Simple and Effective Synthesis of Indoor 3D Scenes

Jing Yu Koh[1,*] Harsh Agrawal[2,†,*] Dhruv Batra[2] Richard Tucker[1] Austin Waters[1]
Honglak Lee[3] Yinfei Yang[4,†] Jason Baldridge[1] Peter Anderson[1]
[1]Google Research [2]Georgia Tech [3]University of Michigan [4]Apple

Abstract

We study the problem of synthesizing immersive 3D indoor scenes from one or more images. Our aim is to generate high-resolution images and videos from novel viewpoints, including those that extrapolate far beyond the input images while maintaining 3D consistency. Existing approaches are highly complex, with many separately trained stages and components. We propose a simple alternative: an image-to-image GAN that maps directly from reprojections of incomplete point clouds to full high-resolution RGB-D images. On the Matterport3D and RealEstate10K datasets, our approach significantly outperforms prior work when evaluated by humans, as well as on FID scores. Further, we show that our model is useful for generative data augmentation. A vision-and-language navigation (VLN) agent trained with trajectories spatially-perturbed by our model improves success rate by up to 1.5% over a state of the art model on the R2R benchmark. For more details, we refer readers to our full paper (`https://arxiv.org/abs/2204.02960`) and video results (`https://youtu.be/lhwwlrRfFp0`).

1. Introduction

We study the problem of synthesizing immersive 3D indoor scenes from one or more context images captured along a trajectory. Our aim is to generate high-resolution images and videos from novel viewpoints, including viewpoints that extrapolate far beyond the context image(s), while maintaining the 3D consistency of the scene. Solving this problem would make photos and videos interactive and immersive, with applications not only to content creation but also robotics and embodied AI. For example, models that can predict around corners could be used by navigation agents as world models [7] for model-based planning in novel environments [5, 10]. Such models could also be used to train agents in interactive environments synthesized from static images and videos.

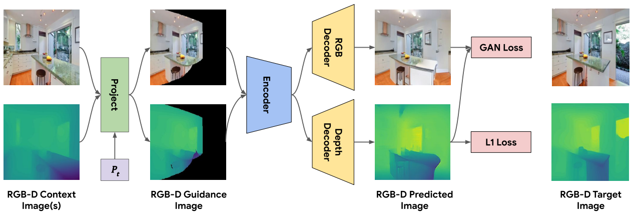

Figure 1. Our lightweight approach to 3D scene synthesis accumulates context images in an RGB point cloud. To generate a new viewpoint, we simply apply an image-to-image GAN to the *guidance image* of the reprojected point cloud. We achieve surprisingly strong results with this simple approach, significantly outperforming more complicated models.

Previous approaches attempting this under large viewpoint changes [10, 15, 18] typically operate on point clouds, which are accumulated from the available context images. The use of point clouds naturally incorporates camera projective geometry into the model and helps maintain the 3D consistency of the scene [12]. To generate novel views, these approaches reproject the point-cloud relative to the target camera pose into an RGB-D *guidance image*. These guidance images are extremely sparse because of missing context and completing them requires extensive inpainting and outpainting. In prior work, Pathdreamer [10] achieves this by assuming the availability of semantic segmentations, and combining a stochastic depth and semantic segmentation (structure) generator with an RGB image generator.On the other hand, PixelSynth [15] creates guidance images using a differentiable point cloud renderer, generates a support set of additional views using PixelCNN++ [16] operating on the latent space of a VQ-VAE [14], combines and refines these images using a GAN [6] similar to SynSin [18].

Approach & Results We propose a simple alternative: an image-to-image GAN that maps directly from guidance images to high-resolution photorealistic RGB-D (see Fig. 1). Compared to Pathdreamer, our simple model forgoes the stochastic structure generator, spatially adaptive

[*] Equal contribution.
[†] Work done while at Google.

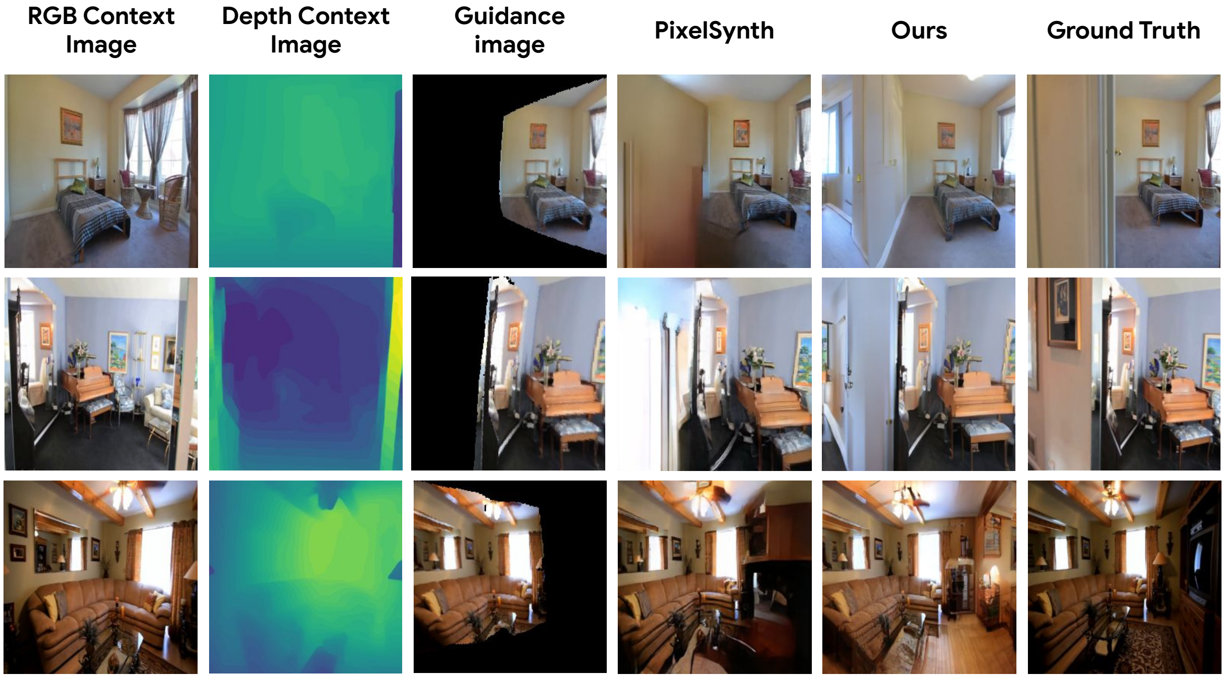

Figure 2. Qualitative comparison of predictions on the RealEstate10K [20] dataset. In these selected examples, our model completes the scene by imagining adjacent rooms (Row 1, Row 2), while keeping wall and carpet colors consistent (Row 1), and introducing new elements like lamps and a wall painting (Row 3).

normalization layers, dependence on semantic segmentations, and multi-step training. By dropping the dependence of semantic segmentation inputs, we unlock training on a much broader range of data, such as commonly available RGB-D datasets [11, 13, 19], and video data such as the RealEstate10K dataset [21]. We eschew many components of PixelSynth: differentiable rendering, support set generation using PixelCNN++and a VQ-VAE, and the multiple sampling and re-ranking procedure. Perhaps surprisingly, with random masking of the guidance images during training, plus other architectural changes supported by thorough ablation studies, our lightweight approach outperforms prior work. In human evaluations of image quality, our model is preferred to Pathdreamer in 60% of comparisons and preferred to PixelSynth in 77%. Our FID scores on 360° panoramic images from Matterport3D [3] improve over Pathdreamer's by 27.9% relatively (from 27.2 to 19.6) on single step viewpoint predictions (an average of 2.2m), and from 65.8 to 58.0 when predicting over longer trajectories of up to 6 novel viewpoints. On RealEstate10K [21] – a collection of real estate video walkthroughs – our FID scores outperform PixelSynth, improving from 25.5 to 23.5 (and 23.6 to 21.5 for indoor images).

VLN Results Motivated by these strong results on image generation, we investigate the usefulness of our model for data augmentation in embodied AI. For this purpose, we focus on the task of vision-and-language navigation (VLN) using the Room-to-Room (R2R) dataset [1], which requires an agent to follow natural language navigation instructions in previously unseen photorealistic environments. Training data for the task consists of instruction-trajectory demonstrations, where each trajectory is defined by a sequence of high-resolution 360° panoramas from the Matterport3D dataset. Inspired by trajectory augmentations with camera hardware for self-driving cars [2, 4], we hypothesize that spatially perturbing the location of the captured images could reduce overfitting to the incidental details of these trajectories, and improve generalization to unseen environments. We first create an improved baseline by upgrading the VLN↻BERT agent [8] to use much stronger MURAL [9] image features (+5% success rate). We then implement synthetic trajectory augmentation by using our model to spatially perturb the location of the training panoramas by up to 1.5m while training the agent. This augmentation improves the agent's success rate in unseen environments by an additional 1% on its own, or 1.5% when combined with renders of spatially-perturbed images from the Habitat [17] simulator – achieving state-of-the-art performance on the R2R test set (success rate of 66%). In contrast, traditional image data augmentation techniques such as cropping, color distortion and blurring do not yield improvements.

We refer readers to our full paper[1] for further details and results. Our code will be released to facilitate future work on data augmentation and embodied AI.

[1] https://arxiv.org/abs/2204.02960

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. 2

[2] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. End to end learning for self-driving cars. *arXiv:1604.07316*, 2016. 2

[3] 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

[4] Felipe Codevilla, Matthias Müller, Antonio López, Vladlen Koltun, and Alexey Dosovitskiy. End-to-end driving via conditional imitation learning. In *ICRA*, 2018. 2

[5] Chelsea Finn and Sergey Levine. Deep visual foresight for planning robot motion. In *ICRA*, 2017. 1

[6] 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. 1

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

[8] Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. A recurrent vision-and-language BERT for navigation. In *ECCV*, 2021. 2

[9] Aashi Jain, Mandy Guo, Krishna Srinivasan, Ting Chen, Sneha Kudugunta, Chao Jia, Yinfei Yang, and Jason Baldridge. Mural: Multimodal, multitask retrieval across languages. In *arXiv:2109.05125*, 2021. 2

[10] Jing Yu Koh, Honglak Lee, Yinfei Yang, Jason Baldridge, and Peter Anderson. Pathdreamer: A world model for indoor navigation. In *ICCV*, 2021. 1

[11] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2041–2050, 2018. 2

[12] Arun Mallya, Ting-Chun Wang, Karan Sapra, and Ming-Yu Liu. World-consistent video-to-video synthesis. *ECCV*, 2020. 1

[13] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 2

[14] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae2. 2019. 1

[15] Chris Rockwell, David F Fouhey, and Justin Johnson. Pixelsynth: Generating a 3d-consistent experience from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14104–14113, 2021. 1

[16] Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P Kingma. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. *ICLR*, 2017. 1

[17] 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. 2

[18] 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

[19] 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. 2

[20] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. 2018. 2

[21] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 37, 2018. 2