BEYOND THE FRUSTUM: SINGLE-VIEW AMODAL 3D SCENE RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Single-image 3D methods either reconstruct the visible surface or generate a plausible room, leaving a gap for metric completion tied to the input. We study *evidence-grounded amodal occupancy completion*: recover the scan-observed geometry of every object supported by a single RGB(-D) view, including occluded and out-of-frustum parts, while predicting no unsupported objects. We present Seed2Scene, a feed-forward framework that back-projects depth into a sparse, DINO-decorated 3D seed and completes it with a 3D U-Net inside a predicted scene-fitted box. A scalar coverage channel, aggressive seed dropout, and a completion-masked BCE–Dice loss are designed to discourage trivial seed copying and to focus learning on unseen voxels. On 3RScan, Seed2Scene substantially outperforms reconstructive, generative, and indoor-occupancy baselines on unseen geometry; explicit in-frustum/out-of-frustum evaluation and a volume-matched dilation control confirm that the gain reflects completion rather than surface inflation. The model also transfers zero-shot to ScanNet. Code and trained models will be made public.

1 INTRODUCTION

A single image of an indoor scene exposes only a thin shell of the world: a half-visible chair, a partially observed wall, the leg of a sofa cropped at the image boundary. Many applications are well served by photorealistic radiance fields (Mildenhall et al., 2020; Kerbl et al., 2023) or by methods that synthesize an entire room from one image. Others need a *hybrid* that completes only what the camera saw: an embodied agent grasping a partly visible mug needs the mug's full shape, and a planner avoiding a sofa cropped at the image boundary needs to know where the sofa ends. Both need the metric occupancy of the real furniture, not a plausible substitute, and today's single-image 3D pipelines provide neither the geometry of partly visible objects nor a faithful occupancy of the surrounding volume.

Mainstream single-image 3D divides into two regimes. *Feed-forward reconstructors* – pointmap regressors such as DUSt3R (Wang et al., 2024), VGGT (Wang et al., 2025b), and MapAnything (Keetha et al., 2026), monocular depth predictors (Ranftl et al., 2021; Bhat et al., 2023; Yang et al., 2024; Ke et al., 2024), and per-pixel primitive predictors (Szymanowicz et al., 2024; 2025a) – recover what the camera sees, but their outputs are 2.5D: silent about occluded or out-of-frustum geometry, and in the single-view case little more than depth back-projection. *Image-conditioned scene generators* (Wang et al., 2025a; Yu et al., 2025; Labs, 2025; Schneider et al., 2025) extrapolate to entire rooms with diffusion-based synthesis (Sargent et al., 2024; Sun et al., 2025; Liu et al., 2024a; Ren et al., 2025; Bahmani et al., 2025), but optimize perceptual realism rather than metric agreement: they *hallucinate* a plausible room, not the one the camera observed.

A narrower line of work bridges this gap. Monocular semantic scene completion (SSC) (Cao & De Charette, 2022; Yao et al., 2023; Li et al., 2023; Jiang et al., 2024; Yu et al., 2024b;a; Wu et al., 2025), density fields behind the visible surface (Wimbauer et al., 2023; Han et al., 2024; Li et al., 2024; Hayler et al., 2024; Jevtić et al., 2025; Wulff et al., 2025), and RGB-D completion (Iwase et al., 2024; Sun et al., 2024; Jin et al., 2024; Reed et al., 2024; Lu et al., 2025) show that learned priors can extend beyond the observed surface. Each covers only part of the picture: SSC is tied to a fixed semantic vocabulary and a fixed volume in front of the camera, density fields target view-synthesis losses rather than metric occupancy, and RGB-D completion works at the object level or needs depth sequences. *None* recovers, from a single RGB(-D) frame, the geometry of every object the camera saw, including the parts behind other objects or outside the field of view.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

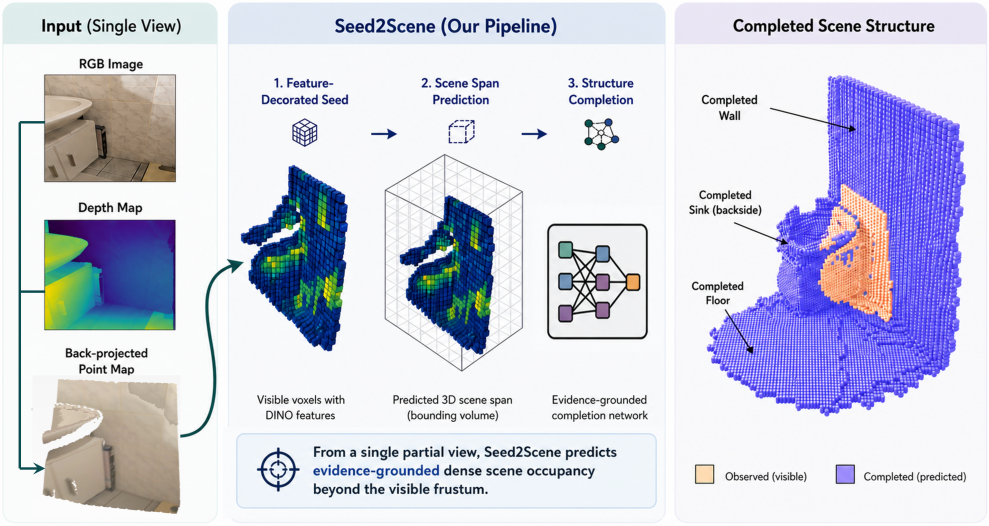

Figure 1: **Single-view amodal 3D scene reconstruction.** From one RGB-D frame, Seed2Scene predicts the metric voxel occupancy of *every object the camera observed*, including occluded backs and the parts of partly visible objects outside the camera frustum, and invents nothing for objects without support in the input. The output sits between the 2.5D shell of feed-forward reconstructors and the unconstrained imagination of image-conditioned scene generators.

Our work. We define *evidence-grounded amodal occupancy completion* (Fig. 1): from a single RGB(-D) frame, predict the metric voxel occupancy of every object with *support* in the view – its visible surface, its *occluded back*, and any part outside the frustum – and predict nothing for objects without support. The target is bounded by what the complete multi-view scan observed, which separates the task from open-ended scene generation. Our method, Seed2Scene, lifts the image and a depth map (from a sensor or a monocular predictor (Bhat et al., 2023; Yang et al., 2024; Ke et al., 2024)) into a sparse 3D *seed* decorated with DINOv2 features (Oquab et al., 2024), regresses a scene-fitted bounding box, and completes the seed into a dense occupancy with a 3D U-Net. Fed the seed directly, a vanilla U-Net can minimize its loss by copying the seed and ignoring unobserved voxels. Three choices counter this. *(i)* A scalar *coverage channel*, broadcast to every voxel, reports what fraction of the seed survives inside the predicted box. *(ii)* Aggressive *seed dropout* erases observed voxels at training time, so the network must learn to complete sparser evidence. *(iii)* A *completion-masked* BCE–Dice loss concentrates supervision on unobserved voxels.

We make the following contributions:

- We define *evidence-grounded amodal occupancy completion* as a single-view task and protocol: recover the scan-observed geometry of every input-supported object, score in-frustum occlusion and out-of-frustum completion separately, and predict nothing for unsupported objects.
- We propose Seed2Scene, a feed-forward pipeline that lifts an RGB-D frame into a feature-decorated seed, regresses a scene-fitted bounding box, and completes the volume with a 3D U-Net trained under a completion-masked loss and aggressive seed dropout.

On 3RScan (Wald et al., 2019), Seed2Scene raises the F-score on unseen voxels from 0.064, for the strongest reconstructive, generative, or occupancy baseline, to 0.266; frustum-split and volume-matched controls attribute the gain to completion rather than surface inflation, and the model transfers zero-shot to ScanNet (Dai et al., 2017).

2 Related Work

Visible-surface 3D from sparse and single views. DUSt3R (Wang et al., 2024), MASt3R (Leroy et al., 2024), and VGGT (Wang et al., 2025b) regress dense pointmaps from image pairs or short sequences, and successors (Yang et al., 2025; Tang et al., 2025; Jang et al., 2025; Wang & Agapito, 2025; Wang et al., 2025c; Keetha et al., 2026) scale this to many views, online operation, and metric reconstruction. Single-image depth predictors (Ranftl et al., 2021; Bhat et al., 2021; 2023; Yang et al., 2024; Ke et al., 2024) give dense per-pixel depth, and per-pixel primitive predictors – MINE (Li

et al., 2021), Splatter Image (Szymanowicz et al., 2024), Flash3D (Szymanowicz et al., 2025a), Bolt3D (Szymanowicz et al., 2025b), and Liao et al. (2025) – predict one or several Gaussians per pixel, the last two with extra depth layers for shallow disocclusion. All model the visible surface only: their outputs are 2.5D, with no commitment to occluded geometry or to anything outside the frustum. Seed2Scene instead reasons in a voxel grid fitted to the observed objects that extends past the visible region and is supervised on geometry behind the seed.

Monocular semantic scene completion. SSC (Song et al., 2017; Dai et al., 2020) jointly predicts occupancy and semantics from depth or partial scans. Outdoor monocular SSC has matured rapidly (Cao & De Charette, 2022; Yao et al., 2023; Li et al., 2023; Jiang et al., 2024; Yu et al., 2024b; Xi et al., 2026); indoor monocular SSC is sparser, with ISO (Yu et al., 2024a) introducing the Occ-ScanNet benchmark, GA-MonoSSC (Li et al., 2025) and DepthSSC (Yao et al., 2025) improving on it, MonoMRN (Wang et al., 2025d) and AdaSFormer (Wang et al., 2026) refining indoor predictions on NYUv2 and Occ-ScanNet, and EmbodiedOcc (Wu et al., 2025) addressing the streaming setting. A self-supervised line predicts volumetric density fields (Wimbauer et al., 2023; Han et al., 2024; Li et al., 2024; Hayler et al., 2024; Jevtić et al., 2025; Wulff et al., 2025), mostly outdoors. Closest in spirit is VisHall3D (Lu et al., 2025), which splits outdoor SSC into a visible-reconstruction stage and an invisible-hallucination stage. We collapse this into a *single-stage* model in which the visible/invisible distinction enters as a scalar coverage channel rather than a staged architecture. Their benchmarks, NYUv2 and Occ-ScanNet, score semantic occupancy in a fixed 4.8 m box in front of the camera. Such a box clips objects that extend beyond it and still scores objects without a single supporting pixel. Our class-agnostic target is instead the union of the *observed instances*, wherever they extend, so we evaluate ISO under our protocol.

Hidden-surface and scene-volume completion from RGB-D. ScanComplete (Dai et al., 2018) and SG-NN (Dai et al., 2020) complete fused depth scans. OctMAE (Iwase et al., 2024) predicts complete object geometry from one RGB-D frame with an octree U-Net and masked-autoencoder pretraining. Behind-the-Veil (Sun et al., 2024) reconstructs occluded indoor surfaces from a depth sequence. 3DFIRES (Jin et al., 2024) recovers hidden surfaces from one or few images via a directed ray distance function. SceneSense (Reed et al., 2024) synthesizes occupancy from partial RGB-D with a diffusion U-Net that preserves free and occupied space. SCube (Ren et al., 2024) builds a voxel-aligned scene from three or more images. We differ in three respects: a *single* RGB-D frame as input, a target volume that extends *beyond the frustum*, and a deterministic 3D U-Net conditioned on an explicit coverage signal, with neither diffusion nor scene-specific octree fitting.

Holistic single-image 3D scene reconstruction. Total3D-style holistic methods (Zhang et al., 2021; Liu et al., 2022; Gkioxari et al., 2022; Yan et al., 2023), panoptic 3D reconstruction (Dahnert et al., 2021; Chu et al., 2023), Gen3DSR (Ardelean et al., 2025), MIDI (Huang et al., 2025b), SceneGen (Meng et al., 2026), CUPID (Huang et al., 2025a), PixARMesh (Zhang et al., 2026), SceneReGen (Tian et al., 2026), and Coherent 3D Scene Diffusion (Dahnert et al., 2024) recover full scenes by retrieving, generating, or regressing per-object meshes. All commit to object-level abstractions (CAD priors, semantic categories, parametric layouts) that struggle with the long tail of real furniture. PixARMesh and SceneReGen generate these meshes from RGB and report quantitative results only on synthetic scenes. Our Seed2Scene predicts class-agnostic, geometry-only occupancy, trained and evaluated on real scans.

Image-conditioned 3D scene generation. VistaDream (Wang et al., 2025a), WonderWorld (Yu et al., 2025), Marble (Labs, 2025), WorldExplorer (Schneider et al., 2025), Wonderland (Liang et al., 2025), and RealmDreamer (Shriram et al., 2025) build full rooms by inpainting novel views and lifting them into 3D, often through diffusion-based multi-view or video synthesizers (Sargent et al., 2024; Tang et al., 2023; Liu et al., 2024b; Sun et al., 2025; Liu et al., 2024a; Ren et al., 2025; Bahmani et al., 2025) and a radiance representation (Mildenhall et al., 2020; Kerbl et al., 2023); object-level generators (Hong et al., 2024; Liu et al., 2023; Xiang et al., 2025) do the same at object scale. All are trained for perceptual realism rather than metric agreement, and so produce *plausible* rooms rather than *faithful* reconstructions.

3 Proposed Method

Problem. Given a single RGB image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$ together with a metric depth map $\mathbf{D} \in \mathbb{R}_{>0}^{H\times W}$ – from a sensor or from a monocular predictor (Bhat et al., 2023; Yang et al., 2024; Ke et al., 2024) – and the camera intrinsics $\mathbf{K}$, our goal is to predict a metric volumetric occupancy of every object the

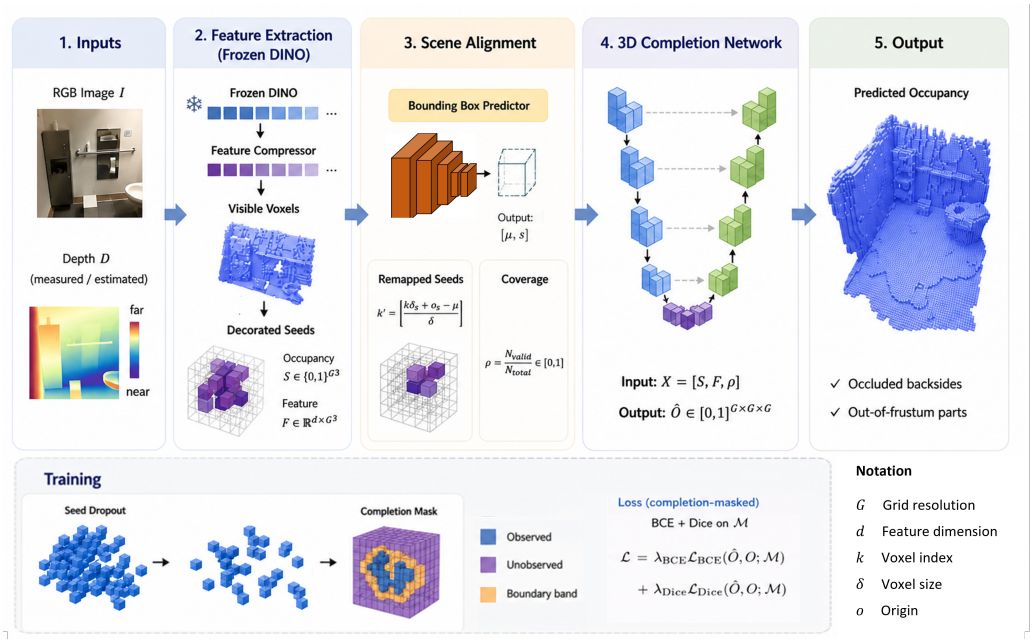

Figure 2: **Overview of Seed2Scene.** *(Left)* RGB image and depth map are unprojected to a point cloud and tagged with DINOv2 features, giving the feature-decorated *visible seed*. *(Scene alignment)* A bounding-box head regresses the scene-fitted target box $\mathbf{B} = (\boldsymbol{\mu}, s)$; the seed is remapped into its voxel grid (Eq. 4), and the retained fraction ρ records how much of the evidence survives. *(Completion)* A 3D U-Net consumes the concatenation $[\mathbf{S}; \mathbf{F}; \rho\mathbf{1}]$ of seed occupancy, per-voxel DINO features, and the broadcast coverage channel, and outputs the dense occupancy $\hat{\mathbf{O}}$, which covers occluded backs and out-of-frustum parts of partly visible objects while staying anchored to the input.

camera observed at a fixed resolution G:

$$\mathcal{F}(\mathbf{I}, \mathbf{D}, \mathbf{K}) \;\mapsto\; (\,\mathbf{B},\, \hat{\mathbf{O}}\,), \qquad \hat{\mathbf{O}} \in [0,1]^{G \times G \times G}, \tag{1}$$

where $\mathbf{B} = (\boldsymbol{\mu}, s) \in \mathbb{R}^3 \times \mathbb{R}_{>0}$ is the scene-fitted bounding box (translation $\boldsymbol{\mu}$ and isotropic scale s) inside which $\hat{\mathbf{O}}$ lives. The box keeps the camera's axes: it is fitted to the extent of the observed objects, not rotated into a world or gravity frame. For each object with at least one supporting pixel, the output covers both occluded surfaces inside the camera frustum and parts that lie outside it. The target is bounded by the dataset reconstruction: all geometry of those instances observed from any view of the complete scan. We neither supervise nor claim recovery of surfaces that no scan view observed. Walls and floors are instances too: a partly visible floor makes its whole scanned extent a target, so completing it extrapolates the room layout, which the input constrains less than an object's shape. Within that bound, the model must commit beyond the input surface while remaining *anchored* to the RGB-D evidence and inventing no unsupported object.

Overview. Fig. 2 visualizes the pipeline. We lift the input into a sparse 3D *seed* tagged with frozen DINOv2 (Oquab et al., 2024) features (Sec. 3.1), predict the scene-fitted box $\mathbf{B}$, remap the seed into it, and complete it into the dense occupancy $\hat{\mathbf{O}}$ with a 3D U-Net conditioned on the features and a scalar *coverage* (Sec. 3.2). A completion-masked loss and aggressive *seed dropout* (Sec. 3.3) focus learning on unobserved voxels and discourage copying of the seed.

3.1 Feature-Lifted Occupancy Seed

Depth back-projection. Each pixel $\mathbf{u} = (u, v)$ with depth $\mathbf{D}(\mathbf{u})$ is unprojected into 3D using $\mathbf{x}(\mathbf{u}) = \mathbf{D}(\mathbf{u})\,\mathbf{K}^{-1}\,[u,\, v,\, 1]^\top$, yielding a point cloud $\mathcal{P} = \{\mathbf{x}(\mathbf{u})\}$ in the camera frame.

Frozen DINO feature tagging. We pass $\mathbf{I}$ through a frozen DINOv2 ViT-L/14 backbone (Oquab et al., 2024), similar to Di Lorenzo et al. (2025), and bilinearly upsample the patch features to image resolution, giving a 1024D descriptor $\mathbf{f}(\mathbf{u}) \in \mathbb{R}^{1024}$ per pixel. The high dimensionality is unnecessary downstream and incompatible with the memory budget of a dense voxel grid, so we compress each

descriptor with a two-layer MLP, ℓ_2-normalized at the input and `LayerNorm`ed at the output:

$$\tilde{\mathbf{f}}(\mathbf{u}) \;=\; \mathrm{LN}\big(\mathbf{W}_2\,\mathrm{SiLU}(\mathbf{W}_1\,\mathbf{f}(\mathbf{u})/\|\mathbf{f}(\mathbf{u})\|_2)\big) \;\in\; \mathbb{R}^d, \qquad d = 64. \tag{2}$$

The ℓ_2-normalization preconditions the input distribution and `LayerNorm` stabilizes the output across scenes; we found both essential for cross-scene generalization.

Voxelization. The decorated point cloud is voxelized on a camera-aligned grid into a sparse 3D seed, which Sec. 3.2 remaps into the predicted scene-fitted box $\mathbf{B}$. Concretely, we let

$$\mathbf{S} \in \{0,1\}^{G^3} \quad \text{(binary occupancy)}, \qquad \mathbf{F} \in \mathbb{R}^{d\times G^3} \quad \text{(per-voxel feature volume)},$$

where $\mathbf{F}$ is obtained by mean-pooling $\tilde{\mathbf{f}}(\mathbf{u})$ over points falling in the same voxel and is zero at unoccupied voxels. The pair $(\mathbf{S}, \mathbf{F})$ is the *evidence seed*: it specifies what the completion stage must remain consistent with.

3.2 Coverage-Conditioned Structure Completion

Two design choices set the completion stage apart from prior monocular semantic scene completion: a *scene-fitted* target volume and a *coverage channel* that exposes the evidence boundary.

Scene-fitted bounding box. The target box $\mathbf{B} = (\boldsymbol{\mu}, s)$ is the tight bounding box around the geometry of all objects the camera observed (Sec. 3.3). A camera-aligned cube would waste capacity behind the camera and along the optical axis where the seed has no support, while clipping the very instances we are trying to complete; predicting $\mathbf{B}$ explicitly avoids this. We regress $\mathbf{B}$ from the seed with a lightweight 3D CNN $\mathcal{F}_{\mathrm{bbox}}$ that operates on the raw seed occupancy and pooled features:

$$(\boldsymbol{\mu}, s) \;=\; \mathcal{F}_{\mathrm{bbox}}(\mathbf{S},\, \mathbf{F}). \tag{3}$$

The seed is then *remapped* from its camera-aligned voxel grid $(\mathbf{o}_s, \delta_s)$ into the canonical grid defined by $\mathbf{B}$, with voxel size $\delta = s/G$ and min corner $\mathbf{b} = \boldsymbol{\mu} - \frac{s}{2}\mathbf{1}$. A seed voxel at index $\mathbf{k}$ maps to canonical index

$$\mathbf{k}' \;=\; \left\lfloor \frac{\mathbf{k}\,\delta_s + \mathbf{o}_s - \mathbf{b}}{\delta} \right\rfloor, \tag{4}$$

and remapped voxels falling outside the canonical grid are *discarded*. This keeps the completion network committed to a specific target volume rather than extrapolating arbitrarily far. $\mathcal{F}_{\mathrm{bbox}}$ is a five-layer stride-2 3D CNN over the seed occupancy followed by global pooling and a two-layer MLP that also receives the log seed extent and the mean-pooled DINO descriptor; it regresses per-axis log-scale ratios and normalized center offsets relative to the seed box and is trained with a Smooth-L1 loss under seed dropout (Appendix A.1). The head anchors the box to per-view evidence rather than dataset statistics (Sec. 4.2); our in-domain completion claim does not depend on it.

Coverage channel. Remapping necessarily discards a fraction of the seed: when the predicted box is smaller than the seed extent, points outside the box are dropped; when it is larger, the seed sparsely populates the canonical grid. The retained ratio

$$\rho \;=\; \frac{N_{\mathrm{valid}}}{N_{\mathrm{total}}} \;\in\; [0,1] \tag{5}$$

summarizes how much of the evidence supports the prediction. We expose ρ to the network as an additional input channel, constant across voxels, allowing the model to attenuate or amplify its reliance on the seed. The full input to the completion U-Net is the channel-wise concatenation

$$\mathbf{X}_{\mathrm{in}} \;=\; \big[\,\mathbf{S}\,;\;\mathbf{F}\,;\;\rho\,\mathbf{1}\,\big] \;\in\; \mathbb{R}^{(1+d+1)\times G^3}. \tag{6}$$

Variable ρ thus encodes the visible/invisible boundary as a continuous quantity rather than architecturally, as in staged approaches (Lu et al., 2025).

3D U-Net. We process $\mathbf{X}_{\mathrm{in}}$ with a five-level 3D U-Net of base width $C = 32$ (channel widths $\{32, 64, 128, 256, 512\}$). Each level is a DoubleConv block of two $3\times3\times3$ convolutions, each followed by InstanceNorm and ReLU; downsampling is $2\times2\times2$ max-pooling, upsampling is transposed convolution, and skip connections concatenate encoder features into the decoder. A final $1\times1\times1$ convolution produces an occupancy logit volume $\hat{\mathbf{L}} \in \mathbb{R}^{G^3}$ from which $\hat{\mathbf{O}} = \sigma(\hat{\mathbf{L}})$. Instance normalization (rather than batch normalization) decouples per-sample feature statistics across scenes and proved important for zero-shot transfer across datasets.

Table 1: **Single-view amodal occupancy completion on 3RScan (Wald et al., 2019) (in-domain).** F_{seen} is computed on voxels supported by the input depth; F_{unseen} on voxels with no support in the seed, and is the main number for amodal completion. *Seed dilation* thickens the observed seed until it occupies the same volume as our prediction; it is a control, not a method. Best result in **bold**.

Method	Prec. ↑	Rec. ↑	F ↑	F_{seen} ↑	F_{unseen} ↑
Flash3D (Szymanowicz et al., 2025a)	0.074	0.122	0.083	0.689	0.016
Nova3R (Chen et al., 2026)	0.073	0.096	0.074	0.523	0.022
MIDI (Huang et al., 2025b)	0.013	0.049	0.018	0.009	0.011
VistaDream (Wang et al., 2025a)	0.021	0.020	0.018	0.151	0.014
SceneSense (Reed et al., 2024)	**0.336**	0.120	0.105	0.650	0.064
ISO (Yu et al., 2024a)	0.051	0.046	0.042	0.256	0.002
PixARMesh (Zhang et al., 2026)	0.055	0.123	0.076	0.431	0.052
Seed2Scene	0.174	**0.580**	**0.268**	**0.773**	**0.266**
Seed dilation (volume-matched control)	–	–	–	–	*0.089*

3.3 TRAINING

We train on indoor-scene datasets that pair fused multi-view reconstructions and per-instance annotations with many calibrated RGB-D frames (Wald et al., 2019; Dai et al., 2017). For each query view $(\mathbf{I}, \mathbf{D}, \mathbf{K})$ we identify the instances with at least one pixel projecting into it, extract *exactly* those instances from the fused scan, and voxelize the recovered geometry at resolution G inside its tight bounding box. The resulting binary occupancy $y \in \{0,1\}^{G^3}$ and box supervise $\hat{\mathbf{O}}$ and $\mathbf{B}$. Thus "complete" means complete relative to what the multi-view scan captured, not an unknowable watertight object: the target includes geometry hidden from the query view but seen elsewhere in the scan, and excludes both unsupported instances and surfaces unseen by every sensor view.

A 3D U-Net trained naively on this input can reach a low loss by copying $\mathbf{S}$: the seed is fed directly, while the unobserved voxels are sparse and easy to ignore. Two mechanisms counter this solution.

Completion-masked supervision. We restrict the loss to unobserved voxels and a thin boundary band via a soft *completion mask*

$$\mathbf{M} \;=\; \mathrm{clip}\Big((1-\mathbf{S}) + 0.3\,\big[\mathrm{MaxPool}_3(\mathbf{S}) - \mathbf{S}\big]^{+},\; 0,\; 1 \Big), \tag{7}$$

where $\mathrm{MaxPool}_3$ is a unit-stride $3\times3\times3$ max-pool. The first term covers everything outside the seed; the second adds a 30% bonus on voxels adjacent to the seed surface so the prediction connects smoothly to the observed geometry. Under $\mathbf{M}$, we combine class-balanced binary cross-entropy with a soft Dice loss (Milletari et al., 2016):

$$\mathcal{L}_{\mathrm{BCE}} = -\frac{1}{\|\mathbf{M}\|_1} \sum_{\mathbf{k}} \mathbf{M}[\mathbf{k}] \Big[\, w_+\, y[\mathbf{k}] \log \hat{o}[\mathbf{k}] + (1 - y[\mathbf{k}]) \log(1 - \hat{o}[\mathbf{k}]) \,\Big], \tag{8}$$

$$\mathcal{L}_{\mathrm{Dice}} = 1 - \frac{2 \sum_{\mathbf{k}} \mathbf{M}[\mathbf{k}]\, \hat{o}[\mathbf{k}]\, y[\mathbf{k}] + \varepsilon}{\sum_{\mathbf{k}} \mathbf{M}[\mathbf{k}] \big(\hat{o}[\mathbf{k}] + y[\mathbf{k}]\big) + \varepsilon}, \tag{9}$$

with $w_+ = \min(N_-/N_+,\, 20)$, where N_+ and N_- count the occupied and empty target voxels, to stabilize training on sparse volumes ($\varepsilon = 10^{-6}$); the total objective is $\mathcal{L} = \mathcal{L}_{\mathrm{BCE}} + \mathcal{L}_{\mathrm{Dice}}$. The mask alone, however, only *de-emphasizes* copying.

Seed dropout. To discourage copying further, we zero each occupied seed voxel and its feature with probability $p_{\mathrm{drop}} = 0.4$ at training time, and recompute ρ on the dropped-out seed so the network is exposed to varying levels of evidence sparsity. At inference, no dropout is applied.

4 EXPERIMENTS

Datasets. We train on **3RScan** (Wald et al., 2019), providing 1,335 RGB-D scans of 432 indoor rooms with per-instance annotations and reconstructed meshes; we follow the standard split (1,178 scans / 385 rooms for training; 157 scans / 47 rooms for evaluation). Each scan provides ∼200 calibrated RGB-D frames, giving thousands of query views with full 3D supervision. To test generalization, we also evaluate *zero-shot* – with no fine-tuning – on **ScanNet** (Dai et al., 2017).

Table 2: **Zero-shot generalization to ScanNet (Dai et al., 2017).** The 3RScan-trained model is evaluated *without fine-tuning*; columns as in Tab. 1. Seed2Scene keeps its lead on F_{unseen} by a margin comparable to the in-domain setting.

Method	Prec. ↑	Rec. ↑	F ↑	F_{seen} ↑	F_{unseen} ↑
Flash3D	0.142	0.156	0.141	0.780	0.025
Nova3R	0.096	0.052	0.063	0.328	0.012
MIDI	0.025	0.024	0.023	0.015	0.016
VistaDream	0.025	0.014	0.017	0.094	0.012
SceneSense	**0.209**	0.153	0.167	0.665	0.087
ISO	0.052	0.039	0.041	0.201	0.002
PixARMesh	0.067	0.142	0.091	0.464	0.051
Seed2Scene	0.109	**0.661**	**0.187**	**0.824**	**0.162**

Table 3: **Unseen voxels split by the input frustum** (3RScan). IF: in-frustum but depth-unobserved voxels (occlusion); OF: out-of-frustum voxels. Nova3R is the only baseline that targets in-frustum completion.

	IF (occluded)			OF (out of frustum)		
Method	Prec.	Rec.	F	Prec.	Rec.	F
Nova3R	0.010	0.005	0.006	0.015	0.032	0.024
Seed2Scene	**0.183**	**0.771**	**0.296**	**0.175**	**0.441**	**0.251**

Query views and ground truth. For each test scene, we sample three frames at random. For each query frame we identify the instances with at least one pixel in the view, extract their geometry from the fused multi-view mesh, and voxelize it at $G{=}128$ inside its tight metric bounding box (Sec. 3.3); the model must recover *exactly these instances*. The ground truth contains surfaces occluded in the query whenever another scan view captured them, but no geometry unseen by every sensor view; we claim nothing beyond this coverage bound.

Metrics. We report *bounding-box IoU* for $\mathbf{B}$ and voxel precision, recall, and F-score at an untuned 0.5 threshold. $\mathbf{F}_{\text{seen}}$ is computed on voxels supported by input depth and $\mathbf{F}_{\text{unseen}}$ on voxels with no seed support. To separate two distinct completion regimes, we further split unseen voxels into *in-frustum but depth-unobserved* (IF: occlusion) and *out-of-frustum* (OF: cropped continuation). Appendix B reports threshold sensitivity and further decompositions.

Baselines. We compare with seven methods from three families. *Reconstructive:* **Flash3D** (Szymanowicz et al., 2025a) predicts layered Gaussians per pixel, which we voxelize on the same $G{=}128$ grid with its 32-pixel boundary padding, and **Nova3R** (Chen et al., 2026) is a recent feed-forward amodal reconstructor. *Generative:* **MIDI** (Huang et al., 2025b) is a state-of-the-art image-conditioned scene generator, **PixARMesh** (Zhang et al., 2026) autoregressively generates a mesh for each segmented object, and **VistaDream** (Wang et al., 2025a) hallucinates an entire room from one view. *Occupancy:* **SceneSense** (Reed et al., 2024) predicts indoor occupancy from RGB-D, and **ISO** (Yu et al., 2024a), the reference method of Occ-ScanNet, predicts semantic occupancy in a fixed 4.8 m box in front of the camera; it is trained on Occ-ScanNet, which is built from ScanNet. Flash3D, Nova3R, and SceneSense receive the same sensor depth as Seed2Scene; VistaDream uses its native UniDepth front-end. Point and Gaussian outputs are voxelized on the evaluation grid, MIDI meshes are surface-sampled first, and SceneSense keeps its native 10-cm grid because resampling hurts it. All methods are scored against the same target occupancy (Appendix A.3). Nova3R targets in-frustum completion and the generative methods optimize other objectives, so they mark task boundaries rather than fail at their own tasks.

Implementation. We use $G = 128$, $d = 64$ compressed DINOv2 features, seed-dropout probability $p_{\text{drop}} = 0.4$, AdamW, and InstanceNorm throughout the U-Net; Appendix A.2 gives the three-stage training schedule. Training runs on a single A100. At inference, a single forward pass returns $\mathbf{B}$ and $\hat{\mathbf{O}}$ in 0.4 seconds on average.

4.1 Single-View Amodal Occupancy Completion

Voxel-level amodal completion. Tab. 1 reports reconstruction quality on 3RScan. Seed2Scene reaches $F_{unseen}{=}\ 0.266$, versus 0.064 for the strongest baseline. Reconstructive methods retain visible-surface quality (Flash3D $F_{seen}{=}\ 0.689$, Nova3R 0.523) but provide little unseen occupancy. ISO likewise recovers visible geometry ($F_{seen}{=}\ 0.256$) but almost no unseen occupancy (0.002). The low scores of MIDI and VistaDream mainly expose metric scale, pose, and instance-set mismatch rather than their native generative quality. PixARMesh aligns its objects far better ($F_{seen}{=}\ 0.463$) yet recovers little unseen occupancy (0.052). Two controls show that the gain is completion, not inflation. Splitting unseen voxels by the frustum (Tab. 3) widens the margin on *in-frustum occluded* voxels, Nova3R's intended regime, and keeps it large out of frustum. Dilating the observed seed on every

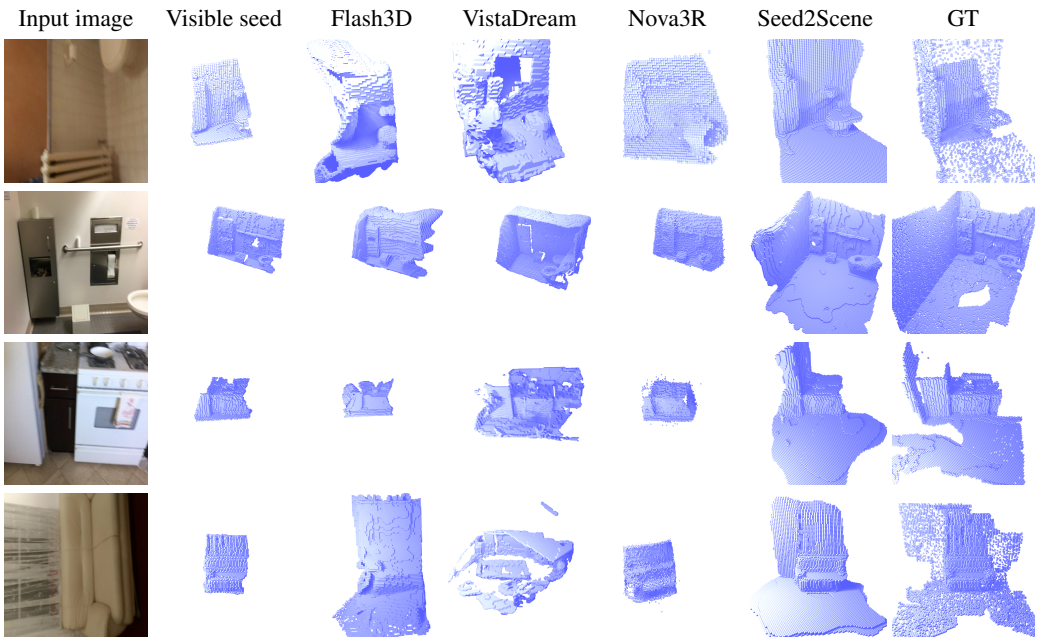

Figure 3: **Qualitative single-view amodal occupancy completion.** From left to right: input image, the lifted feature seed, three single-view baselines (Flash3D (Szymanowicz et al., 2025a), VistaDream (Wang et al., 2025a), Nova3R (Chen et al., 2026)), Seed2Scene, and the ground-truth occupancy of the observed objects.

Table 4: **Ablations on box, depth source, and losses.** *(a) Box localization*: IoU with the ground-truth box and scale ratio $s_{\text{pred}}/s_{\text{gt}}$ (1 is perfect), in-domain on 3RScan and zero-shot on ScanNet. 3RScan mean is the 3RScan training-set mean box, applied unchanged on ScanNet; the italic ScanNet mean is fitted on the target data. The 3RScan mean box collapses across datasets, whereas the learned head transfers. *(b) Depth source* (ScanNet, zero-shot): monocular depth lowers F_{unseen}, yet Seed2Scene with UniDepth (Piccinelli et al., 2024) still beats every baseline in Tab. 2. *(c) Loss components* (3RScan): removing the Dice term $\mathcal{L}_{\text{Dice}}$ or the completion mask $\mathbf{M}$ lowers F_{unseen}, Dice most.

(a) Box localization

	3RScan		ScanNet (0-shot)	
	IoU↑	Ratio	IoU↑	Ratio
3RScan mean	0.401	1.792	0.139	2.242
Predicted	**0.444**	**1.531**	**0.412**	**2.172**
ScanNet mean	–	–	*0.427*	*2.144*

(b) Depth source (ScanNet)

	F↑	F_{seen}↑	F_{unseen}↑
Metric3D v2	0.100	0.752	0.083
UniDepth v2	0.153	0.759	0.132
Sensor depth (GT)	*0.187*	*0.824*	*0.162*

(c) Loss & Mask (3RScan)

$\mathcal{L}_{\text{Dice}}$	M	F↑	F_{seen}↑	F_{unseen}↑
✗	✓	0.198	**0.822**	0.196
✓	✗	0.259	0.725	0.257
✓	✓	**0.268**	0.773	**0.266**

frame until it occupies the same volume as our prediction reaches only $F_{\text{unseen}} = 0.089$ (last row of Tab. 1), so the improvement cannot be explained by thickening or extending the visible surface.

Zero-shot generalization. We apply Seed2Scene trained on 3RScan, with *no* fine-tuning, to ScanNet (Tab. 2). On randomly sampled rather than curated frames, it reaches $F_{\text{unseen}} = 0.162$ versus the strongest baseline's 0.087, and its box head transfers nearly intact (Sec. 4.2).

Qualitative comparison. Fig. 3 compares four scenes. Flash3D yields a thin shell behind the visible pixels, VistaDream plausible furniture in the wrong places, and Nova3R a surface that stops at the frustum boundary. Seed2Scene recovers the occluded backs of partially visible objects *and* their parts outside the frustum, while staying anchored to the seed.

4.2 Ablations

The box head removes dataset dependence. In-domain, a constant training-set mean box is a strong prior: with it, the network still reaches $F_{\text{unseen}} = 0.254$ versus 0.266 with the predicted box (app. Tab. 6), so the completion result does not rest on learned localization. Tab. 4a shows that this strength is dataset-specific regularity. On ScanNet, whose capture hardware and trajectories differ, the IoU

Table 5: **Ablations on input channels and grid resolution.** *(a) Input channels* (3RScan): removing either the coverage channel ρ or the DINO feature volume $\mathbf{F}$ degrades $\mathrm{F}_{\text{unseen}}$. *(b) Grid resolution* G (3RScan): all methods degrade as the grid gets finer, but Seed2Scene's lead over Nova3R (Chen et al., 2026) and SceneSense (Reed et al., 2024) *widens*; at $G{=}128$ (default) both baselines fall below $\mathrm{F}_{\text{unseen}}{=}\,0.025$ while Seed2Scene retains 0.266.

(a) Input channels (3RScan)

ρ	$\mathbf{F}$	F↑	F_{seen}↑	$\mathrm{F}_{\text{unseen}}$↑
✗	✓	0.259	0.733	0.257
✓	✗	0.246	0.693	0.244
✓	✓	**0.268**	**0.773**	**0.266**

(b) Grid resolution G (3RScan)

	$G=32$		$G=64$		$G=128$	
	F↑	$\mathrm{F}_{\text{unseen}}$↑	F↑	$\mathrm{F}_{\text{unseen}}$↑	F↑	$\mathrm{F}_{\text{unseen}}$↑
Nova3R	0.224	0.102	0.139	0.050	0.074	0.022
SceneSense	0.120	0.088	0.026	0.024	0.010	0.005
Seed2Scene	**0.505**	**0.502**	**0.405**	**0.402**	**0.268**	**0.266**

of the 3RScan mean box collapses from 0.401 to 0.139, whereas the 3RScan-trained head transfers nearly intact ($0.444{\to}0.412$) and matches a mean box fitted on ScanNet itself (0.427) without seeing ScanNet data. It transfers since it regresses offsets relative to the observed seed's box, never absolute coordinates, and so anchors to per-view evidence rather than a memorized layout. In-domain, it also guards the tail: when the mean box retains $< 1\%$ of the seed (12 views), prediction raises $\mathrm{F}_{\text{unseen}}$ by 0.257 (app. Tab. 9). The oracle box reaches 0.378, so localization remains the main headroom.

Sensor vs. predicted depth. Tab. 4b replaces the sensor depth used at test time with two off-the-shelf monocular predictors. Predicted depth is noisier, and as expected $\mathrm{F}_{\text{unseen}}$ drops – but Seed2Scene with UniDepth still outperforms *every* baseline in Tab. 2, including SceneSense fed with sensor depth.

Loss components. Tab. 4c disables $\mathcal{L}_{\text{Dice}}$ and the completion mask $\mathbf{M}$ in turn. Removing $\mathcal{L}_{\text{Dice}}$ has the largest single-knob effect ($0.266{\to}0.196$ on $\mathrm{F}_{\text{unseen}}$); removing $\mathbf{M}$, which adds the easily copied seed voxels to the loss, also costs $\sim$3 relative percent. Both terms are necessary.

Input channels. Tab. 5a disables each of the two non-trivial U-Net input channels in turn. Removing the coverage channel ρ has only a modest effect ($0.266{\to}0.257$ on $\mathrm{F}_{\text{unseen}}$). Removing the DINO feature volume $\mathbf{F}$ (leaving only the binary seed and ρ) is more harmful: $\mathrm{F}_{\text{unseen}}$ drops to 0.244 and F_{seen} from 0.773 to 0.693. DINO context thus lets the U-Net extrapolate *consistently* rather than expand the seed noisily.

Grid resolution. Tab. 5b sweeps the resolution from $G{=}32$ to 128 (default). All methods degrade as the grid gets finer, but at different rates. Going from $G{=}32$ to 128, Nova3R retains 22% of its $\mathrm{F}_{\text{unseen}}$ ($0.102{\to}0.022$), SceneSense less than 6% ($0.088{\to}0.005$), while Seed2Scene retains 53% ($0.502{\to}0.266$). Our lead over the strongest baseline thus grows from $\sim5\times$ at $G{=}32$ to $\sim12\times$ at $G{=}128$: Seed2Scene scales *better* than the baselines.

Robustness and over-completion. Appendix B stratifies visibility, injects depth noise, sweeps thresholds, and measures false positives. Per-frame $\mathrm{F}_{\text{unseen}}$ stays within 0.191–0.232 across visibility quintiles, falls only to 0.187 under 5-cm depth noise, and F varies smoothly with the threshold. Of false-positive voxels, 72.5% lie within three voxels of the ground truth and only 0.51% of their volume is detached from it, so errors are local boundary over-completion, not unsupported objects.

5 Conclusion

We introduced Seed2Scene for evidence-grounded single-view amodal 3D completion. It recovers scan-observed occupancy of input-supported objects, including occluded and out-of-frustum geometry, using a scene-fitted volume, completion-masked supervision, and seed dropout. Seed2Scene substantially improves $\mathrm{F}_{\text{unseen}}$ on 3RScan and transfers zero-shot to ScanNet. Frustum splits, matched-volume dilation, and robustness analyses show that the gain is learned completion rather than surface inflation. The code and trained models will be made public.

Limitations. The dense 128^3 grid limits surface detail, the model is geometry-only and single-view, and fused-scan supervision cannot cover surfaces unseen by every sensor. Sparse decoders and multi-view fusion are natural extensions. Fusing three views into the seed already improves completion without retraining (Appendix B). Predicted scene bounds remain the largest numerical bottleneck.

AI USE STATEMENT

We used a large language model to polish the text of the manuscript and to help find related work. It did not contribute to the task definition, the method, the implementation, or the experiments. The authors verified all text and references and take full responsibility for the content.

ETHICS STATEMENT

All authors have read and adhere to the ICLR Code of Ethics. This work uses only the public 3RScan and ScanNet datasets, under their terms of use, and involves no new data collection or human subjects. Like other scene-reconstruction methods, single-view completion could help infer the layout of a private space from one image. Our output is a coarse, geometry-only occupancy of objects the camera already observed, which limits but does not remove this risk. Predicted occupancy can be wrong, especially outside the frustum, so safety-critical uses such as robot navigation should treat it as uncertain rather than measured.

REPRODUCIBILITY STATEMENT

Sec. 4 names the datasets, the standard 3RScan split, and the zero-shot ScanNet protocol, and Sec. 3.3 together with Appendix A.3 describes how query views and ground-truth occupancy are constructed, including the three-frames-per-scene sampling and how baseline outputs are scored on the evaluation grid. Secs. 3.1–3.3 specify the seed construction, the network, the loss, the completion mask, and the seed-dropout rate; Appendices A.1 and A.2 give the bounding-box parameterization and the three-stage training schedule with batch sizes, learning rates, and iteration counts. Code, trained models, and evaluation scripts will be released.

REFERENCES

Andreea Ardelean, Mert Özer, and Bernhard Egger. Gen3DSR: Generalizable 3D scene reconstruction via divide and conquer from a single view. In *International Conference on 3D Vision (3DV)*, pp. 616–626, 2025.

Sherwin Bahmani, Tianchang Shen, Jiawei Ren, Jiahui Huang, Yifeng Jiang, Haithem Turki, Andrea Tagliasacchi, David B Lindell, Zan Gojcic, Sanja Fidler, et al. Lyra: Generative 3D scene reconstruction via video diffusion model self-distillation. *arXiv preprint arXiv:2509.19296*, 2025.

Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. AdaBins: Depth estimation using adaptive bins. In *CVPR*, pp. 4009–4018, 2021.

Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. ZoeDepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023.

Anh-Quan Cao and Raoul De Charette. MonoScene: Monocular 3D semantic scene completion. In *CVPR*, pp. 3991–4001, 2022.

Weirong Chen, Chuanxia Zheng, Ganlin Zhang, Andrea Vedaldi, and Daniel Cremers. NOVA3R: Non-pixel-aligned visual transformer for amodal 3D reconstruction. In *ICLR*, 2026.

Tao Chu, Pan Zhang, Qiong Liu, and Jiaqi Wang. BUOL: A bottom-up framework with occupancy-aware lifting for panoptic 3D scene reconstruction from a single image. In *CVPR*, pp. 4937–4946, 2023.

Manuel Dahnert, Ji Hou, Matthias Nießner, and Angela Dai. Panoptic 3D scene reconstruction from a single RGB image. *NeurIPS*, 34:8282–8293, 2021.

Manuel Dahnert, Angela Dai, Norman Müller, and Matthias Nießner. Coherent 3D scene diffusion from a single RGB image. In *NeurIPS*, 2024.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *CVPR*, pp. 5828–5839, 2017.

Angela Dai, Daniel Ritchie, Martin Bokeloh, Scott Reed, Jürgen Sturm, and Matthias Nießner. ScanComplete: Large-scale scene completion and semantic segmentation for 3D scans. In *CVPR*, 2018.

Angela Dai, Christian Diller, and Matthias Nießner. SG-NN: Sparse generative neural networks for self-supervised scene completion of RGB-D scans. In *CVPR*, pp. 849–858, 2020.

Gaia Di Lorenzo, Federico Tombari, Marc Pollefeys, and Daniel Barath. Object-X: Learning to reconstruct multi-modal 3D object representations. In *NeurIPS*, 2025.

Georgia Gkioxari, Nikhila Ravi, and Justin Johnson. Learning 3D object shape and layout without 3D supervision. In *CVPR*, pp. 1695–1704, 2022.

Keonhee Han, Dominik Muhle, Felix Wimbauer, and Daniel Cremers. Boosting self-supervision for single-view scene completion via knowledge distillation. In *CVPR*, 2024.

Adrian Hayler, Felix Wimbauer, Dominik Muhle, Christian Rupprecht, and Daniel Cremers. S4C: Self-supervised semantic scene completion with neural fields. In *International Conference on 3D Vision (3DV)*, 2024.

Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. LRM: Large reconstruction model for single image to 3D. In *ICLR*, 2024.

Binbin Huang, Haobin Duan, Yiqun Zhao, Zibo Zhao, Yi Ma, and Shenghua Gao. CUPID: Generative 3D reconstruction via joint object and pose modeling. *arXiv preprint arXiv:2510.20776*, 2025a.

Zehuan Huang, Yuan-Chen Guo, Xingqiao An, Yunhan Yang, Yangguang Li, Zi-Xin Zou, Ding Liang, Xihui Liu, Yan-Pei Cao, and Lu Sheng. MIDI: Multi-instance diffusion for single image to 3D scene generation. In *CVPR*, pp. 23646–23657, 2025b.

Shun Iwase, Katherine Liu, Vitor Guizilini, Adrien Gaidon, Kris Kitani, Rares Ambrus, and Sergey Zakharov. Zero-shot multi-object scene completion. In *ECCV*, 2024.

Wonbong Jang, Philippe Weinzaepfel, Vincent Leroy, Lourdes Agapito, and Jerome Revaud. Pow3R: Empowering unconstrained 3D reconstruction with camera and scene priors. In *CVPR*, 2025.

Aleksandar Jevtić, Christoph Reich, Felix Wimbauer, Oliver Hahn, Christian Rupprecht, Stefan Roth, and Daniel Cremers. Feed-forward SceneDINO for unsupervised semantic scene completion. In *ICCV*, 2025.

Haoyi Jiang, Tianheng Cheng, Naiyu Gao, Haoyang Zhang, Tianwei Lin, Wenyu Liu, and Xinggang Wang. Symphonize 3D semantic scene completion with contextual instance queries. In *CVPR*, 2024.

Linyi Jin, Nilesh Kulkarni, and David Fouhey. 3DFIRES: Few image 3D reconstruction for scenes with hidden surfaces. In *CVPR*, 2024.

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *CVPR*, pp. 9492–9502, 2024.

Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, et al. MapAnything: Universal feed-forward metric 3D reconstruction. In *International Conference on 3D Vision (3DV)*, 2026.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian splatting for real-time radiance field rendering. *ACM TOG*, 42(4):139:1–139:14, 2023.

World Labs. Generating bigger and better worlds. `https://www.worldlabs.ai/blog/bigger-better-worlds`, 2025. Accessed 2025-10-15.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3D with MASt3R. In *ECCV*, 2024.

Jiaxin Li, Zijian Feng, Qi She, Henghui Ding, Changhu Wang, and Gim Hee Lee. MINE: Towards continuous depth MPI with NeRF for novel view synthesis. In *ICCV*, pp. 12578–12588, 2021.

Rui Li, Tobias Fischer, Mattia Segu, Marc Pollefeys, Luc Van Gool, and Federico Tombari. Know your neighbors: Improving single-view reconstruction via spatial vision-language reasoning. In *CVPR*, 2024.

Shijie Li, Zhongyao Cheng, Rong Li, Shuai Li, Juergen Gall, Xun Xu, and Xulei Yang. Global-aware monocular semantic scene completion with state space models. In *ICCV*, 2025.

Yiming Li, Zhiding Yu, Christopher Choy, Chaowei Xiao, Jose M. Alvarez, Sanja Fidler, Chen Feng, and Anima Anandkumar. VoxFormer: Sparse voxel transformer for camera-based 3D semantic scene completion. In *CVPR*, 2023.

Hanwen Liang, Junli Cao, Vidit Goel, Guocheng Qian, Sergei Korolev, Demetri Terzopoulos, Konstantinos N. Plataniotis, Sergey Tulyakov, and Jian Ren. Wonderland: Navigating 3D scenes from a single image. In *CVPR*, 2025.

Ziwei Liao, Mohamed Sayed, Steven L. Waslander, Sara Vicente, Daniyar Turmukhambetov, and Michael Firman. Complete Gaussian splats from a single image with denoising diffusion models. *arXiv preprint arXiv:2508.21542*, 2025.

Fangfu Liu, Wenqiang Sun, Hanyang Wang, Yikai Wang, Haowen Sun, Junliang Ye, Jun Zhang, and Yueqi Duan. ReconX: Reconstruct any scene from sparse views with video diffusion model. *arXiv preprint arXiv:2408.16767*, 2024a.

Haolin Liu, Yujian Zheng, Guanying Chen, Shuguang Cui, and Xiaoguang Han. Towards high-fidelity single-view holistic reconstruction of indoor scenes. In *ECCV*, pp. 429–446, 2022.

Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3D mesh in 45 seconds without per-shape optimization. *NeurIPS*, 36:22226–22246, 2023.

Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. SyncDreamer: Generating multiview-consistent images from a single-view image. In *ICLR*, 2024b.

Haoang Lu, Yuanqi Su, Xiaoning Zhang, Longjun Gao, Yu Xue, and Le Wang. VisHall3D: Monocular semantic scene completion from reconstructing the visible regions to hallucinating the invisible regions. In *ICCV*, 2025.

Yanxu Meng, Haoning Wu, Ya Zhang, and Weidi Xie. SceneGen: Single-image 3D scene generation in one feedforward pass. In *International Conference on 3D Vision (3DV)*, 2026.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *International Conference on 3D Vision (3DV)*, pp. 565–571, 2016.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. UniDepth: Universal monocular metric depth estimation. In *CVPR*, 2024.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, pp. 12179–12188, 2021.

Alec Reed, Brendan Crowe, Doncey Albin, Lorin Achey, Bradley Hayes, and Christoffer Heckman. SceneSense: Diffusion models for 3D occupancy synthesis from partial observation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024.

Xuanchi Ren, Yifan Lu, Hanxue Liang, Zhangjie Wu, Huan Ling, Mike Chen, Sanja Fidler, Francis Williams, and Jiahui Huang. SCube: Instant large-scale scene reconstruction using VoxSplats. In *NeurIPS*, 2024.

Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. GEN3C: 3D-informed world-consistent video generation with precise camera control. In *CVPR*, pp. 6121–6132, 2025.

Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, et al. ZeroNVS: Zero-shot 360-degree view synthesis from a single image. In *CVPR*, pp. 9420–9429, 2024.

Manuel-Andreas Schneider, Lukas Höllein, and Matthias Nießner. WorldExplorer: Towards generating fully navigable 3D scenes. In *SIGGRAPH Asia Conference Papers*, 2025.

Jaidev Shriram, Alex Trevithick, Lingjie Liu, and Ravi Ramamoorthi. RealmDreamer: Text-driven 3D scene generation with inpainting and depth diffusion. In *International Conference on 3D Vision (3DV)*, 2025.

Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *CVPR*, pp. 1746–1754, 2017.

Su Sun, Cheng Zhao, Yuliang Guo, Ruoyu Wang, Xinyu Huang, Yingjie Victor Chen, and Liu Ren. Behind the veil: Enhanced indoor 3D scene reconstruction with occluded surfaces completion. In *CVPR*, 2024.

Wenqiang Sun, Shuo Chen, Fangfu Liu, Zilong Chen, Yueqi Duan, Jun Zhang, and Yikai Wang. DimensionX: Create any 3D and 4D scenes from a single image with controllable video diffusion. In *ICCV*, 2025.

Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3D reconstruction. In *CVPR*, pp. 10208–10217, 2024.

Stanislaw Szymanowicz, Eldar Insafutdinov, Chuanxia Zheng, Dylan Campbell, Joao F Henriques, Christian Rupprecht, and Andrea Vedaldi. Flash3D: Feed-forward generalisable 3D scene reconstruction from a single image. In *International Conference on 3D Vision (3DV)*, pp. 670–681, 2025a.

Stanislaw Szymanowicz, Jason Y. Zhang, Pratul Srinivasan, Ruiqi Gao, Arthur Brussee, Aleksander Holynski, Ricardo Martin-Brualla, Jonathan T. Barron, and Philipp Henzler. Bolt3D: Generating 3D scenes in seconds. In *ICCV*, 2025b.

Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. MVDiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. In *NeurIPS*, 2023.

Zhenggang Tang, Yuchen Fan, Dilin Wang, Hongyu Xu, Rakesh Ranjan, Alexander Schwing, and Zhicheng Yan. MV-DUSt3R+: Single-stage scene reconstruction from sparse views in 2 seconds. In *CVPR*, 2025.

Zefan Tian, Yuteng Ye, Yiheng Zhang, Yuhang Yang, Xueqiang Lv, Shizhou Zhang, Le Liu, and Di Xu. SceneReGen: Generative reconstruction of 3D scenes from a single image. *arXiv preprint arXiv:2608.23930*, 2026.

Johanna Wald, Armen Avetisyan, Nassir Navab, Federico Tombari, and Matthias Nießner. RIO: 3D object instance re-localization in changing indoor environments. In *ICCV*, pp. 7658–7667, 2019.

Haiping Wang, Yuan Liu, Ziwei Liu, Wenping Wang, Zhen Dong, and Bisheng Yang. VistaDream: Sampling multiview consistent images for single-view scene reconstruction. In *ICCV*, 2025a.

Hengyi Wang and Lourdes Agapito. 3D reconstruction with spatial memory. In *International Conference on 3D Vision (3DV)*, 2025.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *CVPR*, pp. 5294–5306, 2025b.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3D perception model with persistent state. In *CVPR*, 2025c.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D vision made easy. In *CVPR*, pp. 20697–20709, 2024.

Xuzhi Wang, Xinran Wu, Song Wang, Lingdong Kong, and Ziping Zhao. Monocular semantic scene completion via masked recurrent networks. In *ICCV*, 2025d.

Xuzhi Wang, Xinran Wu, Song Wang, Lingdong Kong, and Ziping Zhao. AdaSFormer: Adaptive serialized transformers for monocular semantic scene completion from indoor environments. In *CVPR*, 2026.

Felix Wimbauer, Nan Yang, Christian Rupprecht, and Daniel Cremers. Behind the scenes: Density fields for single view reconstruction. In *CVPR*, 2023.

Yuqi Wu, Wenzhao Zheng, Sicheng Zuo, Yuanhui Huang, Jie Zhou, and Jiwen Lu. EmbodiedOcc: Embodied 3D occupancy prediction for vision-based online scene understanding. In *ICCV*, 2025.

Philipp Wulff, Felix Wimbauer, Dominik Muhle, and Daniel Cremers. Dream-to-recon: Monocular 3D reconstruction with diffusion-depth distillation from single images. In *ICCV*, 2025.

Zichen Xi, Hao-Xiang Chen, Nan Xue, Hongyu Yan, Qi-Yuan Feng, Levent Burak Kara, Joaquim Jorge, and Qun-Ce Xu. FlowSSC: Universal generative monocular semantic scene completion via one-step latent diffusion. *arXiv preprint arXiv:2601.15250*, 2026.

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3D latents for scalable and versatile 3D generation. In *CVPR*, pp. 21469–21480, 2025.

Kai Yan, Fujun Luan, Miloš Hašan, Thibault Groueix, Valentin Deschaintre, and Shuang Zhao. PSDR-Room: Single photo to scene using differentiable rendering. In *SIGGRAPH Asia 2023 Conference Papers*, pp. 1–11, 2023.

Jianing Yang, Alexander Sax, Kevin J. Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3R: Towards 3D reconstruction of 1000+ images in one forward pass. In *CVPR*, 2025.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, pp. 10371–10381, 2024.

Jiawei Yao, Chuming Li, Keqiang Sun, Yingjie Cai, Hao Li, Wanli Ouyang, and Hongsheng Li. NDC-Scene: Boost monocular 3D semantic scene completion in normalized device coordinates space. In *ICCV*, pp. 9421–9431, 2023.

Jiawei Yao, Jusheng Zhang, Xiaochao Pan, Tong Wu, and Canran Xiao. Depthssc: Monocular 3d semantic scene completion via depth-spatial alignment and voxel adaptation. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2154–2163, 2025.

Hong-Xing Yu, Haoyi Duan, Charles Herrmann, William T Freeman, and Jiajun Wu. WonderWorld: Interactive 3D scene generation from a single image. In *CVPR*, pp. 5916–5926, 2025.

Hongxiao Yu, Yuqi Wang, Yuntao Chen, and Zhaoxiang Zhang. Monocular occupancy prediction for scalable indoor scenes. In *ECCV*, 2024a.

Zhu Yu, Runmin Zhang, Jiacheng Ying, Junchen Yu, Xiaohai Hu, Lun Luo, Si-Yuan Cao, and Hui-Liang Shen. Context and geometry aware voxel transformer for semantic scene completion. In *NeurIPS*, 2024b.

Cheng Zhang, Zhaopeng Cui, Yinda Zhang, Bing Zeng, Marc Pollefeys, and Shuaicheng Liu. Holistic 3D scene understanding from a single image with implicit representation. In *CVPR*, pp. 8833–8842, 2021.

Xiang Zhang, Sohyun Yoo, Hongrui Wu, Chuan Li, Jianwen Xie, and Zhuowen Tu. PixARMesh: Autoregressive mesh-native single-view scene reconstruction. In *CVPR*, 2026.

A Implementation Details

A.1 Bounding-box regressor

The bounding-box head regresses the transformation from seed space to the full ground-truth bounding box. It takes as input the seed occupancy grid $\mathbf{S} \in \{0,1\}^{G^3}$ voxelized in seed-normalized space, the log seed half-extent $\log \bar{\sigma}_{\text{seed}}$ as a scale signal, and mean-pooled DINOv2 features as a semantic size prior. A five-layer stride-2 3D CNN (channel progression $1 \to c \to 2c \to 4c \to 8c \to 16c$, each layer followed by InstanceNorm and ReLU) reduces the 128^3 grid to 4^3, followed by global average pooling. The pooled feature, the scale scalar, and a projected 32-D DINO descriptor enter a two-layer MLP that outputs per-axis log-scale ratios $\Delta \log \boldsymbol{\sigma} = \log(\boldsymbol{\sigma}_{\text{gt}}/\boldsymbol{\sigma}_{\text{seed}})$ and normalized center offsets $\Delta\boldsymbol{\mu} = (\boldsymbol{\mu}_{\text{gt}} - \boldsymbol{\mu}_{\text{seed}})/\boldsymbol{\sigma}_{\text{seed}}$. The zero-initialized output layer starts as the identity transform. At inference, $\boldsymbol{\sigma}_{\text{gt}} = \boldsymbol{\sigma}_{\text{seed}} \odot \exp(\Delta \log \boldsymbol{\sigma})$ and $\boldsymbol{\mu}_{\text{gt}} = \boldsymbol{\mu}_{\text{seed}} + \Delta\boldsymbol{\mu} \odot \boldsymbol{\sigma}_{\text{seed}}$. We use a Smooth-L1 loss on both targets and seed dropout (p=0.4). Tab. 4a reports the localization accuracy of this predictor in-domain and zero-shot, and Tab. 6 reports completion quality on 3RScan with the predicted, the mean, and the oracle box.

Table 6: **Completion quality by box source** on 3RScan, at a fixed completion network. IoU is measured against the ground-truth box; the scale ratio is $s_{\text{pred}}/s_{\text{gt}}$, with 1 being perfect. The oracle uses the ground-truth box.

Box source	IoU ↑	Scale ratio	F ↑	F_{seen} ↑	F_{unseen} ↑
Training-set mean	0.401	1.792	0.256	0.776	0.254
Predicted	**0.444**	**1.531**	0.268	0.773	0.266
Oracle (GT box)	–	–	*0.380*	*0.796*	*0.378*

A.2 Training schedule

Training has three stages: structure-completion learning, bounding-box training, and structure-completion fine-tuning. All stages run on a single A100 GPU with 40 GB of memory and use AdamW with gradient clipping at 0.5; each stage runs for 50k iterations. The completion U-Net is first trained with the ground-truth bounding box for the remapping of seed voxel indices, with batch size 8 and learning rate 4×10^{-4}. The bounding-box head is then trained with batch size 32 and learning rate 10^{-3}. Finally, the completion network is fine-tuned with the frozen bounding-box predictor, with batch size 8 and learning rate 10^{-4}.

A.3 Evaluation protocol

For each dataset (3RScan, ScanNet) we randomly select three frames from every test scene. For point-cloud or Gaussian outputs (Flash3D (Szymanowicz et al., 2025a), Nova3R (Chen et al., 2026), VistaDream (Wang et al., 2025a)) we voxelize the points at the default 128^3 evaluation resolution. For MIDI (Huang et al., 2025b) we uniformly sample 80,000 mesh-surface points before the same voxelization. All methods are scored against the same target occupancy, and a cell counts as occupied if any predicted point, mesh sample, or voxel falls inside it. Surface sampling can leave cells of large MIDI meshes empty and so lower its recall. Rasterizing the triangles would mark a superset of the sampled cells; even if that doubled MIDI's true positives, its per-frame F would at most double, to $F_{\text{unseen}} \leq 0.022$ on 3RScan and ≤ 0.032 on ScanNet. SceneSense (Reed et al., 2024) remains at its native 10-cm grid in Tabs. 1 and 2, because resampling degrades it. Tab. 5b sweeps the evaluation resolution for all methods.

B Additional Analyses

Volume-matched dilation control. As an architecture-independent control for surface inflation, we dilate the observed seed on every frame until it occupies the same volume as our prediction. This volume-matched dilation obtains $F_{\text{unseen}} = 0.089$ versus 0.266 for Seed2Scene (Tab. 1). The frustum

decomposition in Tab. 3 complements it: the margin over Nova3R widens on in-frustum occluded voxels, Nova3R's intended regime, and stays large out of frustum.

Visibility stratification. Tab. 7 groups the test set by the fraction of target geometry visible in the input. F_{unseen} stays in a narrow 0.191–0.232 band even when less than 4.4% of the target is visible. Nova3R stays between 0.019 and 0.025 in all bins, so additional visible evidence does not turn surface extrapolation into dense completion.

Table 7: **Performance by visible fraction** (equal-count bins; per-frame averages within each bin).

Visible fraction	Precision	Recall	F	F_{unseen}	Nova3R F_{unseen}
$[0.000, 0.044)$	0.167	0.323	0.201	0.191	0.020
$[0.044, 0.072)$	0.202	0.382	0.252	0.231	0.022
$[0.072, 0.097)$	0.199	0.428	0.261	0.232	0.025
$[0.097, 0.129)$	0.197	0.446	0.265	0.226	0.024
$[0.129, 0.272]$	0.186	0.510	0.262	0.214	0.019

Depth noise and operating point. Gaussian depth perturbations degrade performance gracefully (Tab. 8a): at 5 cm, $F_{unseen} = 0.187$ remains above every baseline's clean-input value. The occupancy-threshold sweep in Tab. 8b uses a random evaluation subset, so its $t = 0.5$ values differ from the full-set headline. The trade-off is smooth, and the paper's $t = 0.5$ is untuned.

Table 8: **Robustness analyses.** (a) Gaussian depth noise. (b) Occupancy-threshold sweep on a random evaluation subset.

(a) Depth noise

σ (cm)	P	R	F	F_{unseen}
0	0.174	0.580	0.268	0.266
1	0.165	0.589	0.261	0.261
2	0.145	0.594	0.248	0.246
5	0.110	0.532	0.218	0.187

(b) Threshold sweep

t	P	R	F
0.1	0.10	0.57	0.17
0.3	0.15	0.46	0.23
0.5	0.18	0.42	0.25
0.7	0.20	0.38	0.26
0.9	0.27	0.29	0.28

False-positive geometry. False positives concentrate near the target surface: 34.1%, 59.0%, and 72.5% lie within one, two, and three voxels of the ground truth. Detached false-positive components occur 0.50 times per frame on average and zero times at the median, and contain only 0.51% of the false-positive volume. Lower precision therefore reflects mostly over-thickened boundaries and connected extensions rather than unsupported free-standing objects. Because the fused-scan ground truth is incomplete, some counted false positives may be real geometry absent from the target scan.

Where bounding-box prediction matters. Tab. 9 stratifies the gain over a constant training-set mean box by the seed evidence the mean box retains. The predictor behaves as a tail-safety mechanism rather than a source of in-domain average-case gains; its cross-dataset benefit appears in Tab. 4a.

Table 9: **Gain of the predicted box over the constant mean box**, stratified by the seed evidence retained by the mean box. Values are per-sample macro averages; the two small bins are explanatory rather than statistically tight.

Seed retained by mean box	Samples	ΔF	ΔF_{unseen}
$< 1\%$	12	$+0.300$	$+0.257$
1–50%	13	$+0.096$	$+0.076$
$> 50\%$	417	$+0.018$	$+0.009$

Multi-view compatibility. We fuse the back-projected seeds of a reference frame and two nearby frames and run the unmodified single-view model. This experiment uses a camera-conditioned variant

of the box head, which scores F$= 0.280$ and $F_{unseen} = 0.274$ on single views over the full test set, compared with 0.268 and 0.266 for the head used elsewhere. With the seen/unseen split fixed to the reference view, fusion raises F to 0.301 and F_{unseen} to 0.292 without fine-tuning. This validates seed fusion as a direct extension, while a trained sparse multi-view model remains future work.

Representation and alternative pipelines. The dense grid suits metric occupancy queries but has cubic memory growth and loses fine surface detail. Sparse octrees, hash grids, or a completion-then-refinement mesh or point decoder are compatible upgrades. Likewise, the convolutional 3D U-Net is deliberately plain so that gains can be attributed to the task and the supervision; sparse Transformer decoders are a promising orthogonal direction. An instance-segmentation plus geometry-conditioned generator is complementary, but must solve per-instance metric scale and pose, remain registered to the observed depth, and compose objects and structural surfaces into one coherent scene. MIDI and PixARMesh approximate the object-generative route. MIDI's low F_{seen} points to metric registration as one bottleneck. PixARMesh registers its objects far better ($F_{seen} = 0.431$) yet still recovers little unseen occupancy (0.052), partly because it omits walls and floors.

C Additional Qualitative Results

Figs. 4 and 5 show further predictions on the 3RScan and ScanNet test sets.

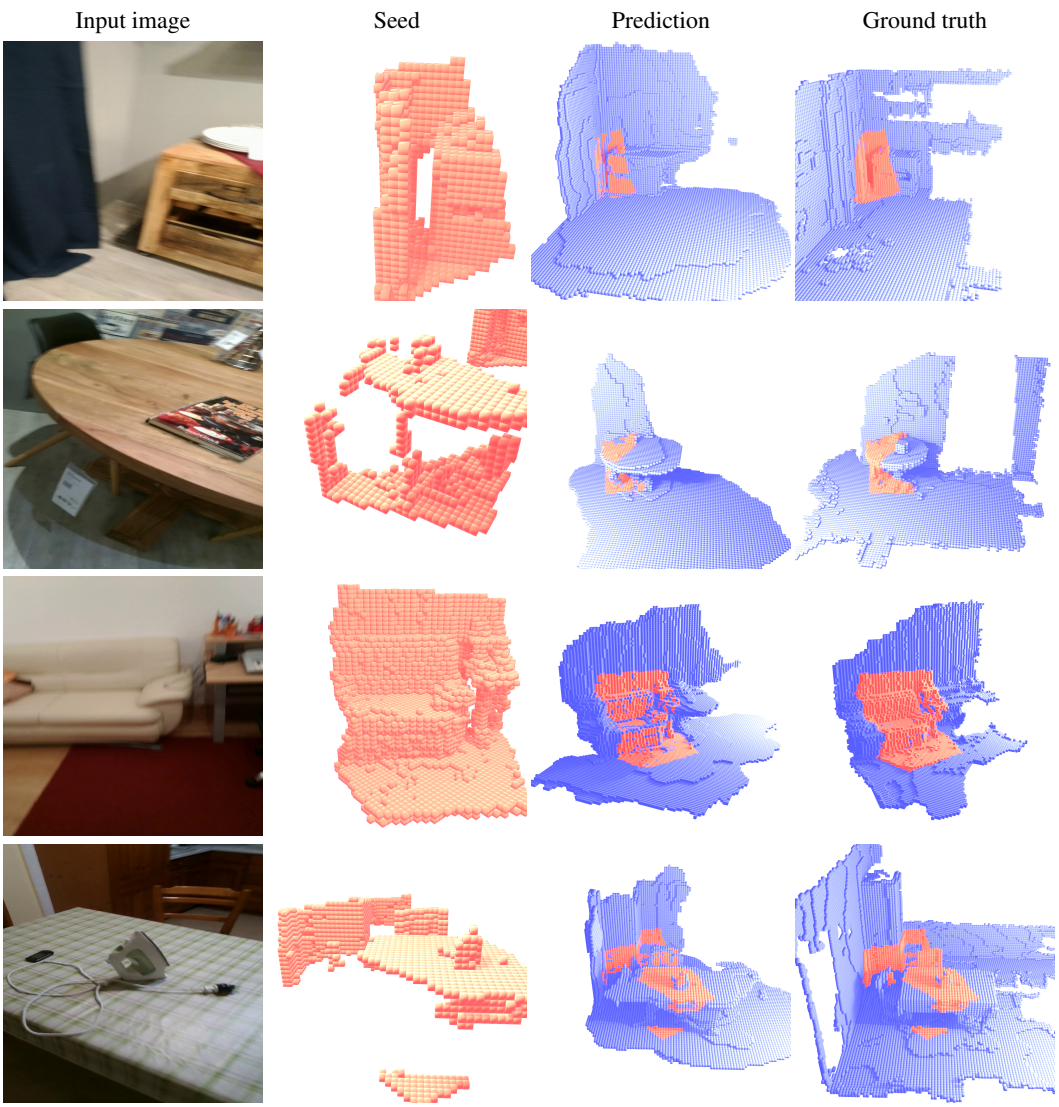

Figure 4: **Additional qualitative results on 3RScan.** From left to right: input image, the lifted feature seed, our prediction, and the ground-truth occupancy of the observed objects. Seed is colorized differently for clearer visualization.

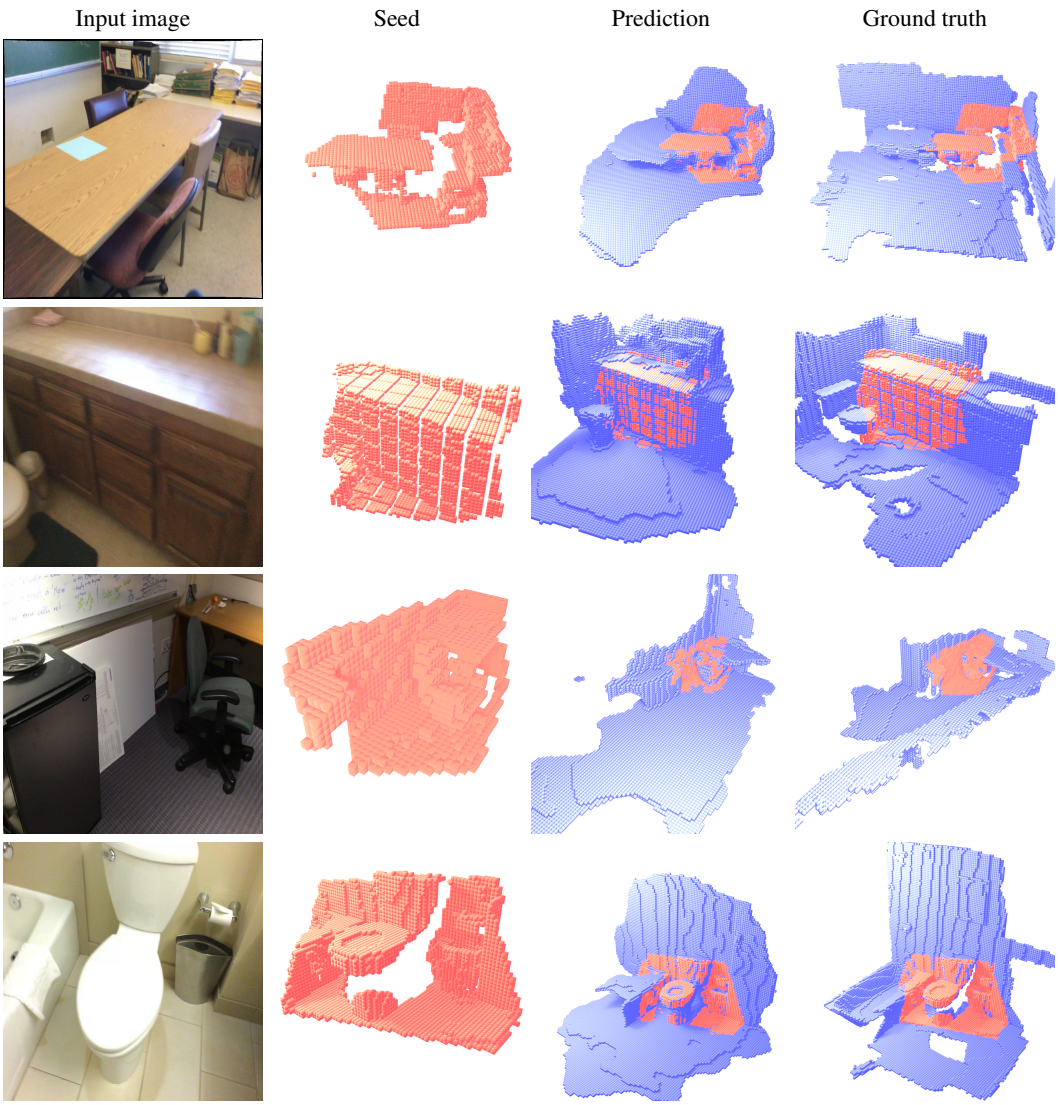

Figure 5: **Additional qualitative results on ScanNet.** From left to right: input image, the lifted feature seed, our prediction, and the ground-truth occupancy of the observed objects. Seed is colorized differently for clearer visualization.