LaRI3R: Bidirectional Ray Intersections for Occluded 3D Scene Reconstruction from Two Views

Anonymous authors
Paper under double-blind review



Figure 1: LaRI3R recovers occluded 3D geometry by leveraging complementary observations across two views in both 2D and 3D spaces. The random color in "Object Results" and the yellow color in "Scene Results" denote the recovered occluded geometry. The proposed bidirectional layered representation and geometry-aware feature aggregation pipeline improve reconstruction of occluded surfaces for both objects and scenes, and outperform existing visible-surface reconstruction and single-view scene reasoning methods. Meanwhile, LaRI3R effectively supports downstream 3D tasks, improving camera pose estimation over geometric foundation models.

Abstract

Recovering complete 3D scene geometry from limited observations remains challenging due to severe occlusions and restricted viewpoints. Although observing a scene from many viewpoints can progressively improve reconstruction completeness, dense multi-view reconstruction often requires costly exploration and computation. Recent single-view feedforward methods address the trade-off between reconstruction quality and efficiency through layered representations, but they still suffer from strong ambiguity in heavily occluded regions. Extending layered reconstruction to two views is also non-trivial because occlusion ordering becomes view-dependent, and visually corresponding evidence can be weak or absent across views. We present LaRI3R, a new feedforward two-view framework for complete 3D scene reconstruction under severe occlusions. Our approach is built on two core designs: 1) We introduce a bidirectional layered representation that predicts geometry in both forward and backward ray orders, which reduces geometric inconsistencies in heavily occluded regions. 2) Instead of relying purely on appearance-based matching, we perform geometry-aware feature aggregation using explicit 3D positional cues. This design enables geometrically related but visually dissimilar regions to communicate across views. Experiments on both object- and scene-level benchmarks show that LaRI3R outperforms representative generative, iterative, and feedforward methods while preserving efficient inference of a feedforward model. We further show that the recovered occluded geometry serves as a strong structural cue for downstream 3D vision tasks such as camera pose estimation with notable improvements over existing large geometric models.

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.

1 INTRODUCTION

Understanding complete 3D scene geometry from limited visual observations remains a fundamental challenge in 3D vision. Due to occlusions and restricted viewpoints, a single image only reveals a subset of the scene, and large portions of the geometry remain unseen. This ambiguity becomes particularly problematic for applications such as robotics and embodied agents, where accurate geometric understanding directly affects navigation, manipulation, and decision-making. Although observing a scene from many viewpoints can progressively resolve occlusions and improve reconstruction completeness, acquiring dense observations requires active exploration and substantial computation. This creates a practical trade-off: single-view reconstruction remains highly ambiguous under severe occlusions, but dense multi-view reconstruction is often expensive and impractical.

Existing approaches lie at two extremes of this trade-off. Generative methods (Xiang et al., 2025; Liu et al., 2023b; Yu et al., 2024; Voleti et al., 2024; Wu et al., 2024b; Lai et al., 2025) infer plausible complete geometry from limited observations using strong learned priors. However, they often rely on large diffusion backbones and iterative sampling, which leads to high computational cost. Many of these methods (Xiang et al., 2025; Huang et al., 2025; Boss et al., 2025) also reconstruct geometry in canonical object coordinates instead of view-aligned geometry, which limits downstream geometric reasoning. At the other extreme, dense multi-view reconstruction methods (Schönberger & Frahm, 2016; Deng et al., 2025a) improve geometric completeness by aggregating observations across many viewpoints. However, these approaches typically require extensive viewpoint coverage, iterative optimization, or costly cross-view processing. Iterative reconstruction pipelines further refine geometry through repeated querying or additional cues such as poses, masks, or user interactions, which often leads to slow and complex pipelines (Wang et al., 2025b; Li et al., 2024b; Duisterhof et al., 2025; Li et al., 2024a; Wimbauer et al., 2023).

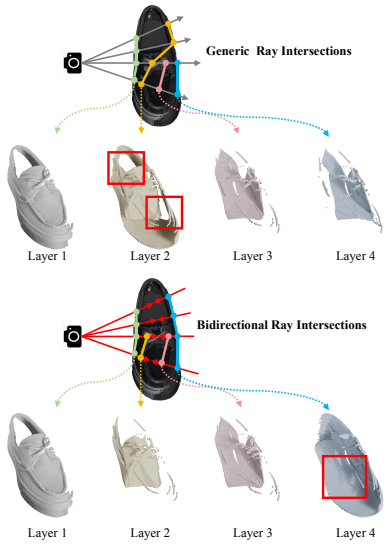

Figure 2: **Generic Ray Intersections vs. Bidirectional Ray Intersections.** Generic ray intersections assign layer points only along a single ray direction, which often produces discontinuous surfaces across layers (e.g., layer 2). The proposed bidirectional ray intersections iteratively extract outermost surfaces from both forward and backward directions to produce more coherent layered geometry (e.g., layer 4), which improves geometric reasoning and network prediction.

Recently, layered geometric representations (Hu et al., 2024b; Li et al., 2025; Shin et al., 2019) have emerged as an efficient feedforward alternative. Inspired by layered depth images (Shade et al., 1998), these methods predict multiple ray-surface intersections instead of only the first visible depth. Particularly, LaRI (Li et al., 2025) demonstrates that layered point maps can efficiently reconstruct complete scene geometry from a single image. However, single-view reconstruction still leaves deeply occluded regions fundamentally ambiguous because the missing geometry lacks sufficient visual evidence. Although incorporating more viewpoints can alleviate this issue, existing dense multi-view pipelines (Wang et al., 2025e; Li et al., 2023a) often require broad scene coverage and substantial computation, which limits their practicality in embodied settings.

In this work, we explore two-view reconstruction as a compromise between reconstruction completeness and practical efficiency. Compared to single-view reconstruction, a second view introduces complementary geometric observations that significantly reduce occlusion ambiguity without requiring full scene exploration. However, extending layered reconstruction to stereo views remains non-trivial. Existing two-view models such as DUSt3R (Wang et al., 2024) primarily rely on appearance-based matching, which becomes unreliable for occluded geometry because unseen surfaces may project to weak, missing, or dissimilar regions across views. Conventional layered representations (Hu et al., 2024b; Li et al., 2025; Shin et al., 2019) also follow a unidirectional ray ordering, which often produces discontinuous and inconsistent layers under heavy occlusion.

To address these challenges, we introduce LaRI3R, a feedforward two-view framework for complete 3D scene reconstruction under severe occlusions. We propose a bidirectional layered representation that predicts ray-surface intersections in both forward and backward orders along each camera ray, which reduces geometric inconsistencies in heavily occluded regions. Furthermore, we perform geometry-aware feature aggregation using explicit 3D positional cues instead of relying purely on appearance matching. Specifically, intermediate bidirectional layered point maps serve as 3D anchors that lift view-specific features into a shared 3D space, where a point-based transformer (Wu et al., 2024c) aggregates spatially grounded features across views. This design enables geometrically related but visually dissimilar observations to communicate effectively for unseen structure recovery.

Our LaRI3R outperforms existing generative and iterative approaches on both object- and scene-level reconstruction while preserving feedforward efficiency. We further show that the predicted geometry serves as a strong structural representation for downstream 3D tasks. Combined with a standard PnP solver, our LaRI3R improves camera pose estimation over large-scale geometric foundation models that only reconstruct visible surfaces.

Our contributions are threefold:

- We propose LaRI3R, a two-view complete scene reconstruction method that improves reconstruction completeness and maintains feedforward efficiency under severe occlusions.
- We introduce a bidirectional layered representation and a geometry-aware feature aggregation scheme that robustly recover occluded scene components, where traditional appearance-based matching may fail.
- We show that LaRI3R serves as a versatile 3D pipeline, delivering strong object- and scene-level reconstruction while also improving downstream tasks such as camera pose estimation.

2 Related Work

Learning-based 3D Reconstruction. Learning-based 3D reconstruction has been extensively studied over the past decade. Early approaches (Cao et al., 2022; Ding et al., 2022; Gu et al., 2020; Wu et al., 2024a; Xu et al., 2022; Zhang et al., 2023; Yao et al., 2018; Li et al., 2023a;b; Wu et al., 2025a) are largely built upon classical multi-view stereo (MVS) pipelines, which estimate dense depth maps from multiple calibrated images and subsequently fuse them into a coherent 3D geometry. These methods often rely on accurate camera calibration (Schönberger & Frahm, 2016) and mainly reconstruct visible surfaces from multi-view observations. Recent methods such as DUSt3R (Wang et al., 2024) have shifted toward end-to-end feedforward reconstruction that directly infers geometry from images. DUSt3R employs cross-attention (Dosovitskiy, 2020) over image patch tokens to aggregate information across views and regress point maps jointly. Subsequent works extend this framework to dynamic scenes (Chen et al., 2025a; Wang et al., 2025b;d; Zhang et al., 2024b; Lu et al., 2025), camera pose estimation (Dong et al., 2025; Deng et al., 2025b), and more efficient inference strategies (Wang et al., 2025a; Yang et al., 2025; Chen et al., 2025b; Deng et al., 2025a; Wang et al., 2025e; Lin et al., 2025). Despite their strong performance, these methods primarily model visible geometry and often struggle under severe occlusions since each pixel predicts only a single surface along the viewing ray. In contrast, our method reconstructs occluded geometry through a multi-layer framework. We introduce a Bidirectional Ray Intersection Representation that predicts multiple surfaces along each ray from both directions, which better preserves object contours, reduces surface splitting, and improves consistency across layered point maps through geometry-aware feature aggregation.

Occluded Scene Reconstruction. Recent works have explored the reconstruction of occluded geometry and incomplete scenes. Generative approaches (Liu et al., 2023b;a; Tang et al., 2024; Hong et al., 2023; Yu et al., 2024; Voleti et al., 2024; Wu et al., 2024b; Xiang et al., 2025; Mao et al., 2023) synthesize novel viewpoints to assist 3D reconstruction, while other methods (Zhang et al., 2024a; Jun & Nichol, 2023; Xiang et al., 2025; Lan et al., 2024; Lai et al., 2025) directly generate geometry in 3D space. However, these methods often suffer from inconsistent synthesized views or hallucinated geometry in unobserved regions. Semantic scene completion methods (Iwase et al., 2024; Li et al., 2023c; Yan et al., 2022; Wu et al., 2023) reconstruct incomplete scenes from RGB-D observations, typically with external depth priors. Other approaches (Wang et al., 2025b; Duisterhof et al., 2025) iteratively complete geometry through virtual viewpoint querying. Layered

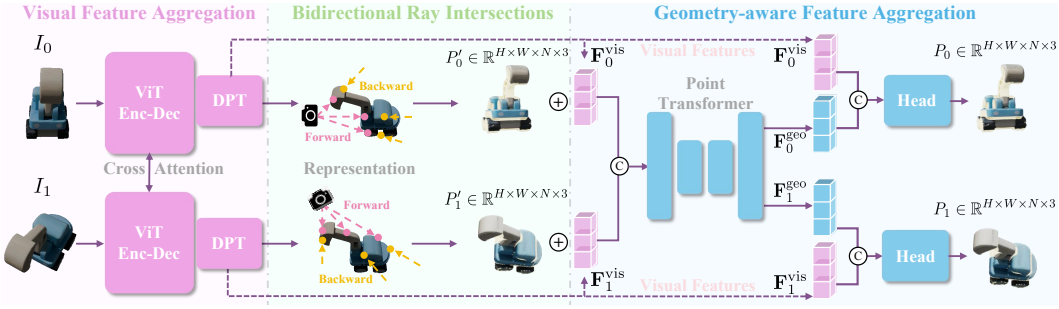

Figure 3: **Overview.** Given a pair of uncalibrated images I_0 and I_1, we first use a ViT encoder-decoder network to extract 2D visual features $\mathbf{F}_i^{vis}$ for each view. These features are used to regress the initial multi-layer point maps P_i' based on the proposed bidirectional ray intersection representation. Next, the point maps P_i' and visual features $\mathbf{F}_i^{vis}$ from both views are jointly processed by a point-based transformer to perform geometry-aware feature aggregation in 3D space, producing geometric features $\mathbf{F}_i^{geo}$. Finally, a 3D convolution head combines $\mathbf{F}_i^{vis}$ and $\mathbf{F}_i^{geo}$ to regress the final bidirectional point maps P_i.

methods (Hu et al., 2024b; Li et al., 2025; Shin et al., 2019) explicitly model multiple ray-scene intersections to reconstruct occluded surfaces. However, these methods often degrade under severe occlusions due to limited geometric reasoning across layers. In contrast, our method introduces a bidirectional ray intersection representation with geometry-aware feature aggregation to better leverage complementary visual and geometric cues for occluded reconstruction.

3 Methodology

Problem setup. Given a pair of images $\{I_0, I_1\}$ captured from two viewpoints, where $I_i \in \mathbb{R}^{H\times W\times 3}$, LaRI3R predicts an N-layer point map $P_i \in \mathbb{R}^{H\times W\times N\times 3}$ for each view. The first layer represents visible surfaces, and the remaining layers represent occluded surfaces. To model occluded geometry consistently across views, we introduce a bidirectional ray intersection representation (Sec. 3.1). We then propose a geometry-aware feature aggregation scheme (Sec. 3.2) that jointly reasons over visual and geometric cues in 3D space.

Overview. Fig. 3 illustrates the overall pipeline. The input images $\{I_0, I_1\}$ are first partitioned into patch tokens and processed by a shared-weight ViT (Dosovitskiy, 2020) encoder. The extracted features are passed through separate transformer decoders with DPT (Ranftl et al., 2021) heads to produce layered 2D feature maps $\mathbf{F}_i^{\text{vis}} \in \mathbb{R}^{H\times W\times N\times C}$. These features are then used to regress the initial bidirectional point maps P_i' in the camera coordinates of I_0 (Wang et al., 2024) (Sec. 3.1). Subsequently, we leverage a point transformer network to jointly encode the point maps and features from both views for geometry-aware feature aggregation in 3D space. The network extracts geometric features $\mathbf{F}_i^{geo} \in \mathbb{R}^{H\times W\times N\times C'}$, which are combined with the 2D patch features $\mathbf{F}_i^{vis}$ to predict the final bidirectional multi-layer point maps P_i (Sec. 3.2).

3.1 Bidirectional Ray Intersections

Existing layered representations (Shin et al., 2019; Li et al., 2025) construct multi-layer geometry by tracing each ray in a single forward direction from the camera center toward the scene. In this formulation, layer assignments depend entirely on the intersection order along each viewing ray. However, under heavy occlusion, different rays often intersect different subsets of surfaces. As a result, points that belong to the same physical structure may be assigned to different layers across neighboring rays (Fig. 2). This inconsistency introduces discontinuities within each layer and makes reasoning over occluded geometry substantially more difficult. To address this limitation, we introduce a bidirectional ray intersection representation for layered geometry construction. Instead of assigning points only according to the forward ray order, our method extracts outermost surfaces from two complementary directions, *i.e.* the forward and backward ray directions. This design encourages more consistent layer assignments for occluded structures and reduces geometric fragmentation across neighboring rays. Algorithm 1

summarizes the procedure. Given a complete point set $\mathcal{Q}$, we iteratively extract the outermost points along the forward and backward directions, denoted as $\mathcal{Q}_k^{\text{fwd}}$ and $\mathcal{Q}_k^{\text{bwd}}$, respectively. The extracted points are assigned to the left and right layers $\hat{P}_i^l$ and $\hat{P}_i^r$ before removal from the remaining point set $\mathcal{Q}_k$. This process progressively constructs the layered representation from both directions simultaneously. The proposed bidirectional construction produces smoother and more coherent surfaces within each layer. Consequently, the network no longer needs to compensate for large intra-layer geometric shifts caused by inconsistent ray ordering. This leads to more stable geometric reasoning and improves reconstruction quality in heavily occluded regions. Importantly, the performance gains arise solely from the redesigned output representation without introducing additional network parameters or computational overhead. As shown in Fig. 2, our representation reconstructs more continuous surfaces for heavily occluded structures such as the bottom of the shoe. In contrast, conventional forward-only strategies split the same structure across multiple layers.

Algorithm 1: Bidirectional Ray Intersections

Input: Complete point set $\mathcal{Q}$, number of iterations K.
Output: Multi-layer point maps for view i: $\hat{P}_i = \{\hat{P}_i^n\}_{n=1}^{2K}$.
Initialize remaining points $\mathcal{Q}_0 \leftarrow \mathcal{Q}$;
Initialize layer indices $l \leftarrow 1, r \leftarrow 2K$;
for $k = 1$ *to* K **do**
  Extract outermost points along forward ray direction: $\mathcal{Q}_k^{\text{fwd}} \subset \mathcal{Q}_{k-1}$;
  Extract outermost points along backward ray direction: $\mathcal{Q}_k^{\text{bwd}} \subset \mathcal{Q}_{k-1}$;
  Assign $\mathcal{Q}_k^{\text{fwd}}$ to layer $\hat{P}_i^l$ and $\mathcal{Q}_k^{\text{bwd}}$ to layer $\hat{P}_i^r$;
  $\mathcal{Q}_k \leftarrow \mathcal{Q}_{k-1} \setminus \big(\mathcal{Q}_k^{\text{fwd}} \cup \mathcal{Q}_k^{\text{bwd}}\big)$;
  $l \leftarrow l + 1, r \leftarrow r - 1$;
end

3.2 Geometry-aware Feature Aggregation

Based on the proposed bidirectional ray intersections, LaRI3R predicts the final 3D geometry through a geometry-aware feature aggregation scheme that jointly integrates visual and geometric cues. Existing two-view models such as DUSt3R (Wang et al., 2024) primarily aggregate features through 2D visual correspondences. Although this strategy works well for visible surfaces, it becomes unreliable for occluded geometry because the same occluded 3D point may project to weak, missing, or visually dissimilar regions across views. To address this issue, LaRI3R performs feature aggregation directly in 3D space, where spatial consistency provides a more stable cue for associating points that belong to the same underlying surface.

Visual-based cross-view aggregation. We first adopt the ViT encoder-decoder architecture from DUSt3R (Wang et al., 2024) to extract dense visual features from the two input views. Given an image pair $\{I_0, I_1\}$, each image is encoded into patch tokens through a shared-weight ViT encoder. A Transformer decoder with cross-attention then exchanges information between the two views to aggregate visual correspondences:

$$\mathbf{F}_0^{\text{2d}} = \text{DPT}\Big(\text{Dec}\big(\text{Enc}(I_0), \text{Enc}(I_1)\big)\Big), \tag{1a}$$

$$\mathbf{F}_1^{\text{2d}} = \text{DPT}\Big(\text{Dec}\big(\text{Enc}(I_1), \text{Enc}(I_0)\big)\Big), \tag{1b}$$

where $\mathbf{F}_i^{\text{2d}} \in \mathbb{R}^{H \times W \times C'}$ denotes the dense 2D feature maps. We reshape these features into multi-layer visual features $\mathbf{F}_i^{\text{vis}} \in \mathbb{R}^{H \times W \times N \times C}$ and regress the initial bidirectional point maps through a shallow 3D convolutional head:

$$[P_i', O_i'] = \text{Head}\big(\mathbf{F}_i^{\text{vis}}\big), \tag{2}$$

where P_i' and O_i' are the initial point and confidence maps of view i.

Geometry-aware cross-view aggregation. Although the visual branch captures appearance correspondences across views, purely visual aggregation remains insufficient for occluded geometry reasoning. The projected features of occluded points often differ significantly across viewpoints due to varying visibility conditions. Consequently, visually similar regions do not always correspond to the same underlying 3D structure. To enable robust cross-view aggregation for occluded surfaces, we introduce a geometry-aware aggregation branch based on a hierarchical point transformer network $G_\theta(\cdot)$ (Wu et al., 2024c). Instead of associating features through only visual similarity, the network aggregates features according to explicit 3D positions. Given the initial point maps (P_0', P_1')

Table 1: **Quantitative comparison on GSO dataset.** Our method achieves the best performance among all compared approaches.

Method	CD ↓	FS@0.1 ↑	FS@0.05 ↑	FS@0.02 ↑
SF3D (Boss et al., 2025)	0.037	0.913	0.738	0.487
SPAR3D (Huang et al., 2025)	0.038	0.912	0.745	0.486
Trellis (Xiang et al., 2025)	0.024	0.973	0.881	0.637
Hunyuan3D (Lai et al., 2025)	0.024	0.970	0.900	0.633
DUSt3R (Wang et al., 2024)	0.059	0.802	0.640	0.382
RaySt3R (Duisterhof et al., 2025)	0.030	0.950	0.862	0.569
LaRI (Li et al., 2025)	0.025	0.966	0.894	0.643
Ours (224×224)	0.020	0.984	0.936	0.700
Ours (512×512)	**0.018**	**0.989**	**0.950**	**0.742**

and visual features $(\mathbf{F}_0^{\text{vis}}, \mathbf{F}_1^{\text{vis}})$, the geometric features are computed as:

$$(\mathbf{F}_0^{\text{geo}}, \mathbf{F}_1^{\text{geo}}) = G_\theta\Big([P'_0, P'_1], [\mathbf{F}_0^{\text{vis}}, \mathbf{F}_1^{\text{vis}}]\Big). \tag{3}$$

The point transformer jointly processes the point maps and features from both views in the same coordinate space. This design enables features from geometrically related but visually dissimilar regions to be aggregated directly in 3D space. As a result, the network learns more stable geometric associations for occluded structures. Finally, we concatenate the geometric features $\mathbf{F}_i^{\text{geo}}$ with the visual features $\mathbf{F}_i^{\text{vis}}$ and regress the final multi-layer point maps through a shallow 3D convolutional head:

$$[P_i, O_i] = \text{Head}\big([\mathbf{F}_i^{\text{vis}}, \mathbf{F}_i^{\text{geo}}]\big), \tag{4}$$

where P_i and O_i are the final point and confidence maps of view i.

Supervision signal. We adopt a scale-invariant 3D regression loss similar to DUSt3R (Wang et al., 2024) and extend it to multi-layer point maps. Given the predicted point map P and the bidirectional ray intersection ground-truth point map $\hat{P}$, the regression loss is defined as:

$$\mathcal{L}_{\text{regr}} = \sum_{(u,v)\in D} \left\| \frac{P(u,v)}{z} - \frac{\hat{P}(u,v)}{z_{\text{gt}}} \right\|_2, \tag{5}$$

where z and z_{gt} denote scale normalizers computed from the valid point set D. We further incorporate the predicted confidence map O to form a confidence-aware regression loss:

$$\mathcal{L}_{\text{conf}} = \sum_{(u,v)\in D} O(u,v)\, \mathcal{L}_{\text{regr}}(u,v) - \alpha \sum_{(u,v)\in D} \log O(u,v), \tag{6}$$

where α is a weighting factor. The final training objective is:

$$\mathcal{L}_{\text{final}} = \sum_{i=0}^{1} \Big(\mathcal{L}_{\text{conf}}(P'_i, O'_i) + \mathcal{L}_{\text{conf}}(P_i, O_i) \Big). \tag{7}$$

4 Experiments

In this section, we evaluate our method on both object-level and scene-level reconstruction tasks and perform extensive ablation studies to validate the effectiveness of each component. We further evaluate the potential of LaRI3R in downstream applications such as pose estimation.

4.1 Datasets

Training data. We use the pre-processed datasets provided by LaRI (Li et al., 2025), which are derived from Objaverse (Deitke et al., 2023), 3D-FRONT (Fu et al., 2021), and ScanNet++ (Yeshwanth et al., 2023). The training data include approximately 190K images from 16K Objaverse scenes (12 views each), 100K images from 18K 3D-FRONT scenes (6 views each), and 48K real-world images from ScanNet++.

Table 2: **Quantitative comparison on the 3D-FRONT and SCRREAM evaluation sets. Bold** and underlined numbers denote the best and second-best results.

	3D-FRONT						SCRREAM					
	Visible		Unseen		Overall		Visible		Unseen		Overall	
Method	CD ↓	FS@0.05 ↑	CD ↓	FS@0.05 ↑	CD ↓	FS@0.05 ↑	CD ↓	FS@0.05 ↑	CD ↓	FS@0.05 ↑	CD ↓	FS@0.05 ↑
Metric3D-v2	0.252	0.118	–	–	0.279	0.123	0.063	0.534	–	–	0.086	0.473
DepthPro	0.050	0.696	–	–	0.103	0.562	0.055	0.603	–	–	0.079	0.535
DUSt3R	0.061	0.651	–	–	0.116	0.525	0.059	0.653	–	–	0.086	0.565
MoGe	0.040	0.788	–	–	0.096	0.621	0.035	0.786	–	–	0.063	0.668
VGGT	0.038	0.847	–	–	0.089	0.630	0.028	0.863	–	–	0.058	0.726
Pi3	0.023	0.911	–	–	0.064	0.741	0.027	0.866	–	–	0.057	0.729
DAv3	**0.021**	**0.913**	–	–	0.062	0.744	**0.022**	**0.920**	–	–	0.053	**0.773**
CUT3R	0.093	0.397	0.271	0.164	0.146	0.314	0.071	0.658	0.192	0.238	0.091	0.543
LaRI	0.050	0.839	0.076	0.739	0.061	0.799	0.057	0.589	0.077	0.494	0.059	0.590
Ours	0.035	0.885	**0.047**	**0.836**	**0.041**	**0.868**	0.043	0.743	**0.064**	**0.589**	**0.049**	0.704

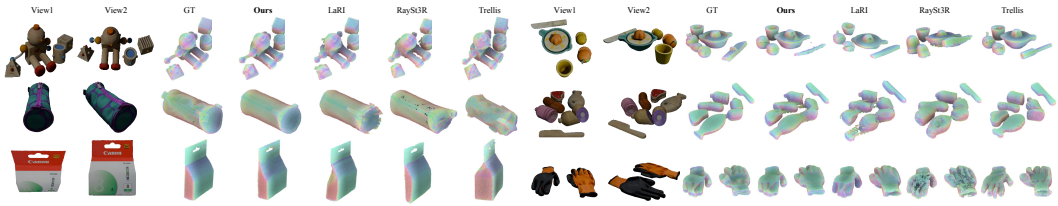

Figure 4: **Reconstruction results on GSO dataset.** Our method predicts results that are more consistent with the input image than existing methods.

Testing data. We use the GSO (Downs et al., 2022) dataset for object-level evaluation, which contains 1,030 objects, each rendered from 12 viewpoints. In addition, we randomly select approximately 2K images from 100 scenes in the Omni3D (Brazil et al., 2023) dataset to evaluate relative camera pose estimation. For scene-level evaluation, we conduct experiments on the validation set of 3D-FRONT and the SCRREAM (Jung et al., 2024) dataset.

Constructing image pairs. Following CroCo (Weinzaepfel et al., 2023), we use ground-truth depth to identify geometrically consistent points across views and select image pairs whose overlap exceeds a predefined threshold. Details are provided in Appendix A.3.

4.2 Evaluation Metrics

We adopt two commonly used metrics, Chamfer Distance (CD) and F-score (FS) to evaluate the multi-layer point maps. Chamfer Distance measures the average closest-point distance between the predicted and ground-truth point sets, which provides a global measure of geometric discrepancy. The F-score evaluates the proportion of predicted points that lie within a specified distance threshold of the ground truth to capture both precision and recall. For the predicted poses, we compute the angular errors in rotation and translation and report the area under the curve (AUC) over thresholds of $\tau = 5°, 20°$, computed on the minimum of the rotation and translation errors.

4.3 Implementation Details

We adopt the ViT encoder–decoder architecture from DUSt3R (Wang et al., 2024) as the backbone of the visual branch and initialize it with the corresponding pre-trained weights. For the geometric branch, we build a network based on PTv3 (Wu et al., 2024c), initialized with self-supervised pre-trained weights from Sonata (Wu et al., 2025b). We train object-level and scene-level models independently using a two-stage training procedure. In the first stage, we train on 224×224 images for 50 epochs, using 200K image pairs for the object-level and scene-level datasets. In the second stage, we fine-tune the models on high-resolution 512×512 images using 100K image pairs. During training, we select image pairs with an overlap ratio between 0.3 and 0.8. The learning rate is set to 1×10^{-4}, with a warm-up period of 15 epochs.

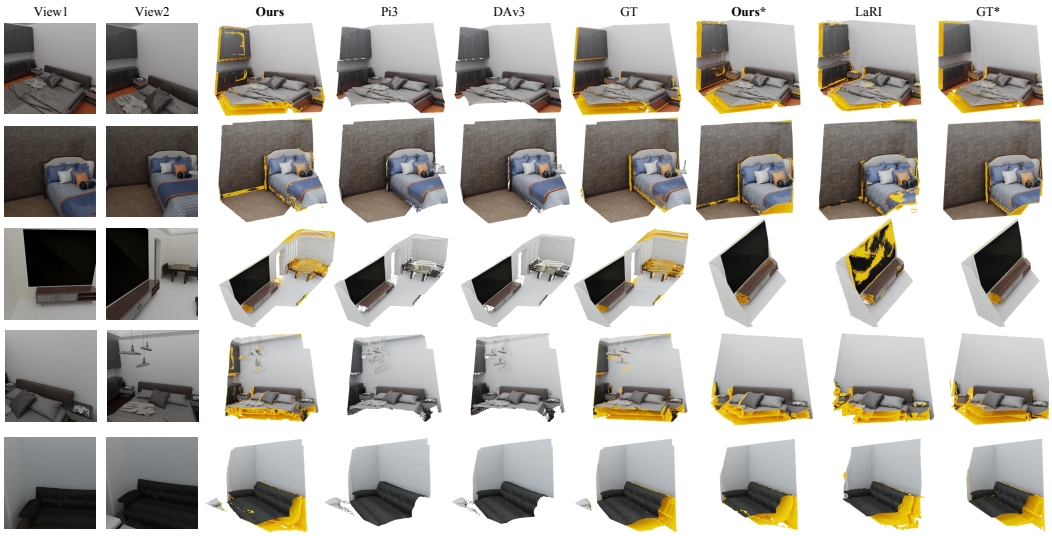

Figure 5: **Qualitative comparisons on scene-level reconstruction.** Our method accurately reconstructs visible surfaces while producing more plausible geometry in occluded regions. (The highlighted points indicate occluded regions, "*" denotes the geometry reconstructed from View 1.)

4.4 Benchmark Comparisons

Reconstruction performance. We first evaluate our method on the GSO dataset and compare it against recent state-of-the-art approaches, including generative (Xiang et al., 2025; Boss et al., 2025; Huang et al., 2025; Lai et al., 2025), iterative (Duisterhof et al., 2025), and layered (Li et al., 2025) reconstruction methods. In addition, we compare with DUSt3R (Wang et al., 2024), which focuses on visible surface reconstruction. For methods that produce reconstructions in a canonical coordinate system (Xiang et al., 2025; Boss et al., 2025; Huang et al., 2025; Lai et al., 2025), we align the predicted geometry to the ground truth via a brute-force search (Li et al., 2025) followed by ICP registration (Chetverikov et al., 2002). For view-aligned reconstruction methods (Duisterhof et al., 2025; Li et al., 2025; Wang et al., 2024), predictions are aligned using a shift-and-scale transformation. For each input image, we construct image pairs by selecting another view with a low overlap ratio in the range of $(0.1, 0.3)$. We report the reconstruction results from view I_0 to align with monocular methods and analyze the fusion of different image branch results in subsequent ablation studies. As shown in Tab. 1, our method achieves the best overall performance, with a 28% reduction in Chamfer Distance and a 15% improvement in FS@0.02 compared to the previous state-of-the-art method (Li et al., 2025). Qualitative results shown in Fig. 4 indicate that our method reconstructs objects with more plausible and complete geometry.

Additionally, we evaluate different methods on the scene-level dataset, reporting metrics separately for seen, unseen, and overall performance. We compare our method with monocular (Wang et al., 2025c; Bochkovskii et al., 2024; Hu et al., 2024a) and multi-view (Wang et al., 2024; 2025a;b;e; Lin et al., 2025) reconstruction approaches, as well as the method (Li et al., 2025) specifically designed for occluded geometry estimation. As shown in Tab. 2, our method consistently outperforms the compared methods on 3D-FRONT across unseen and overall regions. On SCRREAM, it achieves the lowest Chamfer distance in both unseen and overall regions, demonstrating the effectiveness in recovering unobserved geometry in real-world scenes. Fig. 5 provides qualitative comparisons, which show that our method maintains reasonable visible-surface reconstruction while producing more coherent occluded structures for large scene elements such as cabinets, beds, and walls, compared to LaRI (Li et al., 2025).

Relative camera pose estimation. As LaRI3R predicts additional occluded geometry that extends beyond visible surfaces, we further evaluate its practical benefits in downstream tasks such as camera pose estimation. Following DUSt3R, we estimate the relative pose between two views by aligning their predicted multi-layer point maps with the *occluded geometry* included to augment the geometric information. Given corresponding points from the source and target views, P_i^{src} and P_i^{tgt}, we solve for the optimal rotation R, translation t, and scale σ by minimizing the squared error:

$$R^*, t^*, \sigma^* = \arg\min_{R,t,\sigma} \sum_i \left\| \sigma(R P_i^{\text{src}} + t) - P_i^{\text{tgt}} \right\|_2^2. \tag{8}$$

Table 3: **Relative camera pose evaluation.** Using LaRI3R's visible and occluded point maps in pose solving outperforms feedforward methods and visible-only geometry.

Method	GSO		Omni3D	
	AUC@5 ↑	AUC@20 ↑	AUC@5 ↑	AUC@20 ↑
DUSt3R	0.049	0.325	0.051	0.332
Reloc3R	0.050	0.302	0.123	0.573
VGGT	0.266	0.737	0.141	0.672
Ours (vis. only)	0.221	0.642	0.209	0.688
Ours (vis. + occ.)	**0.342**	**0.744**	**0.210**	**0.694**

Table 4: **Ablation on GSO dataset.** BRI and Geo. Agg. denote the Bidirectional Ray Intersection and Geometry-aware Aggregation, respectively. Each module contributes to improvements over the baseline to give the best overall performance.

	Baseline	BRI	Geo. Agg.	CD ↓	FS@0.05 ↑	FS@0.02 ↑
(a)	✓			0.0226	0.909	0.655
(b)	✓	✓		0.0206	0.929	0.685
(c)	✓	✓	✓	**0.0200**	**0.936**	**0.700**

To improve robustness against noise and outliers, a more robust alternative is to use a PnP (Hartley & Zisserman, 2003; Lepetit et al., 2009) solver within RANSAC (Fischler & Bolles, 1981). The results are shown in Tab. 3. Our method demonstrates clear advantages over feedforward approaches. Moreover, as shown in the last two rows, incorporating 3D points from occluded layers further improves performance. This demonstrates that the multi-layer representation provides richer geometric cues beyond visible surfaces.

4.5 Ablation studies

In this section, we conduct experiments to validate the effectiveness of each module. All experiments are performed on 224×224 images. The quantitative results are summarized in Tab. 4. Following LaRI, our baseline employs the DUSt3R backbone to regress unidirectional multi-layer point maps. BRI and Geo. Agg. denote the proposed Bidirectional Ray Intersection (Sec.3.1) and Geometry-aware Feature Aggregation (Sec.3.2), respectively. The introduction of the Bidirectional Ray Intersection Representation produces more coherent multi-layer point maps than the unidirectional baseline and yields clear improvements in both **(a)** and **(b)**. The Geometry-aware Feature Aggregation further enhances feature integration across viewpoints to give more accurate reasoning of multi-layer geometry **(c)**.

Performance on different branches. We evaluate the reconstruction performance of the point maps P_0 and P_1 predicted by the individual branches corresponding to I_0 and I_1, respectively, as well as the results obtained by directly concatenating the point clouds from both views. As shown in Tab. 5, predictions from view I_0 achieve higher accuracy than those from I_1 because the latter requires learning an additional view transformation that can reduce reconstruction accuracy (Wang et al., 2025e). Moreover, since the two views provide complementary information, fusing their predicted point clouds further improves reconstruction performance.

Table 5: **Performance on different branches.**

Settings	CD ↓	FS@0.1 ↑	FS@0.05 ↑	FS@0.02 ↑
View I_0	0.0200	0.984	0.936	0.700
View I_1	0.0226	0.974	0.912	0.658
Fusion	**0.0192**	**0.985**	**0.942**	**0.732**

Table 6: **Ablation on different pair overlap.**

Overlap	CD ↓	FS@0.1 ↑	FS@0.05 ↑	FS@0.02 ↑
Low	0.0200	0.984	0.936	0.700
Medium	**0.0188**	**0.988**	**0.944**	**0.720**
High	0.0196	0.984	0.936	0.709

Effect of different overlap levels. We evaluate reconstruction performance under varying degrees of overlap between image pairs. Specifically, we categorize the overlap into three levels: low (IoU in $[0.1, 0.3]$), medium (IoU in $[0.3, 0.6]$), and high (IoU > 0.6). Tab. 6 shows that medium overlap achieves the best results.

5 Conclusion

We introduce LaRI3R, a two-view 3D reconstruction framework that leverages a bidirectional layered representation and geometry-aware aggregation to jointly reason about visible and occluded geometry. Experiments on both object- and scene-level benchmarks show that LaRI3R produces more coherent and consistent multi-layer point maps and outperforms existing single-view and multi-view reconstruction methods. **Limitation**: LaRI3R is limited by its fixed number of layers, which may be insufficient for complex scenes. The ray-based representation can become sparse around grazing-angle surfaces. Future work could explore adaptive layer allocation to better handle varying geometric complexity. Moreover, the reconstructed visible and occluded geometry could serve as a richer geometric basis for generative models to further complete unobserved regions.

AI USE STATEMENT

We used generative AI tools (large language models) to assist with manuscript editing, figure layout and presentation. We did not use generative AI tools for research methodology, experiment design, method implementation, or result interpretation. All AI-assisted content was reviewed and verified by the authors.

REFERENCES

Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. *arXiv preprint arXiv:2410.02073*, 2024.

Mark Boss, Zixuan Huang, Aaryaman Vasishta, and Varun Jampani. Sf3d: Stable fast 3d mesh reconstruction with uv-unwrapping and illumination disentanglement. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 16240–16250, 2025.

Garrick Brazil, Abhinav Kumar, Julian Straub, Nikhila Ravi, Justin Johnson, and Georgia Gkioxari. Omni3d: A large benchmark and model for 3d object detection in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13154–13164, 2023.

Chenjie Cao, Xinlin Ren, and Yanwei Fu. Mvsformer: Learning robust image representations via transformers and temperature-based depth for multi-view stereo. *arXiv preprint arXiv:2208.02541*, 2022.

Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. Easi3r: Estimating disentangled motion from dust3r without training. *arXiv preprint arXiv:2503.24391*, 2025a.

Zhuoguang Chen, Minghui Qin, Tianyuan Yuan, Zhe Liu, and Hang Zhao. Long3r: Long sequence streaming 3d reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5273–5284, 2025b.

Dmitry Chetverikov, Dmitry Svirko, Dmitry Stepanov, and Pavel Krsek. The trimmed iterative closest point algorithm. In *2002 International Conference on Pattern Recognition*, volume 3, pp. 545–548. IEEE, 2002.

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13142–13153, 2023.

Kai Deng, Zexin Ti, Jiawei Xu, Jian Yang, and Jin Xie. Vggt-long: Chunk it, loop it, align it–pushing vggt's limits on kilometer-scale long rgb sequences. *arXiv preprint arXiv:2507.16443*, 2025a.

Weijian Deng, Dylan Campbell, Chunyi Sun, Jiahao Zhang, Shubham Kanitkar, Matt E Shaffer, and Stephen Gould. Pos3r: 6d pose estimation for unseen objects made easy. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 16818–16828, 2025b.

Yikang Ding, Wentao Yuan, Qingtian Zhu, Haotian Zhang, Xiangyue Liu, Yuanjiang Wang, and Xiao Liu. Transmvsnet: Global context-aware multi-view stereo network with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8585–8594, 2022.

Siyan Dong, Shuzhe Wang, Shaohui Liu, Lulu Cai, Qingnan Fan, Juho Kannala, and Yanchao Yang. Reloc3r: Large-scale training of relative camera pose regression for generalizable, fast, and accurate visual localization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 16739–16752, 2025.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, pp. 2553–2560. IEEE, 2022.

Bardienus P Duisterhof, Jan Oberst, Bowen Wen, Stan Birchfield, Deva Ramanan, and Jeffrey Ichnowski. Rayst3r: Predicting novel depth maps for zero-shot object completion. *arXiv preprint arXiv:2506.05285*, 2025.

Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24 (6):381–395, 1981.

Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, et al. 3d-front: 3d furnished rooms with layouts and semantics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10933–10942, 2021.

Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2495–2504, 2020.

Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

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. *arXiv preprint arXiv:2311.04400*, 2023.

Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024a.

Tao Hu, Wenhang Ge, Yuyang Zhao, and Gim Hee Lee. X-ray: A sequential 3d representation for generation. *Advances in Neural Information Processing Systems*, 37:136193–136219, 2024b.

Zixuan Huang, Mark Boss, Aaryaman Vasishta, James M Rehg, and Varun Jampani. Spar3d: Stable point-aware reconstruction of 3d objects from single images. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 16860–16870, 2025.

Shun Iwase, Katherine Liu, Vitor Guizilini, Adrien Gaidon, Kris Kitani, Rareş Ambruş, and Sergey Zakharov. Zero-shot multi-object scene completion. In *European Conference on Computer Vision*, pp. 96–113. Springer, 2024.

Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023.

HyunJun Jung, Weihang Li, Shun-Cheng Wu, William Bittner, Nikolas Brasch, Jifei Song, Eduardo Pérez-Pellitero, Zhensong Zhang, Arthur Moreau, Nassir Navab, et al. Scrream: Scan, register, render and map: A framework for annotating accurate and dense 3d indoor scenes with a benchmark. *Advances in Neural Information Processing Systems*, 37:44164–44176, 2024.

Zeqiang Lai, Yunfei Zhao, Haolin Liu, Zibo Zhao, Qingxiang Lin, Huiwen Shi, Xianghui Yang, Mingxin Yang, Shuhui Yang, Yifei Feng, et al. Hunyuan3d 2.5: Towards high-fidelity 3d assets generation with ultimate details. *arXiv preprint arXiv:2506.16504*, 2025.

Yushi Lan, Fangzhou Hong, Shuai Yang, Shangchen Zhou, Xuyi Meng, Bo Dai, Xingang Pan, and Chen Change Loy. Ln3diff: Scalable latent neural fields diffusion for speedy 3d generation. In *European Conference on Computer Vision*, pp. 112–130. Springer, 2024.

Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. EPnP: An accurate O(n) solution to the PnP problem. *International journal of computer vision*, 81(2):155–166, 2009.

Rui Li, Dong Gong, Wei Yin, Hao Chen, Yu Zhu, Kaixuan Wang, Xiaozhi Chen, Jinqiu Sun, and Yanning Zhang. Learning to fuse monocular and multi-view cues for multi-frame depth estimation in dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21539–21548, 2023a.

Rui Li, Danna Xue, Shaolin Su, Xiantuo He, Qing Mao, Yu Zhu, Jinqiu Sun, and Yanning Zhang. Learning depth via leveraging semantics: Self-supervised monocular depth estimation with both implicit and explicit semantic guidance. *Pattern Recognition*, 137:109297, 2023b.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9848–9858, 2024a.

Rui Li, Biao Zhang, Zhenyu Li, Federico Tombari, and Peter Wonka. Lari: Layered ray intersections for single-view 3d geometric reasoning. *arXiv preprint arXiv:2504.18424*, 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9087–9098, 2023c.

Zhenyu Li, Mykola Lavreniuk, Jian Shi, Shariq Farooq Bhat, and Peter Wonka. Amodal depth anything: Amodal depth estimation in the wild. *arXiv preprint arXiv:2412.02336*, 2024b.

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. *arXiv preprint arXiv:2511.10647*, 2025.

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. *Advances in Neural Information Processing Systems*, 36:22226–22246, 2023a.

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9298–9309, 2023b.

Jiahao Lu, Tianyu Huang, Peng Li, Zhiyang Dou, Cheng Lin, Zhiming Cui, Zhen Dong, Sai-Kit Yeung, Wenping Wang, and Yuan Liu. Align3r: Aligned monocular depth estimation for dynamic videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22820–22830, 2025.

Qing Mao, Rui Li, Yu Zhu, Jinqiu Sun, and Yanning Zhang. Enhancing 3d-2d representations for convolution occupancy networks. *Pattern Recognition*, 134:109097, 2023.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12179–12188, 2021.

Johannes L Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4104–4113, 2016.

Jonathan Shade, Steven Gortler, Li-wei He, and Richard Szeliski. Layered depth images. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, pp. 231–242, 1998.

Daeyun Shin, Zhile Ren, Erik B Sudderth, and Charless C Fowlkes. 3d scene reconstruction with multi-layer depth and epipolar transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2172–2182, 2019.

Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision*, pp. 1–18. Springer, 2024.

Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In *European Conference on Computer Vision*, pp. 439–457. Springer, 2024.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5294–5306, 2025a.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10510–10522, 2025b.

Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5261–5271, 2025c.

Shizun Wang, Zhenxiang Jiang, Xingyi Yang, and Xinchao Wang. C4d: 4d made from 3d through dual correspondences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7570–7580, 2025d.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20697–20709, 2024.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. *arXiv preprint arXiv:2507.13347*, 2025e.

Philippe Weinzaepfel, Thomas Lucas, Vincent Leroy, Yohann Cabon, Vaibhav Arora, Romain Brégier, Gabriela Csurka, Leonid Antsfeld, Boris Chidlovskii, and Jérôme Revaud. Croco v2: Improved cross-view completion pre-training for stereo matching and optical flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17969–17980, 2023.

Felix Wimbauer, Nan Yang, Christian Rupprecht, and Daniel Cremers. Behind the scenes: Density fields for single view reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9076–9086, 2023.

Chao-Yuan Wu, Justin Johnson, Jitendra Malik, Christoph Feichtenhofer, and Georgia Gkioxari. Multiview compressive coding for 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9065–9075, 2023.

Jiang Wu, Rui Li, Haofei Xu, Wenxun Zhao, Yu Zhu, Jinqiu Sun, and Yanning Zhang. Gomvs: Geometrically consistent cost aggregation for multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20207–20216, 2024a.

Jiang Wu, Rui Li, Yu Zhu, Rong Guo, Jinqiu Sun, and Yanning Zhang. Sparse2dgs: Geometry-prioritized gaussian splatting for surface reconstruction from sparse views. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 11307–11316, 2025a.

Kailu Wu, Fangfu Liu, Zhihan Cai, Runjie Yan, Hanyang Wang, Yating Hu, Yueqi Duan, and Kaisheng Ma. Unique3d: High-quality and efficient 3d mesh generation from a single image. *Advances in Neural Information Processing Systems*, 37:125116–125141, 2024b.

Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler, faster, stronger. In *CVPR*, 2024c.

Xiaoyang Wu, Daniel DeTone, Duncan Frost, Tianwei Shen, Chris Xie, Nan Yang, Jakob Engel, Richard Newcombe, Hengshuang Zhao, and Julian Straub. Sonata: Self-supervised learning of reliable point representations. In *CVPR*, 2025b.

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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 21469–21480, 2025.

Qingshan Xu, Wanjuan Su, Yuhang Qi, Wenbing Tao, and Marc Pollefeys. Learning inverse depth regression for pixelwise visibility-aware multi-view stereo networks. *International Journal of Computer Vision*, 130(8):2040–2059, 2022.

Xingguang Yan, Liqiang Lin, Niloy J Mitra, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Shapeformer: Transformer-based shape completion via sparse representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6239–6249, 2022.

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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 21924–21935, 2025.

Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 767–783, 2018.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12–22. IEEE, 2023.

Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024.

Bowen Zhang, Yiji Cheng, Jiaolong Yang, Chunyu Wang, Feng Zhao, Yansong Tang, Dong Chen, and Baining Guo. Gaussiancube: Structuring gaussian splatting using optimal transport for 3d generative modeling. *CoRR*, 2024a.

Jingyang Zhang, Shiwei Li, Zixin Luo, Tian Fang, and Yao Yao. Vis-mvsnet: Visibility-aware multi-view stereo network. *International Journal of Computer Vision*, 131(1):199–214, 2023.

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. *arXiv preprint arXiv:2410.03825*, 2024b.

A Appendix

A.1 Ablation on Number of Layers

We investigate the impact of the number of layers in the bidirectional representation on performance. Tab. 7 shows the performance saturates at four layers. Notably, even two layers already outperform the baseline since the bidirectional representation captures the outer surfaces of the object from both forward and backward directions. This better preserves the overall contour and reduces surface splitting for more accurate reconstruction.

Table 7: **Effect of the number of layers on reconstruction performance.** Performance saturates at four layers, while even a two-layer bidirectional representation surpasses the baseline by capturing the object's outer surfaces from both directions.

Method	Layer	CD ↓	FS@0.1 ↑	FS@0.05 ↑	FS@0.02 ↑
Baseline	5	0.0226	0.977	0.909	0.655
Ours	2	0.0210	0.984	0.924	0.677
Ours	4	0.0206	0.984	0.929	0.685
Ours	6	0.0206	0.984	0.929	0.686

A.2 Influence of Visible and Unobserved Regions Across Views

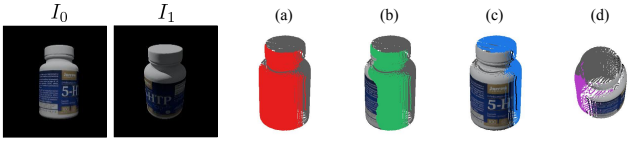

Figure 6: **Visualization of the four view-conditioned visibility regions.** The highlighted areas in the figure correspond to: (a) points visible in I_0, (b) points jointly visible in both I_0 and I_1, (c) points occluded in I_0 but visible in I_1, and (d) points occluded in both views.

We evaluate reconstruction performance across visible and occluded regions under two-view observations. Specifically, we partition the ground-truth point cloud of image I_0 into four regions as shown in Fig. 6: (a) points visible from I_0, (b) points jointly visible from both I_0 and I_1, (c) points occluded in I_0 but visible in I_1, and (d) points occluded in both views. We evaluate reconstruction performance in the four visibility regions for the reconstruction from viewpoint I_0, with quantitative results shown in Table 8. The regions visible in I_0 achieve higher accuracy than occluded regions, and the overlapping region that is (b) visible in both I_0 and I_1 achieves the highest performance. Points that are occluded in I_0 but visible in I_1 exhibit better reconstruction than points occluded in both views, which indicate the effect of complementary observations across views.

Table 8: **Reconstruction performance in different visibility regions.** Quantitative results for reconstruction from viewpoint I_0 across the four view-conditioned visibility regions defined in Fig. 6. The region jointly visible in both I_0 and I_1 achieves the highest reconstruction performance.

Region	CD ↓	FS@0.1 ↑	FS@0.05 ↑	FS@0.02 ↑
Points visible in I_0	0.0211	**0.977**	0.915	0.687
Points jointly visible in both I_0 and I_1	**0.0205**	**0.977**	**0.918**	**0.694**
Points occluded in I_0 but visible in I_1	0.0241	0.968	0.889	0.642
Points occluded in both views	0.0276	0.955	0.858	0.607

A.3 Image Pair Construction Details

We construct image pairs by leveraging shared information between two views. Given two images I_i and I_j with depth maps Z_i and Z_j and camera intrinsics K_i and K_j, each pixel (u, v) in I_i with depth $Z_i(u, v)$ is projected onto the image plane of the second camera using the relative pose (R_{ij}, t_{ij}):

$$\mathbf{x}_i^j = \Pi\big(R_{ij}\, Z_i(u,v)\, K_i^{-1} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} + t_{ij}\big), \tag{9}$$

where $\Pi(\cdot)$ denotes the perspective projection. The depth at $\mathbf{x}_i^j$ is sampled from Z_j, denoted as $\tilde{Z}_j(\mathbf{x}_i^j)$. This sampled depth is then projected back to the first camera to give the reprojected coordinates $\tilde{\mathbf{x}}_i$ and reprojected depth $\tilde{Z}_i$. The corresponding points are identified by checking both coordinate and depth consistency:

$$\|\tilde{\mathbf{x}}_i - \mathbf{x}_i\|_2 < \epsilon_{2D}, \quad |\tilde{Z}_i - Z_i(u,v)| < \epsilon_Z, \tag{10}$$

where ϵ_{2D} and ϵ_Z are thresholds for coordinate and depth consistency, respectively. We measure the fraction of pixels in one view that have consistent correspondences in the other view (Fig. 7), and similarly for the reverse direction. The minimum of these two fractions is taken as the overlap score, which reflects the shared 3D content between the views.

Table 9: **Comparison with Trellis.** Our method achieves better reconstruction quality while reducing GPU memory usage and inference time.

Method	CD ↓	FS@0.05 ↑	FS@0.02 ↑	Memory ↓	Time ↓
Trellis	0.024	0.881	0.637	13.0 GB	3.2 s
Ours	**0.018**	**0.950**	**0.742**	**6.0 GB**	**0.5 s**



Figure 7: Illustration of image pairs, where red regions denote overlapping pixels with consistent correspondences between the two views.

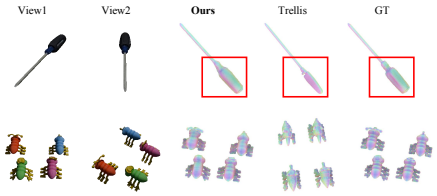

Figure 8: Visual comparison with Trellis.

A.4 Comparison with Generative Method

We compare our method with the generative method Trellis, which natively supports multi-view inputs, in terms of reconstruction quality, memory consumption, and inference efficiency. As shown in Tab.9, our method achieves better performance while reducing GPU memory usage by 54% and inference time by 84%. Qualitative comparisons are presented in Fig. 8. Generative methods tend to produce geometrically plausible shapes, but they are prone to hallucinations. This leads to inconsistencies between the reconstructed geometry and the input images. In contrast, although they may introduce noisy geometry in challenging regions, reconstruction methods preserve image consistency more faithfully. These observations suggest that combining generative priors with reconstruction-based constraints could be a promising direction for achieving more complete and reliable geometry estimation in future work.