Reflect3r: Single-View 3D Stereo Reconstruction Aided by Mirror Reflections

Jing Wu Zirui Wang Iro Laina Victor Adrian Prisacariu
University of Oxford
{jing.wu, iro.laina, victor.prisacariu}@eng.ox.ac.uk, ryan@robots.ox.ac.uk

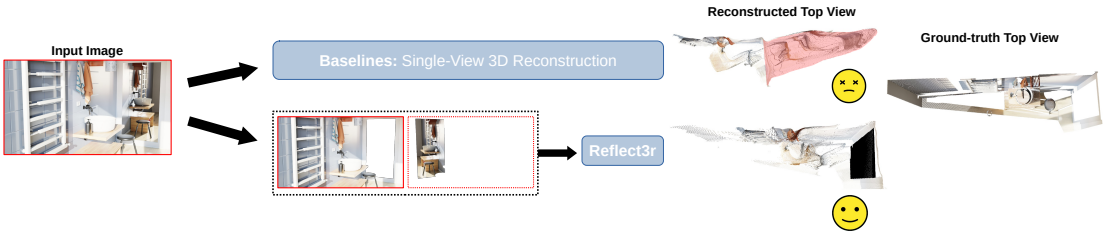

Figure 1. **Single-view 3D reconstruction with mirror reflections.** Given an image containing a mirror, we aim to reconstruct the 3D geometry of the scene. Existing methods cannot recognise the reflective cues and fail by predicting a false geometry for the mirror region, which is highlighted with light red (Top). We reinterpret the mirror reflection as a virtual view captured by a simulated camera, enabling a stereo formulation that leads to more accurate geometry reconstruction (Bottom).

Abstract

Mirror reflections are common in everyday environments and can provide stereo information within a single capture, as the real and reflected virtual views are visible simultaneously. We exploit this property by treating the reflection as an auxiliary view and designing a transformation that constructs a physically valid virtual camera, allowing direct pixel-domain generation of the virtual view while adhering to the real-world imaging process. This enables a multi-view stereo setup from a single image, simplifying the imaging process, making it compatible with powerful feed-forward reconstruction models for generalizable and robust 3D reconstruction. To further exploit the geometric symmetry introduced by mirrors, we propose a symmetric-aware loss to refine pose estimation. Our framework also naturally extends to dynamic scenes, where each frame contains a mirror reflection, enabling efficient per-frame geometry recovery. For quantitative evaluation, we provide a fully customizable synthetic dataset of 16 Blender scenes, each with ground-truth point clouds and camera poses. Extensive experiments on real-world data and synthetic data are conducted to illustrate the effectiveness of our method. Code and Blender scenes are available in: https://github.com/jingwu2121/reflect3r.git.

1. Introduction

Mirror reflections are ubiquitous in our daily environments. They allow multi-view cues, comprising the real scene and its mirrored counterparts, to be inferred from a single capture, naturally forming an epipolar geometry. Unlike a classical multi-view setup, the real–virtual view pairs share the same intrinsic parameters, which simplifies calibration. This property can significantly reduce hardware requirements, removing the need for cross-camera synchronization and shortening scanning time.

Effectively leveraging reflections for 3D reconstruction is non-trivial, as it involves a multi-view relationship derived from only a single image. Early mirror-based methods [24, 26] analytically reconstructed simple polyhedral shapes, but relied on highly controlled settings and thus do not generalize to real-world scenes with complex shapes, materials, and occlusions. More recently, Kawahara et al. [28] explored using water reflections as indirect views by explicitly modeling light transport to recover scene geometry and appearance, reconstructing 3D point coordinates via standard stereo triangulation from correspondences in the direct–reflected pairs. Fang et al. [9] explored human pose reconstruction by leveraging mirror reflections of human bodies, relying on 2D correspondence keypoints for initialisation. However, both approaches are inflexible and degrade significantly when the reflected view has a large angular difference from the real view, as reliable correspondences become sparse. Moreover, they are tailored to highly specific scenarios, such as water reflections of outdoor buildings, simplified objects, or the human body, and lack the ability to generalize to diverse, real-world scenes. Recent feed-forward reconstruction models [17, 32, 34]

achieve impressive accuracy, robustness, and generalisation across diverse settings, yet they lack awareness of reflected virtual views. As a result, they fail to exploit reflections in single-view reconstruction and are often confused by them in this monocular setup. To address this, we explicitly integrate reflection-derived virtual views into feed-forward reconstruction models, treating the problem as a stereo setup, formulating the problem as *reflection-aided single-view 3D reconstruction*.

The goal of this stereo configuration is to reconstruct the full 3D geometry of the scene from a single RGB image capture containing a mirror, leveraging the inherent multi-view constraints introduced by mirror reflections. Our key insight is to **reinterpret reflected virtual views as auxiliary views that contribute complementary geometric and appearance cues**. This is non-trivial, as the reflected views need first to be transformed into physically valid virtual views and then combined with the real view to form a consistent multi-view setup.

Motivated by this, we propose **Reflect3r**, a reflection-aided single-image 3D reconstruction framework capable of operating at the scene level in unconstrained environments. As illustrated in Fig. 1, Reflect3r substantially improves the coverage of 3D reconstruction. It recovers 3D geometry from a single image containing a mirror by treating the image as comprising two complementary views separated by the mirror surface.

To achieve this, we design a multi-view setup configuration process that operates directly in the pixel domain to create a valid virtual view, simulating the real-world imaging process. In real-world scenarios, the real and reflected virtual views often exhibit large angular differences, making reconstruction challenging. Recent large-scale reconstruction models [34] have demonstrated a strong ability to recover geometry even under sparse or wide-angle inputs. Building on this, Reflect3r leverages DUSt3R, which directly predicts dense geometry from multi-view inputs. However, DUSt3R alone fails in the mirror geometry prediction and often hallucinates depth within reflective regions.

To mitigate this, we provide DUSt3R with the pre-built multi-view setup, helping it better capture the spatial configuration. This stereo formulation treats the reflection as an auxiliary camera and introduces a natural symmetry constraint. To exploit this constraint, we design a symmetric-aware loss for post-optimization, refining the estimated camera poses by enforcing symmetry between the real and virtual poses with respect to the mirror plane.

Furthermore, our formulation naturally extends to dynamic scenes: given a video in which each frame contains a mirror reflection, Reflect3r can reconstruct per-frame geometry efficiently and at low capture cost.

Finally, existing 3D datasets with mirrors lack ground-truth for the virtual views. To enable quantitative evaluation, we construct a fully customizable synthetic dataset in Blender [5], consisting of 16 manually created scenes with ground-truth point clouds and real/virtual camera poses. The contributions of this work are as follows:

- We reinterpret mirror reflections as auxiliary viewpoints, enabling a formulation of *single-view stereo* for 3D reconstruction.
- We design a multi-view configuration process that works directly in the pixel domain to create a valid virtual view consistent with the physical imaging process.
- We exploit the inherent symmetry of mirrored scenes and propose a symmetric-aware loss to refine pose estimation.
- To facilitate quantitative evaluation, we contribute a synthetic dataset with 16 ground-truth point clouds and poses, tailored for this task, fully customizable and easily extendable for future research.

Extensive experiments demonstrate that Reflect3r reconstructs a more complete 3D geometry from a single image cue compared to all baselines.

2. Related Works

2.1. Computer Vision with Mirror

Mirror Detection and Segmentation. Mirrors have long been treated as challenging elements in image understanding due to their ambiguous visual cues and deceptive appearances. In static image analysis, several works have focused on detecting and segmenting mirrors. DAM [37], Progressive Mirror Detection [18], and PDNet [21] incorporate geometric and semantic priors to distinguish mirrors from ordinary surfaces. For video, Warren et al. [35] and Lau et al. [38] extend mirror segmentation into the temporal domain, leveraging motion consistency to improve accuracy.

Mirror Generation. Beyond detection, recent generative models have explored how to synthesize realistic reflections. MirrorVerse v1 and v2 [7, 8] tackle this by modelling scene semantics to generate plausible mirror content during image generation.

Mirror Geometry and Depth Estimation. In 3D vision, mirrors are particularly problematic due to their impact on geometry and depth estimation. Classical approaches often misinterpret mirror reflections as real geometry, leading to incorrect scene understanding. To address this in laser scans, [39] proposed a sensor fusion technique for dealing with LiDAR sensor failures on mirror and glass surfaces. [13] detects mirrors and corrects laser-scanned point clouds based on heuristics using known mirror dimensions. Mirror3D [29] corrected depth estimates in mirrored regions of both raw scanned and estimated depth by estimating the 3D mirror plane based on RGB input and surrounding depth context.

Novel View Synthesis. In the context of novel view synthesis and 3D reconstruction, modeling reflective materials has gained attention. Works such as Ref-NeRF [31], NeRFReN [12], NeRSP [41], and 3DGS-DR [15] extend radiance fields [23] and 3D Gaussian Splatting [14] to better handle reflections in surfaces like glass and metal. Closer to our task, MirrorGaussian [19], Mirror-NeRF [42], and MS-NeRF [40] explicitly target mirrors in 3D, aiming to reconstruct scenes with accurate mirror appearance and geometry. [28] explores the 3D reconstruction from water reflections.

Most prior works treat reflections as noise, and attempts to leverage mirrors for reconstruction remain limited, despite the fact that reflections inherently provide a second viewpoint of the scene. In contrast, our work incorporates mirror reflections by reinterpreting them as views captured by virtual cameras, reformulating the single-image cue as a multi-view setup, and enabling more complete 3D reconstruction even in unconstrained settings.

2.2. Single-Image-to-3D Reconstruction

Single-image-to-3D methods can be broadly divided into two lines of work.

Generative. The first line recovers 3D geometry through generation by hallucinating multi-view images of a 3D model and reconstructing the scene from the synthesized dense views. Early approaches, such as RealFusion [22] and Make-It-3D [30], adopt a per-scene optimization strategy, distilling prior knowledge from 2D generative models into 3D representations via the Score Distillation Sampling (SDS) loss [27]. While initially designed for object-level reconstruction, this paradigm has been extended to scene-level synthesis in works like Reconfusion [36]. More recent methods [10, 16, 20, 44] bypass per-scene optimization by directly training diffusion models that, conditioned on an input view and target camera parameters, predict novel views which are then used for 3D reconstruction.

Reconstruction Aided by Reflection. The second line of work reconstructs 3D geometry directly from the given information without hallucination. Early mirror-based methods [24, 26] reconstructed simple polyhedral shapes under highly controlled conditions, but it is unable to generalise to real-world scenes with complex geometry and materials. Kawahara et al. [28] and Fang et al. [9] use the 2D correspondences between the real and the reflected water/mirror view to reconstruct a scene or a human body, respectively. However, these methods are inflexible and degrade significantly when the reflected view has a large angular difference from the real view, as reliable correspondences become sparse. Moreover, they are tailored to highly specific scenarios, such as water reflections of outdoor buildings, simplified objects, or the human body, and cannot generalize to diverse, real-world scenes.

In contrast, our work extends image-to-3D reconstruction by leveraging real visual cues, specifically reflections, treating them as auxiliary viewpoints to enhance single-image 3D reconstruction without relying on hallucinated geometry. Furthermore, we reformulate a valid multi-view setup by simulating the physical imaging process, ensuring compatibility with the backbone model and improving the quality of single-image reconstruction, thereby enabling generalization to real-world scenes.

3. Method

Given an input image $I \in \mathbb{R}^{H\times W\times 3}$ containing mirrors $\mathcal{F} = \{F_1, F_2, \dots\}$, our goal is to reconstruct the 3D geometry of the scene as a point cloud. We achieve this by decomposing the single input image into a set of incomplete multi-views $\mathcal{I} = \{I_{\text{real}}, I_{\text{vir}_1}, I_{\text{vir}_2}, \dots, I_{\text{vir}_N}\}$, where I_{real} is the real view and I_{vir_i} is the mirror reflection derived virtual view, and N is the number of the virtual views. Each virtual view provides complementary information about the scene.

Specifically, for each real-virtual view pair $(I_{\text{real}}, I_{\text{vir}_i})$, there exists a mirror plane $\mathbf{M}_i = (\mathbf{n}_i, \mathbf{p}_i)$, where $\mathbf{n}_i \in \mathbb{R}^{3\times 1}$ is the normal vector and $\mathbf{p}_i \in \mathbb{R}^{3\times 1}$ is a point on the plane, such that the corresponding real and virtual camera poses $\mathbf{C}_{\text{real}} = [\mathbf{R}_{\text{real}}|\mathbf{t}_{\text{real}}] \in \mathbb{R}^{4\times 4}$ and $\mathbf{C}_{\text{vir}_i} = [\mathbf{R}_{\text{vir}_i}|\mathbf{t}_{\text{vir}_i}] \in \mathbb{R}^{4\times 4}$ are related by the corresponding reflection transformation matrix $\mathbf{T}_{\text{reflect}_i}$:

$$\mathbf{C}_{\text{vir}_i} = \mathbf{T}_{\text{reflect}_i}\mathbf{C}_{\text{real}}. \tag{1}$$

This geometric constraint ensures that the real and virtual cameras share a symmetric configuration relative to the mirror plane $\mathbf{M}_i$, with $\mathbf{T}_{\text{reflect}_i}$ describing the transformation induced by the mirror reflection, which not only enables multi-view geometry reconstruction using reflected appearances but also provides strong geometric priors for estimating virtual camera poses.

To this end, we formulate the task from a single-view 3D reconstruction problem to a multi-view 3D reconstruction problem, which can be effectively addressed using modern 3D reconstruction models, such as DUSt3R, which is designed to handle 3D reconstruction in a general setting. By leveraging our virtual view design, we adapt DUSt3R to perform 3D reconstruction under the single-view mirror-assisted setup.

We first present the theory underlying the design of $\mathbf{T}_{\text{reflect}_i}$ in Section 3.1, which simulates the physical imaging process of a virtual camera, ensuring that pixel-domain operations are equivalent to real-world image formation. We then describe the full Reflect3r pipeline in Section 3.2 and introduce a symmetric-aware loss for post-optimisation in Section 3.3, which leverages reflection constraints to further refine the estimated poses.

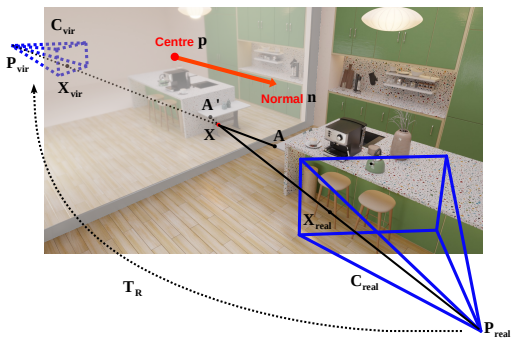

Figure 2. Physical imaging process of a scene containing a mirror. The reflection plane is shown as semi-transparent to reveal the virtual camera.

3.1. Virtual View Imaging Principle

We first introduce the theory of simulating the virtual view imaging to obtain a multi-view setup, leveraging a mirror reflection to simplify 3D reconstruction. This approach allows us to directly flip the view in the pixel domain to create a valid virtual view, effectively transforming a single image containing mirrors into a multi-view input. The resulting view pairs can then be processed by advanced feed-forward 3D reconstruction models.

The physical imaging process is illustrated in Fig. 2. We define the reflection matrix $\mathbf{T}_{\text{reflect}}$, which transforms a real camera $\mathbf{C}_{\text{real}}$ (solid line) into its symmetric valid virtual camera $\mathbf{C}_{\text{vir}}$ (dashed line) across a mirror plane $\mathbf{M}$. The reflection matrix is extended from the Householder matrix [11]:

$$\mathbf{T}_{\text{reflect}} = \mathbf{diag}(-1,1,1,1) \cdot \begin{bmatrix} \mathbf{I} - 2\mathbf{n}\mathbf{n}^\top & 2(\mathbf{n}^\top \mathbf{p})\mathbf{n} \\ \mathbf{0}^\top & 1 \end{bmatrix}; \tag{2}$$

$$= \mathbf{diag}(-1,1,1,1) \cdot \begin{bmatrix} \mathbf{R}_{\text{reflect}} & \mathbf{t}_{\text{reflect}} \\ \mathbf{0}^\top & 1 \end{bmatrix}, \tag{3}$$

where $\mathbf{n}, \mathbf{p}$ are the normal and the point on the mirror plane (centre in Fig. 2), respectively. $\mathbf{diag}(-1,1,1,1)$ flips the x-axis of the coordinate system. Using this matrix, the virtual camera $\mathbf{C}_{\text{vir}}$ is derived as:

$$\mathbf{C}_{\text{vir}} = \mathbf{T}_{\text{reflect}} \cdot [\mathbf{R}_{\text{real}} | \mathbf{t}_{\text{real}}]. \tag{4}$$

For a 3D point $\mathbf{A} \in \mathbb{R}^{3\times 1}$ visible to both cameras, the real camera observes both the point $\mathbf{A}$ and its mirror reflection $\mathbf{A}'$, while the virtual camera observes only $\mathbf{A}$. The light rays $\mathbf{P}_{\text{real}}\mathbf{A}'$ and $\mathbf{P}_{\text{vir}}\mathbf{A}$ intersect at the same point on the mirror plane. Let $\mathbf{X} \in \mathbb{R}^{3\times 1}$ denote this intersection. The projection of $\mathbf{A}$ onto the image plane of $\mathbf{C}_{\text{real}}$ is therefore equivalent to projecting $\mathbf{X}$ onto $\mathbf{C}_{\text{real}}$, given by:

$$\mathbf{X}_{\text{real}} = \begin{bmatrix} u_{\text{real}} \\ v_{\text{real}} \\ 1 \end{bmatrix} = \mathbf{K} \cdot [\mathbf{R}_{\text{real}} | \mathbf{t}_{\text{real}}] \cdot \mathbf{X}, \tag{5}$$

where $\mathbf{K}$ is the intrinsic parameter of the real camera. Since the real and virtual cameras share the intrinsic parameter $\mathbf{K}$, the projection of $\mathbf{A}$ on the virtual camera is:

$$\mathbf{X}_{\text{vir}} = \begin{bmatrix} u_{\text{vir}} \\ v_{\text{vir}} \\ 1 \end{bmatrix} = \mathbf{K} \cdot \mathbf{T}_{\text{reflect}} \cdot [\mathbf{R}_{\text{real}} | \mathbf{t}_{\text{real}}] \cdot \mathbf{X}. \tag{6}$$

Since the real and virtual cameras are symmetric with respect to the mirror plane, and $\mathbf{diag}(-1,1,1,1)$ flips the x-axis, thus:

$$u_{\text{real}} + u_{\text{vir}} = W; \tag{7}$$

$$v_{\text{real}} = v_{\text{vir}}, \tag{8}$$

where W is the width of the camera pixel plane.

By flipping the real view horizontally, we ensure that the projection on the virtual camera is a flipped view of the real camera, maintaining the stereo setup needed for 3D reconstruction, enabling us to further leverage the reflection information as an auxiliary view.

3.2. Reflect3r Pipeline

In this section, we introduce the pipeline of Reflect3r, which is illustrated in Fig. 3.

Mirror Detection and Multi-View Setup. We first apply mirror detection to identify the reflection regions $\mathcal{F}$ in the image, using Detect Any Mirror (DAM) [37]. The detected regions are then flipped horizontally to simulate the imaging process of the virtual camera. This allows the single input image I to be decomposed into a real view and its corresponding reflected views.

Initial Prediction. We adopt DUSt3R as the backbone to generate the initial point cloud from the virtual-real pair, which processes one image pair at a time. It first constructs a pairwise graph $\mathcal{E} = \{e_i = (R, V_i) \mid i = 1, 2, \dots\}$, where R denotes the real view and V_i the i-th virtual view, and then predicts the corresponding point cloud set $\mathcal{S} = \{S_e \mid e \in \mathcal{E}\}$.

Post-optimisation. We follow DUSt3R to perform the post-optimisation. To align point clouds into a global unified pointmap $\mathcal{U} = \{U_i \mid i = \{R, V_1, V_2, ...\}$, DUSt3R adopts a pair-wise loss. For an input image pair $e = (R, V_j) \in \mathcal{E}$, DUSt3R first predicts the pointmaps $S_e = (S_{R;(R,V_j)}, S_{\text{vir}_j;(R,V_j)})$ and the confidence map $O_e = (O_{R;(R,V_j)}, O_{\text{vir}_j;(R,V_j)})$. DUSt3R introduces a pair-wise pose $P_e \in \mathbb{SE}(3)$ and a scaling factor $\sigma_e \in \mathbb{R}$ for each pair e to rotate all pairwise predictions to align them to a shared global coordinate system X_{vir}. The alignment loss is formulated as:

$$\mathcal{L}_{\text{pairwise}} = \sum_{e \in \mathcal{E}} \sum_{v \in e} \sum_{j=1}^{HW} O_v^j ||U_v^j - \sigma_e P_e S_{v;e}^j||. \tag{9}$$

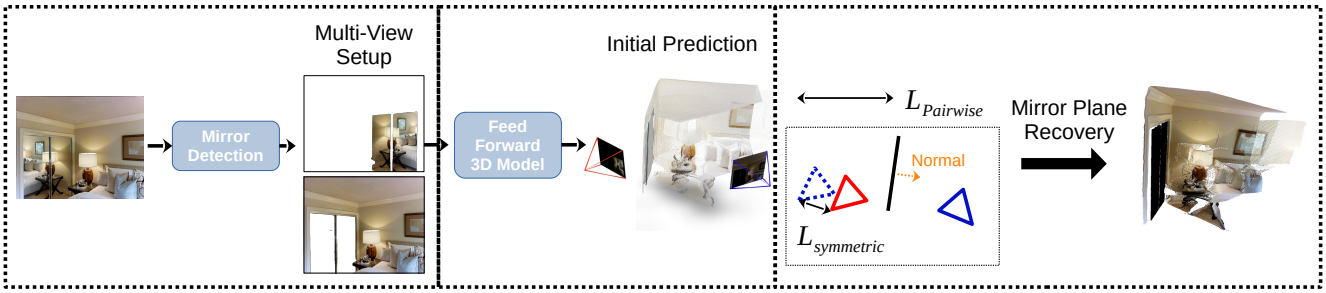

Figure 3. Overview of the proposed Reflect3r pipeline. Reflect3r reconstructs 3D scenes from a single-view image by leveraging mirror reflections. A reflection transformation is designed to ensure that flipping the real view in the pixel domain, simulating a virtual camera imaging, enables seamless integration with modern feed-forward models. Following the initial prediction, the reflection transformation is used as a geometric constraint to refine pose optimization.

We extend the post-optimisation process with a **symmetric-aware loss** $\mathcal{L}_{\text{sym}}$ that constrains pose estimation and incorporates the geometric prior from our virtual camera design. The final loss term $\mathcal{L}$ adopted for the post-optimization is

$$\mathcal{L} = \mathcal{L}_{\text{pairwise}} + \mathcal{L}_{\text{sym}}. \tag{10}$$

We recover the mirror plane from the result point cloud by estimating its normal and a point on the plane from the result point cloud. The details of the mirror plane recovery are covered in the supplementary.

3.3. Symmetric-Aware Loss

For each pair e, the estimated camera poses $(\mathbf{C}_{\text{real}}, \mathbf{C}_{\text{vir}_j})$ are constrained to be symmetric with respect to a mirror plane $\mathbf{M}_e$, estimated from the real view's point cloud U_{real} using the mirror mask M_{real}. Specifically, we estimate the normal $\mathbf{n}_e$ by performing Principal Component Analysis (PCA) of the mirror part of the point cloud $U_{\text{real}} \cdot M_{\text{real}}$. We obtain the point $\mathbf{p}_e$ on the plane by computing the centroid of the masked point cloud $U_{\text{real}} \cdot M_{\text{real}}$.

$$\mathbf{n}_e = \text{PCA}(U_{\text{real}} \cdot M_{\text{real}}); \quad \mathbf{p}_e = \overline{U_{\text{real}} \cdot M_{\text{real}}}. \tag{11}$$

The symmetry constraint enforces that the virtual camera $\mathbf{C}_{\text{vir}_j}$ and the real camera $\mathbf{C}_{\text{real}}$ satisfy the relation in Eq. (4), as defined by the reflection transformation $\mathbf{T}_{\text{reflect}}$ in Eq. (2).

To enforce this constraint, we formulate an optimization problem that minimizes the discrepancy between the transformed real camera pose $\mathbf{C}'_{\text{real}} = \mathbf{T}_{\text{reflect}} \cdot \mathbf{C}_{\text{real}}$ and the estimated virtual camera pose $\mathbf{C}_{\text{vir}_j}$ for all pairs in $\mathcal{E} = \{(R, V_j) \mid j = 1, 2, \dots\}$. The optimisation problem is defined as:

$$\min_{\mathbf{C}_{\text{real}}, \mathbf{C}_{\text{vir}}} \sum_{j=1}^{N} \sum_{(R,V_j)\in\mathcal{E}} ||\mathbf{C}'_{\text{real}} \ominus \mathbf{C}_{\text{vir}_j}|| \tag{12}$$

$$\text{s.t.} \quad \mathbf{C}'_{\text{real}} = \mathbf{T}_{\text{reflect}} \cdot \mathbf{C}_{\text{real}}, \quad \forall (R, V_j) \in \mathcal{E}, \tag{13}$$

Figure 4. Thumbnails of the dataset, where each image represents a fully customizable Blender scene.

where $\ominus$ denotes the difference between the transformed and estimated virtual camera poses.

We then compute the symmetric-aware loss by decomposing it into two components: the rotational difference and the translational difference:

$$\mathcal{L}_{\text{sym}} = \mathcal{L}_{\text{rot}} + \mathcal{L}_{\text{trans}}. \tag{14}$$

The rotation term penalizes the angular difference between the rotation of $\mathbf{C}_{\text{vir}_j}$ and $\mathbf{C}'_{\text{real}}$, measured using quaternions:

$$\mathcal{L}_{\text{rot}} = \sum_{(D,V_j)\in\mathcal{E}} 1 - \mathbf{q}'^{\top}_{\text{real}} \mathbf{q}_{\text{vir}_j}; \qquad ||\mathbf{q}|| = 1, \tag{15}$$

where $\mathbf{q}'_{\text{real}}$ and $\mathbf{q}_{\text{vir}_j}$ are the unit quaternions corresponding to the rotation matrices of $\mathbf{C}_{\text{vir}_j}$ and $\mathbf{C}'_{\text{real}}$, respectively.

The translation term penalizes the Euclidean distance between the translation vectors:

$$\mathcal{L}_{\text{trans}} = \sum_{(D,V_j)\in\mathcal{E}} ||\mathbf{t}'_{\text{real}} - \mathbf{t}_{\text{vir}_j}||_2^2. \tag{16}$$

This symmetric-aware formulation enforces consistency between the camera poses of the real and virtual views, guiding the optimisation toward a geometrically plausible configuration.

4. Dataset

We construct a reflection-aided single-view reconstruction dataset of 16 synthetic scenes sourced from Blender Demo [4], BlenderKit [1], and CGTrader [2], manually modelled in Blender [5]. Each scene is remodelled by us by adding a mirror surface positioned in a plausible location. For some

scenes, we also manually model additional details to ensure consistent richness across the dataset, avoiding cases where certain scenes lack sufficient complexity or objects. Example scenes are shown in Fig. 4, with a full dataset overview provided in the Appendix.

To support reflection-based reconstruction, we provide a Blender toolkit running in Blender scripting that simulates the virtual camera imaging process described in Section 3.1. Our dataset provides fully customizable Blender files with ground-truth point clouds for both real and mirror-reflected regions, along with real and virtual camera poses, while the editable nature of these scenes facilitates dataset extension for larger and more diverse future benchmarks.

5. Experiments

5.1. Experiment Setup

Backbone. Reflect3r builds on DUSt3R [34], a feed-forward architecture for stereo-based 3D reconstruction. Given an image pair, a shared Vision Transformer (ViT) encoder extracts tokens, which are refined through cross-attentive transformer decoders and decoded into pointmaps assigning 3D coordinates to each pixel, along with confidence maps estimating prediction reliability.

Baselines. We compare Reflect3r with three representative feed-forward 3D reconstruction models: DUSt3R [34], MASt3R [17], and VGGT [32], and one monocular depth estimation model MoGe [33]. For DUSt3R and MASt3R, which do not accept a single input image, we duplicate the prompt image as input, while for VGGT, we directly use the prompt image.

Dataset. Since mirrors are often small and may not provide meaningful reflections, we focus on scenes where reflections are informative. We collect 16 real scenes with significant mirror reflections from the Mirror3D [29], ScanNet [6], Matterport3D [3], and NYUv2 [25] datasets to evaluate our method. In addition to real data, we also report results on our proposed synthetic dataset, which provides accurate ground truth for computing quantitative metrics and performing ablation studies.

Evaluation Metrics. We evaluate reconstruction quality using 4 metrics: **completeness**, **accuracy**, **F1 score**, and **chamfer distance**. Accuracy and completeness measure the percentage of reconstruction-to-ground-truth and ground-truth-to-reconstruction distances below a 1 cm threshold, respectively. The F1 score is computed as the harmonic mean of accuracy and completeness. Chamfer Distance measures the similarity between two point sets by computing the average nearest-neighbour distance from each point in one set to the other, ensuring both sets are close in 3D space. The math behind the metrics is provided in the supplementary.

Figure 5. Qualitative results of Reflect3r and all the baselines running on the **real-world** data, where the predicted geometries corresponding to the mirror area are highlighted with light red.

5.2. Results

Qualitative Results. We compare reconstructed 3D point clouds on both real and synthetic datasets in Fig. 5 and Fig. 6, where mirror regions are highlighted in light red for all baselines. In our method, the mirror region is correctly identified and represented as a black plane. By reinterpreting reflections as virtual views captured by a virtual camera, Reflect3r introduces stereo information that significantly improves reconstruction coverage and completeness. In contrast, DUSt3R, MASt3R, and VGGT all fail on both real and synthetic scenes: DUSt3R and VGGT misinterpret the mirror as false geometry (with the exception of VGGT on synthetic Scene 2), while MASt3R collapses to flat, degenerate geometry in all cases. MoGe performs better at handling mirror regions compared to these baselines, yet it still produces erroneous geometry (e.g., real Scene 3 and synthetic Scene 3).

Moreover, the inability of the baselines to exploit stereo information prevents them from recovering occluded regions. In contrast, our stereo formulation enables more complete reconstruction of hidden areas, for example, the back of the lamp (Fig. 5, Scene 1), the rear surface of the

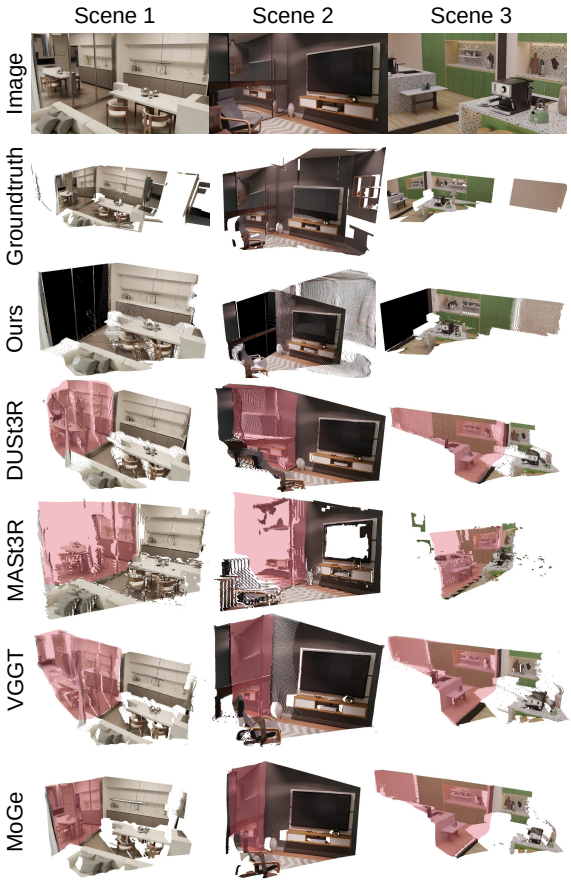

Figure 6. Qualitative results of Reflect3r and all the baselines running on the **synthetic** data, where the predicted geometries corresponding to the mirror area are highlighted with light red.

Method	Comp. % ↑	Accuracy % ↑	F1 ↑	Chamfer ↓
Dust3r	68.20%	71.84%	69.82%	0.1021
Mast3r	37.80%	62.62%	45.42%	0.1189
VGGT	76.70%	78.98%	77.73%	0.0906
MoGe	79.50%	88.51%	83.69%	0.0641
Reflect3r	**89.37%**	**96.64%**	**92.81%**	**0.0261**

Table 1. Quantitative comparison of point cloud coverage between Reflect3r and all baselines on the **synthetic** dataset. Comp. and Chamfer denote completeness and chamfer distance, respectively.

towel (Fig. 5, Scene 2), and the basin interior (Fig. 5, Scene 3), all of which are effectively recovered by Reflect3r but missed by other methods.

Quantitative Results. We report quantitative comparisons against baseline methods in Tab. 1, computing the mean across all the synthetic scenes for each metric.

Reflect3r consistently outperforms all baselines across all metrics. It achieves a significant gain in completeness, exceeding the second-best baseline, MoGe (79.50%), by roughly 10 percentage points. Notably, Reflect3r simultaneously achieves high completeness, high accuracy, and low chamfer distance, demonstrating its ability to recover a larger portion of the scene while preserving geometric fidelity. In contrast, MASt3R performs poorly due to its limited capability in sparse-view scenarios, and both DUSt3R and VGGT are adversely affected by the ambiguity introduced by mirrors, often misinterpreting reflections as real geometry.

These results demonstrate the effectiveness of reformulating the single-view input containing mirrors as a stereo setup: by leveraging the mirrored information, Reflect3r achieves denser and more accurate reconstructions. Reflected regions with overlapping content are reconstructed with higher confidence, while even non-overlapping or occluded areas, such as regions behind the camera or outside the field of view, benefit indirectly through improved global consistency and scene completion.

5.3. Ablation Study on Symmetric-Aware Loss

We conduct an ablation study to evaluate the effectiveness of the symmetry-aware loss in refining camera pose optimisation. As shown in Tab. 2, we compare Reflect3r on the synthetic dataset with and without the symmetry-aware loss, using translation error (T_{err}) and rotation error (R_{err}) between predicted and ground-truth poses as evaluation metrics. The results show a clear improvement in both the pose position (translation) and the pose orientation (rotation).

Setting	T_{err} ↓	R_{err} (°) ↓
Without Sym-aware loss	19.76%	8.1624
With Sym-aware loss	**13.58%**	**5.8705**

Table 2. Ablation study on the **synthetic** dataset evaluating the effectiveness of the symmetric-aware loss. We report translation and rotation errors between the estimated and ground-truth poses, where translation error is the Euclidean distance between pose translations, and rotation error is the angular difference between pose rotations.

In Figure 8, we visualise a representative synthetic scene, comparing the optimised camera poses with and without the symmetry-aware loss. The figure shows three sets of real–virtual camera pairs: ground-truth, estimates without the symmetry-aware loss, and estimates with the symmetry-aware loss. For clarity, all real camera poses are aligned to a common reference pose (green), making it easier to observe how the loss improves the pose estimation.

The translation error and the rotation error both decrease notably, showing the effectiveness of the symmetric-aware loss. With this loss, the predicted virtual camera poses (pink) are more symmetric and better aligned with the ground-truth (violet), while without it, the estimated reflected poses (cyan) deviate significantly. This demonstrates that the symmetry-aware constraint helps enforce geometric consistency between the real and virtual views, leading to improved camera optimisation.

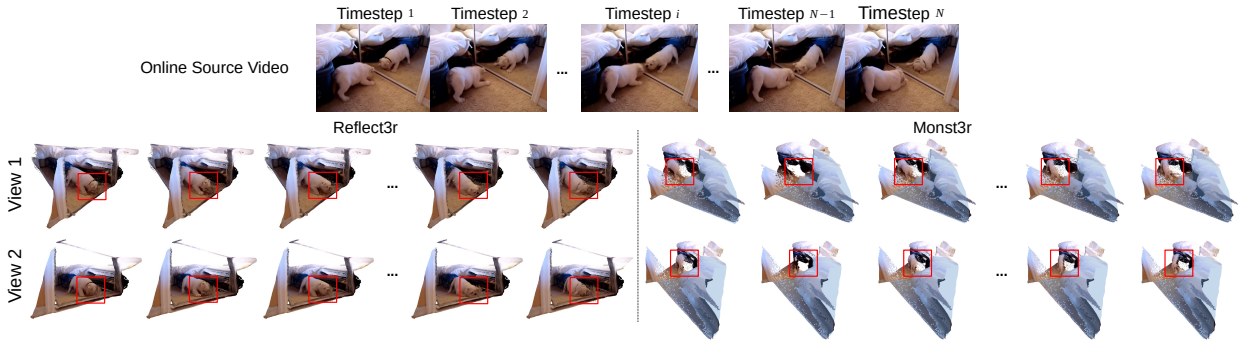

Figure 7. Dynamic reflection-aided 3D reconstruction. Comparison of single-view dynamic reconstruction results between Reflect3r and MonST3R on an online YouTube video.

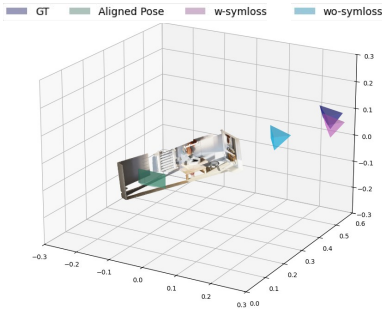

Figure 8. Comparison of pose estimation results with and without the symmetric-aware loss. Violet indicates the ground-truth virtual pose, while pink and cyan show the estimated poses with and without the symmetric-aware loss, respectively.

5.4. Beyond the Static Scenes

While our method is demonstrated on static scenes, the mirror-assisted reconstruction paradigm naturally extends to dynamic scene reconstruction. Traditional dynamic 3D reconstruction requires synchronised multi-view images at each time step, which demands expensive hardware and precise camera synchronisation.

In contrast, mirrors offer a low-cost, scalable alternative, allowing multiple views to be captured simultaneously using a single camera. Calibration is also simplified, as virtual cameras induced by mirrors inherently share the same intrinsic parameters as the real camera.

We demonstrate this concept using a dynamic version of Reflect3r, built upon MonST3R [43], a recent method designed for dynamic 3D reconstruction from a monocular video. While MonST3R operates with only single-view information per frame, we augment it with virtual mirror views to provide multi-view cues at each time step. The qualitative comparison between our dynamic Reflect3r and MonST3R is shown in Fig. 7, where we take an online YouTube video and perform the dynamic reconstruction for both methods.

Dynamic Reflect3r first detects the mirror in each frame and formulates the corresponding virtual view. Following MonST3R, we construct video graph pairs within a temporal window, and additionally introduce **spatial pairs** between the real and virtual views at each time step. As virtual views tend to be more fragmented and have limited coverage, we construct temporal pairs exclusively between real views to ensure reliable motion estimation. Specifically, for a video $\mathbf{V} = [\mathbf{I}^0, ..., \mathbf{I}^T]$, and a temporal window w, we define the view pair set at time step t as, $\mathbf{W}^t = \{(\mathbf{I}^i_{\text{real}}, \mathbf{I}^i_{\text{vir}}), (\mathbf{I}^a_{\text{real}}, \mathbf{I}^b_{\text{real}}) | (a, b, i) \in [t, ..., t + w]\}$. Similar to DUSt3R, these pairs are then fed into MonST3R to obtain the initial point cloud prediction. In the post-optimisation process, we keep MonST3R's temporal losses ($\mathcal{L}_{\text{smooth}}$, $\mathcal{L}_{\text{flow}}$) and incorporate our proposed symmetry-aware loss to improve spatial consistency between real and virtual views ($\mathbf{I}^i_{\text{real}}, \mathbf{I}^i_{\text{vir}}$) inside each frame.

At each time step, our method produces dual-view reconstructions using only a single camera and a mirror, enabling temporally dense and geometrically complete 3D reconstruction with minimal calibration effort.

In contrast, relying only on monocular input, MonST3R fails to recover occluded regions such as the dog's face (highlighted in red in Fig. 7), and misinterprets the mirror, producing incorrect geometry (highlighted in light blue).

6. Conclusion

In this paper, we reformulate reflection-aided single-view reconstruction as stereo reconstruction by introducing a simple pixel-domain operation to leverage mirror reflections as auxiliary views. This design enables a low-cost and easily deployable solution for both static and dynamic scene reconstructions. We propose Reflect3r, a pipeline that takes a single image as input, formulates virtual views via reflection, and reconstructs a more complete 3D point cloud via joint optimisation of geometry and camera poses. Extensive experiments on real and synthetic data demonstrate significant improvements over existing baselines. Additionally, we contribute a synthetic dataset of fully editable Blender scenes tailored for this setting to support future research.

References

[1] BlenderKit. Blenderkit. 5

[2] CGTrader. Cgtrader. 5

[3] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017. 6

[4] Blender Online Community. Blender demo. 5

[5] Blender Online Community. Blender - a 3d modelling and rendering package, 2018. 2, 5

[6] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 6

[7] Ankit Dhiman, Manan Shah, and R Venkatesh Babu. Mirrorverse: Pushing diffusion models to realistically reflect the world. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025. 2

[8] Ankit Dhiman*, Manan Shah*, Rishubh Parihar, Yash Bhalgat, Lokesh R Boregowda, and R Venkatesh Babu. Reflecting reality: Enabling diffusion models to produce faithful mirror reflections. In *3DV*, 2025. 2

[9] Qi Fang, Qing Shuai, Junting Dong, Hujun Bao, and Xiaowei Zhou. Reconstructing 3d human pose by watching humans in the mirror. In *CVPR*, 2021. 1, 3

[10] Ruiqi Gao*, Aleksander Holynski*, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul P. Srinivasan, Jonathan T. Barron, and Ben Poole*. Cat3d: Create anything in 3d with multi-view diffusion models. *Advances in Neural Information Processing Systems*, 2024. 3

[11] Gene H Golub and Charles F Van Loan. *Matrix computations*. JHU Press, 2012. 4

[12] Yuan-Chen Guo, Di Kang, Linchao Bao, Yu He, and Song-Hai Zhang. Nerfren: Neural radiance fields with reflections. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[13] Peter Käshammer and Andreas Nüchter. Mirror identification and correction of 3d point clouds. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2015. 2

[14] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 3

[15] Ye Keyang, Hou Qiming, and Zhou Kun. 3d gaussian splatting with deferred reflection. *ACM SIGGRAPH Conference Proceedings, Denver, CO, United States, July 28 - August 1, 2024*, 2024. 3

[16] Xin Kong, Shikun Liu, Xiaoyang Lyu, Marwan Taher, Xiaojuan Qi, and Andrew J Davison. Eschernet: A generative model for scalable view synthesis. *CVPR*, 2024. 3

[17] Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3d with mast3r, 2024. 1, 6

[18] Jiaying Lin, Guodong Wang, and Rynson W.H. Lau. Progressive mirror detection. In *Proc. CVPR*, 2020. 2

[19] Jiayue Liu, Xiao Tang, Freeman Cheng, Roy Yang, Zhihao Li, Jianzhuang Liu, Yi Huang, Jiaqi Lin, Shiyong Liu, Xiaofei Wu, et al. Mirrorgaussian: Reflecting 3d gaussians for reconstructing mirror reflections. *ECCV*, 2024. 3

[20] 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. *ICCV*, 2023. 3

[21] Haiyang Mei, Bo Dong, Wen Dong, Pieter Peers, Xin Yang, Qiang Zhang, and Xiaopeng Wei. Depth-aware mirror segmentation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[22] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Realfusion: 360 reconstruction of any object from a single image. *CVPR*, 2023. 3

[23] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *ECCV*, 2020. 3

[24] H. Mitsumoto, S. Tamura, K. Okazaki, N. Kajimi, and Y. Fukui. 3-d reconstruction using mirror images based on a plane symmetry recovering method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1992. 1, 3

[25] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 6

[26] S.A. Nene and S.K. Nayar. Stereo with mirrors. *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, 1998. 1, 3

[27] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 3

[28] Kawahara Ryo, Kuo Meng-Yu, Nobuhara Shohei, and Nishino Ko. Appearance and shape from water reflection. *WACV*, 2020. 1, 3

[29] Jiaqi Tan, Weijie Lin, Angel X Chang, and Manolis Savva. Mirror3D: Depth refinement for mirror surfaces. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 6

[30] Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. *arXiv*, 2023. 3

[31] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T. Barron, and Pratul P. Srinivasan. Ref-NeRF: Structured view-dependent appearance for neural radiance fields. *CVPR*, 2022. 3

[32] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 1, 6

[33] 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. *CVPR*, 2025. 6

[34] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. *CVPR*, 2024. 1, 2, 6

[35] Alex Warren, Ke Xu, Jiaying Lin, Gary K.L. Tam, and Rynson W.H. Lau. Effective video mirror detection with inconsistent motion cues. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17244–17252, 2024. 2

[36] Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P. Srinivasan, Dor Verbin, Jonathan T. Barron, Ben Poole, and Aleksander Holynski. Reconfusion: 3d reconstruction with diffusion priors. *arXiv*, 2023. 3

[37] Zhaohu Xing, Lihao Liu, Yijun Yang, Hongqiu Wang, Tian Ye, Sixiang Chen, Wenxue Li, Guang Liu, , and Lei Zhu. Detect any mirrors: Boosting learning reliability on large-scale unlabeled data with an iterative data engine. *CVPR*, 2025. 2, 4

[38] Ke Xu, Tsun Wai Siu, and Rynson W.H. Lau. Zoom: Learning video mirror detection with extremely-weak supervision. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38, 2024. 2

[39] Shao-Wen Yang and Chieh-Chih Wang. Dealing with laser scanner failure: Mirrors and windows. *2008 IEEE International Conference on Robotics and Automation*, 2008. 2

[40] Ze-Xin Yin, Peng-Yi Jiao, Jiaxiong Qiu, Ming-Ming Cheng, and Bo Ren. Ms-nerf: Multi-space neural radiance fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 3

[41] Han Yufei, Guo Heng, Fukai Koki, Santo Hiroaki, Shi Boxin, Okura Fumio, Ma Zhanyu, and Jia Yunpeng. Nersp: Neural 3d reconstruction for reflective objects with sparse polarized images. *CVPR*, 2024. 3

[42] Junyi Zeng, Chong Bao, Rui Chen, Zilong Dong, Guofeng Zhang, Hujun Bao, and Zhaopeng Cui. Mirror-nerf: Learning neural radiance fields for mirrors with whitted-style ray tracing. *Proceedings of the 31st ACM International Conference on Multimedia*, 2023. 3

[43] 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*, 2024. 8

[44] Chuanxia Zheng and Andrea Vedaldi. Free3d: Consistent novel view synthesis without 3d representation. *CVPR*, 2024. 3