation from an overhead view where the state of the mobile robot at any given time is denoted by $s \in \mathbb{R}^2$. The robot aims to travel from its current state $s_{\text{curr}} \in \mathbb{R}^2$ to a goal state $s_{\text{goal}} \in \mathbb{R}^2$ along a continuous *trajectory* $\tau(t) : [0,1] \to \mathbb{R}^2$. Each location $s \in \mathbb{R}^2$ is assumed to belong to one of $C$ semantic classes $\mathbb{C} \coloneqq \{1, \dots, C\}$. For example, in outdoor navigation, these classes could correspond to "road", "grass", "building" etc. We associate a semantic label $x(s) : \mathbb{R}^2 \to \mathbb{C}$ with each state that is a property of the environment the robot navigates.

The *cost* of a semantic label $c_{\text{sem}} : \mathbb{C} \to \mathbb{R}^+$ denotes the difficulty of traversing a region with that label. For example, the cost of traversing a "road" would be lower than traversing over tall "grass". This mapping from semantic labels to costs is usually defined by domain experts or learned from data. We can extend the cost of semantic labels to an entire trajectory as $c_{\text{traj}}(\tau) : \mathbb{T} \to \mathbb{R}^+ = \int_0^1 c_{\text{sem}}(x(\tau(t)))\,\mathrm{d}t$, where $\mathbb{T}$ is the space of all trajectories. Optimal path planning can be expressed as a constrained optimization problem:

$$\tau^* = \arg\min_{\tau \in \mathbb{T}} \int_0^1 c_{\text{sem}}(x(\tau(t)))\,\mathrm{d}t \qquad (1)$$
$$\text{s.t. } \tau(0) = s_{\text{curr}},\ \tau(1) = s_{\text{goal}}$$

In order to assist a global planner to plan trajectories that minimize costs over a long horizon, we predict semantics over a "far-field" rectangular region around the robot: $\mathbf{\Omega}_{\text{ffield}} \coloneqq \{s \in \mathbb{R}^2 \,|\, \|s - s_{\text{curr}}\|_1 \le w_{\text{ffield}}\}$. In a-priori unknown environments, the robot infers semantic labels using onboard sensors such as cameras and LiDARs to observe its surroundings, combined with perception algorithms that predict semantic labels from sensor data. However, due to limited range of onboard sensing, the robot can only detect semantic labels in a localized circular region around the robot's current position $\mathbf{\Omega}_{\text{sense}} \coloneqq \{s \in \mathbb{R}^2 \,|\, \|s - s_{\text{curr}}\|_2 \le r_{\text{sense}}\}$ that is much smaller than the far-field i.e. $r_{\text{sense}} \ll w_{\text{ffield}}$. This divides the far-field space $\mathbf{\Omega}_{\text{ffield}}$ into two sets: observed states $\mathbf{\Omega}_{\text{obs}} \subseteq \mathbf{\Omega}_{\text{ffield}}$ where an estimated semantic mapping $x(s)$ has been obtained, and unobserved states $\mathbf{\Omega}_{\text{unobs}} = \mathbf{\Omega}_{\text{ffield}} \setminus \mathbf{\Omega}_{\text{obs}}$ whose semantic labels have not been observed until the robot explores the environment more.

We assume that the robot is able to perfectly estimate the semantics of observed regions $x(\mathbf{\Omega}_{\text{obs}})$, and needs to infer the semantics of unobserved states $x(\mathbf{\Omega}_{\text{unobs}})$. The unobserved semantics, however, are not independent from the semantics of observed states. Strong spatial correlations are common in outdoor environments. For example, marshes are often close to bodies of water and roads are usually continuous and extend linearly. Therefore, the main objective of this paper is to accurately predict the conditional probability distribution $p(x(\mathbf{\Omega}_{\text{unobs}}) \,|\, x(\mathbf{\Omega}_{\text{obs}}))$. These predictions will then be used to plan trajectories that minimize navigation costs over the far-field region.

## III. Methods

Our uncertainty-aware global navigation framework is composed of three stages. First, given the observations $x(\mathbf{\Omega}_{\text{obs}})$, we produce an estimate of $p(x(\mathbf{\Omega}_{\text{unobs}}) \,|\, x(\mathbf{\Omega}_{\text{obs}}))$. Next, we accumulate the estimates over time into a cost map belief. Finally, we deploy a planner that computes the least-cost path from the current robot state $s_{\text{curr}}$ to the goal state $s_{\text{goal}}$ using the maximum a-posteriori (MAP) cost map estimate.

### A. Data pre-processing and notation

In order to leverage diffusion models (described in Sec. III-B) that operate in continuous spaces, we convert discrete semantic maps into continuous images. We associate each semantic category with a unique 3D RGB color tuple $\text{RGB}(c) : \mathbb{C} \to [0,1]^3$. Then, we define *continuous* semantic images for observed, unobserved and the entire far-field region as

$$\mathbf{x}^{\text{obs}}/\mathbf{x}^{\text{unobs}}/\mathbf{x}^{\text{ffield}} \coloneqq \{\text{RGB}(x(s)) \,|\, s \in \mathbf{\Omega}_{\text{obs}}/\mathbf{\Omega}_{\text{unobs}}/\mathbf{\Omega}_{\text{ffield}}\}$$

respectively. Examples of such images are shown in Fig. 1. We define the binary observation mask that indexes the observed pixels as $\mathbf{m}^{\text{obs}} \coloneqq \{\mathbb{I}(s \in \mathbf{\Omega}_{\text{obs}}) \,|\, s \in \mathbf{\Omega}_{\text{ffield}}\}$.

### B. Preliminaries: denoising diffusion probabilistic models

Diffusion models [20, 8, 21] are a recently-proposed, highly performant class of generative models that can learn and sample from a given *continuous* data distribution $\mathbf{x}_0 \in \mathbb{R}^D \sim p(\mathbf{x}_0)$. They have shown to be extremely effective in learning complex, multi-modal image distributions. In this work, we use Denoising Diffusion Probabilistic Models (DDPM) [8] to learn distributions over continuous far-field semantic images $\mathbf{x}_0 \equiv \mathbf{x}^{\text{ffield}}$. DDPM introduces $T$ latent variables $\mathbf{x}_1, \dots, \mathbf{x}_T$ by the "forward diffusion process" $q(\mathbf{x}_t \,|\, \mathbf{x}_{t-1}) \coloneqq \mathcal{N}\left(\mathbf{x}_t\,;\, \sqrt{1-\beta_t}\,\mathbf{x}_{t-1}, \beta_t \mathbf{I}\right)$ that progressively adds small amounts of *i.i.d.* Gaussian noise with variance $\beta_t > 0$ at each diffusion timestep $t \in \{1, \dots, T\}$.

**Algorithm 1** Unconditional diffusion: inpainting as inference

**Require:** Unconditional diffusion model $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t)$, $\mathbf{x}^{\text{obs}}$, $\mathbf{m}^{\text{obs}}$

1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $t = T, \dots, 1$ **do**
▷ *Standard reverse diffusion step* [8]
3: $\boldsymbol{\mu}_\theta(\mathbf{x}_t) \coloneqq \frac{1}{\sqrt{\alpha_t}} \left( \mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\overline{\alpha}_t}} \boldsymbol{\epsilon}_\theta(\mathbf{x}_t) \right)$
4: $\mathbf{x}_{t-1} \sim \mathcal{N}(\boldsymbol{\mu}_\theta(\mathbf{x}_t), \tilde{\beta}_t \mathbf{I})$
▷ *In-painting step* [21]
5: $\mathbf{x}^{\text{obs}}_{t-1} \sim \mathcal{N}\left(\sqrt{\overline{\alpha}_{t-1}}\mathbf{x}^{\text{obs}}, (1-\overline{\alpha}_{t-1})\mathbf{I}\right)$ ▷ *Noised* $\mathbf{x}^{\text{obs}}$
6: $\mathbf{x}_{t-1}\left[\mathbf{m}^{\text{obs}}\right] \coloneqq \mathbf{x}^{\text{obs}}_{t-1}$ ▷ *Overwrite observed region*
▷ *MCMC mixing steps to improve consistency* [21, 5]
7: **for** $k = 1, \dots, M_t$ **do**
8: $\mathbf{s}_\theta(\mathbf{x}_{t-1}) \coloneqq -\frac{1}{\sqrt{1-\overline{\alpha}_{t-1}}} \boldsymbol{\epsilon}_\theta(\mathbf{x}_{t-1})$ ▷ *Score function*
9: $\mathbf{x}_{t-1} \sim \mathcal{N}\left(\mathbf{x}_{t-1} + \lambda \mathbf{s}_\theta(\mathbf{x}_{t-1}), 2\lambda \mathbf{I}\right)$ ▷ *MCMC step*
10: **end for**
11: **end for**
12: **return** $\mathbf{x}_0$

With an appropriate noise schedule $\beta_{1:T}$, the final latent variable $\mathbf{x}_T$ is approximately normally distributed i.e. $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Because forward diffusion is Gaussian, latent variables from intermediate timesteps $\mathbf{x}_t$ can be directly sampled from $\mathbf{x}_0$ as $\mathbf{x}_t = \sqrt{\overline{\alpha}_t}\,\mathbf{x}_0 + \sqrt{(1-\overline{\alpha}_t)}\,\boldsymbol{\epsilon}$ where $\overline{\alpha}_t = \prod_{s=1}^{t}(1-\beta_t)$ [20, 8] and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$ is noise sampled from the standard Normal distribution. DDPM learns to denoise $\mathbf{x}_t$. It trains a neural network $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t)$ with weights $\theta$ to regress $\mathbf{x}_t$ to the noise $\boldsymbol{\epsilon}$ that generated $\mathbf{x}_t$ from $\mathbf{x}_0$. This allows DDPM to perform "reverse diffusion" that progressively denoises $\mathbf{x}_T$ to $\mathbf{x}_0$. In the limit of small $\beta_t$, the reversal of forward diffusion $q(\mathbf{x}_{t-1} \,|\, \mathbf{x}_t)$ becomes Gaussian [20] and can be approximated as $q(\mathbf{x}_{t-1} \,|\, \mathbf{x}_t) \approx p_\theta(\mathbf{x}_{t-1} \,|\, \mathbf{x}_t) = \mathcal{N}\left(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t), \tilde{\beta}_t\right)$, where $\boldsymbol{\mu}_\theta$ is a reparametrization of $\boldsymbol{\epsilon}_\theta$ (line 3 of Alg. 1). See Ho et al. [8] for more details.

Given a dataset of far-field semantic maps $\{\mathbf{x}^{\text{ffield}}_{(n)}\}_{n=1}^{N}$ obtained from [16], and observed semantics $\mathbf{x}^{\text{obs}}$ provided at test-time, we wish to predict the distribution $p(\mathbf{x}^{\text{unobs}} \,|\, \mathbf{x}^{\text{obs}})$ where $\mathbf{x}^{\text{ffield}} \equiv (\mathbf{x}^{\text{unobs}}, \mathbf{x}^{\text{obs}})$. In the navigation domain, the observed region is small relative to the unobserved region that is to be predicted. As a result, empirically, direct application of methods such as RePaint [13] fail to lead to plausible completions. See Appendix V for more detail. Additionally, we are sensitive to the compute time required to produce completions. To address these problems, we consider two approaches using diffusion models.

### C. Unconditional diffusion model: inpainting as inference

This is the conventional approach of fitting a diffusion model to fully-observed images $\{\mathbf{x}^{\text{ffield}}_{(n)}\}_{n=1}^{D}$ and learning $p_\theta(\mathbf{x}^{\text{ffield}})$. The diffusion model is "unconditional" because its noise function $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t)$ only takes the output of the previous diffusion step as input.

Given $(\mathbf{x}^{\text{obs}}, \mathbf{m}^{\text{obs}})$ at test time, we treat sampling from $p_\theta(\mathbf{x}^{\text{unobs}} \,|\, \mathbf{x}^{\text{obs}})$ as a test-time inference problem with respect to the learned joint model $p_\theta(\mathbf{x}^{\text{obs}}, \mathbf{x}^{\text{unobs}})$. In the image generation literature, this is called "inpainting" unknown pixels given known pixels. Unconditional diffusion models are known to perform test-time inpainting even if not explicitly trained to do so [20, 21, 13]. We find that vanilla inpainting struggles to generate samples $\mathbf{x}^{\text{unobs}}$ that are consistent with $\mathbf{x}^{\text{obs}}$, especially when the observed region is small. We overcome this problem by extending the inpainting approach to perform additional Langevin Markov Chain Monte Carlo (MCMC) steps at each noise level [21, 5]. MCMC steps encourage mixing of the sample at intermediate distributions $p_\theta(\mathbf{x}_t)$ of the reverse diffusion process resulting in increased consistency between $\mathbf{x}^{\text{obs}}_t$ and $\mathbf{x}^{\text{unobs}}_t$. Our detailed algorithm is outlined in Alg. 1.

**Algorithm 2** Conditional diffusion: inpainting mask as input

**Require:** Conditional model $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \mathbf{x}^{\text{obs}}, \mathbf{m}^{\text{obs}})$, $\mathbf{x}^{\text{obs}}$, $\mathbf{m}^{\text{obs}}$

1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $t = T, \dots, 1$ **do**
▷ *Conditional reverse diffusion step*
3: $\boldsymbol{\mu}_\theta(\mathbf{x}_t, \mathbf{x}^{\text{obs}}, \mathbf{m}^{\text{obs}}) \coloneqq \frac{1}{\sqrt{\alpha_t}} \left( \mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\overline{\alpha}_t}} \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \mathbf{x}^{\text{obs}}, \mathbf{m}^{\text{obs}}) \right)$
4: $\mathbf{x}_{t-1} \sim \mathcal{N}\left(\boldsymbol{\mu}_\theta(\mathbf{x}_t, \mathbf{x}^{\text{obs}}, \mathbf{m}^{\text{obs}}), \tilde{\beta}_t \mathbf{I}\right)$
5: **end for**
6: **return** $\mathbf{x}_0$

### D. Conditional diffusion model: inpainting mask as input

While test-time inference on unconditional diffusion models can generate high-fidelity inpaintings consistent with observed regions, the critical MCMC procedure can be very slow. Therefore, we also investigate a faster approach using *conditional* diffusion models [7, 22] that takes the observed data $(\mathbf{x}^{\text{obs}}, \mathbf{m}^{\text{obs}})$ as additional inputs. Specifically, we modify the U-Net [17] architecture of the diffusion model's noise function $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t)$ to take four additional input channels — three RGB channels corresponding to $\mathbf{x}^{\text{obs}}$ and one binary channel corresponding to $\mathbf{m}^{\text{obs}}$. The functional form of the *conditional* noise function now becomes $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \mathbf{x}^{\text{obs}}, \mathbf{m}^{\text{obs}})$.

This requires us to modify the training procedure and generate a dataset of pairs of semantic images and observation masks $\{(\mathbf{x}^{\text{ffield}}_{(n)}, \mathbf{m}^{\text{obs}}_{(n)})\}_{n=1}^{N}$. During training, we randomly generate masks that correspond to simple, linear robot trajectories. First, we sample a pixel on the perimeter of the image and create a line segment between the selected pixel and the center of the image; the line segment corresponds to a hypothetical robot trajectory. The observed region constitutes all pixels that are within the sensor range ($r_{\text{sense}}$) from some point along the trajectory. The mask $\mathbf{m}^{\text{obs}}$ is set to 1 in the observed region and 0 in the unobserved regions. Once the conditional diffusion model is trained, the test-time inpainting procedure is a straightforward extension of the standard reverse diffusion process outlined in Alg. 2 where $(\mathbf{x}^{\text{obs}}, \mathbf{m}^{\text{obs}})$ are passed as additional inputs.

### E. Fusing observations

The estimator maintains a belief over the cost of traversing through a location $s$ given previous observations,

$p(c_{\text{sem}}(s)|\mathbf{x}^{\text{obs}}_{1:t})$. The area is discretized into a grid and we assume that the cost belief for each cell is a Gaussian which is independent of all other cells. The samples from the diffusion model are converted into sampled semantic maps by assigning to each pixel the class with the nearest associated color in the RGB color space. For each pixel in the sampled semantic maps, we form Gaussian observations by selecting a mean and a measurement noise. The mean is selected as the cost associated with the most commonly occurring class type at that location. The measurement noise is computed as the fraction of samples that disagree with the most commonly occurring class type. The cost belief at each cell is updated using the Kalman Filter update equations.

## IV. Experiments

**Environment and trials:** We explore the characteristics of our approach outlined in Sec. III in simulated long-horizon navigation problems. For training, a total of $25,000$ semantic maps are sampled evenly from five counties mapped by the Chesapeake Bay Land Use and Land Cover (LULC) project [16], which has mapped land cover and land use over the 250,000 square kilometers that makes up the Chesapeake Bay watershed at meter-per-pixel resolution. Each semantic map is scaled to be a $128\times128$ semantic image corresponding to 256 meter $\times$ 256 meter area ($w_{\text{ffield}} = 128m$). For testing, we sampled ten 5x5 km maps from counties where training data was not drawn from. In each map, five feasible start-goal pairs were sampled. We translate each semantic class into a manually picked traversal speed. These maps capture real semantic distributions across the northeastern United States. An example map is shown in the appendix.

**Baseline and oracles:** We assume that the robot can sense within a radius $r_{\text{sense}} = 50$ meters. We compare our diffusion approaches to a baseline (*Base-50m*) that observes the ground truth costs also within 50 meter radius of $s_{\text{curr}}$. To benchmark performance, we also report the performance of "oracle" policies: *Oracle-60m*, *Oracle-70m*, and *Oracle-100m* that have farther sensing capabilities. We evaluate two variants of our approach: *Uncond-Diff* using unconditional diffusion described in Sec. III-C, and *Cond-Diff* using conditional diffusion described in Sec. III-D.

In each problem, the robot iterates through a traditional sense-plan-act cycle. Costmaps are updated with local and diffusion observations if applicable. Replanning is performed on the updated costmaps with D*lite [12]. Finally the agent navigates 20m along the most recent trajectory before repeating. Trials are terminated when the agent reaches the goal. Diffusion observations are only employed every five observation cycles.

**Performance:** As of writing only a preliminary investigation with 50 trials has been completed and is reported in Appendix III. Since the trails were limited, the results we obtained were not statistically significant. Furthermore, we were only able to compute average case performance in our preliminary investigation where our methods didn't show statistically significant positive or negative trends. We intend to perform a more focused study to characterize

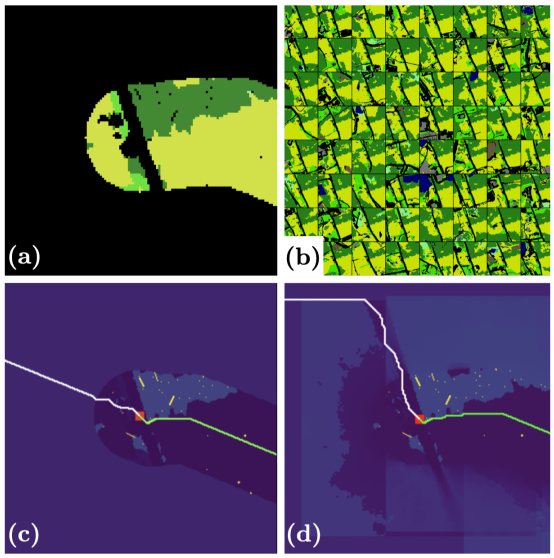


Fig. 2: Visualization of an instance where *Cond-Diff* significantly improved navigation. **(a)** The robot's semantic observation history. **(b)** Samples from the diffusion model conditioned on observed semantics. **(c)** Baseline plan that doesn't use far-field predictions. The red square is the robot's position, green line the robot's state history, and white line the robot's future plan. The robot's goal lies up and left from its current position. The baseline using a nominal unobserved cost drives the robot into the field. **(d)** Costmap incorporating semantics from diffusion model and corresponding plan. The inferred costs direct the robot towards the road instead of the field. This decision reduces the total accumulated cost by 23 minutes in a trial that took the baseline 59 minutes to complete.

situations where diffusion-based inpainting outperforms or underperforms the baseline. However, our explorations and qualitative results suggest that diffusion-based inpainting is a promising research direction that can potentially improve navigation performance; we will work on more thorough quantitative evaluation after the deadline.

**Qualitative example**: Figure 2 demonstrates a case in a trial where using learned environment structure improves navigation, reducing the accumulated cost by $39\%$ for the trial. Please see caption for details. We provide an additional example of inpainting benefiting navigation, and structure completion with uncertainty in the Appendix.

## V. Future Work

The proposed approach has many limitations we are interested in exploring. Diffusing samples is expensive, and intelligently allocating compute to predict environment structure when it is likely to be most impactful would free up platform resources. It would also be interesting to deploy this approach at multiple scales. Instead of inpainting in the current robot's state, clever selection of trajectory history and inpainting window could enable our approach to capture influential environmental structure at much coarser (or finer) spatial scales. We observed that the number of MCMC steps required to produce globally consistent inpainting samples (Sec. III-C) can vary significantly depending on the size and properties of $\mathbf{x}^{\text{obs}}$. Adaptively tuning the number of MCMC steps to trade-off speed and accuracy in a principled manner is another promising direction for future work.

# Appendix

## Learning semantic traversability priors using diffusion models for uncertainty-aware global path planning

### Appendix I
### Related Works

Safely and efficiently navigating through a priori unknown natural environments requires richer environment representations than occupancy maps traditionally used for indoor settings.

To form these richer representations recent approaches often learn to map sensor data directly to traversability maps or costmaps [2, 19, 6]. Approaches also often leverage semantics [14] (usually extracted from images [3] or lidar [23]) and map these semantics to costs either manually, or via learning from experience [1].

To fuel these richer representations, dense sensor data is usually required, which limits the effective sensing range to tens of meters. With this narrow view of its surroundings, myopic behavior is common which puts the robot at higher risk and increases navigation cost. Even with recent approaches to extend this range [4], extending effective observation range in settings with predictable environment structure can still benefit navigation performance.

Previous works for outdoor navigation have leveraged inpainting to produce dense maps from features extracted from sensor data that may not provide dense coverage of a region [19, 15]. However, these works don't focus on extending the effective range of their mapping techniques and do not estimate confidences over inpainting decisions, which is critical to extend inference beyond regions with strong observational support.

Closer in spirit to our work, [18] investigates learning topological structure of subterranean environments to aid with the task of goal selection for efficient exploration of unknown environments. Our work differs in application to navigation in more natural environments where topological maps are not sufficient for safe and efficient navigation.

In [9] the authors more explicitly look at predicting environment structure beyond the range of the robot's sensors by inpainting occupancy grids beyond the sensors effective range. In extensions to this work the authors added uncertainty over the inpainted maps by looking at inter-batch variation [10], and demonstrate their approach can help high-speed vehicles extend local planning horizons beyond traditional sensor ranges, improving performance [11]. In this work, we focus more explicitly on using environment structure to inform global navigation with richer costmaps. Additionally, we leverage the uncertainty produced by our approach to fuse observations over time.

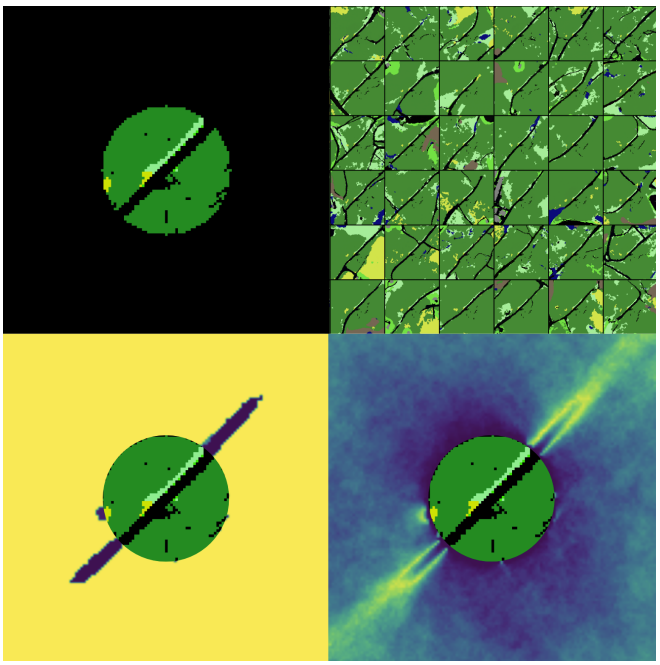

Fig. 3: Example of road completion. Given the semantic observation shown in the top left, the diffusion model proposes environment samples shown in the top right. These samples produce the costmap in the bottom left and the uncertainty map in the bottom right. In both darker color indicates lower values.)

### Appendix II
### Additional Qualitative Navigation Results

One example of learned structure can be seen in Figure 3. Here, the robot sits on a road surrounded by forest. The strong structure in roads results in many of the sampled environments shown in the top right continuing the observed road, resulting in the low-cost continuation of the road in the produced costmap and low uncertainty region under the predicted continuation of the road. Around this low uncertainty region lies the highest uncertainty in the image. This is caused by the sampled roads diverging in direction.

With less structure, the uncertainty increases with distance from observed semantics faster around the forest than the road, as sampled worlds diverge further from the known semantics. Though uncertainty increases, the dominant class remains forest across the samples, resulting in the high cost surrounding the road in the costmap.

### Appendix III
### Preliminary quantitative results

In this submission, we conduct a preliminary investigation using 50 trials. We report:

1) The percentage of total trials that improved (over baseline).

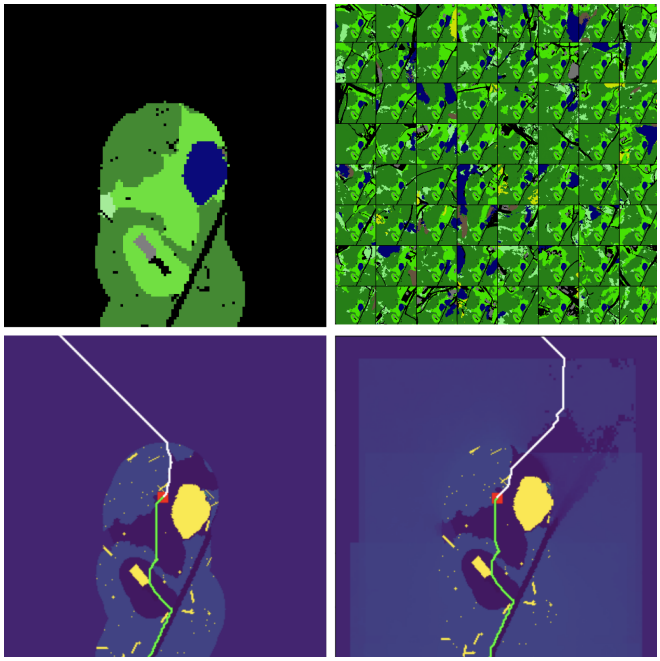

Fig. 4: Visualization of an instance where *Cond-Diff* significantly improved navigation. The top left image shows the robot's semantic observation history, used to condition the samples shown in the top right. These samples are reduced and added to the costmap shown in the bottom right, where the red square is the robot's position, green line the robot's state history, and white line the robot's future plan. The robot's goal lies up and left from its current position. The inferred costs direct the robot towards the road instead of towards the field as the baseline does in the bottom left image. This decision reduces the total accumulated cost by 6 minutes in a trial that took the baseline 106 minutes to complete

2) Average percent difference of cost (time, compared to baseline cost) across trials.

We report these results in Table I.

As of writing with the limited number of trials the results we obtained were not statistically significant. Furthermore, we were only able to compute average case performance in our preliminary investigation where our methods didn't show statistically significant positive or negative trends. We intend to perform a more focused study to characterize situations where diffusion-based inpainting outperforms or underperforms the baseline. However, our explorations and qualitative results suggest that diffusion-based inpainting is a promising research direction that can potentially improve navigation performance.

| Method | Percentage of trials improved | Average percent of cost increase |
|---|---|---|
| *Cond-Diff* | $52\% \pm 14\%$ | $1.1\% \pm 4.9\%$ |
| *Uncond-Diff* | $27\% \pm 15\%^*$ | $9.3\% \pm 5.9\%^*$ |
| *Oracle-60m* | $70\% \pm 12\%$ | $1.1\% \pm 3.8\%$ |
| *Oracle-70m* | $70\% \pm 12\%$ | $-0.81\% \pm 3.8\%$ |
| *Oracle-100m* | $84\% \pm 10\%$ | $-8.7\% \pm 3.5\%$ |

TABLE I: Results of preliminary evaluation. Intervals indicate 95% confidence. * indicates that only 33/50 trials were completed. Negative differences indicate reduction in cost.

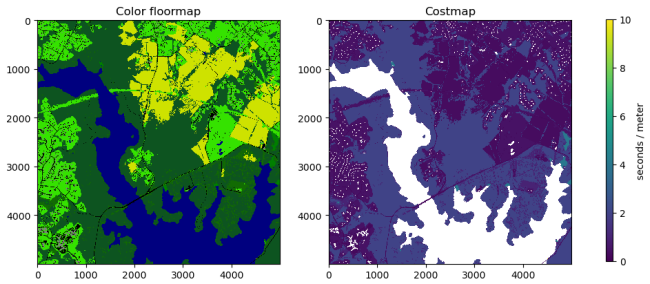


Fig. 5: An example of a semantic map and its corresponding costmap sampled from [16].

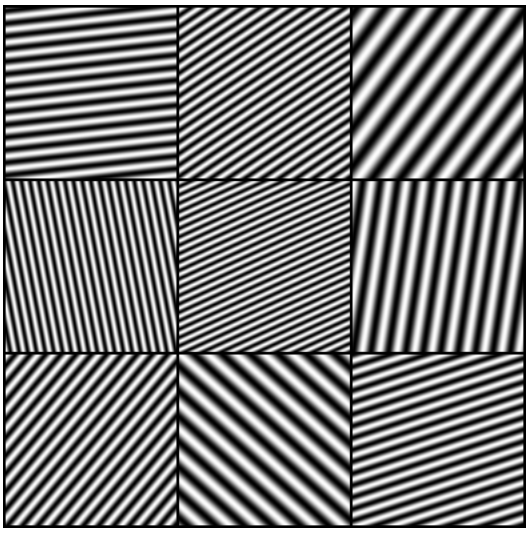

Fig. 6: Example datapoints from the STRIPES dataset.

## Appendix IV
## Example of Evaluation Environment

An example evaluation environment is shown in Figure 5. Semantics from this 5km $\times$ 5km map are translated into costs (traversal times, in seconds) manually, producing the costmap on the right of the figure.

## Appendix V
## Diffusion Sampling Analysis on Toy Dataset

Judging the quality of completions can be difficult when working with real data. We first choose to examine the completions for a simplified dataset and then extrapolate our learnings to our dataset of interest. We start by introducing the *STRIPES dataset*, which contains 25,000 images of stripes of varying thickness and orientation. The STRIPES dataset is generated by randomly sampling spatial sinusoids phases and angles spanning in $[0, 2\pi]$, and wavelengths spanning from $[5, 20]$ pixels. For some examples, see Figure 6. This dataset contains global structure that a diffusion model must learn to accurately generate, akin to global spatial dependencies present in real terrain images. With this dataset, implausible completions are very easy to judge. Using the observations shown in Figure 7, we show the generated completions.

We consider two types of masks shown in Fig. 7. Fig. 7 (*left*) contains a smaller observed mask with larger

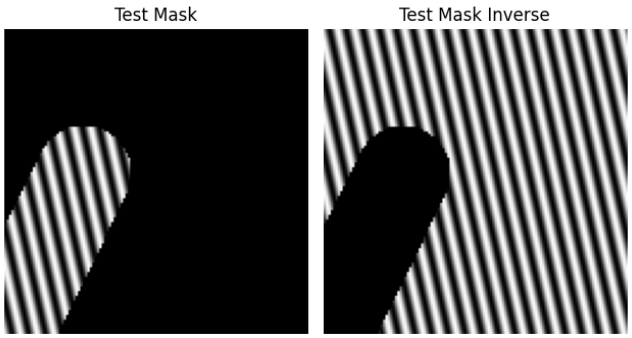


Fig. 7: Masks used to generate completions on STRIPES dataset.

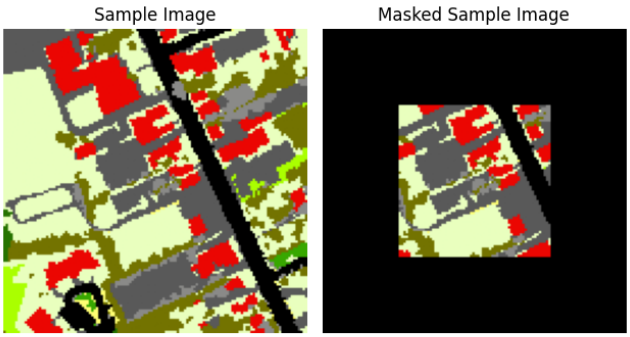


Fig. 9: Sample datapoint and mask generate completions on LULC dataset.

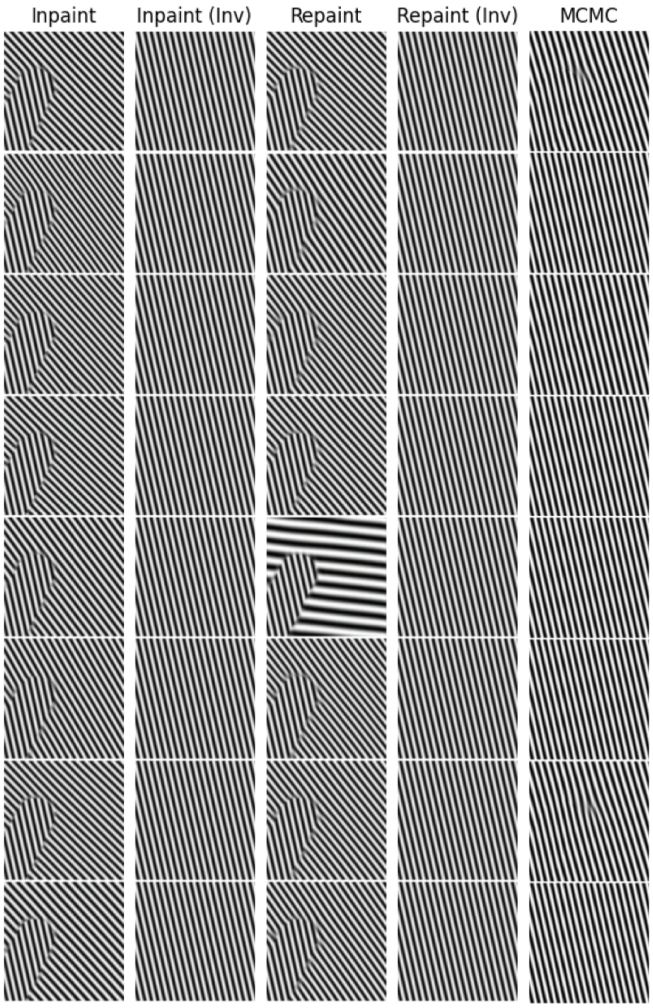


Fig. 8: Example completions from different methods.

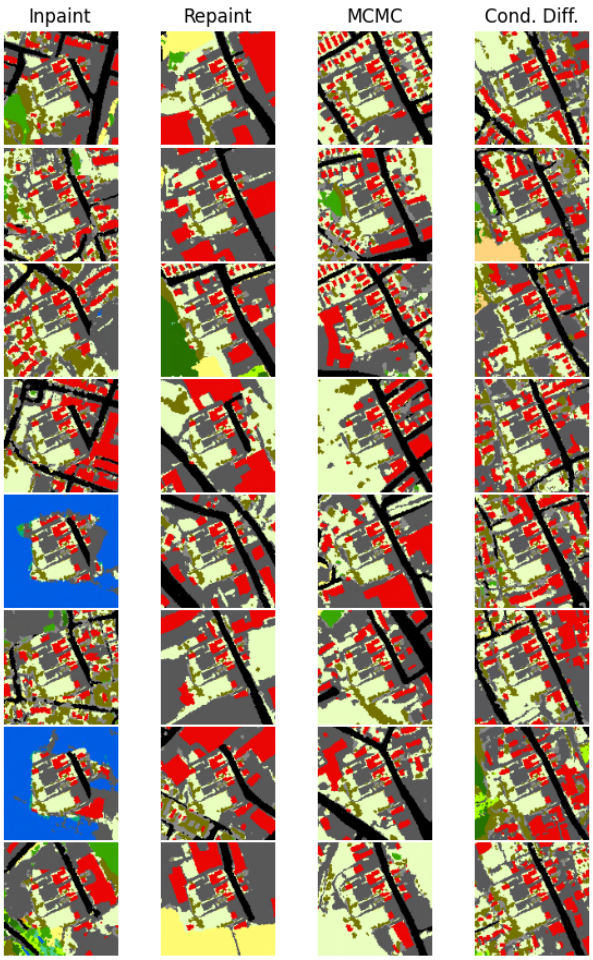


Fig. 10: Example completions on LULC dataset.

unobserved region; this is a harder task to infill. Fig. 7 (*right*), which call the "inverse" mask contains a larger observed region mask with smaller unobserved region; this is an easier task to infill. We find in Fig. 8 that the vanilla inpainting method and RePaint [13] are able to solve the easy problem but struggle with the hard problem.

Our insight is that by performing additional MCMC updates at each denoising timestep as described in Alg. 1, the intermediate distributions are able to mix better and the reverse diffusion process produces inpainting that are consistent with the observed regions.

We make a similar observation with the LULC dataset [16] in Fig. 10. When the problem is hard (i.e. the observed region is small), vanilla inpainting and RePaint [13] produce inpaintings that are essentially independent of the observed regions. Both (1) the MCMC-based reverse diffusion (on the unconditioned model) as well as (2) the conditional diffusion model that is explicitly trained to perform inpainting, are both able to produce inpaintings consistent with the observed regions.

The number of MCMC steps required to produce globally consistent inpainting samples (Sec. III-C, Alg. 1) is low for "easy" problems, and high for "hard" problems. Since the required number of MCMC steps can vary significantly depending on the size and properties of $\mathbf{x}^{\text{obs}}$, estimating how challenging a given inference problem at test-time is, and adaptively tuning the number of MCMC steps to trade-off speed and accuracy in a principled manner is a promising direction for future work.

## Acknowledgments

Research was sponsored by the Army Research Office and was accomplished under Cooperative Agreement Number W911NF-21-2-0150. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. This research was supported by the Office of Naval Research under Contract W911NF-17-2-0181, the Low Cost Autonomous Navigation and Semantic Mapping in the Littorals program and their support is gratefully acknowledged. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein. The lead author was supported by the Department of Defense (DoD) through the National Defense Science & Engineering Graduate (NDSEG) Fellowship Program.

## References

[1] Xiaoyi Cai, Michael Everett, Jonathan Fink, and Jonathan P How. Risk-aware off-road navigation via a learned speed distribution map. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2931–2937. IEEE, 2022.

[2] Mateo Guaman Castro, Samuel Triest, Wenshan Wang, Jason M Gregory, Felix Sanchez, John G Rogers, and Sebastian Scherer. How does it feel? Self-supervised costmap learning for off-road vehicle traversability. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 931–938. IEEE, 2023.

[3] Ping Chao, Chao-Yang Kao, Yu-Shan Ruan, Chien-Hsiang Huang, and Youn-Long Lin. HarDNet: A low memory traffic network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3552–3561, 2019.

[4] Eric Chen, Cherie Ho, Mukhtar Maulimov, Chen Wang, and Sebastian Scherer. Learning-On-The-Drive: Self-supervised adaptation of visual offroad traversability models. *arXiv preprint arXiv:2306.15226*, 2023.

[5] Yilun Du, Conor Durkan, Robin Strudel, Joshua B Tenenbaum, Sander Dieleman, Rob Fergus, Jascha Sohl-Dickstein, Arnaud Doucet, and Will Sussman Grathwohl. Reduce, Reuse, Recycle: Compositional generation with energy-based diffusion models and MCMC. In *International Conference on Machine Learning (ICML)*, pages 8489–8510. PMLR, 2023.

[6] Jonas Frey, Matias Mattamala, Nived Chebrolu, Cesar Cadena, Maurice Fallon, and Marco Hutter. Fast traversability estimation for wild visual navigation. *arXiv preprint arXiv:2305.08510*, 2023.

[7] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[8] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems (NeurIPS)*, 33:6840–6851, 2020.

[9] Kapil Katyal, Katie Popek, Chris Paxton, Joseph Moore, Kevin Wolfe, Philippe Burlina, and Gregory D Hager. Occupancy map prediction using generative and fully convolutional networks for vehicle navigation. *arXiv preprint arXiv:1803.02007*, 2018.

[10] Kapil Katyal, Katie Popek, Chris Paxton, Phil Burlina, and Gregory D Hager. Uncertainty-aware occupancy map prediction using generative networks for robot navigation. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 5453–5459. IEEE, 2019.

[11] Kapil D Katyal, Adam Polevoy, Joseph Moore, Craig Knuth, and Katie M Popek. High-speed robot navigation using predicted occupancy maps. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5476–5482. IEEE, 2021.

[12] Sven Koenig and Maxim Likhachev. D* lite. In *Eighteenth national conference on Artificial intelligence*, pages 476–483, 2002.

[13] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. RePaint: Inpainting Using Denoising Diffusion Probabilistic Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11461–11471, June 2022.

[14] Daniel Maturana, Po-Wei Chou, Masashi Uenoyama, and Sebastian Scherer. Real-Time Semantic Mapping for Autonomous Off-Road Navigation. In *Field and Service Robotics (FSR)*, pages 335–350, Cham, 2018. Springer International Publishing. ISBN 978-3-319-67361-5.

[15] Xiangyun Meng, Nathan Hatch, Alexander Lambert, Anqi Li, Nolan Wagener, Matthew Schmittle, JoonHo Lee, Wentao Yuan, Zoey Chen, Samuel Deng, et al. TerrainNet: Visual Modeling of Complex Terrain for High-speed, Off-road Navigation. In *Proceedings of Robotics: Science and Systems*, Daegu, Republic of Korea, July 2023.

[16] Chesapeake Bay Program. Chesapeake Bay Land Use and Land Cover (LULC) Database 2022 Edition. *Data Release*, 2023. doi: 10.5066/P981GV1L.

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

[18] Manish Saroya, Graeme Best, and Geoffrey A Hollinger. Online exploration of tunnel networks leveraging topological cnn-based world predictions. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6038–6045. IEEE, 2020.

[19] Amirreza Shaban, Xiangyun Meng, JoonHo Lee, Byron Boots, and Dieter Fox. Semantic terrain classification for off-road autonomous driving. In Aleksandra Faust, David Hsu, and Gerhard Neumann, editors, *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine*

*Learning Research*, pages 619–629. PMLR, 08–11 Nov 2022. URL https://proceedings.mlr.press/v164/shaban22a.html.

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

[21] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-Based Generative Modeling through Stochastic Differential Equations. In *9th International Conference on Learning Representations (ICLR), 2021, Virtual Event, Austria, May 3-7, 2021*, 2021.

[22] Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. CSDI: Conditional score-based diffusion models for probabilistic time series imputation. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[23] Xinge Zhu, Hui Zhou, Tai Wang, Fangzhou Hong, Yuexin Ma, Wei Li, Hongsheng Li, and Dahua Lin. Cylinder3D: Cylindrical and asymmetrical 3D convolution networks for LiDAR segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9939–9948, June 2021.