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

Rio Aguina-Kang

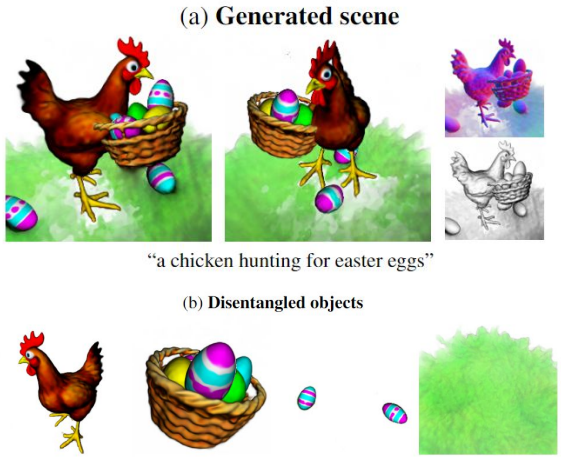
Current Directions in Scene Generation

LLM Scene Generation

- Examples often setting restricted (e.g. indoor scenes)
- Rely on LLM domain knowledge

HOLODECK (2023)

Diffusion/NeRF-Based Methods



- Great results, but panorama based
- Unstructured representation

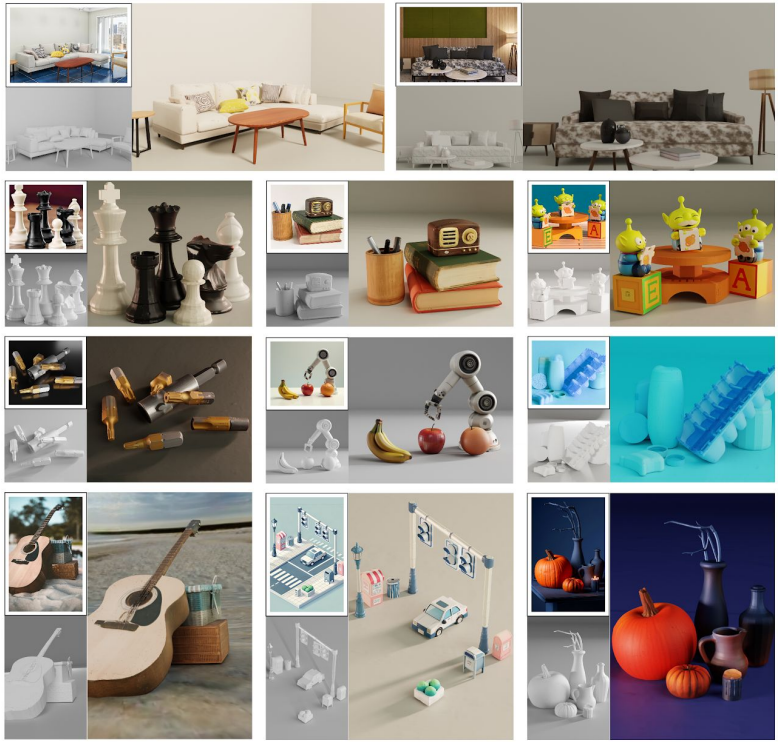
Disentangled 3D Scene Generation with Layout Learning (2023)

Single-View Scene Synthesis

- Often limited, examples either:
  - Require additional training
  - Rely on limited 3D object databases for retrievals
  - Aren't open-vocabulary

Diorama (2024), CAST (2025), MIDI (2025)

Why is single-view scene synthesis limited?



Smaller scale scenes

Poor Fitting

Why is single-view scene synthesis limited?

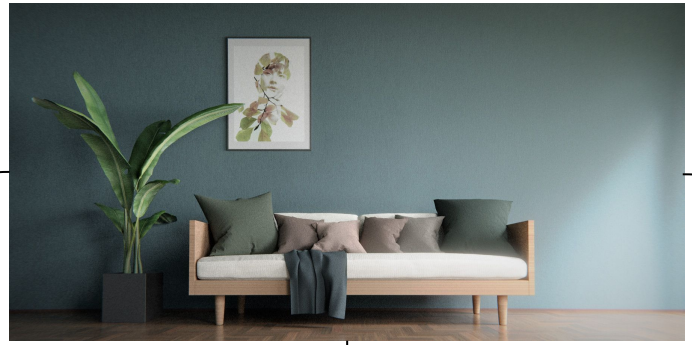
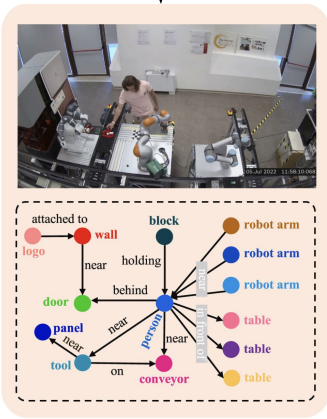


Image Segmentation

Depth Estimation



VLM Annotations

Why is single-view scene synthesis limited?

Why is single-view scene synthesis limited?

Why is single-view scene synthesis limited?

Why is single-view scene synthesis limited?

A

Why is single-view scene synthesis limited?

A

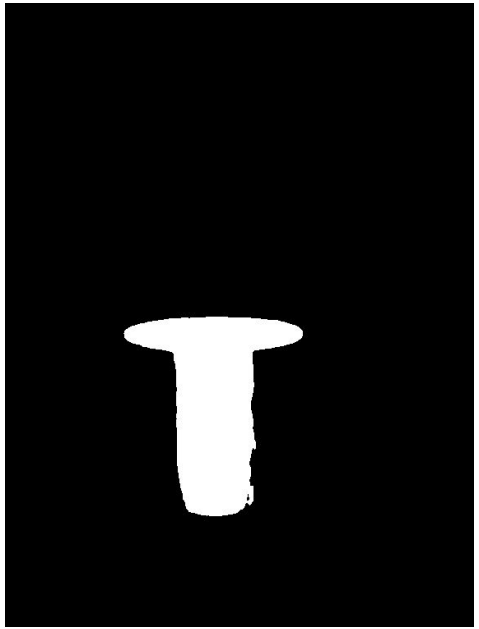
B

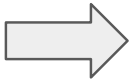
Automating Object Removal?

Promising Results in Object Removal

SmartEraser (2025)

Caveats in Object Removal





Automating Object Removal?

Promising Results in Object Removal

SmartEraser (2025)

Tables are symmetrically placed in the room; each table should have two chairs on opposite sides of the table facing each other, ready for dining …

All the tables aligned to form a line, diving the room up into two halves, place all the chairs on one side of the line and the buffets on the other side

Spatial Reasoning in VLMs?

LayoutVLM (2025)

Method Overview

Input Image

Automated Iterative Inpainting

Object Generation & Fitting

Structured 3D Reconstruction

Method Overview

Input Image

Automated Iterative Inpainting

Object Generation & Fitting

Structured 3D Reconstruction

Method: Automated Inpainting Pipeline

Input Image

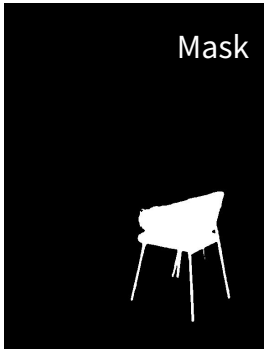
Method: Automated Inpainting Pipeline

Input Image

“Closest object?”
“Child objects?”

VLM

Grounded-SAM



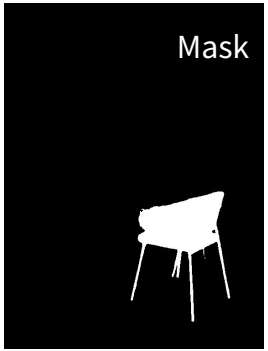
Method: Automated Inpainting Pipeline

Input Image

"Closest object?"
"Child objects?"

VLM

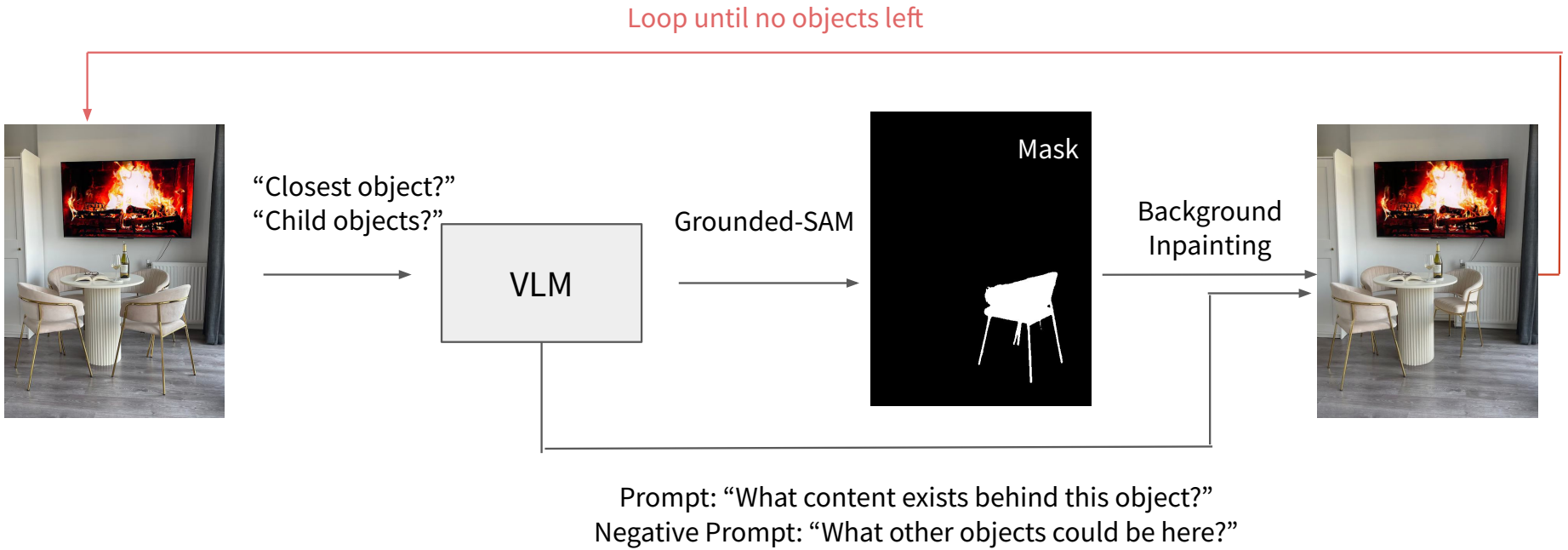
Grounded-SAM



Background Inpainting

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

Method: Automated Inpainting Pipeline



Method Overview

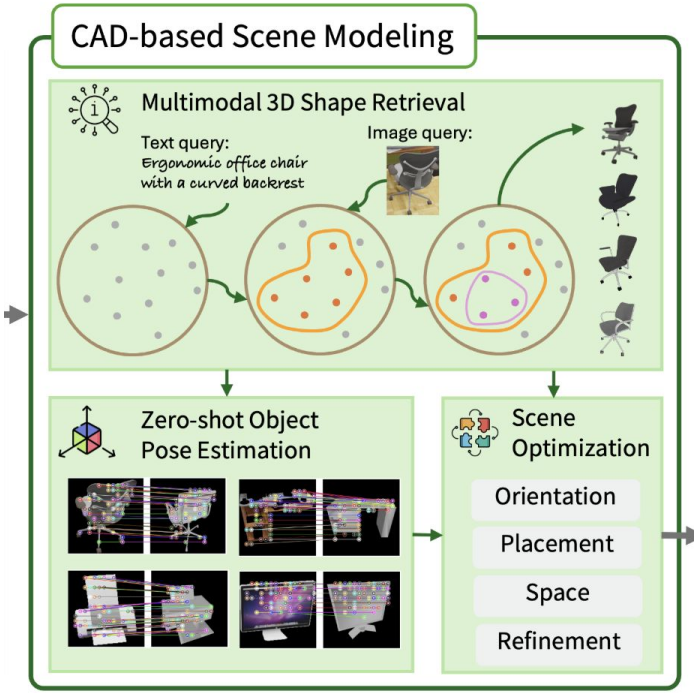
Input Image

Automated Iterative Inpainting

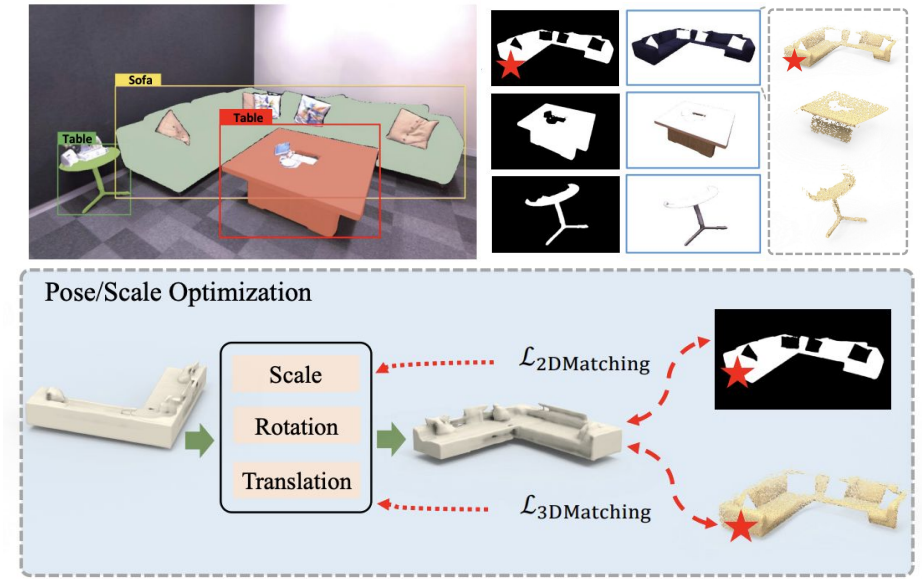
Object Generation & Fitting

Structured 3D Reconstruction

Prior work: Object Generation & Fitting



Diorama
(2024)

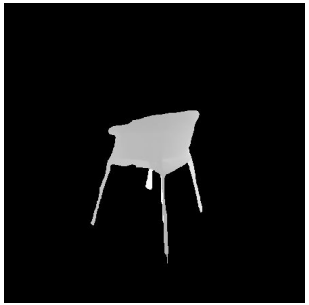


DeepPriorAssembly
(2024)

Method: Object Generation & Fitting

Amodal Depth Segmentations

Camera-view centric 3D reconstruction

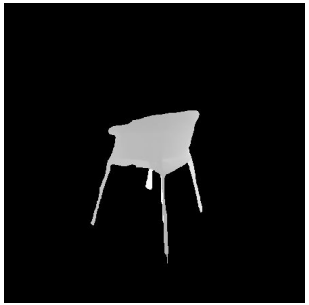


Depth renders

Method: Object Generation & Fitting

Amodal Depth Segmentations

Camera-view centric 3D reconstruction



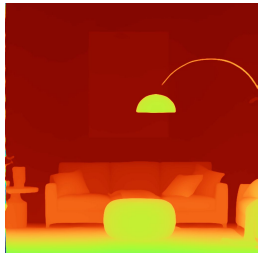
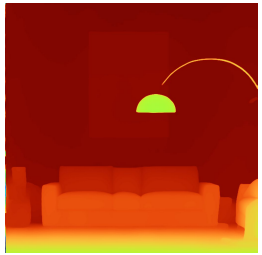
Depth renders

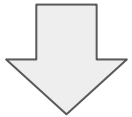
Fitting

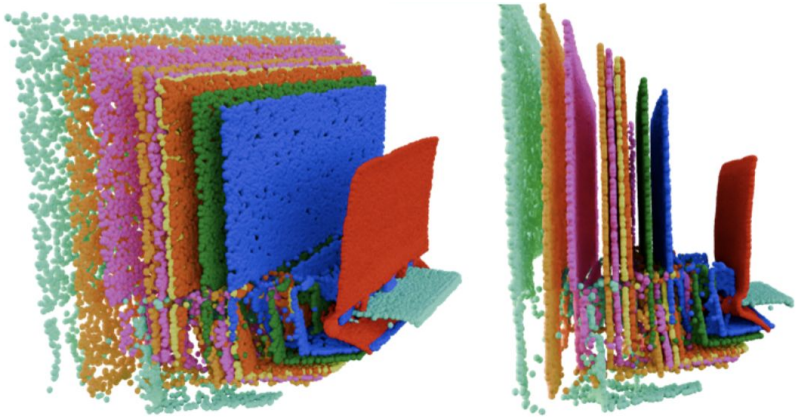
Final Reconstruction

Object fitting becomes a simplified 4 DoF (uniform scale & translation) RANSAC problem!

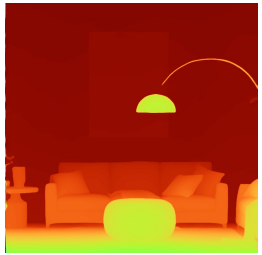
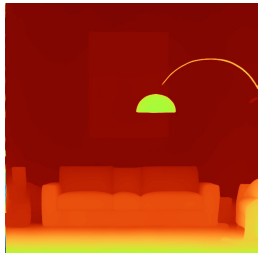
Method: Depth Alignment

$i = 1$  , $i = 2$  , $\cdots$ $i = n$ 



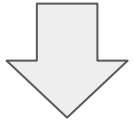


Method: Depth Alignment

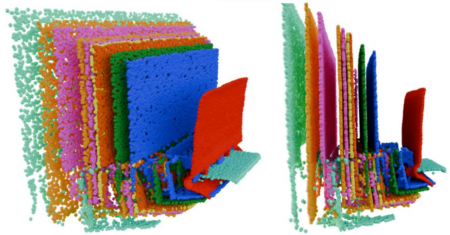
$i = 1$  , $i = 2$  , $\cdots$ $i = n$ 

Fixed Reference

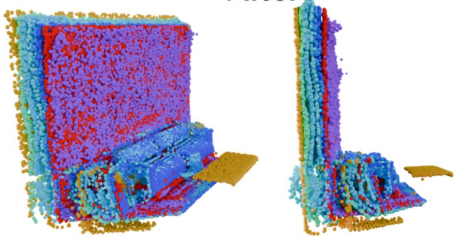
$$\mathcal{L}(\theta) = \sum_{n=1}^{N-1} \sum_{x,y} \left(1 - M_n(x,y)\right) \left| \mathcal{D}'_n(x,y) - \mathcal{D}'_{n+1}(x,y) \right|.$$

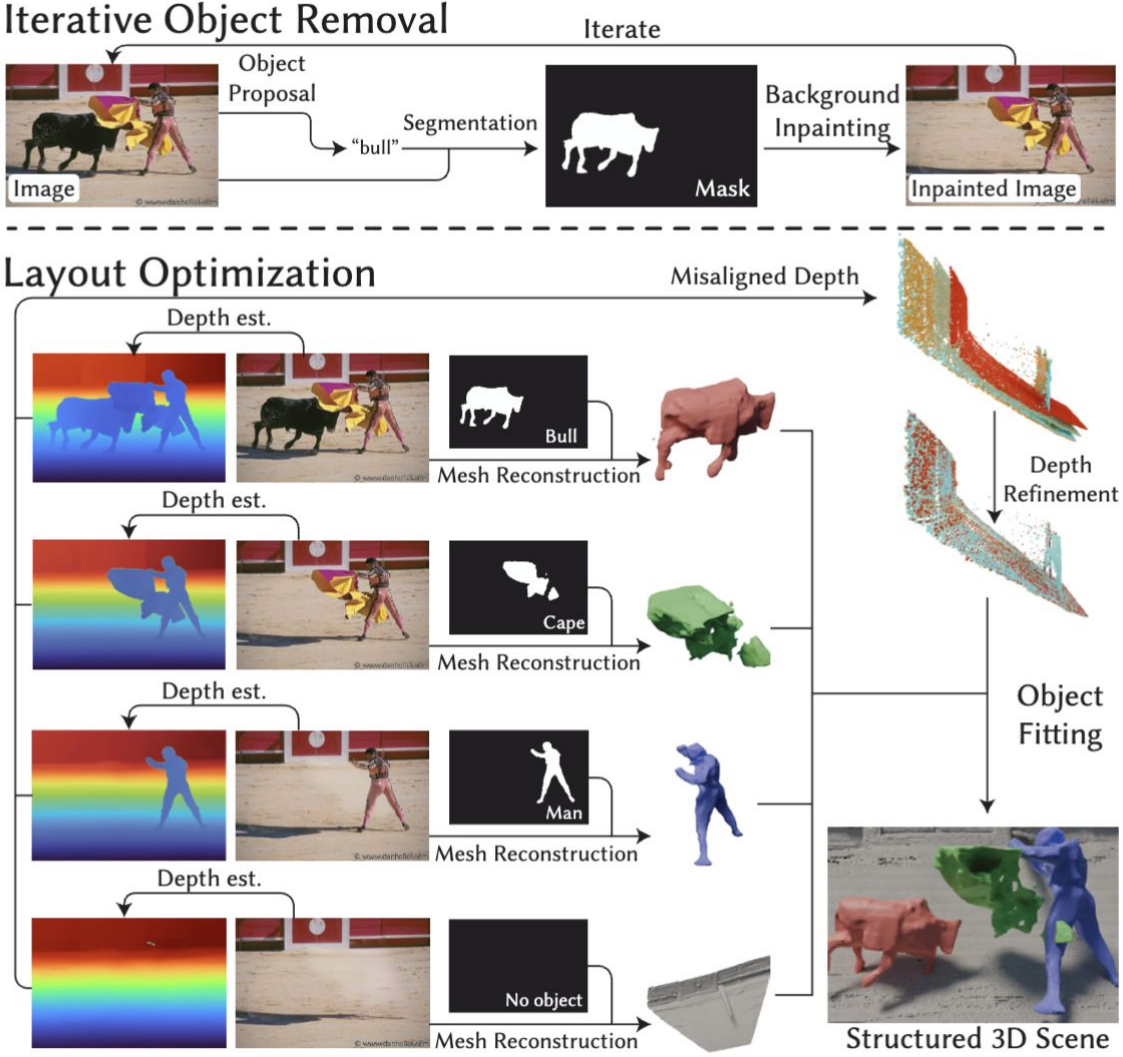


Before



After




Iterative Object Removal
Iterate
Object Proposal
"bull"
Segmentation
Background Inpainting
Image
Mask
Inpainted Image
Layout Optimization
Misaligned Depth
Depth est.
Bull
Mesh Reconstruction
Depth Refinement
Cape
Man
No object
Object Fitting
Structured 3D Scene

Results: Existing Benchmarks

3D FRONT

- Synthetic Indoor Scenes w/ 3D GT
- Well-used benchmark for single-view reconstruction

MIT Scene Parsing Benchmark

- Scene segmentation benchmark w/ semantic annotation (e.g. things vs. stuff)
- No 3D GT, strictly segmentation

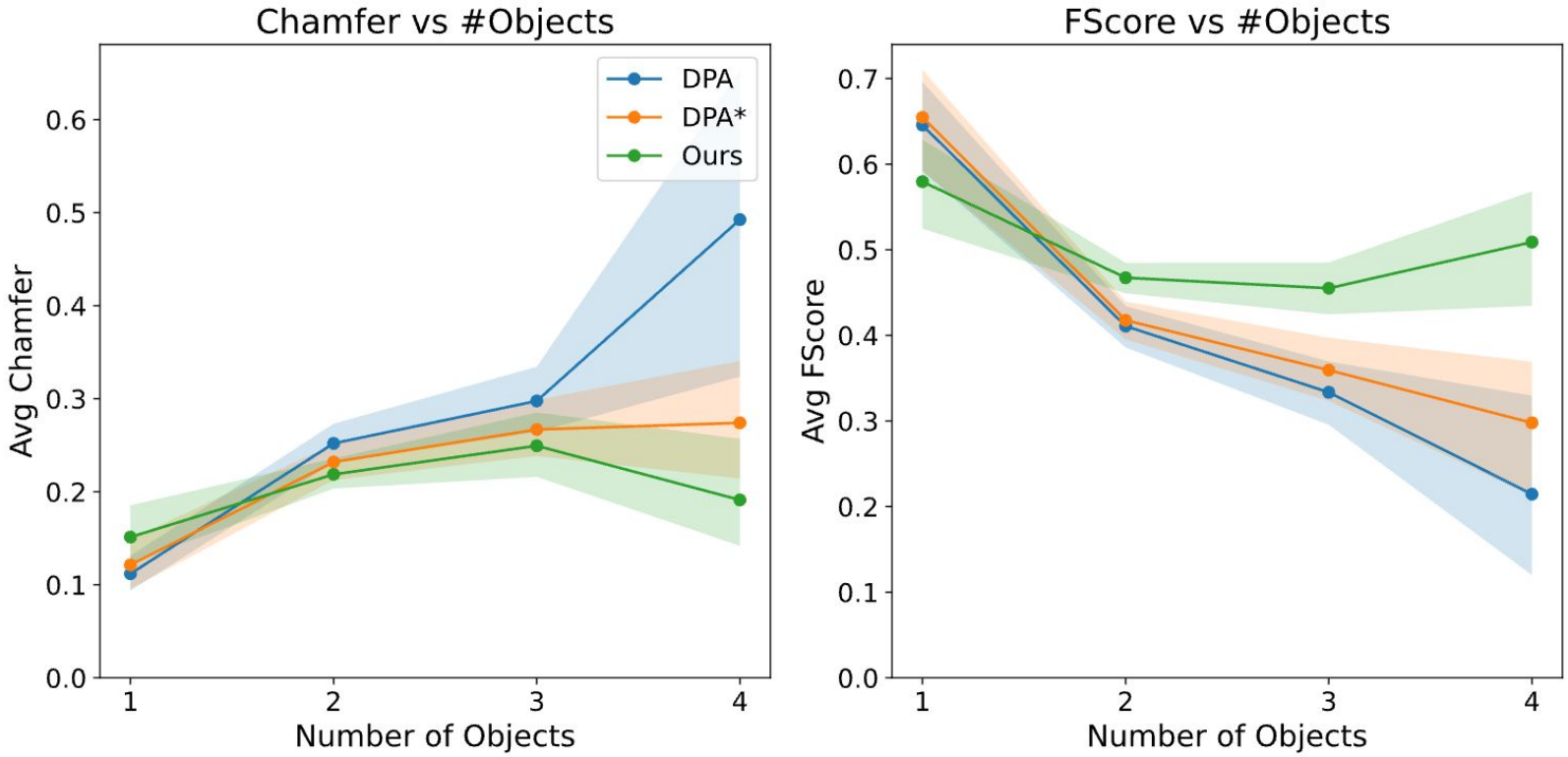
Results: Indoor Scenes (3D Front)

Baselines:

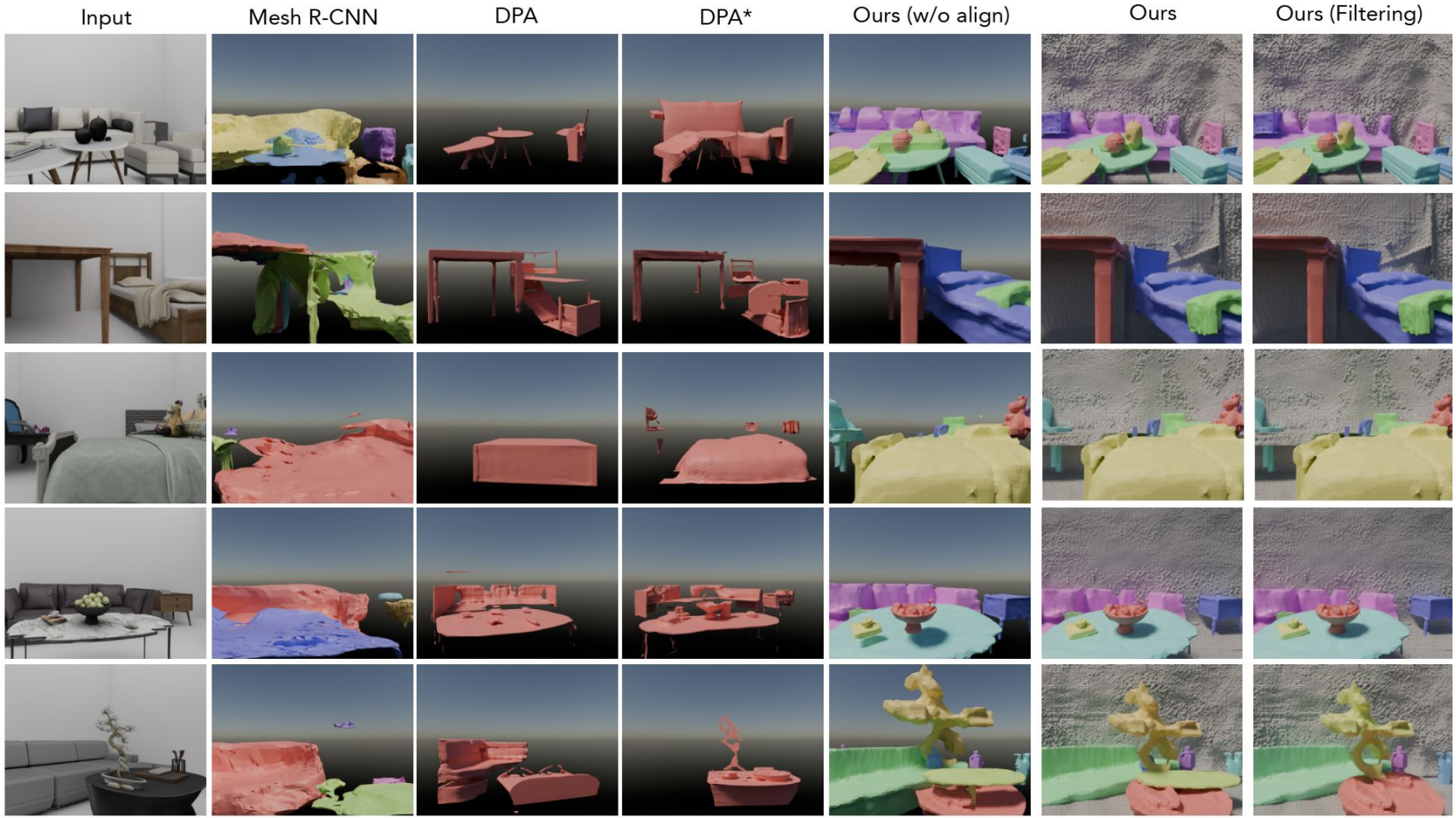
- DeepPriorAssembly (or DPA; closed vocabulary)
- DPA* (augmented w/ VLM labels)
- Ours (without depth alignment)
- Ours (w/ object artifact filtering)

Model	CD ↓	FS ↑	Obj-FS ↑	Depth ↓	IoU ↑	M-IoU ↑
DPA	24.90	42.49	9.357	0.287	0.726	0.213
DPA*	23.04	42.73	9.037	0.286	0.751	0.205
ours (w/o depth ref.)	23.38	45.36	11.21	0.095	0.787	0.529
ours (w/ object filt.)	21.80	47.67	**12.70**	**0.082**	**0.817**	0.534
ours	**21.66**	**48.07**	12.53	0.085	**0.817**	**0.539**

Results: Indoor Scenes (3D Front)



Results: Indoor Scenes (3D Front)

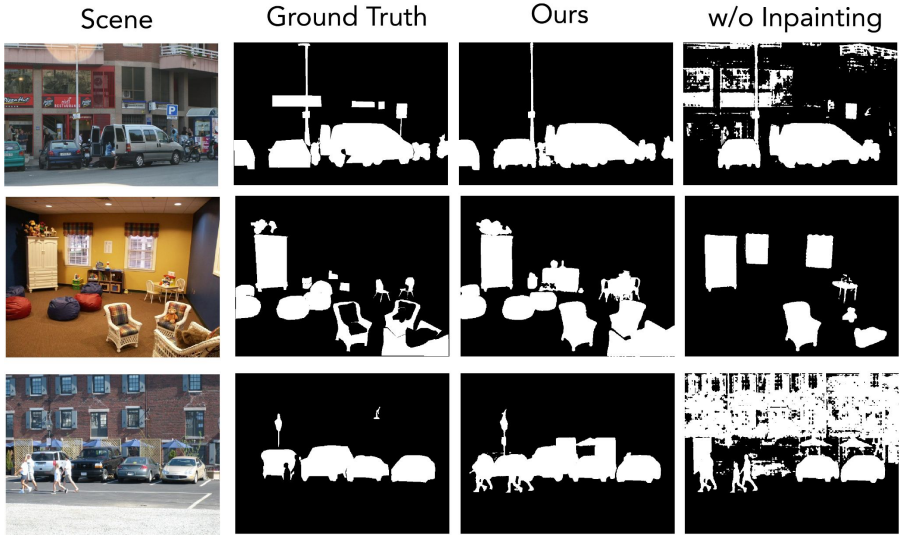


Results: Outdoor Scenes (ADE20K)

Baselines:

- Ours (w/o iterative object removal)

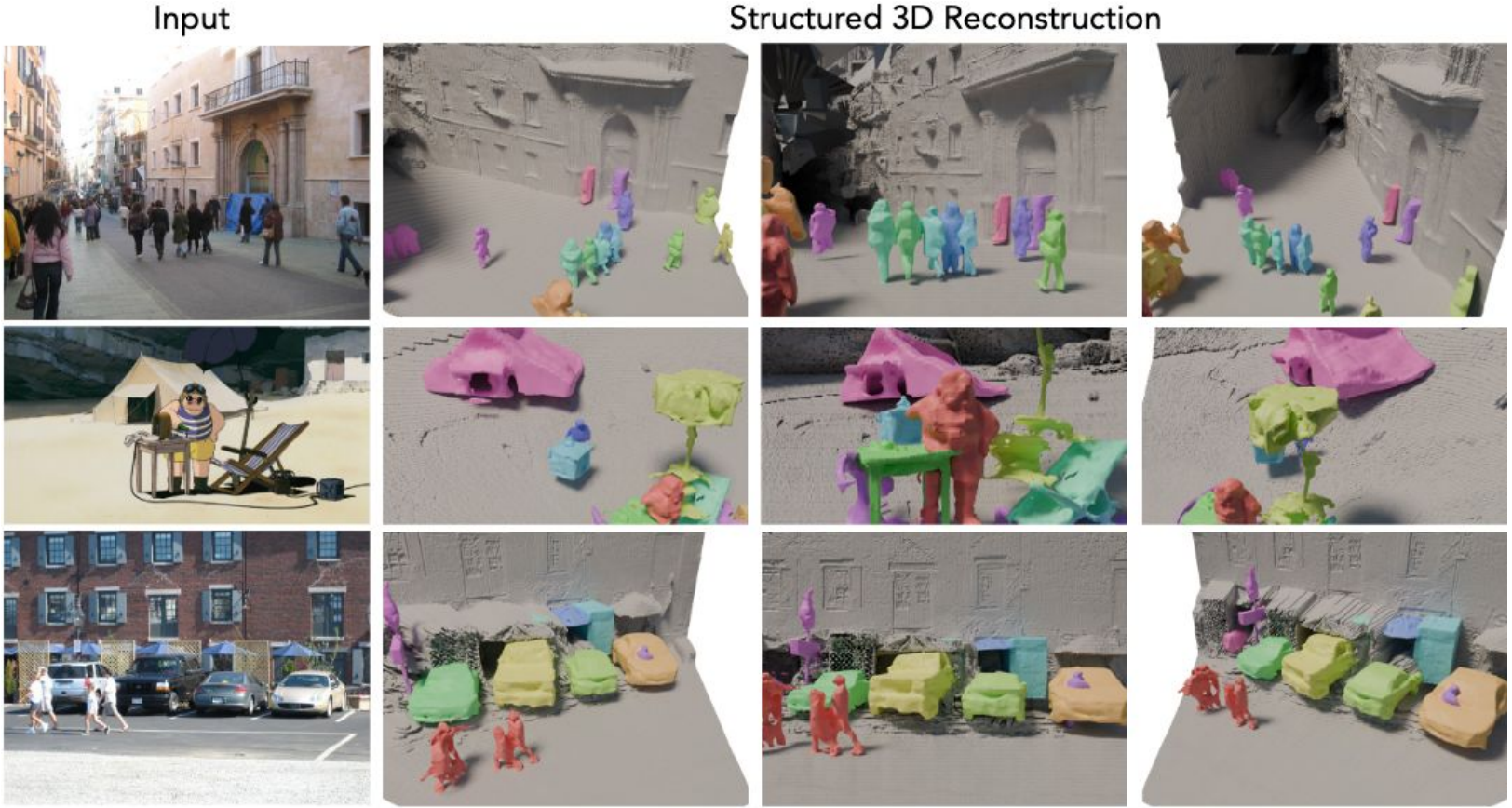
Method	Average IoU
Ours (things)	**0.33**
w/o obj. removal (things)	0.29
Ours (things + stuff*)	**0.28**
w/o obj. removal (things + stuff*)	0.24



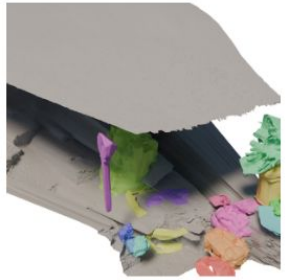
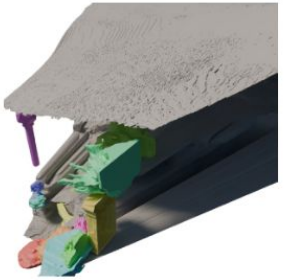
Qualitative Results: Indoor Scenes

Qualitative Results: Indoor Scenes

Qualitative Results: Outdoor Scenes

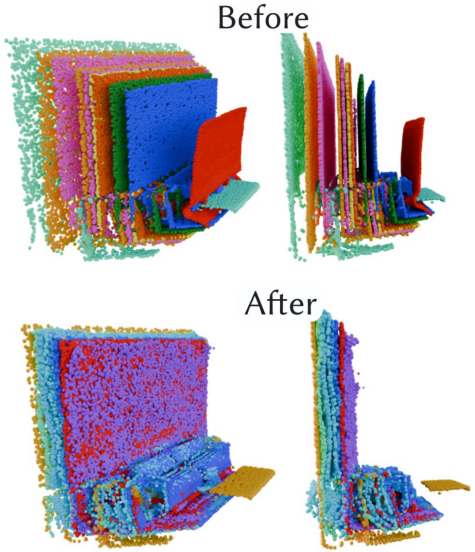


Qualitative Results: Text2Img

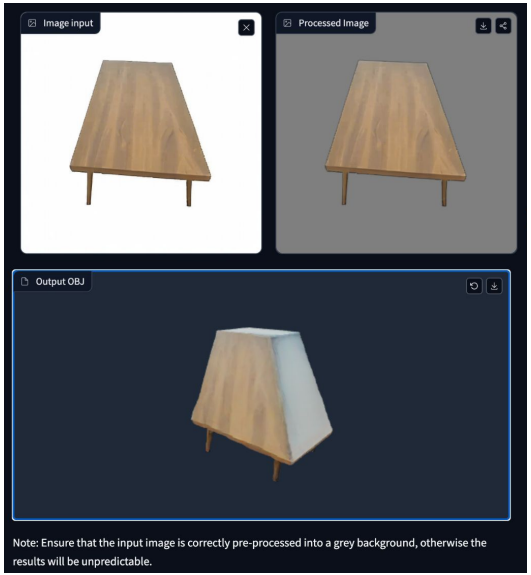
Prompt	Input	Structured 3D Reconstruction		
Moody fantasy illustration of a wizard and his ancient library.				
Underwater ruins draped in coral and kelp, a carved sea-god relief partly obscured, shafts of turquoise light				
Cyberpunk neon alley at midnight, two hooded operatives exchanging data behind steaming vents and tangled chrome cables				

Limitations & Future Work

Artifacts Produced by Inpainting



Depth alignment
==
Heuristic solution



Camera-view centric object generation

A quick thank you to my collaborators & mentors!

Matheus, Thibault, Kevin, Vova and other mentors at Adobe I met along the way… :)