Learning Graph Variational Autoencoders with Constraints and Structured Priors for Conditional Indoor 3D Scene Generation

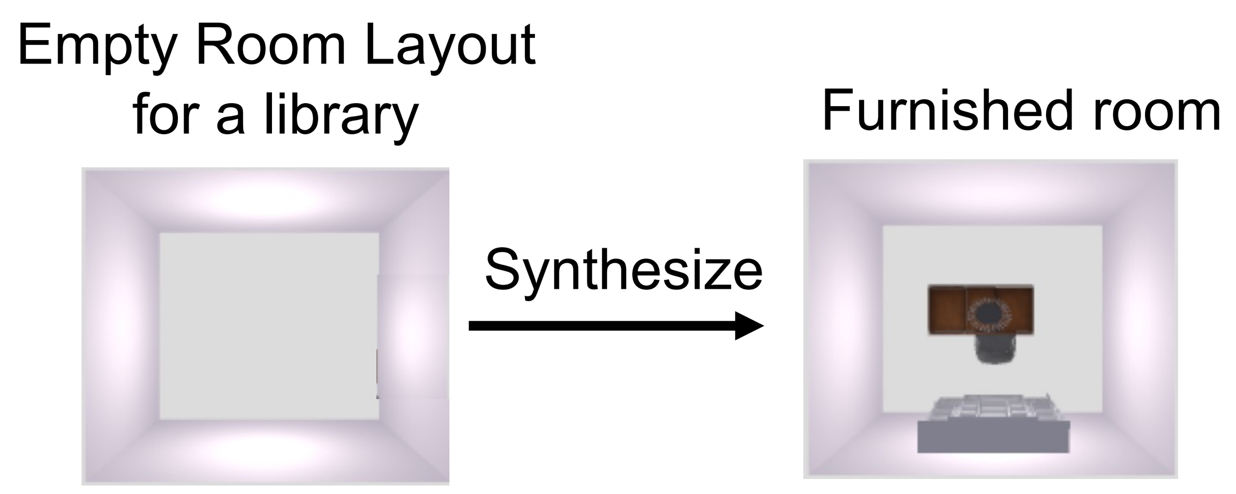
Johns Hopkins Mathematical Institute for Data Science

amazon | science

WACV Waikoloa Hawaii Jan 3-7 • 2023

Aditya Chattopadhyay[1], Xi Zhang[2], David Paul Wipf[2], Himanshu Arora[2], René Vidal[2]
[1]Johns Hopkins University, MD, USA, [2]Amazon.com, Inc

Problem Statement

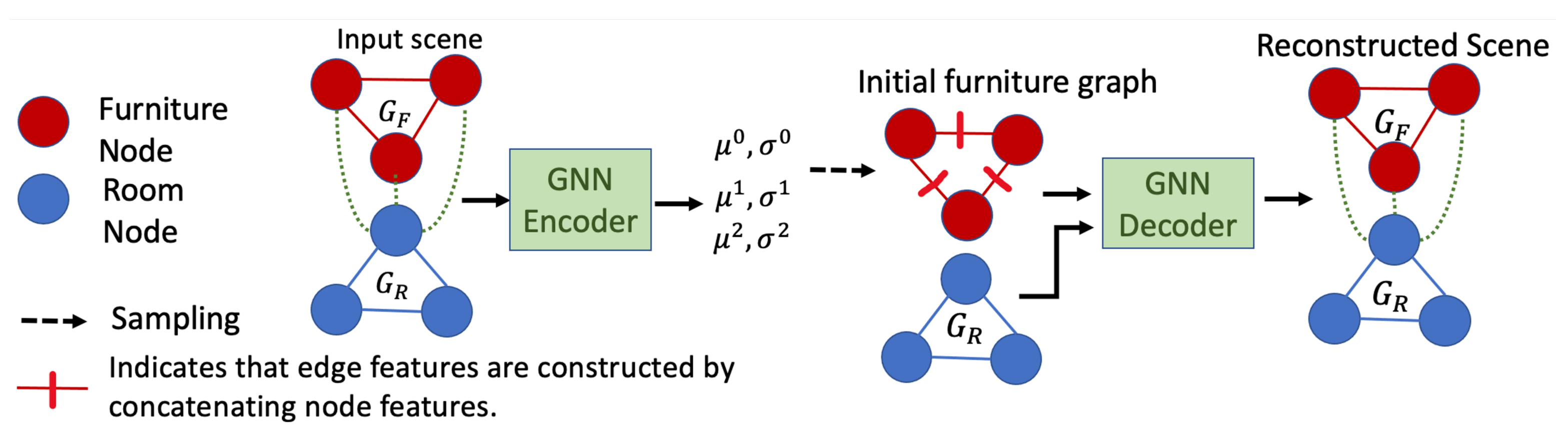


- Given an empty room layout and room type, can we synthesize diverse sets of furniture items consistent with the layout and room type?
- **Challenge:** This requires furniture items to simultaneously satisfy multiple constraints,
  - Each furniture item must be inside the room.
  - Two objects cannot occupy the same volume.
  - Some objects tend to co-occur in particular orientations relative to the room layout.
- Most existing works rely on hand designed heuristics, not scalable!

Proposed Approach

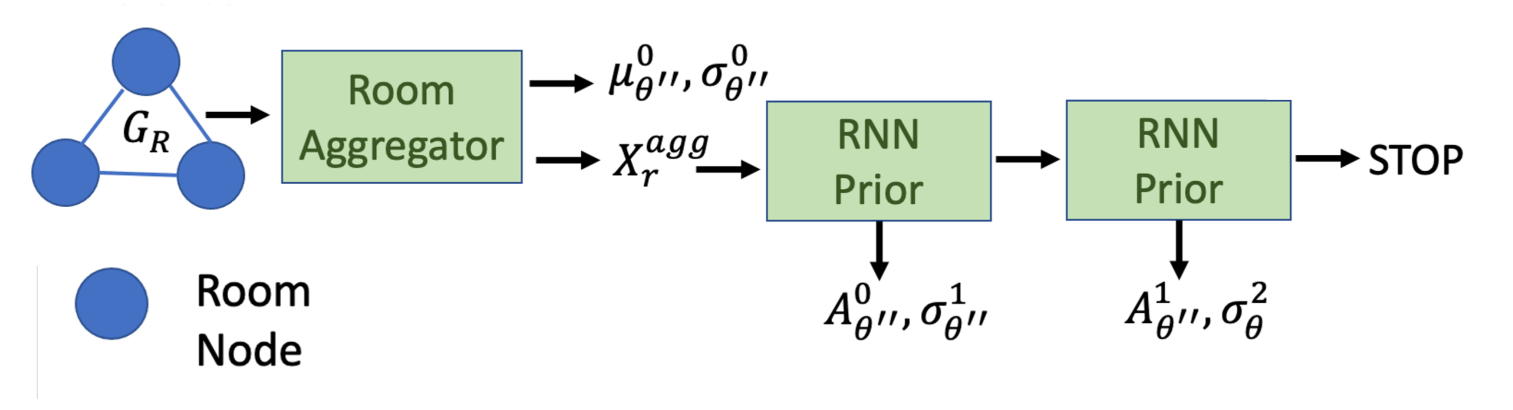
- We represent room and furniture layouts as attributed graphs.
- We propose a VAE model with a Graph Neural Network (GNN) backbone.
- We propose a novel structured autoregressive prior based on linear gaussian models which effectively captures the structure in 3D scenes.
- To learn our proposed VAE model using this autoregressive prior, we propose an efficient way to compute the KL divergence term in the VAE objective.
- We qualitatively and quantitatively evaluate the efficacy of the proposed model on the 3D-FRONT dataset[2].

Proposed VAE Model



- Input scene graph includes furniture and room sub-graphs and is complete.
- GNN encoder predicts mean and variance per furniture node.
- This mean and variance are used to sample latents for the GNN decoder which reconstructs scene.

Proposed VAE autoregressive prior



- The prior $P(Z_0, Z_1, \ldots, Z_{n_F} \mid G_R, T, n_F)$ is conditioned on the room graph G_R, type T and # furnitures n_F.
- The first latent, $Z_0 \sim \mathcal{N}(\mu_{\theta''}(G_R, T), \sigma^0_{\theta''}(G_R, T))$
  - $\mu_{\theta''}(G_R, T), \sigma^0_{\theta''}(G_R, T)$ is computed by a Room Aggregator network which is a GNN.
- Subsequent latents, $Z_i \sim \mathcal{N}(\sum_{k<i} A^k_{\theta''}(G_R, T)Z^k, \sigma^i_{\theta''}(G_R, T))$
  - $A^{i-1}_{\theta''}(G_R, T), \sigma^i_{\theta''}(G_R, T)$ is computed recursively using the RNN Prior network.

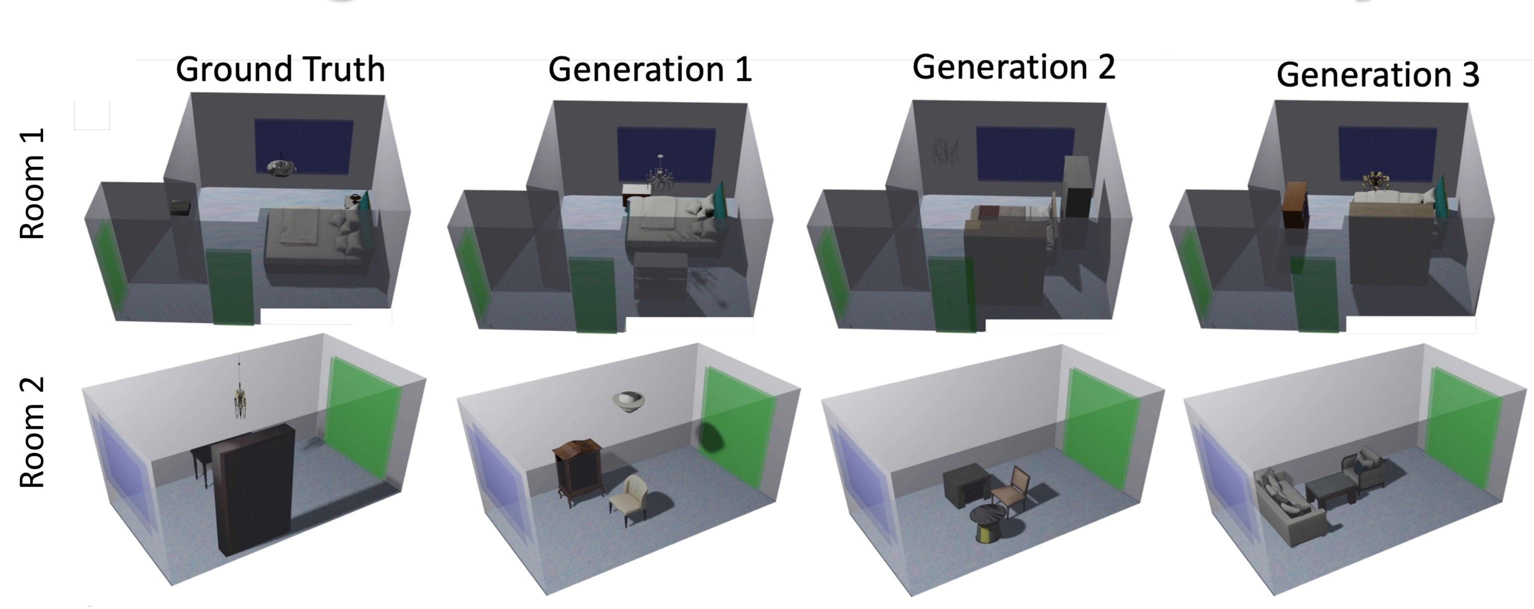
Computing KL divergence term

- VAE objective = Reconstruction Loss + KL div.
- But *apriori*, the KL div. term cannot be computed
  - The autoregressive prior induces an ordering over the latent variables.
  - Thus, we need to match these latents to the latents obtained from the posterior distribution.
  - How can we compute this matching?
- We prove that the proposed autoregressive prior reduces this matching problem to a quadratic assignment problem (QAP) which can be "approximately" solved efficiently!

Qualitative comparison with Prior work

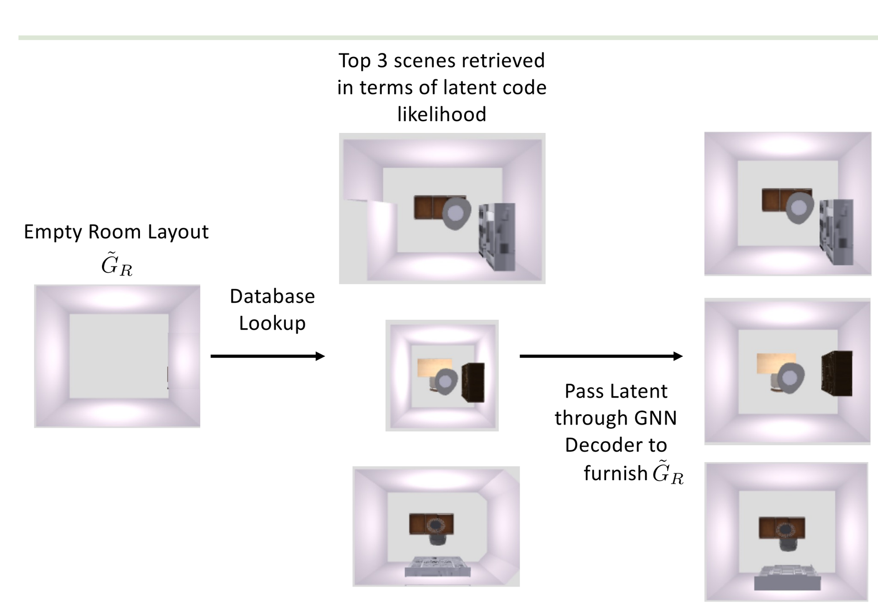
- Baseline 1 is a VAE is the standard normal i.i.d prior.
- Baseline 2 is a VAE where the prior is normal i.i.d, but with mean and variance parameters depend on the room graph.

Diverse generations for same room layout



Utility of the learned latent space

Case 1: editing scenes by latent space interpolation, for example, morphing the top cabinet into a chair.



Case 2: Ranking and finding the most appropriate furnished room from a database for a given empty floorplan using the learned latent space.

References

1. Paschalidou, Despoina, et al. "Atiss: Autoregressive transformers for indoor scene synthesis." *Advances in Neural Information Processing Systems* 34 (2021): 12013-12026.
2. Fu, Huan, et al. "3d-front: 3d furnished rooms with layouts and semantics." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.