

Fast Generative DeOcclusion for Vision and Robotics

Jieneng Chen* Tiezheng Zhang* Xiwei Xuan Ju He Yifan Yin
Haojun Shi Suyu Ye Xinyi Li Ruisheng Yuan Tianmin Shu Alan Yuille
Johns Hopkins University

Abstract

Occlusion remains a core challenge in vision, as projecting a 3D world into 2D inevitably hides much of the scene geometry. We present *FoundationDeOcclusion*, a fast generative framework for fast occlusion recovery that reconstructs hidden geometry and appearance from partial visual observations. *FoundationDeOcclusion* first identifies occluded objects from monocular image sequences using *Grounded-SAM*, then matches them across views using depth cues estimated by 3D reconstruction models. We introduce a novel geometry-aware linear de-occlusion *Transformer (GL-DoT)* that synthesizes the missing regions, which are then integrated into the 3D scene through depth-aware fusion. As a result, *FoundationDeOcclusion* improves 3D scene reconstruction under occlusion. Notably, *GL-DoT* attains strong performance with only four denoising steps and runs at near real-time speed (6 FPS), challenging the prevailing belief that generative models are impractical for time-critical robotic tasks. Finally, we demonstrate the effectiveness of *FoundationDeOcclusion* in real-robot navigation and manipulation. Without bells and whistles, it boosts the prior-art *AnyGrasp* by 43.8% in success rate without task-specific tuning, setting a new state of the art for mobile manipulation under occlusion.

1. Introduction

Recovering the full 3D geometry of real-world scenes from monocular RGB image sequences remains a long-standing challenge in computer vision [1, 56, 57, 61]. Despite advances in monocular reconstruction [37, 64, 65, 67], these models still struggle with occlusion—a fundamental limitation caused by projecting a 3D world onto 2D images, which inevitably hides substantial portions of scene geometry. In realistic settings such as robotics or autonomous navigation [5, 68], full 360° coverage is rarely possible, leading to incomplete reconstructions that hinder visual planning.

As shown in Fig. 1, even the advanced methods like

*Equal technical contribution.

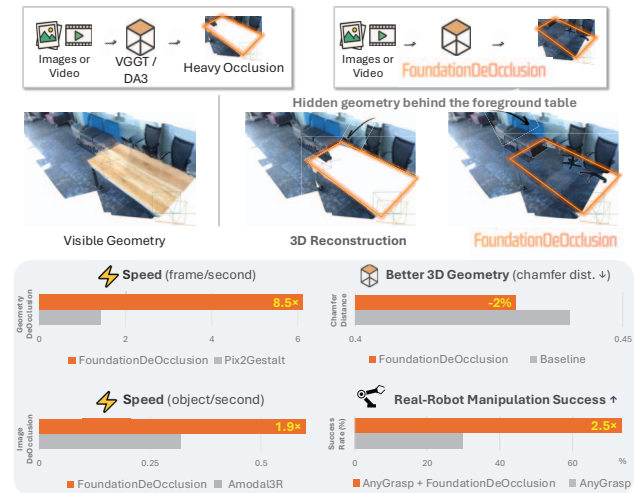


Figure 1. Hidden scene geometry motivates *FoundationDeOcclusion*, which reconstructs occluded regions and restores completeness (top). The comparisons (bottom) show that our method achieves state-of-the-art speed, recovers more accurate geometry than the baseline, and significantly improves real-robot manipulation.

VGGT [64] and Depth Anything 3 [39] can reconstruct visible surfaces accurately but **entirely miss hidden parts**, such as chair legs or wheels under a table, resulting in perceptually incomplete scenes. In contrast, the human visual system naturally performs de-occlusion — inferring occluded structure based on context and prior knowledge [34?]. Equipping robots with this capability is key to robust 3D perception and interaction.

Existing approaches to scene-level de-occlusion or amodal reasoning remain limited. Object-centric methods [22, 30, 48] or heavy point-cloud optimizations do not scale to full scenes, while 2D amodal completion [48] lacks geometric grounding. Moreover, diffusion-based methods [48] are slow, requiring many denoising steps — impractical for time-critical robotic tasks such as navigating to an ambulance or assisting object loading.

We introduce *FoundationDeOcclusion*, a fast, generative framework that augments existing 3D reconstruction

models by recovering occluded geometry and appearance in a training-free manner. Given monocular RGB image sequences, our method (1) identifies occluded regions via Grounding-SAM and matches corresponding instances across views using heuristic depth cues from 3D reconstruction models, and (2) synthesizes occluded regions by leveraging the proposed geometry-aware linear De-Occlusion Transformer (GL-DoT). The recovered regions are then fused back into the scene through depth-aware alignment, yielding perceptually complete reconstructions with spatial coherence.

As a result, FoundationDeOcclusion improves 3D scene reconstruction under occlusion and the proposed GL-DoT achieves near real-time speed (6 FPS) with only four denoising steps. When applied to downstream tasks such as robot navigation and manipulation, it boosts the AnyGrasp baseline by 43.8% in success rate, demonstrating the practical value of fast occlusion recovery.

In summary, our contribution is four-fold:

- We propose FoundationDeOcclusion, a fast generative framework that augments existing 3D reconstruction models to recover occluded geometry and appearance.
- We design a novel geometry-aware linear De-Occlusion Transformer (GL-DoT) that enables fast, generative occlusion recovery with only a few denoising steps.
- We demonstrate strong real-world performance on robot navigation and manipulation, highlighting the potential of generative vision for robotics.

2. Related Work

We briefly review related work in 3D reconstruction, de-occlusion, and robot mobile manipulation.

3D scene reconstruction. Monocular 3D reconstruction has evolved from SfM [1, 56, 57] and SLAM [6, 13, 16] to learning-based paradigms that directly infer depth, pose, and structure from RGB inputs [59, 61]. Recent methods adopt geometry-first formulations with point-cloud-based representations, such as DUST3R [67], CUT3R [65], and MonST3R [85], enabling dense reconstruction and view integration. Scalable systems like MAST3R-SLAM [47] and Transformer-based models like VGGT [64] and follow-ups [89] further improve reconstructions. However, existing methods focus solely on visible surfaces, leaving occluded geometry unaddressed. We bridge this gap by recovering occluded geometry to improve perceptual completeness while preserving global spatial consistency of the scene.

De-occlusion, point-cloud completion, and amodal perception. Recovering structures hidden by occlusion has been explored across 3D completion and 2D amodal reasoning. Point cloud completion methods [49, 73, 77, 81, 82] reconstruct missing geometry from partial scans, but most are object-centric and rely on synthetic supervision or category-specific training. Training-free approaches [22, 30] leverage

diffusion-based test-time optimization for open-set shapes, yet require minutes per object and do not exploit global scene context. In parallel, 2D amodal perception—including inpainting [9, 11, 14, 42, 48, 80, 83], detection [25, 29], and segmentation [31, 52, 58, 84]—infers invisible regions directly in image space but lacks spatial geometric grounding. Depth-aware extensions [26, 36, 38], NERF-based approach [70], and single-/novel-view scene synthesis [15, 18, 23, 55, 78] can improve geometric cues, though often at the cost of runtime. Our method is complementary to these efforts. Integrating de-occlusion directly into VGGT-style scene reconstruction remains largely open, and we focus specifically on achieving fast, scene-level completion suitable for real-world robotic settings — where timely decisions are essential and long diffusion or optimization loops are often impractical.

Robot navigation and manipulation. Mobile manipulation requires coordinating mapping, navigation, and interaction under partial observability. Occlusion is often handled by seeking better viewpoints or delaying action, while belief-driven search [4, 63, 87] selects informative views but can involve long exploration and fails when targets stay hidden. In open-vocabulary settings, many systems assume static voxel or feature-field maps [8, 24, 32], requiring re-mapping as objects move or occlusions change [40, 44, 69, 79]. Dynamic spatio-semantic memories [41, 76] mitigate this but may still spend time re-exploring before updating occluded regions. Occlusion also impacts manipulation: grasp proposals [43, 45, 46] and 6D pose estimation [35, 51, 74] degrade when contact surfaces are hidden, and mechanical search [12] adds extra steps and risk. Over long horizons, uncertainty compounds in task-and-motion planning [19, 28] and vision-language-action systems [2, 33, 60]. In contrast, FoundationDeOcclusion offers fast and cross-view-consistent de-occlusion that (i) increase information gain per step, (ii) update semantic and freespace maps precisely at occluded regions, and (iii) restore graspable, collision-safe surfaces — improving navigation and manipulation in real-world settings.

3. FoundationDeOcclusion

We first revisit reconstruction, formulate the problem of occlusion, and present our FoundationDeOcclusion.

Preliminaries of 3D Scene Reconstruction. Given a sequence of RGB images $\mathcal{I} = (I_v)_{v=1}^N$ of a static scene, a reconstruction model \mathcal{G} (e.g., VGGT and Depth Anything 3) predicts for each view a depth map $D_v \in \mathbb{R}^{H \times W}$, camera intrinsics/extrinsics K_v , and a corresponding point cloud $\mathcal{P}_v \in \mathbb{R}^{3 \times H \times W}$. Formally,

$$\mathcal{G}(\mathcal{I}) = \mathcal{G}((I_v)_{v=1}^N) = (\mathcal{P}_v, D_v, K_v)_{v=1}^N. \quad (1)$$

\mathcal{P} is simply the aggregation of per-view point clouds, pro-

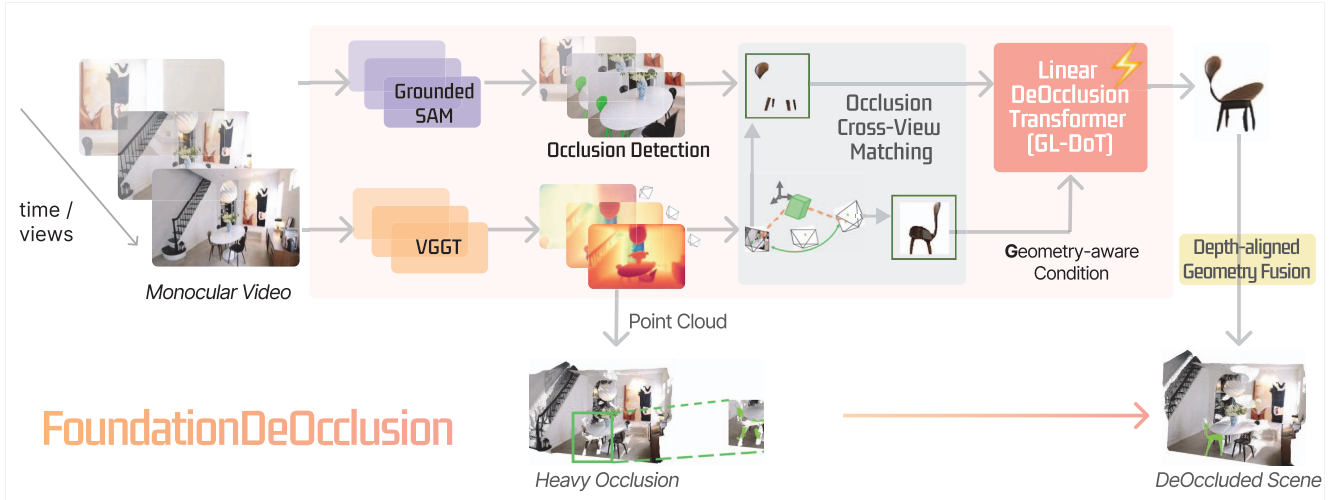


Figure 2. **Overview of the FoundationDeOcclusion framework.** FoundationDeOcclusion first identifies occluded objects from monocular image sequences using Grounded SAM, then matches them across views using heuristic depth cues estimated by 3D reconstruction models (§3.1). We introduce a geometry-aware linear de-occlusion Transformer (GL-DoT, §3.2) that leverages geometry-aware cross-view condition to synthesize the occluded regions, which are then integrated into the 3D scene through depth-aware fusion (§3.3).

viding a deterministic lifting of visible pixels into 3D. While it accurately reconstructs observed surfaces, any geometry hidden in the input views \mathcal{I} is inevitably missing.

Problem formulation. A complete 3D scene $\hat{\mathcal{P}}$ consists of both visible geometry \mathcal{P} reconstructed by \mathcal{G} and the missing, occluded geometry $\mathcal{P}_{occluded}$. Our objective is to estimate $\mathcal{P}_{occluded}$ and fuse it with the visible point cloud \mathcal{P} , yielding a perceptually complete reconstruction $\hat{\mathcal{P}}$, defined by:

$$\hat{\mathcal{P}} = \mathcal{P} \cup \mathcal{P}_{occluded} = \mathcal{G}(\mathcal{I}) \cup \mathcal{G}(f(\mathcal{I})), \quad (2)$$

where $f(\cdot)$ is a system for recovering occluded regions.

As illustrated in Fig. 2, we propose **FoundationDeOcclusion** as an effective framework to tackle occlusion through three modules: (i) Occlusion Detection and Matching (§3.1), (ii) a Geometry-aware Linear De-Occlusion Transformer (§3.2), and (iii) depth-aligned geometry fusion (§3.3). Together, these components provide a fast, generative solution for completing occluded scene geometry.

3.1. Occlusion Detection and Matching

The first module of FoundationDeOcclusion identifies all unique object instances in the scene. A single object instance, such as the chair in Fig. 2, may appear as several fragmented 2D regions. Because each unique object instance may also appear in multiple views, we gather its masks across viewpoints, detect occlusions using depth, and match masks belonging to the same 3D object via spatial distances.

Occlusion detection. We segment each image $I_v \in \mathcal{I}$ using Grounded-SAM [54] for instance masks and RAM [86] for semantic labels, yielding the semantic mask M_v^k for k -th instance at view v . Using the depth map D_v and local mask ad-

jacency, we determine per-view occlusion. When two masks (M_v^i, M_v^j) share a boundary, the one $M_v^o \in \{M_v^i, M_v^j\}$ with the larger depth (*i.e.*, farther from the camera) is marked as occluded for that viewpoint. We assign this mask an occlusion label $L_v^o = 1$, which is later aggregated at the 3D-instance level.

Occlusion cross-view matching. We establish the cross-view association to group 2D masks into unique 3D instances. For each mask, we extract its corresponding 3D geometry from \mathcal{P}_v to compute its centroid in 3D. For “occlusion cross-view matching” in Fig. 2, we measure the Euclidean distance between centroids of masks from different views. Two masks are grouped as the same 3D instance if their distance falls below a threshold τ . We design τ as an adaptive threshold derived from the distribution of all such centroid distances to ensure robust grouping across diverse scenes, and the corresponding ablation can be found in the experiment section. As a result, each 3D instance is associated with its 2D masks across views and a per-mask occlusion status.

3.2. Geometry-Aware Linear DeOcclusion Transformer (GL-DoT)

After detecting unique 3D instances and matching the occlusion across views, this section focuses on completing occluded regions while preserving scene consistency. We build our fast generative model upon the linear DiT architecture [75] (Fig. 3 (a)), whose diffusion Transformer employs linear attention [20, 75], cross-attention for text conditioning, and an FNN for noise prediction. Below, we describe how these components evolve into our design.

Linear DeOcclusion Transformer (L-DoT). We propose

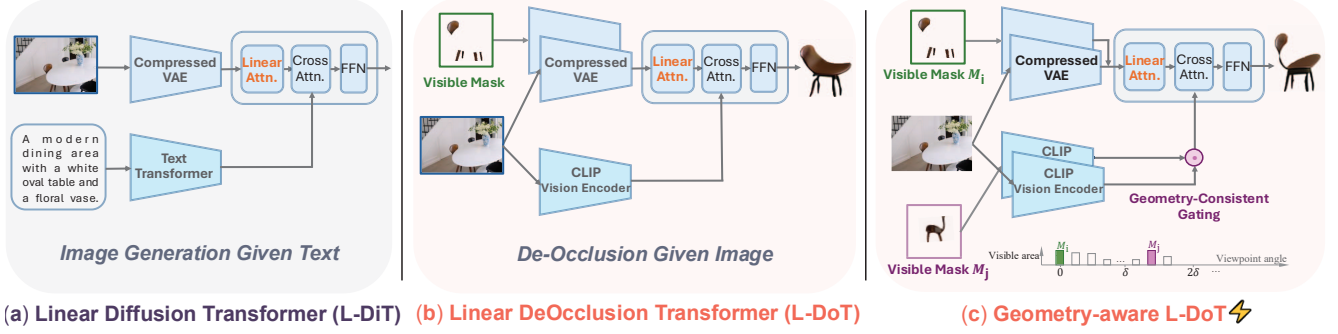


Figure 3. **Design Instantiation from a modern Linear DiT to our Geometry-Consistent Linear DoT.** (a) We start from the linear DiT from SANA [75], which employs linear attention for efficient text-to-image generation. (b) We introduce the Linear DeOcclusion Transformer (L-DoT) to enable efficient object de-occlusion. (c) We further extend this to the Geometry-Consistent Linear DoT (GL-DoT), maximizing information gain across multiple views. Finally, we apply few-step denoising at inference to achieve near real-time de-occlusion.

the novel L-DoT (Fig. 2) for generative de-occlusion. To improve efficiency, we adopt linear attention [75] in our diffusion Transformer, replacing the standard quadratic attention used in DiT [50] and conventional U-Net-based diffusion models like Stable Diffusion 2.1 in Pix2Gestalt [48]. Unlike L-DiT [75], which targets text-to-image generation, we propose the Linear DeOcclusion Transformer (L-DoT), a generator specifically tailored for object-level de-occlusion.

Specifically, let $\mathcal{M} = \{M_v\}_{v=1}^k$ be the selected visible-region masks for an object instance, and let I_v denote the corresponding RGB images. To complete the occluded region in M_v , we synthesize the missing content directly in image space. Let $\mathcal{C}(\cdot)$ be our L-DoT module and $\phi(\cdot)$ the CLIP [53] vision encoder providing conditioning features. The de-occlusion of a single-view is then expressed as

$$\hat{I}_v = \mathcal{C}(M_v | \phi(I_v)), \quad (3)$$

where \hat{I}_v is the completed image at view v . However, relying solely on single-view conditioning overlooks cross-view cues from the same 3D instance, leading to ungrounded or inconsistent completions across viewpoints.

For training L-DoT, we use the synthetic dataset from [48], which provides occluded objects paired with their complete counterparts. Following [75], we adopt rectified flow for noise prediction. During inference, while typical diffusion models (e.g., [48]) degrade sharply when reducing denoising steps, L-DoT sustains high performance with as few as four inference steps. This design delivers strong de-occlusion quality while enabling near-real-time speed.

Geometry-aware Linear DeOcclusion Transformer (GL-DoT). Thus, to enforce consistency across views of the same 3D instance, we extend L-DoT with explicit cross-view conditioning, yielding our *Geometry-aware Linear DeOcclusion Transformer* (GL-DoT), as illustrated in Fig. 3. The idea is simple but effective: L-DoT conditions on the visible regions from other viewpoints to guide completion. Specifically, let

$\mathcal{N}_v \subset \{1, \dots, k\} \setminus \{v\}$ be the indices of the additional conditioning views, and define their aggregated representation:

$$\bar{\phi}_v = \frac{1}{|\mathcal{N}_v|} \sum_{j \in \mathcal{N}_v} \phi(I_j). \quad (4)$$

We then form a convex combination of the current-view and cross-view CLIP vision features:

$$\hat{I}_v = \mathcal{C}(M_v | (1 - \rho) \phi(I_v) + \rho \bar{\phi}_v), \quad (5)$$

where $\rho \in [0, 1]$ balances the contributions from the current view and the cross-view features. This design encourages the synthesized completion to be both semantically plausible and perceptually consistent.

Selecting informative conditioning views is non-trivial. Among all masks of the same instance, we first pick the mask with the largest visible area, then compute pairwise viewpoint angles and retain only the largest mask every δ degrees. For example, with $\delta = 90^\circ$ for a roughly horizontal camera path, at most four representative views are selected. This selection provides broad coverage of 3D surfaces while keeping the computation manageable.

In short, by conditioning on cross-view features, the GL-DoT is able to perform generative de-occlusion with cross-view geometry consistency. The generations are then passed to the next stage for depth-aligned geometric fusion.

3.3. Depth-Aligned Geometry Fusion

For each 3D instance and each viewpoint v , we begin with its visible mask M_v , the de-occluded image \hat{I}_v , and the completed object mask \hat{M}_v . A straightforward strategy is to predict a depth map for the newly synthesized region in $\hat{I}_v \odot \hat{M}_v$ and back-project it into 3D. However, because this de-occluded region is synthesized, its predicted depth \hat{D}_v often has scale or pose inconsistencies, leading to misalignment when fused with the original point cloud \mathcal{P} .

To address this issue, we introduce a simple but effective depth-alignment scheme. The core idea is to use the overlapping visible region as a geometric anchor: by aligning the

predicted depth to the true depth on this shared region, the synthesized portion naturally inherits the correct scale and pose. Let \widehat{D}_v be the depth predicted from the de-occluded image, with $\widehat{D}_v \odot \widehat{M}_v$ representing the completed object (including both visible and generated regions). Let $D_v \odot M_v$ denote the original depth on the visible region. The depth of sampling point $d_v \in D_v \odot M_v$ in the visible region acts as a trusted anchor, and each has a corresponding predicted sample $\widehat{d}_v \in \widehat{D}_v \odot M_v$ at the same 2D position.

We align the two by estimating a per-view scale and shift (α, β) via robust linear regression:

$$(\alpha, \beta) = \arg \min_{\alpha, \beta} \sum_{p \in M_v} \xi \left(d_v(p) - (\alpha \widehat{d}_v(p) + \beta) \right), \quad (6)$$

where $\xi(\cdot)$ is the Huber loss. The estimated α and β are then applied to the full predicted depth map \widehat{D}_v to obtain an aligned depth map. Since this regression only involves pixels in a single object mask and a 1D affine model, it is highly efficient and introduces negligible computational overhead.

Using the aligned depth, we back-project the completed instance and safely fuse it into the original 3D geometry, and repeat this process across views for multi-view geometry fusion. In this way, we improve the completeness of each 3D instance without distorting the global scene structure.

4. Experiments

In this section, we first evaluate the effectiveness of FoundationDeOcclusion in improving 3D reconstruction under occlusion across both indoor and in-the-wild settings (§4.1). We then demonstrate that the de-occluded outputs produced by FoundationDeOcclusion substantially benefit downstream tasks, especially robotic manipulation (§4.2). We explore the data scaling for the thin objects (§4.3) and provide an efficiency analysis (§4.4) to better understand the design choices that enable FoundationDeOcclusion to achieve efficient and effective de-occlusion.

4.1. 3D Reconstruction

Datasets. Evaluating scene de-occlusion requires ground-truth geometry for invisible regions, which is available only in dense scans (while we acknowledge that achieving full 360° coverage is still infeasible). To the best of our knowledge, only a few datasets support this setting: the ScanNet [10] and SceneNN [21]. These datasets provide dense ground-truth point clouds, RGB-D image streams, and per-point instance annotations in both 2D and 3D, enabling precise evaluation of occlusion-aware scene reconstruction. We emphasize that only RGB images are used as input to our method; all additional modalities are used solely for benchmarking. Similar to COCO-A, we construct SceneNN-A, a subset comprising roughly one quarter of SceneNN [21], as

a benchmark for studying occlusion-aware geometry reconstruction.

Evaluation details. We follow standard evaluation protocols for 3D reconstruction and report three metrics: (1) *Chamfer Distance* (CD), and (2) *Completeness*, the mean distance from each ground-truth point to its nearest predicted point. Following prior works [64, 66], we align predictions to the ground-truth point cloud using the Umeyama [62] and ICP [3] algorithms.

Experimental details. FoundationDeOcclusion is a plug-and-play framework that can be seamlessly integrated into any existing 3D reconstruction pipeline that takes only image frames as input. We evaluate it on recently released methods with diverse architectures and input assumptions. To the best of our knowledge, no 3D scene-level occlusion dataset currently exists for evaluation (*i.e.*, there is no 3D scene-level counterpart to COCO-A [88]). We therefore construct an amodal evaluation set by sampling from the testset of ScanNet and SceneNN. During testing, we follow [64] and randomly sample 10 frames per scene.

Results. Quantitative results are reported in Table 1. FoundationDeOcclusion improves reconstructions. The gains are especially pronounced in occluded region, since the introduced synthetic geometry preserves structural integrity without degrading overall reconstructions.

Method	Chamfer Distance ↓	Completeness ↓
VGGT	0.0475	0.0564
+ Ours	0.0471 (+0.9%)	0.0545 (+3.4%)
Depth Anything 3	0.0435	0.0580
+ Ours	0.0427 (+1.8%)	0.0552 (+4.8%)

Table 1. Our FoundationDeOcclusion improves the state-of-the-art baselines VGGT [64] and Depth Anything 3 [39] evaluated on the SceneNN-A.

Qualitative examples of occlusion detection and matching

Our occlusion detection and matching aims at identifying unique 3D instances in the scene and determining whether they are occluded in each input view. The input of this module is the original point cloud constructed by a 3D reconstruction model, and the set of input images capturing the scene. The outputs are unique 3D instances present in the scene, each corresponding to a group of its 2D masks labeled with their occlusion status, *i.e.*, masks annotated as “occluded” indicating they are not fully visible in the current view. We provide visualizations of the outputs of this module in Figs. 4. These examples illustrate how our method correctly merges instance masks across different views and accurately detects occlusion.

Qualitative results of 3D reconstruction. Figure 5 shows qualitative comparisons on indoor benchmarks and in-the-wild scenes. While baseline methods reliably reconstruct

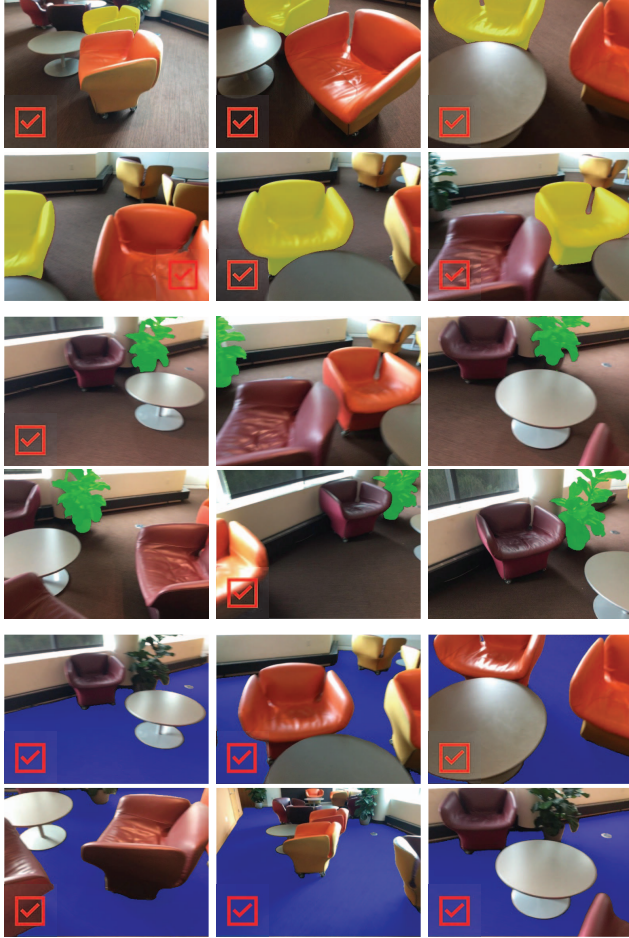


Figure 4. **Qualitative examples of occlusion detection and matching** from a single scene. For each unique 3D instance, we have grouped its corresponding 2D masks, with occluded ones highlighted by a red check mark.

visible surfaces, they consistently fail to recover geometry in occluded regions—an issue that becomes evident as foreground occluders are progressively removed. In contrast, FoundationDeOcclusion produces completions that plausibly recover the missing structures while preserving surface continuity, spatial relationships, and global scene alignment. These results demonstrate that FoundationDeOcclusion enhances both local completeness and global geometric consistency.

4.2. Real-Robot Mobile Manipulation

We evaluate FoundationDeOcclusion on a Stretch mobile manipulator in a mock apartment lab, focusing on two mobile manipulation tasks that stress occlusion handling.

Task 1: Table-top pick-up with occlusion. In this task, the robot must grasp a target (e.g., a bottle) hidden behind a large occluder (e.g., boxes). A single head-mounted RGB-D observation and the target’s text label are fed to the de-

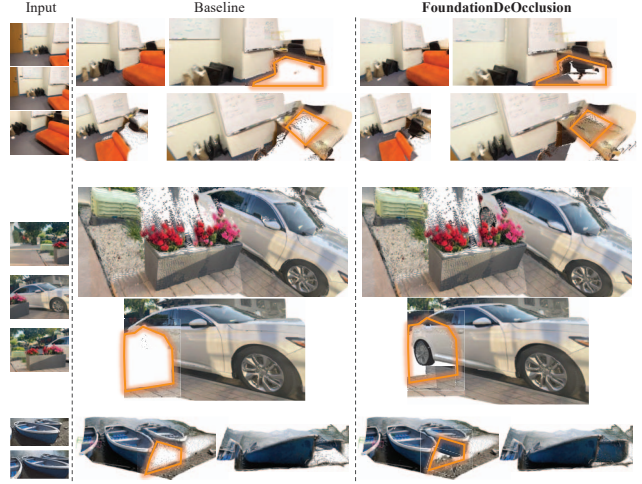


Figure 5. **Qualitative results** on indoor and in-the-wild scenes. We compare reconstructions from baseline and FoundationDeOcclusion. Our method consistently recovers occluded geometry across diverse inputs, preserving structural integrity while maintaining alignment within the scene.

Method	Avg. SR	Light Occlusion		Heavy Occlusion	
		SR	IoU	SR	IoU
Visual Servoing [7]	20.6	29.7	81.4	11.7	51.4
AnyGrasp [17]	29.7	29.7	81.4	29.7	51.4
AnyGrasp+BrushNet [27]	44.1	47.1	84.2	41.1	56.7
AnyGrasp+Ours	73.5	82.3	96.3	64.7	95.7
Δ	+43.8	+52.6	+14.9	+35.0	+44.3

Table 2. FoundationDeOcclusion significantly improves the success rate (SR) and affordance IoU of robot manipulation under varying occlusion levels on the task of **Table-top manipulation**.

occluder, which returns a fused point cloud in which the target’s amodal geometry is reconstructed and integrated. We then invoke AnyGrasp [17] on this completed point cloud to propose stable grasp poses and execute the highest-scoring grasp. This pipeline eliminates the need to first plan and acquire multiple viewpoints for a full reconstruction: the de-occluder provides integrated surface affordances directly from partial views. In contrast, baselines such as visual servoing or vanilla AnyGrasp [17] often fail to localize the target under heavy occlusion or propose grasps on incomplete geometry, resulting in unstable or suboptimal executions (see Table 2). Comparisons with other de-occlusion models, such as Amodal3R [71], are included in the supplementary material due to hardware and time constraints.

Task 2: Mobile place-down with occlusion. In this task, the robot starts at a distance while holding an object and must place it on a tabletop whose free surface is largely occluded by other items in the initial view. Traditional approaches

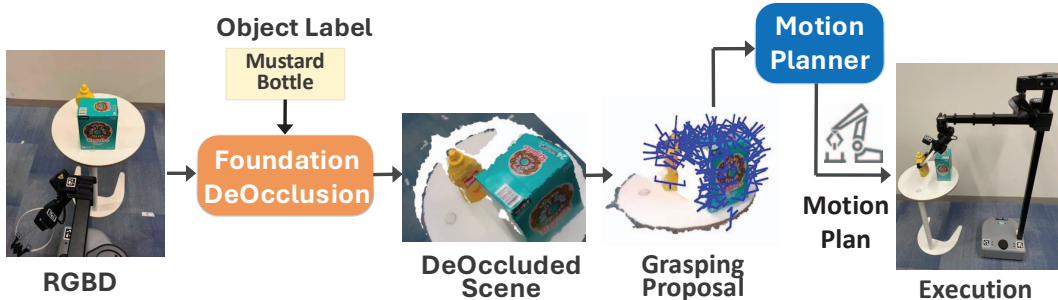


Figure 6. **Real-robot table-top pick-up.** A target object (e.g., a bottle) is hidden behind a large occluder. From a single RGB-D observation, FoundationDeOcclusion reconstructs the amodal geometry of the target and produces a completed point cloud. AnyGrasp [17] then proposes grasps on the recovered surface, yielding stable and well-positioned grasp poses that would be unavailable from the partial observation alone.

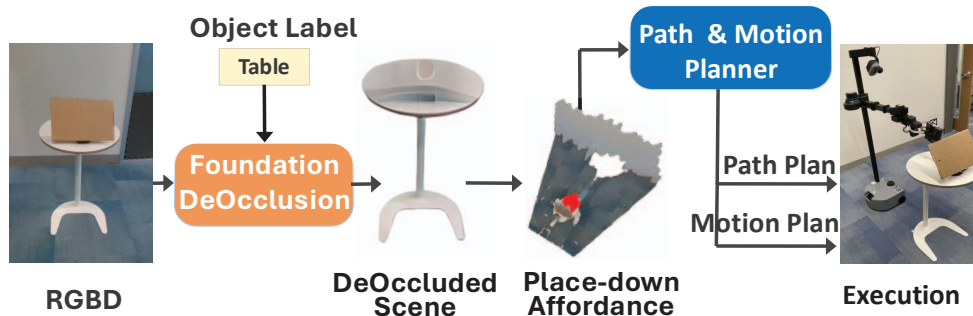


Figure 7. **Real-robot mobile place-down task.** The robot must place a held object onto a tabletop whose free surface is largely occluded by clutter. FoundationDeOcclusion completes the occluded table geometry from a single initial view, enabling the planner to identify collision-free placement locations behind the occluders without requiring additional exploratory viewpoints.

Method	SCT	IoU
DynaMem [41]	27.2	43.6
FoundationDeOcclusion	44.4	82.1

Table 3. FoundationDeOcclusion significantly improves performance for **Mobile place-down**. Metrics include success rate weighted by completion time (SCT), and affordance (IoU).

require navigating close, selecting next-best-views, and actively exploring around the table to reconstruct sufficient geometry before estimating place affordances. Our method instead takes a single RGB-D frame (or a short sequence) at the outset, performs amodal de-occlusion of the table, and fuses the result with the observed point cloud; the planner then selects a collision-free, high-quality placement location that may lie entirely behind occluders and be absent from the original input. We report Success weighted by Completion Time (SCT), and affordance metric (IoU), and observe that *FoundationDeOcclusion* delivers substantial gains in SCT and IoU, indicating more efficient decision making and fewer exploratory detours (see Table 3).

4.3. Data Scaling for Thin Objects

We investigate how scaling the training data impacts de-occlusion performance, particularly for challenging object categories with complex geometry.

Experimental Setup. We compare two models: (1) the original FoundationDeOcclusion model, and (2) a FoundationDeOcclusion model finetuned on a curated dataset of 32K samples. This finetuning dataset consists of 16K samples from the original Pix2Gestalt dataset and 16K synthetic samples rendered via Blender, featuring 3D chairs as occludees and various other 3D assets as occluders.

Results on COCO-A. As shown in Table 4, the finetuned model demonstrates consistent performance gains across the COCO-A benchmark, with particularly significant improvements in the chair category. The overall mIoU improves from 81.2% to 83.6%, while the chair-specific mIoU increases from 68.9% to 74.4%—a gain of 5.5% points.

Evaluation on Challenging Chair-Legs Scenarios. We observed that most chair samples in the COCO-A subset exhibit relatively low occlusion rates, failing to represent the complexity of real-world scenarios where chair legs are frequently hidden behind foreground objects. To rigorously evaluate model robustness, we curated a more challenging

Model	Avg.		Chair		Chair-Legs	
	mIoU \uparrow	Δ	mIoU \uparrow	Δ	mIoU \uparrow	Δ
Our base model	81.2	–	68.9	–	58.3	–
+16K samples	83.6	+2.4	74.4	+5.5	62.2	+4.0

Table 4. **Effect of data scaling on de-occlusion performance.** We compare our proposed FoundationDeOcclusion base model (L-DoT) with a version finetuned on 32K curated samples (16K from Pix2Gestalt + 16K synthetic). The finetuned model shows consistent improvements, especially on chair categories with complex leg geometry.

synthetic test set specifically focused on chair-leg completion. We collected 45 high-quality 3D chair assets and rendered them in Blender. To ensure benchmark difficulty, we filtered the synthetic data with a strict constraint: the bottom 40% of the occludee must have at least a 30% occlusion ratio. On this challenging subset, the baseline model achieves 58.3% mIoU, while the finetuned model reaches 62.2% mIoU—an improvement of 3.9 percentage points. These results demonstrate that targeted data augmentation with synthetic occlusion scenarios effectively improves de-occlusion performance on geometrically complex objects.

4.4. Efficiency Comparison

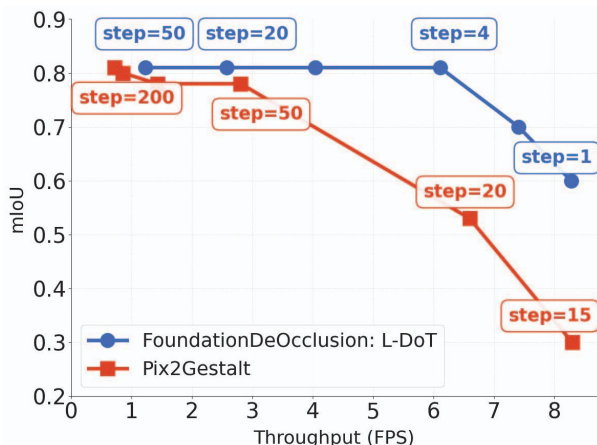


Figure 8. **Efficiency** (mIoU vs. FPS) of FoundationDeOcclusion’s L-DoT and Pix2Gestalt [48] under varying denoising steps.

We evaluate the accuracy-efficiency trade-off of L-DoT on the Amodal COCO (COCO-A [88]) dataset, following the protocol of pix2gestalt [48]. As shown in Figure 8, reducing the number of denoising steps from 50 to 4 leaves mIoU virtually unchanged (0.81 mIoU) while increasing throughput from 1.23 to 6.11 fps. Compared to pix2gestalt, L-DoT achieves higher mIoU with substantially higher throughput, highlighting the advantage of our linear DoT design.

While performance comparisons with image-to-3D-mesh methods [71] are not directly comparable, their efficiency remains comparable. Figure 1 further reports efficiency of

system (reconstruction+generation): our FoundationDeOcclusion pipeline processes 0.62 objects per second, nearly twice as fast as Amodal3R [71] at 0.32 objects, and significantly faster than TRELIS [72] at 0.24 objects per second.

Implementation details. All experiments are conducted on a single H100 GPU. The robot communicates with FoundationDeOcclusion through a high-speed intranet API. The proposed L-DoT is built on the lite DiT [75] with 590M parameters trained with rectified flow, and more hyperparameters for training are detailed in supplementary.

5. Conclusion

We presented FoundationDeOcclusion, a fast generative framework that restores hidden geometry to improve scene-level 3D reconstruction. Central to our design is the geometry-aware linear De-Occlusion Transformer, which enables rapid and spatially consistent occlusion recovery. By consistently improving state-of-the-art 3D reconstruction and increasing AnyGrasp’s real-robot success rate by 43.8%, FoundationDeOcclusion shows that efficient and effective generation can significantly strengthen reconstruction and robotic perception. These results highlight a practical path toward more complete 3D understanding and demonstrate the real-world potential of generative vision in robotics.

Limitations

Our method has several limitations. First, non-convex objects such as chairs with hollow backs or tables with open frames can produce fragmented masks with inconsistent centroids across views, leading to unreliable instance association. Integrating richer geometric priors or learned cross-view matching may help address this. Second, real-robot performance degrades under nighttime lighting, reflecting a daytime bias in the training data that could be mitigated by diversifying the training distribution. Third, the community currently lacks a dedicated 3D reconstruction benchmark for occlusion-aware evaluation—to the best of our knowledge upon submission, there is no 3D counterpart to COCO-A [88] that provides dense ground-truth geometry for occluded regions at scale. This limits systematic and standardized assessment of scene-level de-occlusion methods. In this work, we therefore focus on empirical findings and real-robot experiments to validate the practical impact of occlusion recovery, and we hope our efforts motivate future benchmark development in this direction.

Acknowledgement

We acknowledge discussions with Zihan Qin and Wenxin Ma in the early stages of this work. This work was supported by ONR N00014-23-1-2641 and from National Eye Institute (NEI) with Award ID: R01EY037193.

References

- [1] Sameer Agarwal, Noah Snavely, Steven M Seitz, and Richard Szeliski. Bundle adjustment in the large. In *ECCV*, 2010. 1, 2
- [2] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022. 2
- [3] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, 1992. 5
- [4] Andreas Bircher, Mina Kamel, Kostas Alexis, Helen Oleynikova, and Roland Siegwart. Receding horizon" next-best-view" planner for 3d exploration. In *ICRA*, 2016. 2
- [5] Tara Boroushaki, Junshan Leng, Ian Clester, Alberto Rodriguez, and Fadel Adib. Robotic grasping of fully-occluded objects using rf perception. In *ICRA*, 2021. 1
- [6] Cesar Cadena, Luca Carlone, Henry Carrillo, Yasir Latif, Davide Scaramuzza, José Neira, Ian Reid, and John J Leonard. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on robotics*, 2016. 2
- [7] Francois Chaumette, Seth Hutchinson, and Peter Corke. Visual servoing. In *handbook of robotics*. Springer, 2016. 6
- [8] Boyuan Chen, Fei Xia, Brian Ichter, Kanishka Rao, Keerthana Gopalakrishnan, Michael S Ryoo, Austin Stone, and Daniel Kappler. Open-vocabulary queryable scene representations for real world planning. *arXiv preprint arXiv:2209.09874*, 2022. 2
- [9] Seunggeun Chi, Enna Sachdeva, Pin-Hao Huang, and Kwonjoon Lee. Contact-aware amodal completion for human-object interaction via multi-regional inpainting. In *ICCV*, 2025. 2
- [10] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017. 5
- [11] Aneel Damaraju, Dean Hazineh, and Todd Zickler. Cobl: Toward zero-shot ordinal layering without user prompting. In *ICCV*, 2025. 2
- [12] Michael Danielczuk, Andrey Kurenkov, Ashwin Balakrishna, Matthew Matl, David Wang, Roberto Martín-Martín, Animesh Garg, Silvio Savarese, and Ken Goldberg. Mechanical search: Multi-step retrieval of a target object occluded by clutter. In *ICRA*, 2019. 2
- [13] Andrew J Davison, Ian D Reid, Nicholas D Molton, and Olivier Stasse. Monoslam: Real-time single camera slam. *TPAMI*, 2007. 2
- [14] Kiana Ehsani, Roozbeh Mottaghi, and Ali Farhadi. Segan: Segmenting and generating the invisible. In *CVPR*, 2018. 2
- [15] Noam Elata, Bahjat Kawar, Yaron Ostrovsky-Berman, Miriam Farber, and Ron Sokolovsky. Novel view synthesis with pixel-space diffusion models. In *CVPR*, 2025. 2
- [16] Jakob Engel, Thomas Schöps, and Daniel Cremers. Lsd-slam: Large-scale direct monocular slam. In *ECCV*, 2014. 2
- [17] Hao-Shu Fang, Chenxi Wang, Hongjie Fang, Minghao Gou, Jirong Liu, Hengxu Yan, Wenhai Liu, Yichen Xie, and Cewu Lu. Anygrasp: Robust and efficient grasp perception in spatial and temporal domains. *IEEE Transactions on Robotics*, 2023. 6, 7
- [18] Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T Barron, and Ben Poole. Cat3d: Create anything in 3d with multi-view diffusion models. *NeurIPS*, 2024. 2
- [19] Caelan Reed Garrett, Rohan Chitnis, Rachel Holladay, Beomjoon Kim, Tom Silver, Leslie Pack Kaelbling, and Tomás Lozano-Pérez. Integrated task and motion planning. *Annual review of control, robotics, and autonomous systems*, 2021. 2
- [20] Dongchen Han, Xuran Pan, Yizeng Han, Shiji Song, and Gao Huang. Flatten transformer: Vision transformer using focused linear attention. In *ICCV*, 2023. 3
- [21] Binh-Son Hua, Quang-Hieu Pham, Duc Thanh Nguyen, Minh-Khoi Tran, Lap-Fai Yu, and Sai-Kit Yeung. Scenenn: A scene meshes dataset with annotations. In *3DV*, 2016. 5
- [22] Tianxin Huang, Zhiwen Yan, Yuyang Zhao, and Gim Hee Lee. Compc: Completing a 3d point cloud with 2d diffusion priors. In *ICLR*, 2025. 1, 2
- [23] 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*, 2025. 2
- [24] Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd Omama, Tao Chen, Alaa Maalouf, Shuang Li, Ganesh Iyer, Soroush Saryazdi, Nikhil Keetha, et al. Conceptfusion: Open-set multimodal 3d mapping. *RSS*, 2023. 2
- [25] Sunshine Jiang, Siddharth Ancha, Travis Manderson, Laura Brandt, Yilun Du, Philip R Osteen, and Nicholas Roy. Anomalies-by-synthesis: Anomaly detection using generative diffusion models for off-road navigation. In *ICRA*, 2025. 2
- [26] Seong-Uk Jo, Du Yeol Lee, and Chae Eun Rhee. Occlusion-aware amodal depth estimation for enhancing 3d reconstruction from a single image. *IEEE Access*, 2024. 2
- [27] Xuan Ju, Xian Liu, Xintao Wang, Yuxuan Bian, Ying Shan, and Qiang Xu. Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. In *ECCV*, 2024. 6
- [28] Leslie Pack Kaelbling and Tomás Lozano-Pérez. Integrated task and motion planning in belief space. *The International Journal of Robotics Research*, 2013. 2
- [29] Abhishek Kar, Shubham Tulsiani, Joao Carreira, and Jitendra Malik. Amodal completion and size constancy in natural scenes. In *ICCV*, 2015. 2
- [30] Yoni Kasten, Ohad Rahamim, and Gal Chechik. Point cloud completion with pretrained text-to-image diffusion models. *NeurIPS*, 2023. 1, 2
- [31] Lei Ke, Yu-Wing Tai, and Chi-Keung Tang. Deep occlusion-aware instance segmentation with overlapping bilayers. In *CVPR*, 2021. 2
- [32] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *ICCV*, 2023. 2

- [33] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024. 2
- [34] Wolfgang Köhler. Gestalt psychology. *Psychologische forschung*, 31(1):XVIII–XXX, 1967. 1
- [35] Yann Labbé, Justin Carpentier, Mathieu Aubry, and Josef Sivic. Cosypose: Consistent multi-view multi-object 6d pose estimation. In *ECCV*, 2020. 2
- [36] Hyunmin Lee and Jaesik Park. Instance-wise occlusion and depth orders in natural scenes. In *CVPR*, 2022. 2
- [37] Vincent Leroy, Johann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *ECCV*, 2024. 1
- [38] Zhenyu Li, Mykola Lavreniuk, Jian Shi, Shariq Farooq Bhat, and Peter Wonka. Amodal depth anything: Amodal depth estimation in the wild. In *ICCV*, 2025. 2
- [39] Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. In *ICLR*, 2026. 1, 5
- [40] Peiqi Liu, Yaswanth Orru, Jay Vakil, Chris Paxton, Nur Muhammad Mahi Shafiullah, and Lerrel Pinto. Ok-robot: What really matters in integrating open-knowledge models for robotics. *arXiv preprint arXiv:2401.12202*, 2024. 2
- [41] Peiqi Liu, Zhanqiu Guo, Mohit Warke, Soumith Chintala, Chris Paxton, Nur Muhammad Mahi Shafiullah, and Lerrel Pinto. Dynamem: Online dynamic spatio-semantic memory for open world mobile manipulation. In *ICRA*, 2025. 2, 7
- [42] Zhengzhe Liu, Qing Liu, Chirui Chang, Jianming Zhang, Daniil Pakhomov, Haitian Zheng, Zhe Lin, Daniel Cohen-Or, and Chi-Wing Fu. Object-level scene deocclusion. In *SIGGRAPH*, 2024. 2
- [43] Jeffrey Mahler, Jacky Liang, Sherdil Niyaz, Michael Laskey, Richard Doan, Xinyu Liu, Juan Aparicio Ojea, and Ken Goldberg. Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics. *arXiv preprint arXiv:1703.09312*, 2017. 2
- [44] Andrew Melnik, Michael Büttner, Leon Harz, Lyon Brown, Gora Chand Nandi, Arjun PS, Gaurav Kumar Yadav, Rahul Kala, and Robert Haschke. Uniteam: Open vocabulary mobile manipulation challenge. *arXiv preprint arXiv:2312.08611*, 2023. 2
- [45] Qiwei Meng, Jason Gu, and Yun-Hui Liu. Gpd: Learning geometric primitive deformation for unseen object pose estimation. *IEEE Transactions on Automation Science and Engineering*, 2024. 2
- [46] Arsalan Mousavian, Clemens Eppner, and Dieter Fox. 6-dof graspnet: Variational grasp generation for object manipulation. In *ICCV*, 2019. 2
- [47] Riku Murai, Eric Dexheimer, and Andrew J Davison. Mast3r-slam: Real-time dense slam with 3d reconstruction priors. In *CVPR*, 2025. 2
- [48] Ege Ozguroglu, Ruoshi Liu, Dídac Surís, Dian Chen, Achal Dave, Pavel Tokmakov, and Carl Vondrick. pix2gestalt: Amodal segmentation by synthesizing wholes. In *CVPR*, 2024. 1, 2, 4, 8
- [49] Stuti Pathak, Prashant Kumar, Dheeraj Baiju, Nicholas Mboga, Gunther Steenackers, and Rudi Penne. Revisiting point cloud completion: Are we ready for the real-world? In *ICCV*, 2024. 2
- [50] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. 4
- [51] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. Pynet: Pixel-wise voting network for 6dof pose estimation. In *CVPR*, 2019. 2
- [52] Lu Qi, Li Jiang, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Amodal instance segmentation with kins dataset. In *CVPR*, 2019. 2
- [53] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 4
- [54] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 3
- [55] 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*, 2024. 2
- [56] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 1, 2
- [57] Noah Snavely, Steven M Seitz, and Richard Szeliski. Modeling the world from internet photo collections. *IJCV*, 2008. 1, 2
- [58] Yihong Sun, Adam Kortylewski, and Alan Yuille. Amodal segmentation through out-of-task and out-of-distribution generalization with a bayesian model. In *CVPR*, 2022. 2
- [59] Keisuke Tateno, Federico Tombari, Iro Laina, and Nassir Navab. Cnn-slam: Real-time dense monocular slam with learned depth prediction. In *CVPR*, 2017. 2
- [60] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024. 2
- [61] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *NeurIPS*, 2021. 1, 2
- [62] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *TPAMI*, 1991. 5
- [63] Arthur Wandzel, Yoonseon Oh, Michael Fishman, Nishanth Kumar, Lawson LS Wong, and Stefanie Tellex. Multi-object search using object-oriented pomdps. In *ICRA*, 2019. 2
- [64] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *CVPR*, 2025. 1, 2, 5
- [65] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *CVPR*, 2025. 1, 2

- [66] Qianqian Wang*, Yifei Zhang*, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *CVPR*, 2025. 5
- [67] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024. 1, 2
- [68] Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, et al. Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai. In *CVPR*, 2024. 1
- [69] Abdelrhman Werby, Chenguang Huang, Martin Büchner, Abhinav Valada, and Wolfram Burgard. Hierarchical open-vocabulary 3d scene graphs for language-grounded robot navigation. In *ICRA Workshop on Vision-Language Models for Navigation and Manipulation*, 2024. 2
- [70] Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P Srinivasan, Dor Verbin, Jonathan T Barron, Ben Poole, et al. Reconfusion: 3d reconstruction with diffusion priors. In *CVPR*, 2024. 2
- [71] Tianhao Wu, Chuanxia Zheng, Frank Guan, Andrea Vedaldi, and Tat-Jen Cham. Amodal3r: Amodal 3d reconstruction from occluded 2d images. In *ICCV*, 2025. 6, 8
- [72] 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*, 2025. 8
- [73] Peng Xiang, Xin Wen, Yu-Shen Liu, Yan-Pei Cao, Pengfei Wan, Wen Zheng, and Zhizhong Han. Snowflakenet: Point cloud completion by snowflake point deconvolution with skip-transformer. In *ICCV*, 2021. 2
- [74] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*, 2017. 2
- [75] Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, et al. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. In *ICLR*, 2025. 3, 4, 8
- [76] Zhijie Yan, Shufei Li, Zuoxu Wang, Lixiu Wu, Han Wang, Jun Zhu, Lijiang Chen, and Jihong Liu. Dynamic open-vocabulary 3d scene graphs for long-term language-guided mobile manipulation. *RAL*, 2025. 2
- [77] Yaoqing Yang, Chen Feng, Yiru Shen, and Dong Tian. Fold-ingnet: Point cloud auto-encoder via deep grid deformation. In *CVPR*, 2018. 2
- [78] Kaixin Yao, Longwen Zhang, Xinhao Yan, Yan Zeng, Qixuan Zhang, Lan Xu, Wei Yang, Jiayuan Gu, and Jingyi Yu. Cast: Component-aligned 3d scene reconstruction from an rgb image. *TOG*, 2025. 2
- [79] Sriram Yenamandra, Arun Ramachandran, Karmesh Yadav, Austin Wang, Mukul Khanna, Theophile Gervet, Tsung-Yen Yang, Vidhi Jain, Alexander William Clegg, John Turner, et al. Homerobot: Open-vocabulary mobile manipulation. *arXiv preprint arXiv:2306.11565*, 2023. 2
- [80] Ahmet Burak Yildirim, Vedat Baday, Erkut Erdem, Aykut Erdem, and Aysegül Dunder. Inst-inpaint: Instructing to remove objects with diffusion models. *arXiv preprint arXiv:2304.03246*, 2023. 2
- [81] Xumin Yu, Yongming Rao, Ziyi Wang, Zuyan Liu, Jiwen Lu, and Jie Zhou. Pointr: Diverse point cloud completion with geometry-aware transformers. In *ICCV*, 2021. 2
- [82] Wentao Yuan, Tejas Khot, David Held, Christoph Mertz, and Martial Hebert. Pcn: Point completion network. In *3DV*, 2018. 2
- [83] Xiaohang Zhan, Xingang Pan, Bo Dai, Ziwei Liu, Dahua Lin, and Chen Change Loy. Self-supervised scene de-occlusion. In *CVPR*, 2020. 2
- [84] Bowen Zhang, Qing Liu, Jianming Zhang, Yilin Wang, Liyang Liu, Zhe Lin, and Yifan Liu. Amodal scene analysis via holistic occlusion relation inference and generative mask completion. In *AAAI*, 2024. 2
- [85] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. In *ICLR*, 2025. 2
- [86] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. Recognize anything: A strong image tagging model. In *CVPR*, 2024. 3
- [87] Kaiyu Zheng, Yoonchang Sung, George Konidaris, and Stefanie Tellex. Multi-resolution pomdp planning for multi-object search in 3d. In *IROS*, 2021. 2
- [88] Yan Zhu, Yuandong Tian, Dimitris Metaxas, and Piotr Dollar. Semantic amodal segmentation. In *CVPR*, 2017. 5, 8
- [89] Dong Zhuo, Wenzhao Zheng, Jiahe Guo, Yuqi Wu, Jie Zhou, and Jiwen Lu. Streaming 4d visual geometry transformer. *arXiv preprint arXiv:2507.11539*, 2025. 2