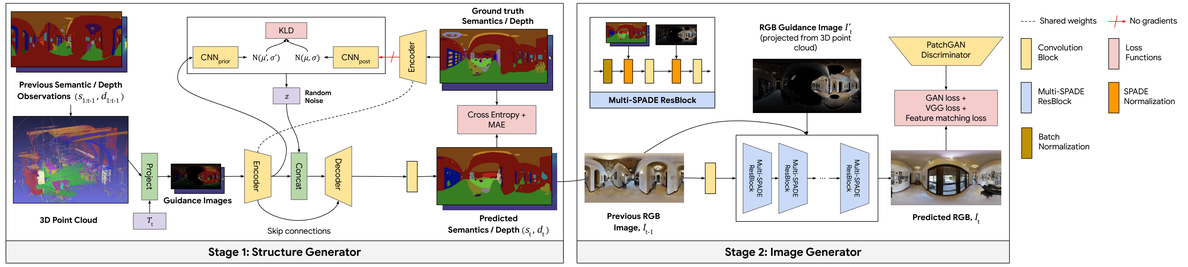
Pathdreamer: A World Model for Indoor Navigation

Jing Yu Koh[1], Honglak Lee[2], Yinfei Yang[1], Jason Baldridge[1], Peter Anderson[1]
[1]Google Research [2]University of Michigan

A 3D Photorealistic World

Pathdreamer is a world model *(Ha and Schmidhuber, 2018)* that generates high-resolution visual observations from a trajectory of future viewpoints in buildings it has never observed. A future trajectory may traverse unseen areas of the environment, requiring the model to in-fill minor object occlusions, or imagine entire room reveals.

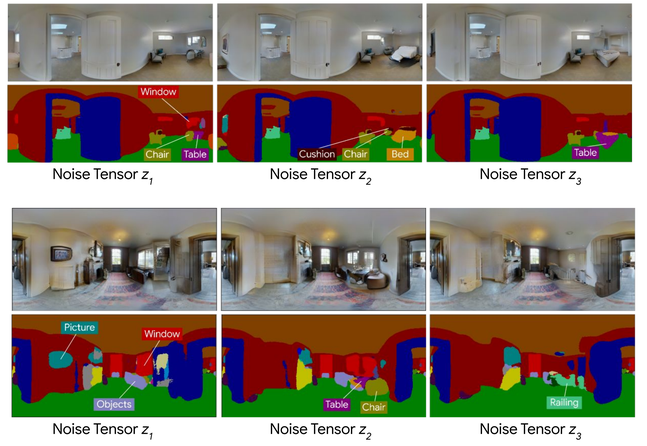
Pathdreamer Model



Pathdreamer is a hierarchical, stochastic, two-stage model for addressing this challenge.

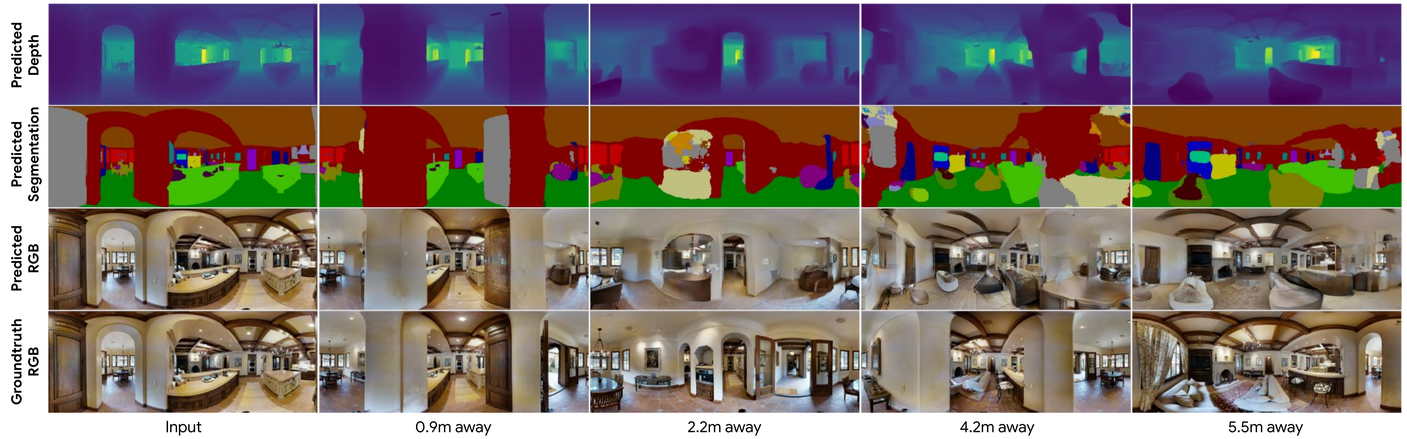
- The input is a sequence of previous observations consisting of RGB images $I_{1:t-1}$, semantic segmentation images $s_{1:t-1}$, and depth images $d_{1:t-1}$.
- Pathdreamer uses a latent noise tensor z_t to capture the stochastic information about the next observation (e.g. the layout of an unseen room).

Diverse Generation



When predicting around corners, Pathdreamer can sample diverse and semantically plausible scene layouts, which are closely reflected in the RGB outputs of the model. Results from three random noise tensors are shown above for two distinct environments.

Long Distance Generation Results

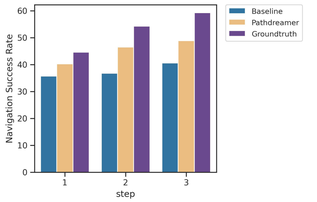


Pathdreamer is capable of synthesizing realistic and diverse 360 degree panoramic images for unseen trajectories in simulations of real buildings. Predictions remain high in quality up to 4-5 metres into the future (e.g., 5.5m in the figure above). Pathdreamer can also be conditioned to generate continuous video sequences[1], by interpolating between points within the environment.

[1]Video generation results: https://youtu.be/HNAmsdk7IJ4

Downstream VLN Results

We evaluate whether Pathdreamer can improve performance on the downstream Vision-and-Language Navigation (VLN) task using the R2R dataset (Anderson et al., 2018). Pathdreamer closes nearly half the gap between a baseline setting and when groundtruth information is accessed.



References

- Ha, David and Schmidhuber, Jürgen. "World models." NeurIPS, 2018.
- Anderson, Peter, et al. "Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments." CVPR, 2018.

goo.gle/pathdreamer