Pathdreamer: A World Model for Indoor Navigation

Jing Yu Koh[1] Honglak Lee[2] Yinfei Yang[1] Jason Baldridge[1] Peter Anderson[1]

[1]Google Research [2]University 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. In the downstream task of Vision-and-Language Navigation (VLN), planning ahead with Pathdreamer provides about half the benefit of looking ahead at unobserved parts of the environment.

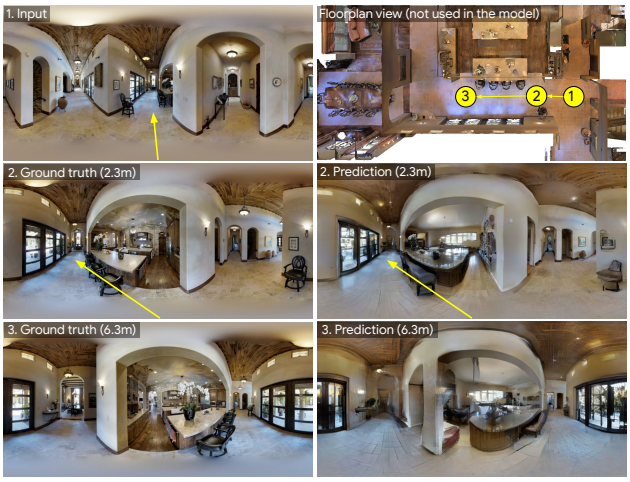

Figure 1: Generating photorealistic 360° visual observations from an imagined 6.3m trajectory in a previously unseen building. Observations also include depth and segmentations (not shown here).

1. Introduction

World models [9], or models of environments [19], 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 [6, 14], learning policies directly from the model (i.e., learning in a dream) [7, 9, 17, 10], and for counterfactual reasoning [3]. Model-based approaches such as these also typically improve the sample efficiency of deep reinforcement learning [19, 15]. However, world models that generate high-dimensional visual observations (i.e., images) have typically been restricted to relatively simple environments, such as Atari games [15] and tabletops [6].

Our goal is to develop a generic visual world model for agents navigating in indoor environments. Specifically, 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 [2], instruction-guided navigation [1, 18, 12] and dialog-guided navigation [20, 11]. 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 [8, 21], 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 Figure 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 [4]. A third problem is temporal consistency: predictions of unseen building re-

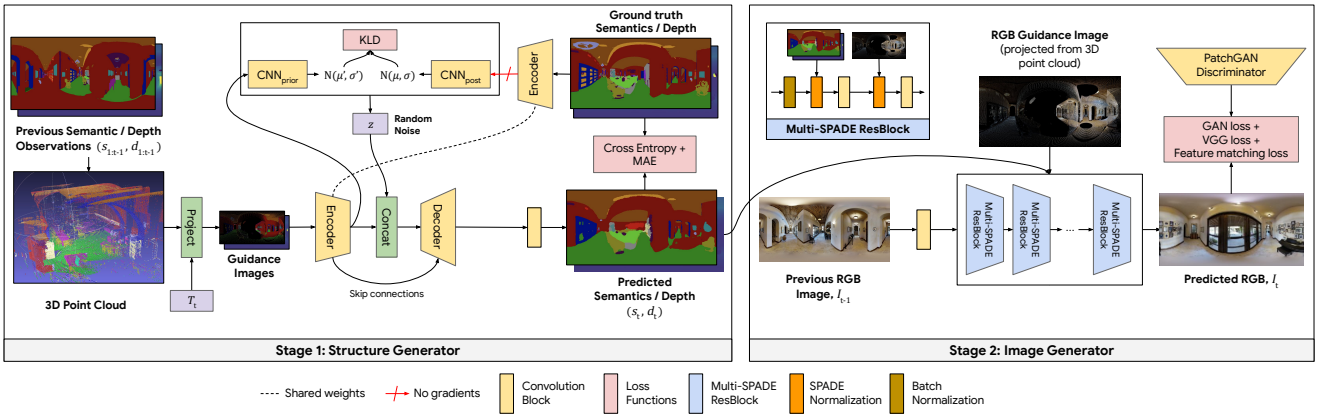

Figure 2: Given a history of visual observations (RGB, depth and semantics) and a trajectory of future viewpoints, the Structure Generator conditions on a sampled noise tensor before generating semantic and depth outputs to provide a high-level structural representation. Realistic RGB images are synthesized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

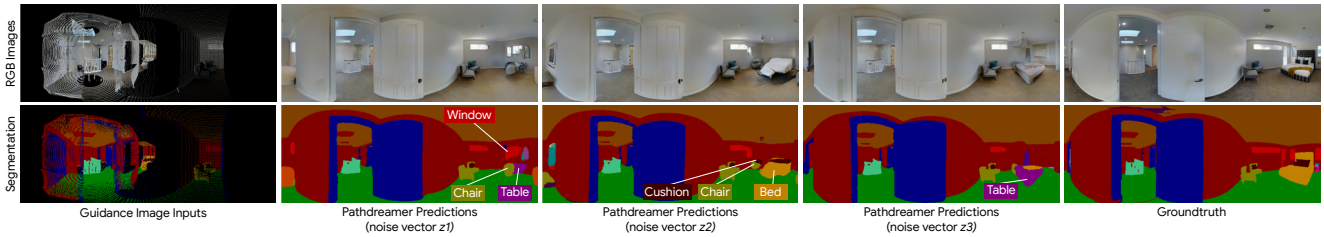

Figure 3: When predicting around corners, the Structure Generator can sample diverse and semantically plausible scene layouts which are closely reflected in the RGB output of the Image Generator, shown here for an example input (left column; unseen areas are indicated by solid black regions). Three alternative *room reveals* and the groundtruth are shown.

gions should ideally be stochastic (capturing the full distribution of possible outcomes), but revisited regions should be rendered in a consistent manner to previous observations.

2. Pathdreamer

Towards this goal, we introduce Pathdreamer– a world model that generates high-resolution visual observations from a trajectory of future viewpoints in buildings it has never observed. Given one or more visual observations (consisting of RGB, depth and semantic segmentation for panoramas), Pathdreamer synthesizes high-resolution visual observations along a trajectory through future viewpoints using a hierarchical two-stage approach (Figure 2).

Pathdreamer's first stage, Structure Generator, generates depth and semantic segmentations. Inspired by work in video prediction [5], 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 [16, 13]. 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 [13]. We assume that the future trajectory may traverse unseen areas of environment, requiring the model to not only in-fill minor object dis-occlusions, but also to imagine diverse outputs for entire room reveals (Figure 3). Note that we generate depth and segmentation because these modalities are useful in many downstream tasks, and modeling them improves the quality of the RGB outputs. As illustrated in Figure 1, Pathdreamer can generate plausible views of previously unseen scenes under large viewpoint changes, while also addressing the *room reveal* problem – in this case correctly hypothesizing that a room resembling a kitchen at position 2.

Empirically, using the Matterport3D dataset [4] 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. Finally, we evaluate whether Pathdreamer predictions can improve performance on the VLN task [1]. We rank Pathdreamer generation results using an instruction-trajectory compatibility model [22] to assess which trajectory best matches the instruction. The agent executes the first action from the top-ranked trajectory before repeating the process, which improves performance.

References

[1] 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

[2] 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] Lars Buesing, Theophane Weber, Yori Zwols, Sebastien Racaniere, Arthur Guez, Jean-Baptiste Lespiau, and Nicolas Heess. Woulda, coulda, shoulda: Counterfactually-guided policy search. In *ICLR*, 2019. 1

[4] 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. 1, 2

[5] Emily Denton and Rob Fergus. Stochastic video generation with a learned prior. 2018. 2

[6] Chelsea Finn and Sergey Levine. Deep visual foresight for planning robot motion. In *ICRA*, pages 2786–2793. IEEE, 2017. 1

[7] 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

[8] John Flynn, Ivan Neulander, James Philbin, and Noah Snavely. Deepstereo: Learning to predict new views from the world's imagery. In *CVPR*, 2016. 1

[9] David Ha and Jürgen Schmidhuber. World models. *NeurIPS*, 2018. 1

[10] Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. *arXiv preprint arXiv:2010.02193*, 2020. 1

[11] 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

[12] 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

[13] Arun Mallya, Ting-Chun Wang, Karan Sapra, and Ming-Yu Liu. World-consistent video-to-video synthesis. *ECCV*, 2020. 2

[14] 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

[15] 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

[16] 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

[17] 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

[18] 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

[19] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. A Bradford Book, Cambridge, MA, USA, 2018. 1

[20] 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

[21] 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

[22] 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. 2