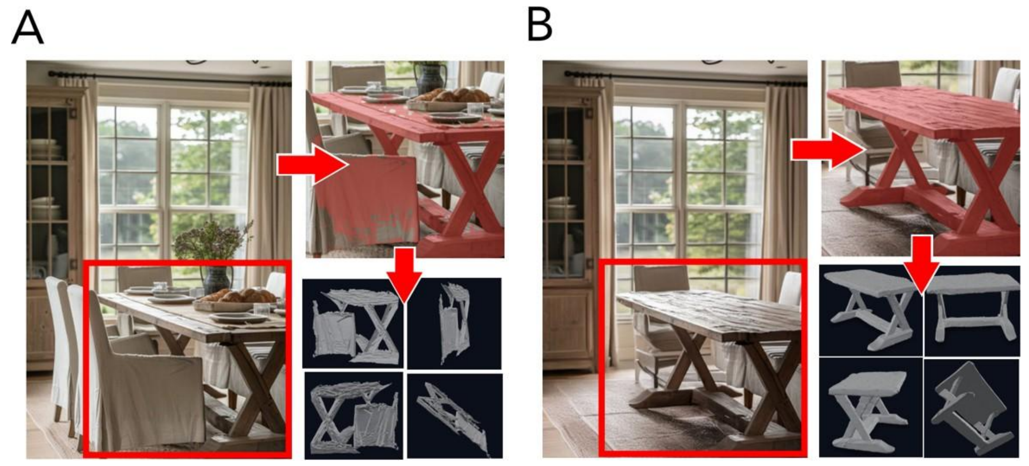
Seeing Through Clutter:
Structured 3D Scene Reconstruction via Iterative Object Removal

Rio Aguina-Kang[1], Kevin James Blackburn-Matzen[2], Thibault Groueix[2], Vladimir Kim[2], Matheus Gadelha[2]

1 University of California, San Diego; 2 Adobe Research

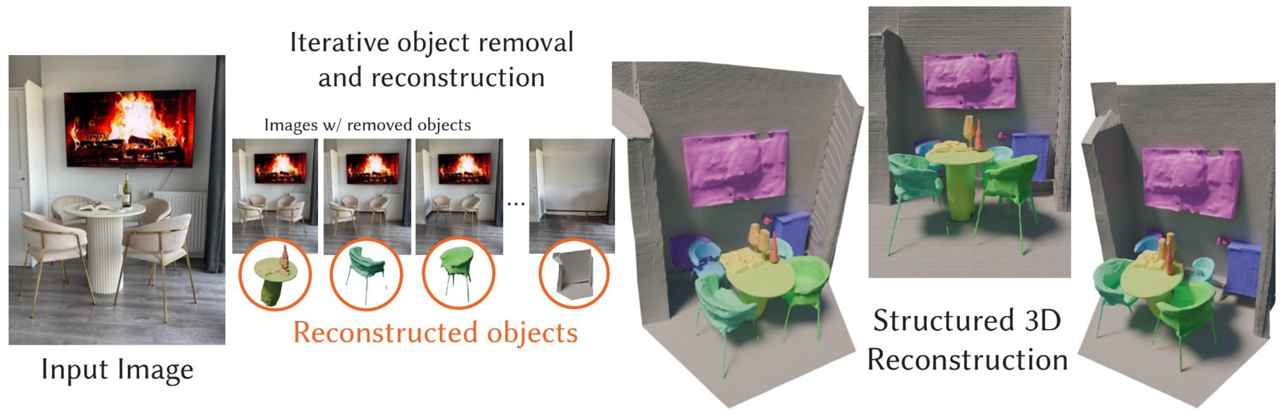
Motivation

We reconstruct structured 3D scenes from a single image. Occlusion and clutter degrade segmentation and depth, producing fragmented masks and unreliable geometry. Our insight: iteratively remove foreground objects, transforming a difficult global problem into a sequence of tractable local ones.



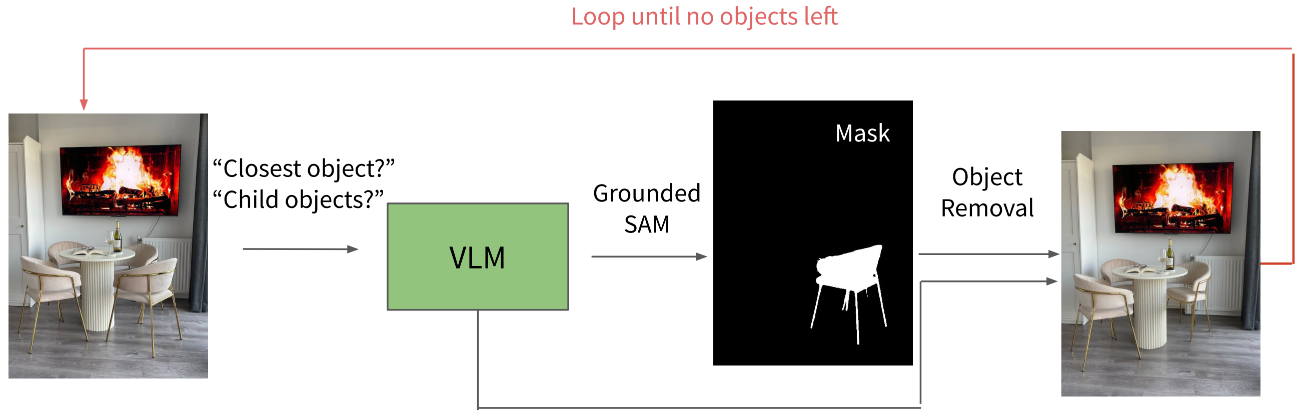
Pipeline Overview

SeeingThroughClutter is training-free. Identify the closest unoccluded object → segment → remove via inpainting → repeat. This yields progressively decluttered images and amodal masks that feed into layout optimization for coherent 3D reconstruction.


Iterative object removal and reconstruction
Images w/ removed objects
...
Input Image
Reconstructed objects
Structured 3D Reconstruction

Stage 1: Automated Inpainting

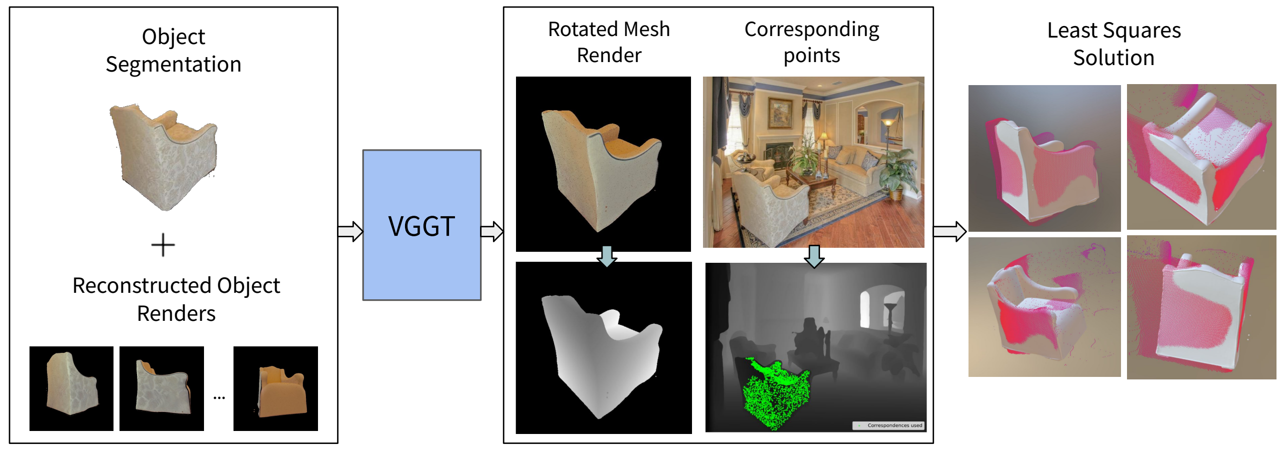
The VLM selects the closest fully visible object (and any secondary objects on top). Grounded-SAM segments it; Flux Kontext or inpainting removes it with VLM-guided prompts for plausible background fill. Loop until no objects remain.



Prompt: "What content exists behind this object?"
Negative Prompt: "What other objects could be here?"

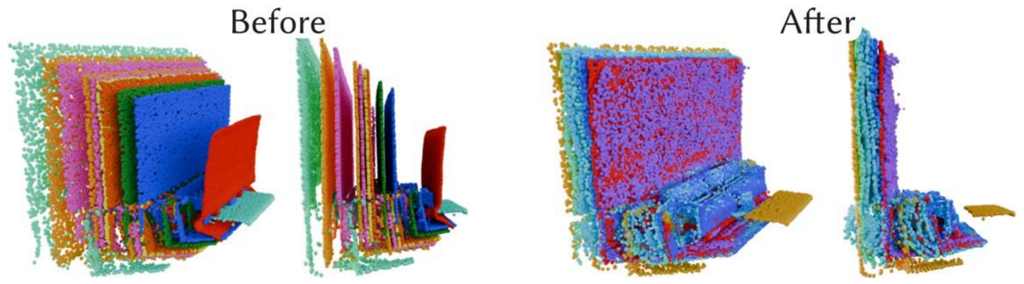
Stage 2: Object Generation & Fitting

For each removed object, reconstruct a 3D mesh and estimate monocular depth on the decluttered image. VGGT finds coarse rotation and correspondences that allow us to optimize for fit directly.



Depth Alignment

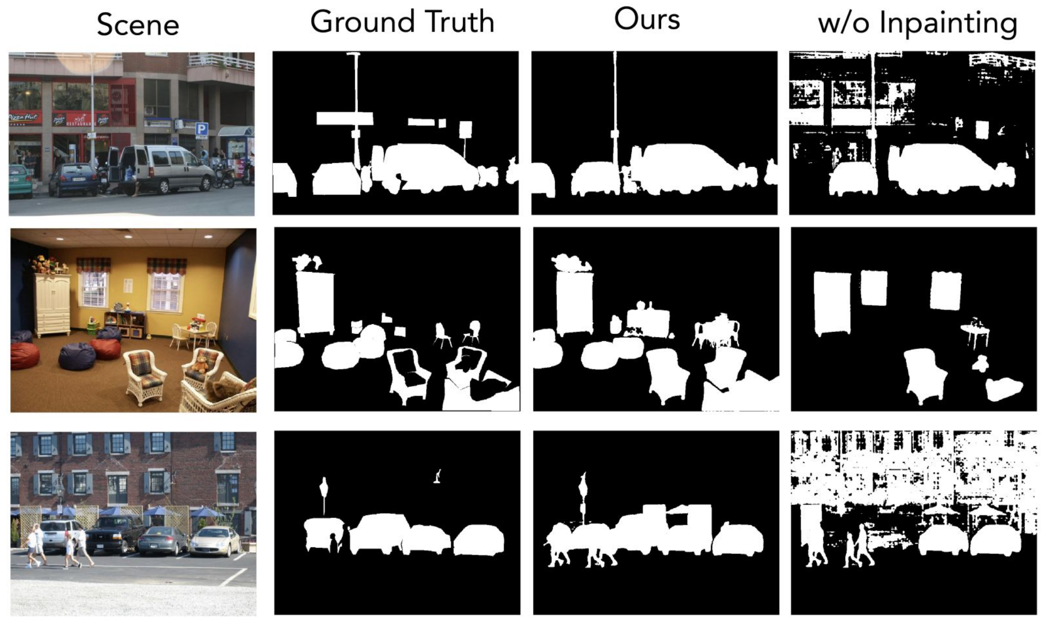
Independent depth estimates across the image sequence are incongruent. A coordinate-based MLP refines each layer's disparity using the original image as a fixed reference, with a consistency loss over non-occluded regions between adjacent layers.



Results: Segmentation

On ADE20K/MIT Scene Parsing: with object removal, IoU reaches 0.44 vs. 0.35 without. RandIdx improves from 0.20 to 0.29. Iterative removal is especially effective for occluded and overlapping objects.

Method	IoU	RandIdx
Ours (w/ obj. removal - Kontext [20])	**0.44**	0.27
Ours (w/ obj. removal - inpainting [30])	**0.44**	**0.29**
Ours (w/o obj. removal)	0.35	0.20
Gen3DSR	0.41	0.16

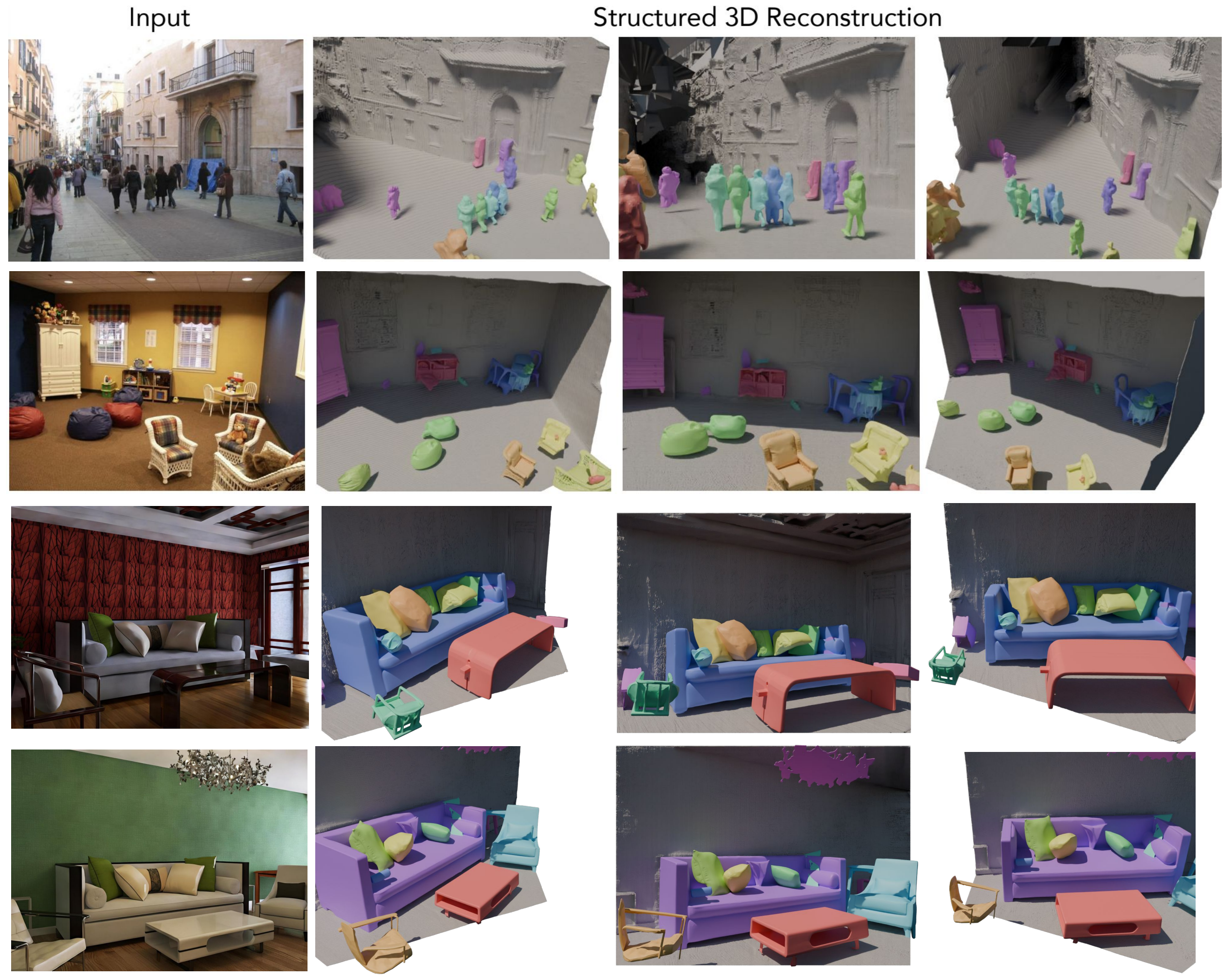


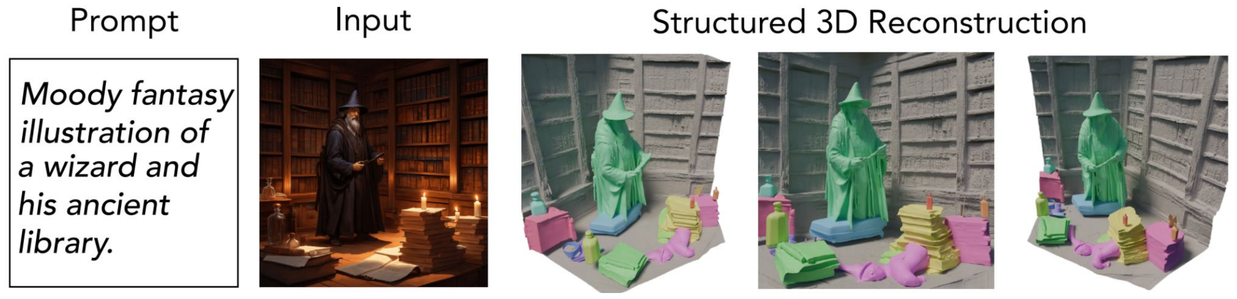
Results: 3D Reconstruction

On 3D-Front: 0.11 Chamfer Distance, 71.65% F1@0.1 — beating Gen3DSR (0.12, 70.18%) and MIDI* (0.24, 49.36%). Training-free design generalizes broadly.

Model	Chamfer ↓	P@0.1 ↑	R@0.1 ↑	F1@0.1 ↑
Gen3DSR	0.12	67.54	74.50	70.18
Gen3DSR (w/ backgrounds)	0.21	59.11	53.31	55.48
MIDI*	0.24	51.95	48.67	49.36
Ours	0.11	72.58	73.38	71.65
Ours (w/ backgrounds)	0.17	72.58	55.56	61.69
Ours (filtered)	0.12	71.03	69.90	68.85
Ours (filtered, w/ backgrounds)	0.17	70.96	54.01	60.10

In-the-Wild





Moody fantasy illustration of a wizard and his ancient library.

Limitations & Takeaways

- Inpainting artifacts can propagate through later iterations
- Inaccurate VGGT correspondences cause poor fits
- Runtime scales with scene complexity

Our training-free method shows that iterative object removal enables cleaner segmentations even in highly occluded scenes, and benefits directly from ongoing advances in foundation models.