

Pathdreamer: A World Model for Indoor Navigation

Jing Yu Koh¹Honglak Lee²Yinfei Yang¹Jason Baldridge¹Peter Anderson¹¹Google Research²University of Michigan

Abstract

People navigating in unfamiliar buildings take advantage of myriad visual, spatial and semantic cues to efficiently achieve their navigation goals. Towards equipping computational agents with similar capabilities, we introduce Pathdreamer, a visual world model for agents navigating in novel indoor environments. Given one or more previous visual observations, Pathdreamer generates plausible high-resolution 360° visual observations (RGB, semantic segmentation and depth) for viewpoints that have not been visited, in buildings not seen during training. In regions of high uncertainty (e.g. predicting around corners, imagining the contents of an unseen room), Pathdreamer can predict diverse scenes, allowing an agent to sample multiple realistic outcomes for a given trajectory. We demonstrate that Pathdreamer encodes useful and accessible visual, spatial and semantic knowledge about human environments by using it in the downstream task of Vision-and-Language Navigation (VLN). Specifically, we show that planning ahead with Pathdreamer brings about half the benefit of looking ahead at actual observations from unobserved parts of the environment. We hope that Pathdreamer will help unlock model-based approaches to challenging embodied navigation tasks such as navigating to specified objects and VLN.

1. Introduction

World models [23], or models of environments [72], are an appealing way to represent an agent’s knowledge about its surroundings. An agent with a world model can predict its future by ‘imagining’ the consequences of a series of proposed actions. This capability can be used for sampling-based planning [16, 57], learning policies directly from the model (i.e., learning in a dream) [17, 23, 64, 25], and for counterfactual reasoning [6]. Model-based approaches such as these also typically improve the sample efficiency of deep reinforcement learning [72, 62]. However, world models that generate high-dimensional visual observations (i.e., images) have typically been restricted to relatively simple environments, such as Atari games [62] and tabletops [16].

Our goal is to develop a generic visual world model for agents navigating in indoor environments. Specifically,



Figure 1: Generating photorealistic 360° visual observations from an imagined 6.3m trajectory in a previously unseen building. Observations also include depth and segmentations (not shown here).

given one or more previous observations and a proposed navigation action sequence, we aim to generate plausible high-resolution visual observations for viewpoints that have not been visited, and do so in buildings not seen during training. Beyond applications in video editing and content creation, solving this problem would unlock model-based methods for many embodied AI tasks, including navigating to objects [5], instruction-guided navigation [3, 66, 40] and dialog-guided navigation [74, 26]. For example, an agent asked to find a certain type of object in a novel building, e.g. ‘find a chair’, could perform mental simulations using the world model to identify navigation trajectories that are most likely to include chair observations – without moving.

Building such a model is challenging. It requires synthesizing completions of partially visible objects, using as few as one previous observation. This is akin to novel view synthesis from a single image [19, 80], but with potentially unbounded viewpoint changes. There is also the related but considerably more extreme challenge of predicting around corners. For example, as shown in Fig. 1, any future navigation trajectory passing the entrance of an unseen room requires the model to plausibly imagine the entire contents

of that room (we dub this the *room reveal* problem). This requires generalizing from the visual, spatial and semantic structure of previously explored environments—which in our case are photo-realistic 3D captures of real indoor spaces in the Matterport3D dataset [7]. A third problem is temporal consistency: predictions of unseen building regions should ideally be stochastic (capturing the full distribution of possible outcomes), but revisited regions should be rendered in a consistent manner to previous observations.

Towards this goal, we introduce Pathdreamer. Given one or more visual observations (consisting of RGB, depth and semantic segmentation for panoramas) of an indoor scene, Pathdreamer synthesizes high-resolution visual observations (RGB, depth and semantic segmentations) along a specified trajectory through future viewpoints, using a hierarchical two-stage approach. Pathdreamer’s first stage, Structure Generator, generates depth and semantic segmentations. Inspired by video prediction [11], these outputs are conditioned on a latent noise tensor capturing the stochastic information about the next observation (such as the layout of an unseen room) that cannot be predicted deterministically. The second stage’s Image Generator renders the depth and semantic segmentations as realistic RGB images using modified Multi-SPADE blocks [63, 51]. To maintain long-term consistency in the generated observations, both stages use back-projected 3D point cloud representations which are re-projected into image space for context [51].

Pathdreamer can generate plausible views for unseen indoor scenes under large viewpoint changes (see Figure 1), while also addressing the *room reveal* problem – in this case correctly hypothesizing that the unseen room revealed at position 2 will most likely resemble a kitchen. Empirically, using the Matterport3D dataset [7] and 360° observations, we evaluate both stages of our model against prior work and reasonable baselines and ablations. We find that the hierarchical structure of the model is essential for predicting over large viewpoint changes, that maintaining both RGB and semantic context is required, and that prediction quality degrades gradually when we evaluate with trajectory rollouts of up to 13m (with viewpoints 2.25m apart on average).

Encouraged by these results, we investigate whether Pathdreamer’s RGB predictions can improve performance on a downstream task: Vision-and-Language Navigation (VLN), using the R2R dataset [3]. VLN requires agents to interpret and execute natural language navigation instructions in a photorealistic 3D environment. A robust finding from previous research is that performance improves dramatically when agents can look ahead at unobserved parts of the environment while following an instruction [50]. We show that replacing look-ahead observations with Pathdreamer predictions maintains around half of the gains, a finding we expect to have significant implications for VLN research. In summary, our main contributions include:

- Proposing the study of visual world models for generic indoor environments and defining evaluation protocols and baselines for future work.
- Pathdreamer, a stochastic hierarchical visual world model combining multiple, independent threads of previous work on video prediction [11], semantic image synthesis [63] and video-to-video synthesis [51].
- Extensive experiments characterizing the performance of Pathdreamer and demonstrating improved results on the downstream VLN task [3].

2. Related Work

Video Prediction Our work is closely related to the task of video prediction, which aims to predict the future frames of a video sequence. While some video prediction methods predict RGB video frames directly [76, 1, 41, 44], many others use hierarchical models to first predict an intermediate representation (such as semantic segmentation) [47, 35, 77, 82, 42], which improves the fidelity of long-term predictions [42]. Several approaches have also incorporated 3D point cloud representations, using projective camera geometry to explicitly infer aspects of the next frame [75, 51, 43]. Inspired by this work, we adopt and combine both the hierarchical two-stage approach and 3D point cloud representations. Further, since our interest is in action-conditional world models, we provide a trajectory of future viewpoints to the model rather than assuming a constant frame rate and modeling camera motion implicitly, which is more typical in video generation [44, 42].

Action-Conditional Video Prediction Conditional video prediction to improve agent reasoning and planning has been explored in several tasks. This includes video prediction for Atari games conditioned on control inputs [60, 10, 62, 25] and 3D game environments like Doom [23]. In robotics, action-conditional video prediction has been investigated for object pushing in tabletop settings to improve generalization to novel objects [15, 16, 14]. This work has been restricted to simple environments and low-resolution images, such as 64×64 images of objects in a wooden box. To the best of our knowledge, we are the first to investigate action-conditional video prediction in building-scale environments with high-resolution (1024×512) images.

World Models and Navigation Priors World models [23] are an appealing way to summarize and distill knowledge about complex, high-dimensional environments. However, world models can differ in their outputs. While Pathdreamer predicts visual observations, there is also a vast literature on world models that predict compact latent representations of future states [38, 24, 25] or other task-specific measurements [13] or rewards [61]. This includes recent work attempting to learn statistical regularities and other priors for indoor navigation—for example, by min-

ing spatial co-occurrences from real estate video tours [8], learning to predict top-down belief maps over room characteristics [58], or learning to reconstruct house floor plans using audio and visual cues from a short video sequence [65]. In contrast to these approaches, we focus on explicitly predicting visual observations (i.e., pixels) which are generic, human-interpretable, and apply to a wide variety of downstream tasks and applications. Further, recent work identifies a close correlation between image prediction accuracy and downstream task performance in model-based RL [4].

Embodied Navigation Agents High-quality 3D environment datasets such as Matterport3D [7], StreetLearn [56, 53], Gibson [81] and Replica [71] have triggered intense interest in developing embodied agents that act in realistic human environments [2]. Tasks of interest include Object-Nav [5] (navigating to an instance of a particular kind of object), and Vision-and-Language Navigation (VLN) [3], in which agents must navigate according to natural language instructions. Variations of VLN include indoor navigation [3, 33, 66, 40], street-level navigation [9, 53], vision-and-dialog navigation [59, 74, 26], VLN in continuous environments [39], and more. Notwithstanding considerable exploration of pretraining strategies [46, 27, 50, 87], data augmentation approaches [20, 21, 73], agent architectures and loss functions [86, 48, 49], existing work in this space considers only model-free approaches. Our aim is to unlock model-based approaches to these tasks, using a visual world model to encode prior commonsense knowledge about human environments and thereby relieve the burden on the agent to learn these regularities. Underscoring the potential of this direction, we note that using the ground-truth environment for planning with beam search typically improves VLN success rates on the R2R dataset by 17-19% [20, 73].

Novel View Synthesis We position our work with respect to novel view synthesis [19, 37, 29, 18, 70, 85, 54]. Methods for representing 3D scenes include point cloud representations [80], layered depth images [12], and mesh representations [68]. Recently, neural radiance fields (NeRF) [55, 52, 83] achieved impressive results by capturing volume density and color implicitly with a neural network. NeRF models synthesize very high quality 3D scenes, but a significant drawback for our purposes is that they require a large number of input views to render a single scene (e.g., 20–62 images per scene in [55]). More importantly, these models are typically trained to represent a single scene, and do not yet generalize well to unseen environments. In contrast, our problem demands generalization to unseen environments, using as little as one previous observation.

3. Pathdreamer

Pathdreamer is a world model that generates high-resolution visual observations from a trajectory of future

viewpoints in buildings it has never observed. The input to Pathdreamer is a sequence of previous observations consisting of RGB images $I_{1:t-1}$, semantic segmentation images $s_{1:t-1}$, and depth images $d_{1:t-1}$ (where the depth and segmentations could be ground-truth or estimates from a model). We assume that a corresponding sequence of camera poses $T_{1:t-1}$ is available from an odometry system, and that the camera intrinsics are known or estimated. Our goal is to generate realistic RGB, semantic segmentation and depth images for a trajectory of future poses T_t, T_{t+1}, \dots, T_T , which may be provided up front or iteratively by some agent interacting with the returned observations. Note that we generate depth and segmentation because these modalities are useful in many downstream tasks. We assume that the future trajectory may traverse unseen areas of the environment, requiring the model to not only in-fill minor object dis-occlusions, but also to imagine entire room reveals (Figure 1).

Figure 2 shows our proposed hierarchical two-stage model for addressing this challenge. It uses a latent noise tensor z_t to capture the stochastic information about the next observation (e.g. the layout of an unseen room) that cannot be predicted deterministically. Given a sampled noise tensor z_t , the first stage (Structure Generator) generates a new depth image \hat{d}_t and segmentation image \hat{s}_t to provide a plausible high-level semantic representation of the scene, using as context the previous semantic and depth images $s_{1:t-1}, d_{1:t-1}$. In the second stage (Image Generator), the predicted semantic and depth images \hat{s}_t, \hat{d}_t are rendered into a realistic RGB image \hat{I}_t using previous RGB images $I_{1:t-1}$ as context. In each stage, context is provided by accumulating previous observations as a 3D point cloud which is re-projected into 2D using T_t .

3.1. Structure Generator: Segmentation & Depth

Pathdreamer’s first stage is the Structure Generator, a stochastic encoder-decoder network for generating diverse, plausible segmentation and depth images. Like [51], to provide the previous observation context, we first back-project the previous segmentations $s_{1:t-1}$ into a unified 3D semantic point cloud using the depth images $d_{1:t-1}$ and camera poses $T_{1:t-1}$. We then re-project this point cloud back into pixel space using T_t to create sparse segmentation and depth guidance images s'_t, d'_t which reflect the current pose.

The input to the encoder is a one-hot encoding of the semantic guidance image $s'_t \in \mathbb{R}^{W \times H \times C}$, concatenated with the depth guidance image $d'_t \in \mathbb{R}^{W \times H \times 1}$. The architecture of the encoder-decoder model is based on RedNet [34] – a ResNet-50 [28] architecture designed for indoor RGB-D semantic segmentation. RedNet uses transposed convolutions for upsampling in the decoder and skip connections between the encoder and decoder to preserve spatial information. Since the input contains a segmentation image, and segmentation classes differ across datasets, the encoder-

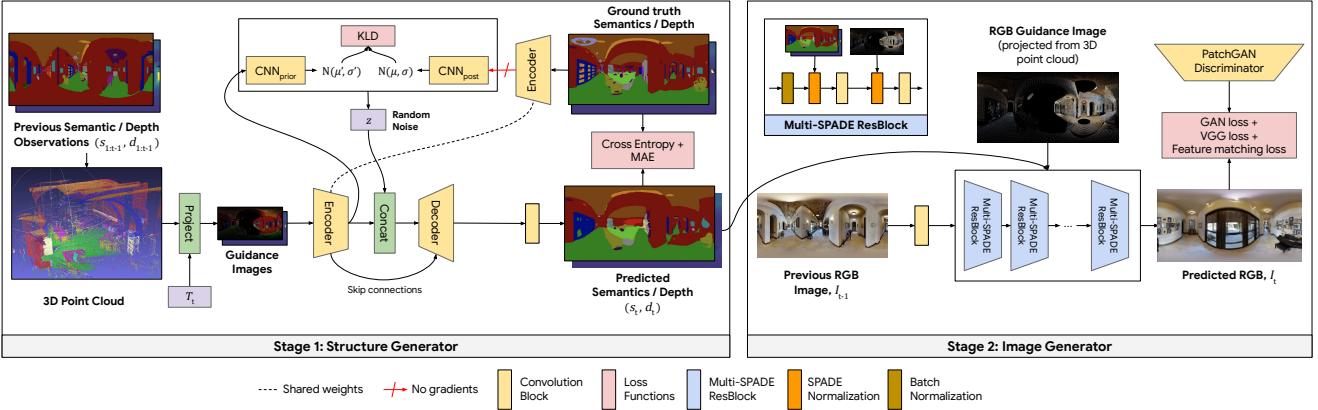


Figure 2: Pathdreamer model architecture at step t . Given a history of visual observations (RGB, depth and semantics) and a trajectory of future viewpoints, the Structure Generator conditions on a sampled noise tensor before generating semantic and depth outputs to provide a high-level structural representation of the scene. Realistic RGB images are synthesized by the Image Generator in the second stage.

decoder is not pretrained. We introduce the latent spatial noise tensor $z_t \in \mathbb{R}^{H' \times W' \times 32}$ into the model by concatenating it with the feature map between the encoder and the decoder. The final output of the encoder-decoder model is a segmentation image \hat{s}_t and a depth image \hat{d}_t , with segmentation predictions generated by a C -way softmax and depth outputs normalized in the range $(0, 1)$ and generated via a sigmoid function. At each step during inference, the segmentation prediction \hat{s}_t is back-projected and added to the point cloud to assist prediction in future timesteps.

To generate the noise tensor z_t , we take inspiration from SVG [11] and learn a conditional prior noise distribution $p_\psi(z_t|s'_t, d'_t)$. Intuitively, there are many possible scenes that may be generated for an unseen building region. We would like z_t to carry the stochastic information about the next observation that the deterministic encoder cannot capture, and we would like for the decoder to make good use of that information. During training, we encourage the first outcome by using a KL-divergence loss to force the prior distribution $p_\psi(z_t|s'_t, d'_t)$ to be close to the posterior distribution $q_\phi(z_t|s_t, d_t)$ which is conditioned on the ground-truth segmentation and depth images. We encourage the second outcome by providing the decoder with sampled z_t values from the posterior distribution q_ϕ (conditioned on the ground-truth outputs) during training. During inference, the latent noise z_t is sampled from the prior distribution p_ψ and the posterior distribution q_ϕ is not used. Both distributions are modeled using 3-layer CNNs that take their input from the encoder and output two channels representing μ and σ to parameterize a multivariate Gaussian distribution $\mathcal{N}(\mu, \sigma)$. As shown in Figure 3, the noise is useful in encoding diverse, plausible representations of unseen regions.

Overall, the Structure Generator is trained to minimize a joint loss consisting of a cross-entropy loss L_{ce} for semantic predictions, a mean absolute error term for depth predic-

tions, and the KL-divergence term for the noise tensor:

$$\begin{aligned} \mathcal{L}_{\text{Structure}} = & \lambda_{ce} L_{ce}(s_t, \hat{s}_t) \\ & + \lambda_d \|d_t - \hat{d}_t\|_1 \\ & + \lambda_{KL} D_{KL}(q_\phi(z_t|s_t, d_t), p_\psi(z_t|s'_t, d'_t)) \end{aligned} \quad (1)$$

where λ_{ce} , λ_d , and λ_{KL} are weights determined by a grid search. We set these to 1, 100, and 0.5 respectively.

3.2. Image Generator: RGB

The Image Generator is an image-to-image translation GAN [22, 78] that converts the semantic and depth predictions \hat{s}_t, \hat{d}_t from the first stage into a realistic RGB image \hat{I}_t . Our model architecture is based on SPADE blocks [63] that use spatially-adaptive normalization layers to insert context into multiple layers of the network. As with our Structure Generator, we maintain an accumulating 3D point cloud containing all previous image observations. This provides a sparse RGB guidance image I'_t when re-projected. Similar to Multi-SPADE [51], we insert two SPADE normalization layers into each residual block: one conditioned on the concatenated semantic and depth inputs $[\hat{s}_t, \hat{d}_t]$, and one conditioned on the RGB guidance image I'_t . The sparsity of the RGB guidance image is handled by applying partial convolutions [45]. In total Image Generator consists of 7 Multi-SPADE blocks, preceded by a single convolution block.

Following SPADE [63], the model is trained with the GAN hinge loss, feature matching loss [78], and perceptual loss [36] from a pretrained VGG-19 [69] model. During training, the generator is provided with the ground-truth segmentation image s_t and ground-truth depth image d_t . Our discriminator architecture is based on PatchGAN [32], and takes as input the concatenation of the ground-truth image I_t or generated image \hat{I}_t , the ground-truth depth image d_t and the ground-truth semantic image s_t . The losses for

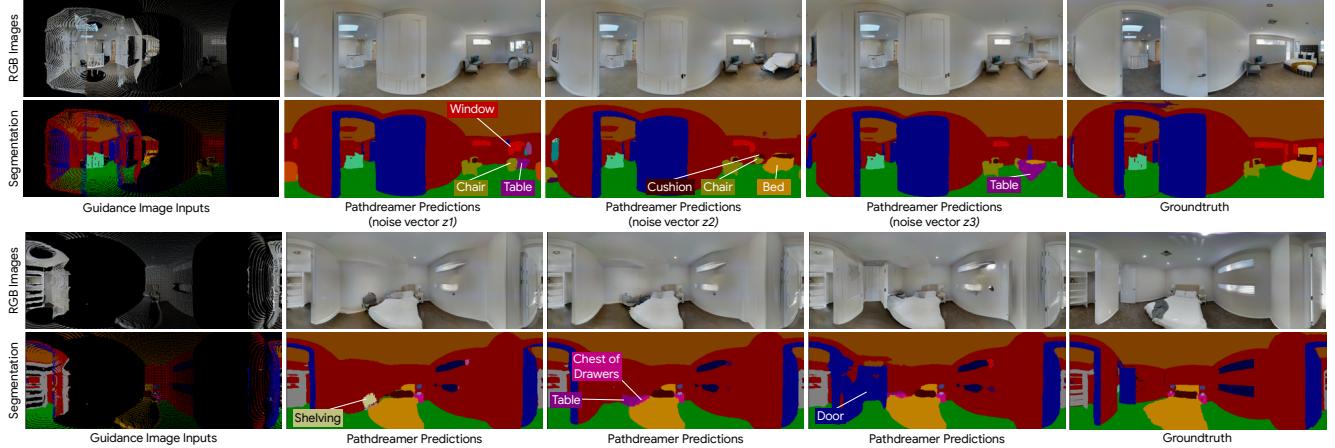


Figure 3: When predicting around corners, the Structure Generator can sample diverse and semantically plausible scene layouts which are closely reflected in the RGB output of the Image Generator, shown here for two guidance image inputs (left columns; unseen areas are indicated by solid black regions). Each example shows three alternative *room reveals* and the groundtruth. In the bottom example, the model considers various completions for a bedroom but fails to anticipate the groundtruth’s matching lamp on the opposite side of the bed.

the generator G and the discriminator D are:

$$\begin{aligned} \mathcal{L}_G = & -\lambda_{\text{GAN}} \mathbb{E}_{x_t} [D(G(x_t))] \\ & + \lambda_{\text{VGG}} \sum_{i=1}^n \frac{1}{n} \left\| \phi^{(i)}(I_t) - \phi^{(i)}(G(x_t)) \right\|_1 \\ & + \lambda_{\text{FM}} \sum_i \frac{1}{n} \left\| D^{(i)}(I_t) - D^{(i)}(G(x_t)) \right\|_1 \quad (2) \end{aligned}$$

$$\begin{aligned} \mathcal{L}_D = & -\mathbb{E}_{x_t} [\min(0, -1 + D(I_t))] \\ & -\mathbb{E}_{x_t} [\min(0, -1 - D(G(x_t)))] \quad (3) \end{aligned}$$

where $x_t = (s_t, d_t, I'_t)$ denotes the complete set of inputs to the generator, $\phi^{(i)}$ denotes the output of the i^{th} layer of the pretrained VGG-19 model, $D^{(i)}$ denotes the output of the discriminator’s i -th layer (conditioning inputs s_t, d_t to the discriminator have been dropped to save space). Like the Structure Generator, the Image Generator is not pretrained.

3.3. Training and Inference

Dataset For training and evaluation we use Matterport3D [7], a dataset of 10.8k RGB-D images from 90 building-scale indoor environments. For each environment, Matterport3D also includes a textured 3D mesh which is annotated with 40 semantic classes of objects and building components. To align with downstream VLN tasks, in all experiments the RGB, depth and semantic images are 360° panoramas in equirectangular format.

Trajectories To train Pathdreamer, we sampled 400k trajectories from the Matterport3D training environments. To define feasible trajectories, we used the navigation graphs from the Room-to-Room (R2R) dataset [3], in which nodes correspond to panoramic image locations, and edges define

navigable state transitions. For each trajectory 5–8 panoramas were sampled, choosing the starting node and the edge transitions uniformly at random. On average the viewpoints in these trajectories are 2m apart. Training with relatively large viewpoint changes is desirable, since the model learns to synthesize observations with large viewpoint changes in a single step (without the need to incur the computational cost of generating intervening frames). However, this does not preclude Pathdreamer from generating smooth video outputs at high frame rates¹.

Training The first and second stages of the model are trained separately. For the Image Generator, we use the Matterport3D RGB panoramas as training targets at 1024×512 resolution. We use the Habitat simulator [67] to render ground-truth depth and semantic training inputs and stitch these into equirectangular panoramas. We perform data augmentation by randomly cropping and horizontally rolling the RGB panoramas, which we found essential due to the limited number of panoramas available.

To train the Structure Generator, we again used Habitat to render depth and semantic images. Since this stage does not require aligned RGB images for training, in this case we performed data augmentation by perturbing the viewpoint coordinates with a random Gaussian noise vector drawn from $\mathcal{N}(0, 0.2\text{m})$ independently along each 3D axis. The Structure Generator was trained with equirectangular panoramas at 512×256 resolution.

Inference To avoid heading discontinuities during inference, we use circular padding on the image x-axis for both the Structure Generator and the Image Generator.

¹See <https://youtu.be/StklIENGqs0> for our video generation results.

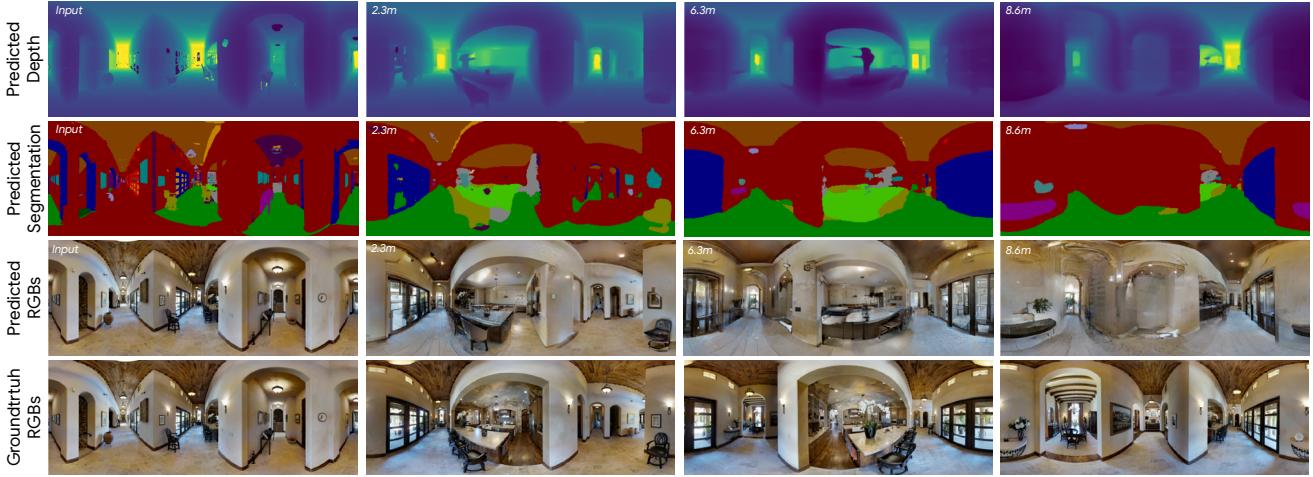


Figure 4: Example full prediction sequence beginning with one observation (depth, semantics, RGB) as context and generating observations for 3 new viewpoints traversing a corridor. At 2.3m the model completes a room reveal, imagining a kitchen-like space. After 8.6m the model’s predictions degrade. More examples are provided in the supplementary.

The 512×256 resolution semantic and depth outputs of the Structure Generator are upsampled to 1024×512 using nearest neighbor interpolation before they are passed to the Image Generator. In quantitative experiments, we set the Structure Generator noise tensor z_t to the mean of the prior.

4. Experiments

For evaluation we use the paths from the Val-Seen and Val-Unseen splits of the R2R dataset [3]. Val-Seen contains 340 trajectories from environments in the Matterport3D training split. Val-Unseen contains 783 trajectories in Matterport3D environments not seen in training. Since R2R trajectories contain 5-7 panoramas and at least 1 previous observation is given as context, we report evaluations over 1–6 steps, representing predictions over trajectory rollouts of around 2–13m (panoramas are 2.25m apart on average). See Figure 4 for an example rollout over 8.6m. We characterize the performance of Pathdreamer in comparison to baselines, ablations and in the context of the downstream task of Vision-and-Language Navigation (VLN).

4.1. Pathdreamer Results

Semantic Generation A key feature of our approach is the ability to generate semantic segmentation and depth outputs, in addition to RGB. We evaluate the generated semantic segmentation images using mean Intersection-Over-Union (mIOU) and report results for:

- **Nearest Neighbor:** A baseline without any learned components, using nearest-neighbor interpolation to fill holes in the projected semantic guidance image s'_t .
- **Ours (Teacher Forcing):** Structure Generator trained using the ground truth semantic and depth images as the previous observation at every time step.
- **Ours (Recurrent):** Structure Generator trained while feeding back its own semantic and depth predictions as

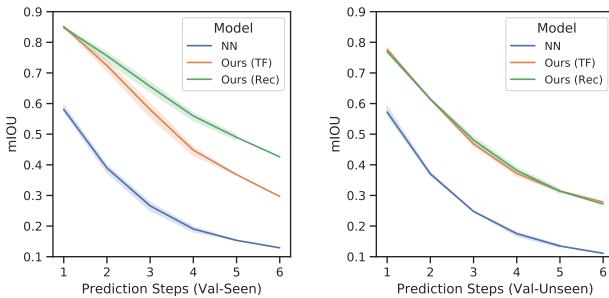
previous observations for the next step prediction. This reduces train-test mismatch and may allow the model to compensate for errors when doing longer roll-outs.

We also tried training the hierarchical convolutional LSTM from [42], but found that it frequently collapsed to a single class prediction. We attribute this to the large viewpoint changes and heavy occlusion in the training sequences; we believe this can be more effectively modeled with point cloud geometry than with a geometry-unaware LSTM.

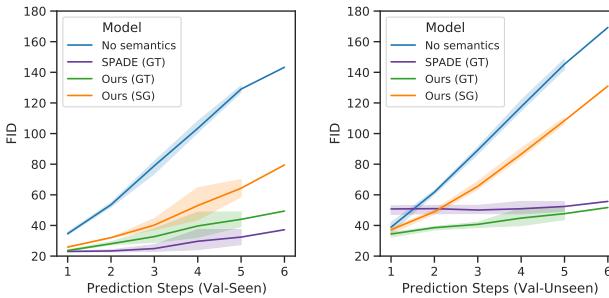
As illustrated in Table 1, Pathdreamer performs far better than the Nearest Neighbor baseline regardless of the number of steps in the rollout or the number of previous observations used as context. As expected, performance in seen environments is higher than unseen. Perhaps surprisingly, in Figure 5a we show that Recurrent training improves results during longer rollouts in the training environments (Val-Seen), but this does not improve results on Val-Unseen, perhaps indicating that the error compensation learned by the Image Generator does not easily generalize.

In addition to accurate predictions, we also want generated results to be *diverse*. Figure 3 shows that our model can generate diverse semantic scenes by interpolating the noise tensor z_t , and that the RGB outputs closely reflect the generated semantic image. This allows us to generate multiple plausible alternatives for the same navigation trajectory.

RGB Generation To evaluate the quality of RGB panoramas generated by the Image Generator, we compute the Fréchet Inception Distance (FID) [30] between generated and real images for each step in the paths. We report results using the semantic images generated by the Structure Generator as inputs (i.e., our full model). To quantify the potential for uplift with better Structure Generators, we also report results using ground truth semantic segmentations as input. We compare to two ablated versions of our model:



(a) Semantic segmentation mean-IOU (\uparrow). [TF]: Teacher Forcing. [Rec]: Recurrent.



(b) RGB generation FID (\downarrow). [GT]: Ground truth semantic inputs. [SG]: Structure Generator predictions.

Figure 5: Pathdreamer semantic segmentation mean-IOU (above) and RGB generation FID (below). Results are shown for Val-Seen (left) and Val-Unseen (right). Confidence intervals indicate the range of outcomes with 1, 2 or 3 previous observations as context.

Model	Context	Val-Seen		Val-Unseen	
		1 Step	1–6 Steps	1 Step	1–6 Steps
Nearest Neighbor	1	59.5	32.0	59.1	30.6
Ours (Teacher Forcing)	1	84.9	59.2	78.3	50.8
Ours (Recurrent)	1	84.7	65.9	77.5	50.9
Nearest Neighbor	2	57.4	35.2	56.5	33.8
Ours (Teacher Forcing)	2	85.4	64.6	77.4	55.5
Ours (Recurrent)	2	85.1	70.2	76.6	55.7
Nearest Neighbor	3	57.4	38.7	56.1	37.7
Ours (Teacher Forcing)	3	85.1	68.5	77.3	60.4
Ours (Recurrent)	3	84.6	72.7	76.8	60.8

Table 1: Mean-IOU (\uparrow) for generated semantic segmentations with varying context and prediction steps.

- **No Semantics:** The semantic and depth inputs s_t, d_t are removed from the Multi-SPADE blocks.
- **SPADE:** An ablation of the RGB inputs to the model, comprising the previous RGB image I_{t-1} and the re-projected RGB guidance image I'_t . The semantic image s_t replaces I_{t-1} as input to the model and the I'_t input layers are removed from the Multi-SPADE blocks, making this effectively the SPADE model [63].

As shown in Table 2, SPADE performs the best in Val-Seen, indicating that the model has the capacity to mem-

Model	Context	Inputs		Val-Seen		Val-Unseen	
		Obs	Sem	RGB	1 Step	1–6 Steps	1 Step
No Semantics	1	-			34.1	81.7	35.2
SPADE	1	GT	✓		23.3	24.9	47.3
Ours	1	GT	✓	✓	24.1	31.3	32.4
Ours	1	SG	✓	✓	26.2	41.7	70.4
No Semantics	2	-			34.0	69.0	39.5
SPADE	2	GT	✓		22.8	25.3	52.3
Ours	2	GT	✓	✓	23.5	31.2	35.3
Ours	2	SG	✓	✓	25.8	38.4	61.0
No Semantics	3	-			35.6	60.4	41.9
SPADE	3	GT	✓		23.1	26.2	52.8
Ours	3	GT	✓	✓	23.2	31.7	35.5
Ours	3	SG	✓	✓	25.6	36.7	52.9

Table 2: FID scores (\downarrow) for generated RGB images with varying context and prediction steps, using either ground truth semantics (GT) or Structure Generator predictions (SG) as input.

orize the training environments. In this case, RGB inputs are not necessary. However, our model performs noticeably better in Val-Unseen, highlighting the importance of maintaining RGB context in unseen environments (which is our focus). Performance degrades significantly in the No Semantics setting in both Val-Seen and Val-Unseen. We observed that without semantic inputs, the model is unable to generate meaningful images over longer horizons, which validates our two-stage hierarchical approach. These results are reflected in the FID scores, as well as qualitatively (Figure 6); Image Generator’s outputs are significantly crisper, especially over longer horizons. Due to the benefit of guidance images, the Image Generator’s textures are also generally better matched with the unseen environment, while SPADE tends to wash out textures, usually creating images of a standard style. Figure 5b plots performance for every setting step-by-step. FID of the Image Generator improves substantially when using ground truth semantics, particularly for longer rollouts, highlighting the potential to benefit from improvements to the Structure Generator.

4.2. VLN Results

Finally, we evaluate whether predictions from Pathdreamer can improve performance on a downstream visual navigation task. We focus on Vision-and-Language Navigation (VLN) using the R2R dataset [3]. Because reaching the navigation goal requires successfully grounding natural language instructions to visual observations, this provides a challenging task-based assessment of prediction quality.

In our inference setting, at each step while moving through the environment we use a baseline VLN agent based on [79] to generate a large number of possible future trajectories using beam search. We then rank these alternative trajectories using an instruction-trajectory compatibility model [84] to assess which trajectory best matches the



Figure 6: Visual comparison of ablated Image Generator outputs on Val-Unseen with *ground truth* segmentation and depth. Both RGB and semantic context are important for best performance.

instruction. The agent then executes the first action from the top-ranked trajectory before repeating the process. We consider three different planning horizons, with future trajectories containing 1, 2 or 3 forward steps.

The instruction-trajectory compatibility model is a dual-encoder that separately encodes textual instructions and trajectories (encoded using visual observations and path geometry) into a shared latent space. To improve performance on incomplete paths, we introduce truncated paths into the original contrastive training scheme proposed in [84]. The compatibility model is trained using only ground truth observations. However, during inference, RGB observations for future steps are drawn from three different sources:

- **Ground truth:** RGB observations from the actual environment, i.e., look-ahead observations.
- **Pathdreamer:** RGB predictions from our model.
- **Repeated pano:** A simple baseline in which the most recent RGB observation is repeated in future steps.
- **Blank pano:** A simple baseline in which blank images are provided as future observations.

Note that in all cases the geometry of the future trajectories is determined by the ground truth R2R navigation graphs. In Table 3, we report Val-Unseen results for this experiment using standard metrics for VLN: navigation error (NE), success rate (SR), shortest path length (SPL), normalized Dynamic Time Warping (nDTW) [31], and success weighted by normalized Dynamic Time Warping (sDTW) [31].

Consistent with prior work [20, 73], we find that looking ahead using ground truth visual observations provides a robust performance boost, e.g., success rate increases from 44.6% with 1 planning step (top panel) to 59.3% with 3 planning steps (bottom panel). At the other extreme, the Repeated pano baseline is weak, with a success rate of just

	Observations	Plan Steps	NE ↓	SR ↑	SPL ↑	nDTW ↑	sDTW ↑
Repeated pano	1	6.75	35.7	33.8	52.0	31.2	
Blank pano	1	7.29	35.9	33.7	50.9	31.5	
Pathdreamer	1	6.55	39.9	38.3	54.6	35.2	
Ground truth	1	5.80	44.6	42.7	58.9	39.4	
Repeated pano	2	6.76	36.8	34.0	51.8	31.7	
Blank pano	2	6.65	40.0	37.2	53.3	34.8	
Pathdreamer	2	5.8	46.5	43.9	59.1	41.2	
Ground truth	2	4.95	54.3	51.3	64.9	48.3	
Repeated pano	3	6.25	40.6	37.7	55.6	35.2	
Blank pano	3	6.48	41.9	38.8	54.2	36.1	
Pathdreamer	3	5.32	50.4	47.3	61.8	44.4	
Ground truth	3	4.44	59.3	55.8	67.9	52.7	

Table 3: VLN Val-Unseen results using an instruction-trajectory compatibility model to rank alternative future trajectories with planning horizons of 1, 2 or 3 steps.

35.7% with 1 planning step (top row). The Blank pano baseline is similar, with a success rate of 35.9%. This is not surprising: these baselines deny the compatibility model any useful visual representation of the next action, which is crucial to performance [20, 73]. However, increasing the planning horizon does improve performance even for the Repeated/Blank pano baselines, since the compatibility model is able to compare the geometry of alternative future trajectories. Finally, we observe that using Pathdreamer’s visual observations closes about half the gap between the Repeated pano baseline and the ground truth observations, e.g., 50.4% success with Pathdreamer vs. 40.6% and 59.3% respectively for the others. We conclude that using Pathdreamer as a visual world model can improve performance on downstream tasks, although existing agents still rely on using a navigation graph to define the feasible action space at each step. Pathdreamer is complementary to current SOTA model-based approaches, and a combination would likely lead to further boosts in VLN performance, which is worth investigating in future work.

5. Conclusion

Pathdreamer is a stochastic hierarchical visual world model that can synthesize realistic and diverse 360° panoramic images for unseen trajectories in real buildings. As a visual world model, Pathdreamer also shows strong promise in improving performance on downstream tasks, which we show with VLN. Most notably, we show that Pathdreamer captures around half the benefit of looking ahead at actual observations from the environment. The efficacy of Pathdreamer in the VLN task may be attributed to its ability to model fundamental constraints in the real world, and thus relieve agents from having to learn the geometry and visual and semantic structure of buildings. Applying Pathdreamer to other embodied navigation tasks such as Object-Nav [5], VLN-CE [39] and street-level navigation [9, 53] are natural directions for future work.

References

- [1] Sandra Aigner and Marco Körner. Futuregan: Anticipating the future frames of video sequences using spatio-temporal 3d convolutions in progressively growing gans. *arXiv preprint arXiv:1810.01325*, 2018. 2
- [2] Peter Anderson, Angel Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, and Amir R. Zamir. On evaluation of embodied navigation agents. *arXiv preprint arXiv:1807.06757*, 2018. 3
- [3] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *CVPR*, pages 3674–3683, 2018. 1, 2, 3, 5, 6, 7
- [4] Mohammad Babaeizadeh, Mohammad Taghi Saffar, Danijar Hafner, Harini Kannan, Chelsea Finn, Sergey Levine, and Dumitru Erhan. Models, pixels, and rewards: Evaluating design trade-offs in visual model-based reinforcement learning. *arXiv preprint arXiv:2012.04603*, 2020. 3
- [5] Dhruv Batra, Aaron Gokaslan, Aniruddha Kembhavi, Oleksandr Maksymets, Roozbeh Mottaghi, Manolis Savva, Alexander Toshev, and Erik Wijmans. Objectnav revisited: On evaluation of embodied agents navigating to objects. *arXiv preprint arXiv:2006.13171*, 2020. 1, 3, 8
- [6] Lars Buesing, Theophane Weber, Yori Zwols, Sébastien Racanière, Arthur Guez, Jean-Baptiste Lespiau, and Nicolas Heess. Woulda, coulda, shoulda: Counterfactually-guided policy search. In *ICLR*, 2019. 1
- [7] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from RGB-D data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017. 2, 3, 5
- [8] Matthew Chang, Arjun Gupta, and Saurabh Gupta. Semantic visual navigation by watching youtube videos. *NeurIPS*, 2020. 3
- [9] Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *CVPR*, 2019. 3, 8
- [10] Silvia Chiappa, Sébastien Racanière, Daan Wierstra, and Shakir Mohamed. Recurrent environment simulators. *ICLR*, 2017. 2
- [11] Emily Denton and Rob Fergus. Stochastic video generation with a learned prior. 2018. 2, 4
- [12] Helisa Dhamo, Keisuke Tateno, Iro Laina, Nassir Navab, and Federico Tombari. Peeking behind objects: Layered depth prediction from a single image. *Pattern Recognition Letters*, 125:333–340, 2019. 3
- [13] Alexey Dosovitskiy and Vladlen Koltun. Learning to act by predicting the future. In *ICLR*, 2017. 2
- [14] Frederik Ebert, Chelsea Finn, Alex X Lee, and Sergey Levine. Self-supervised visual planning with temporal skip connections. *Conference on Robot Learning (CoRL)*, 2017. 2
- [15] Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. In *NeurIPS*, pages 64–72, 2016. 2
- [16] Chelsea Finn and Sergey Levine. Deep visual foresight for planning robot motion. In *ICRA*, pages 2786–2793. IEEE, 2017. 1, 2
- [17] Chelsea Finn, Xin Yu Tan, Yan Duan, Trevor Darrell, Sergey Levine, and Pieter Abbeel. Deep spatial autoencoders for visuomotor learning. In *ICRA*, pages 512–519. IEEE, 2016. 1
- [18] John Flynn, Michael Broxton, Paul Debevec, Matthew DuVall, Graham Fyffe, Ryan Overbeck, Noah Snavely, and Richard Tucker. Deepview: View synthesis with learned gradient descent. In *CVPR*, pages 2367–2376, 2019. 3
- [19] John Flynn, Ivan Neulander, James Philbin, and Noah Snavely. Deepstereo: Learning to predict new views from the world’s imagery. In *CVPR*, 2016. 1, 3
- [20] Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. Speaker-follower models for vision-and-language navigation. *NeurIPS*, 2018. 3, 8
- [21] Tsu-Jui Fu, Xin Eric Wang, Matthew F Peterson, Scott T Grafton, Miguel P Eckstein, and William Yang Wang. Counterfactual vision-and-language navigation via adversarial path sampler. In *ECCV*, pages 71–86, 2020. 3
- [22] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *NeurIPS*, 2014. 4
- [23] David Ha and Jürgen Schmidhuber. World models. *NeurIPS*, 2018. 1, 2
- [24] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *ICLR*, 2019. 2
- [25] Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. *arXiv preprint arXiv:2010.02193*, 2020. 1, 2
- [26] Meera Hahn, Jacob Krantz, Dhruv Batra, Devi Parikh, James M. Rehg, Stefan Lee, and Peter Anderson. Where are you? localization from embodied dialog. 2020. 1, 3
- [27] Weituo Hao, Chunyuan Li, Xiujun Li, Lawrence Carin, and Jianfeng Gao. Towards learning a generic agent for vision-and-language navigation via pre-training. In *CVPR*, June 2020. 3
- [28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 3
- [29] Philipp Henzler, Volker Rasche, Timo Ropinski, and Tobias Ritschel. Single-image tomography: 3d volumes from 2d cranial x-rays. In *Computer Graphics Forum*, volume 37, pages 377–388. Wiley Online Library, 2018. 3
- [30] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 6

- [31] Gabriel Ilharco, Vihan Jain, Alexander Ku, Eugene Ie, and Jason Baldridge. General evaluation for instruction conditioned navigation using dynamic time warping. *NeurIPS Workshop on Visually Grounded Interaction and Language (ViGIL)*, 2019. 8
- [32] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017. 4
- [33] Vihan Jain, Gabriel Magalhaes, Alexander Ku, Ashish Vaswani, Eugene Ie, and Jason Baldridge. Stay on the path: Instruction fidelity in vision-and-language navigation. 2019. 3
- [34] Jindong Jiang, Lunan Zheng, Fei Luo, and Zhijun Zhang. Rednet: Residual encoder-decoder network for indoor rgbd semantic segmentation. *arXiv preprint arXiv:1806.01054*, 2018. 3
- [35] Xiaojie Jin, Huixin Xiao, Xiaohui Shen, Jimei Yang, Zhe Lin, Yunpeng Chen, Zequn Jie, Jiashi Feng, and Shuicheng Yan. Predicting scene parsing and motion dynamics in the future. In *NeurIPS*, pages 6915–6924, 2017. 2
- [36] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711, 2016. 4
- [37] Abhishek Kar, Christian Häne, and Jitendra Malik. Learning a multi-view stereo machine. In *NeurIPS*, pages 365–376, 2017. 3
- [38] Maximilian Karl, Maximilian Soelch, Justin Bayer, and Patrick Van der Smagt. Deep variational bayes filters: Unsupervised learning of state space models from raw data. *ICLR*, 2017. 2
- [39] Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. Beyond the nav-graph: Vision and language navigation in continuous environments. In *ECCV*, 2020. 3, 8
- [40] Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. *EMNLP*, 2020. 1, 3
- [41] Yong-Hoon Kwon and Min-Gyu Park. Predicting future frames using retrospective cycle gan. In *CVPR*, pages 1811–1820, 2019. 2
- [42] Wonkwang Lee, Whie Jung, Han Zhang, Ting Chen, Jing Yu Koh, Thomas Huang, Hyungsuk Yoon, Honglak Lee, and Seunghoon Hong. Revisiting hierarchical approach for persistent long-term video prediction. In *ICLR*, 2021. 2, 6
- [43] Zuoyue Li, Zhaopeng Cui, and Martin R Oswald. Street-view panoramic video synthesis from a single satellite image. *arXiv preprint arXiv:2012.06628*, 2020. 2
- [44] Andrew Liu, Richard Tucker, Varun Jampani, Ameesh Makadia, Noah Snavely, and Angjoo Kanazawa. Infinite nature: Perpetual view generation of natural scenes from a single image. *arXiv preprint arXiv:2012.09855*, 2020. 2
- [45] Guilin Liu, Fitzsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *ECCV*, pages 85–100, 2018. 4
- [46] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, pages 13–23, 2019. 3
- [47] Pauline Luc, Natalia Neverova, Camille Couprie, Jakob Verbeek, and Yann LeCun. Predicting deeper into the future of semantic segmentation. In *ICCV*, pages 648–657, 2017. 2
- [48] Chih-Yao Ma, Jiasen Lu, Zuxuan Wu, Ghassan AlRegib, Zsolt Kira, Richard Socher, and Caiming Xiong. Self-monitoring navigation agent via auxiliary progress estimation. *ICLR*, 2019. 3
- [49] Chih-Yao Ma, Zuxuan Wu, Ghassan AlRegib, Caiming Xiong, and Zsolt Kira. The regretful agent: Heuristic-aided navigation through progress estimation. In *CVPR*, pages 6732–6740, 2019. 3
- [50] Arjun Majumdar, Ayush Shrivastava, Stefan Lee, Peter Anderson, Devi Parikh, and Dhruv Batra. Improving vision-and-language navigation with image-text pairs from the web. *ECCV*, 2020. 2, 3
- [51] Arun Mallya, Ting-Chun Wang, Karan Sapra, and Ming-Yu Liu. World-consistent video-to-video synthesis. *ECCV*, 2020. 2, 3, 4
- [52] Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. In *CVPR*, 2021. 3
- [53] Harsh Mehta, Yoav Artzi, Jason Baldridge, Eugene Ie, and Piotr Mirowski. Retouchdown: Adding touchdown to streetlearn as a shareable resource for language grounding tasks in street view. *EMNLP Workshop on Spatial Language Understanding (SpLU)*, 2020. 3, 8
- [54] Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019. 3
- [55] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *ECCV*, 2020. 3
- [56] Piotr Mirowski, Andras Banki-Horvath, Keith Anderson, Denis Teplyashin, Karl Moritz Hermann, Mateusz Malinowski, Matthew Koichi Grimes, Karen Simonyan, Koray Kavukcuoglu, Andrew Zisserman, et al. The streetlearn environment and dataset. *arXiv preprint arXiv:1903.01292*, 2019. 3
- [57] Anusha Nagabandi, Gregory Kahn, Ronald S Fearing, and Sergey Levine. Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning. In *ICRA*, pages 7559–7566. IEEE, 2018. 1
- [58] Medhini Narasimhan, Erik Wijmans, Xinlei Chen, Trevor Darrell, Dhruv Batra, Devi Parikh, and Amanpreet Singh. Seeing the un-scene: Learning amodal semantic maps for room navigation. In *ECCV*, pages 513–529. Springer, 2020. 3
- [59] Khanh Nguyen and Hal Daumé III. Help, anna! visual navigation with natural multimodal assistance via retrospective curiosity-encouraging imitation learning. *EMNLP*, 2019. 3

- [60] Junhyuk Oh, Xiaoxiao Guo, Honglak Lee, Richard L Lewis, and Satinder Singh. Action-conditional video prediction using deep networks in atari games. In *NeurIPS*, pages 2863–2871, 2015. 2
- [61] Junhyuk Oh, Satinder Singh, and Honglak Lee. Value prediction network. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *NeurIPS*, 2017. 2
- [62] Blazej Osinski, Chelsea Finn, Dumitru Erhan, George Tucker, Henryk Michalewski, Konrad Czechowski, Lukasz Mieczyslaw Kaiser, Mohammad Babaeizadeh, Piotr Kozakowski, Piotr Milos, Roy H Campbell, Afroz Mohiuddin, Ryan Sepassi, and Sergey Levine. Model-based reinforcement learning for atari. In *ICLR*, 2020. 1, 2
- [63] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, pages 2337–2346, 2019. 2, 4, 7
- [64] Aj Piergiovanni, Alan Wu, and Michael S Ryoo. Learning real-world robot policies by dreaming. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7680–7687. IEEE, 2019. 1
- [65] Senthil Purushwalkam, Sebastian Vicenc Amengual Gari, Vamsi Krishna Ithapu, Carl Schissler, Philip Robinson, Abhinav Gupta, and Kristen Grauman. Audio-visual floorplan reconstruction. *arXiv preprint arXiv:2012.15470*, 2020. 3
- [66] Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. Reverie: Remote embodied visual referring expression in real indoor environments. In *CVPR*, 2020. 1, 3
- [67] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A Platform for Embodied AI Research. In *ICCV*, 2019. 5
- [68] Meng-Li Shih, Shih-Yang Su, Johannes Kopf, and Jia-Bin Huang. 3d photography using context-aware layered depth inpainting. In *CVPR*, June 2020. 3
- [69] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015. 4
- [70] Pratul P Srinivasan, Richard Tucker, Jonathan T Barron, Ravi Ramamoorthi, Ren Ng, and Noah Snavely. Pushing the boundaries of view extrapolation with multiplane images. In *CVPR*, pages 175–184, 2019. 3
- [71] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 3
- [72] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. A Bradford Book, Cambridge, MA, USA, 2018. 1
- [73] Hao Tan, Licheng Yu, and Mohit Bansal. Learning to navigate unseen environments: Back translation with environmental dropout. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2610–2621, 2019. 3, 8
- [74] Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. Vision-and-dialog navigation. In *Conference on Robot Learning (CoRL)*, pages 394–406. PMLR, 2020. 1, 3
- [75] Suhani Vora, Reza Mahjourian, Soeren Pirk, and Anelia Angelova. Future semantic segmentation leveraging 3d information. *ECCV 3D Reconstruction meets Semantics Workshop*, 2018. 2
- [76] Jacob Walker, Abhinav Gupta, and Martial Hebert. Patch to the future: Unsupervised visual prediction. In *CVPR*, pages 3302–3309, 2014. 2
- [77] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. *NeurIPS*, 2018. 2
- [78] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, pages 8798–8807, 2018. 4
- [79] Xin Wang, Qiu Yuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In *CVPR*, pages 6629–6638, 2019. 7
- [80] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. In *CVPR*, pages 7467–7477, 2020. 1, 3
- [81] Fei Xia, Amir R. Zamir, Zhi-Yang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson env: real-world perception for embodied agents. In *CVPR*. IEEE, 2018. 3
- [82] Jingwei Xu, Bingbing Ni, Zefan Li, Shuo Cheng, and Xiaokang Yang. Structure preserving video prediction. In *CVPR*, pages 1460–1469, 2018. 2
- [83] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. *arXiv preprint arXiv:2012.02190*, 2020. 3
- [84] Ming Zhao, Peter Anderson, Vihan Jain, Su Wang, Alex Ku, Jason Baldridge, and Eugene Ie. On the evaluation of vision-and-language navigation instructions. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2021. 7, 8
- [85] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *SIGGRAPH*, 2018. 3
- [86] Fengda Zhu, Yi Zhu, Xiaojun Chang, and Xiaodan Liang. Vision-language navigation with self-supervised auxiliary reasoning tasks. In *CVPR*, pages 10012–10022, 2020. 3
- [87] Wanrong Zhu, Xin Wang, Tsu-Jui Fu, An Yan, Pradyumna Narayana, Kazoo Sone, Sugato Basu, and William Yang Wang. Multimodal text style transfer for outdoor vision-and-language navigation. *EACL*, 2021. 3