Text2Room: Extracting Textured 3D Meshes from 2D Text-to-Image Models

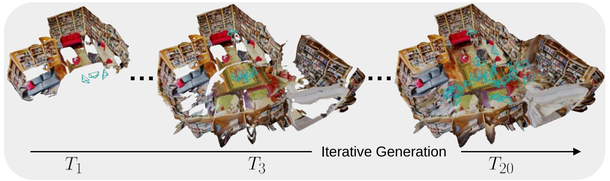
Lukas Höllein[1*], Ang Cao[2*], Andrew Owens[2], Justin Johnson[2], Matthias Nießner[1]
[1]Technical University of Munich, [2]University of Michigan
*joint first authorship

Code & Paper

ICCV23 PARIS

Introduction

Text2Room generates textured scene meshes from a given text prompt:

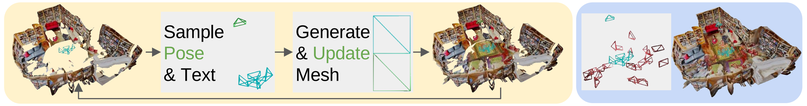


"a living room with lots of bookshelves, couches, and small tables"

progressive mesh generation

rendered images

We iteratively generate meshes from (pose, text) input in two stages:

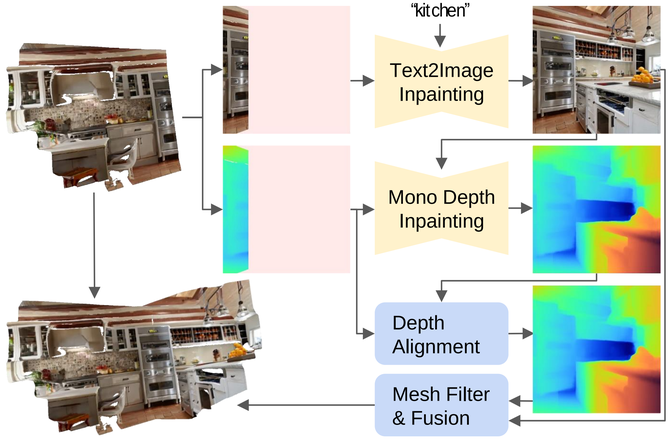


Scene Generation

Scene Completion

Iterative Scene Generation

For each pose, we use inpainting models to complete both rendered RGBs and depths. Then, we perform depth alignment and mesh filtering to get a next mesh patch, that is finally fused with the existing geometry.



Generated 3D Scenes

Editorial Style Photo, Coastal Bathroom, Clawfoot Tub, Seashell, Wicker, Mosaic Tile, Blue and White

Editorial Style Photo, Modern Living Room, Large Window, Leather, Glass, Wood Paneling, Apartment

A living room with a lit furnace, couch, and cozy curtains, bright lamps

Two-Stage Viewpoint Selection

Generation: create the main parts of the scene following predefined camera trajectories and prompts, eventually covering the whole room.

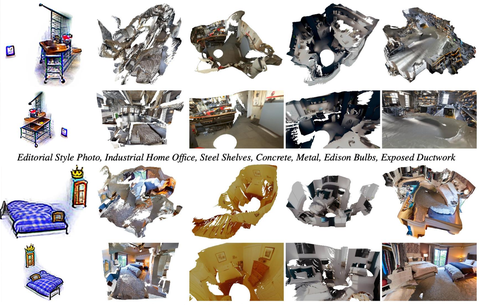
Completion: inpaint the remaining holes by sampling additional poses.

Ablations

No alignment | No filter

No completion | full pipeline

Baselines



Spatially Varying Scene Generation

Mixed Prompts

Layout Guidance