Occlusion-Robust SfM in Construction Sites via Geometry-Guided Foreground Segmentation

Changjiang Yin[1], Shaoming Zhang[1], Qin Ye[1,2*] and Junqi Luo[1]

[1] College of Surveying and Geo-Informatics, Tongji University, 200092, Shanghai, China
[2] Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Natural Resources, 518000, Shenzhen, China

Keywords: Dynamic Occluders, Outlier Detection, Structure-from-Motion, Construction Sites, Prompt-based Segmentation.

Abstract

Accurate 3D reconstruction is a key enabler for construction progress monitoring and digital-twin maintenance. However, in tower-crane imagery, persistent dynamic occluders such as hooks and slings violate the static-scene assumption of conventional Structure-from-Motion (SfM), leading to feature mismatches and degraded reconstruction consistency. In this paper, we present a geometry-guided occlusion-handling pipeline for crane-mounted construction-site SfM. Our approach leverages geometric cues from reprojection errors and depth inconsistencies to identify outlier observations, clusters them into spatially coherent prompts, and uses these to guide a foundation segmentation model (SAM2). The resulting per-frame masks are integrated into mask-constrained SfM optimization, ensuring that only static background contributes to reconstruction. Experiments on three real-world crane-mounted sequences (30 m, 45 m, and 120 m) show consistent reductions in mean reprojection error relative to the unmasked baseline. In the most challenging case, the error decreases from 0.962 to 0.872 pixels (9.4%). Compared with a fixed rectangular masking strategy, the proposed masks yield similar reprojection errors while better preserving valid observations and sparse-point completeness. These results indicate that the proposed framework provides a practical geometry-guided strategy for improving internal reconstruction consistency in crane-mounted construction environments.

1. Introduction

Accurate and reliable 3D reconstruction is increasingly critical for construction progress monitoring, safety management, and digital-twin maintenance. Compared with LiDAR-based approaches, image-based Structure-from-Motion (SfM) offers a low-cost and easily deployable alternative, and has been widely adopted in construction workflows (Schonberger and Frahm, 2016, Sami Ur Rehman et al., 2022, Pal et al., 2023). However, standard SfM pipelines assume a static scene. In crane-mounted downward-looking imagery, persistent dynamic foregrounds, such as hooks, slings, and lifting accessories, would frequently intrude into the field of view. Their motion (e.g., swinging due to wind or load) violates multi-view consistency, leading to mismatches, erroneous triangulations, and elevated reprojection errors that degrade both camera pose and scene estimation (Saputra et al., 2018, Nietiedt et al., 2024). Although robust estimators such as RANSAC can filter random mismatches, they are ineffective against persistent, spatially coherent foreground motions, which often exhibit locally consistent geometry and thus survive standard outlier rejection.

Existing ways to cope with such occlusions each have drawbacks in construction settings. Engineering practice sometimes removes a fixed rectangular region covering the hook’s swing envelope; yet this rigid exclusion either under-masks (when the hook drifts outside the box) or over-masks (removing many valid static features), harming completeness and track continuity. Learned detectors/segmenters include supervised models trained on construction datasets (e.g., SODA (Duan et al., 2022)) as well as open-vocabulary detectors (e.g., Grounding DINO (Liu et al., 2024)). However, they often struggle to generalize across sites due to differences in viewpoint, device appearance, and lattice-like crane structures, causing missed or spurious detections. In addition, dynamic occluders in construction sites (e.g., hooks) share similar appearance and texture with surrounding machinery or scaffolds, and their shapes vary significantly across views, making it difficult for appearance-based detectors to distinguish them reliably. Foundation segmentation models (e.g., SAM (Kirillov et al., 2023) and SAM2 (Ravi et al., 2024)) bring strong cross-domain generalization, yet remain sensitive to prompt quality and spatial coverage (Ji et al., 2024, Wang et al., 2024). Purely geometry-driven filtering (e.g., pointwise thresholds on reprojection errors or depth consistency) can suppress some outliers but tends to produce spatially fragmented selections that do not translate into coherent masks for downstream SfM optimization (Saputra et al., 2018). Moreover, while standard epipolar filtering and bundle adjustment can reject many random mismatches, persistent dynamic foregrounds in crane imagery may still leave spatially coherent residual outliers, motivating an explicit mask-based refinement strategy.

We propose a geometry-guided occlusion-handling pipeline for crane-mounted construction-site SfM that combines geometric cues with promptable segmentation. Specifically, we first detect geometric outliers by intersecting reprojection-error and depth-inconsistency statistics from an initial SfM, then enforce spatial coherence via density-based clustering (DBSCAN (Ester et al., 1996)) to obtain compact and reliable prompt points. These prompts guide a foundation segmentation model (SAM2) to generate per-frame masks of dynamic occluders, which are subsequently used to constrain bundle adjustment within a standard SfM toolchain (e.g., COLMAP (Schonberger and Frahm, 2016)). The resulting optimization is driven primarily by static background features, improving multi-view consistency while preserving point-cloud completeness. On three real-world crane sequences (30 m/45 m/120 m), our approach reduces mean reprojection error by 2–10% versus an unmasked baseline, while avoiding the over-masking side effects of a fixed rectangular region. Our contributions include:

(i) We present a geometry-guided occlusion-handling frame-

* Corresponding author (yeqin@tongji.edu.cn)

work for crane-mounted, downward-looking construction-site SfM. It explicitly addresses persistent foreground motion in tower-crane imagery, where conventional SfM often degrades due to recurring dynamic occluders.

(ii) We propose a geometry-guided prompting mechanism for SAM2 segmentation. By intersecting reprojection-error and depth outliers and applying clustering, we obtain coherent prompts that guide foundation segmentation without task-specific training.

(iii) We demonstrate seamless integration and practical benefits. Our masks are compatible with standard SfM pipelines and improve accuracy–completeness trade-offs on crane-mounted sequences compared to fixed-region and unmasked baselines.

2. Related Works

2.1 Visual Reconstruction in Construction Sites

Vision-based reconstruction and perception in construction have been extensively studied for progress monitoring, inspection, and digital-twin generation. Common data sources include UAVs, fixed surveillance cameras, and crane-mounted cameras, which provide convenient site coverage. These image streams are typically processed using SfM/MVS or SLAM to recover scene geometry and camera motion (Sami Ur Rehman et al., 2022, Pal et al., 2023, Reja et al., 2022). Despite demonstrated feasibility, most pipelines implicitly assume static scenes or focus on higher-level monitoring tasks, leaving persistent dynamic occlusions underexplored in SfM itself. Recent works in construction perception emphasize dataset curation and detection under domain variability (e.g., SODA (Duan et al., 2022)), but detector-based foreground handling alone remains brittle when appearance and viewpoint shift across sites. Our work complements these efforts by keeping the SfM backbone unchanged while injecting *geometry-guided* masks that adapt per frame and per scene, thus mitigating foreground bias without requiring additional training.

2.2 Handling Dynamic Occluders in SfM/SLAM

Dynamic environments pose challenges to visual SLAM/SfM, motivating numerous methods that attempt to separate dynamic from static content before optimization (Saputra et al., 2018, Al-Tawil et al., 2024). Recent systems address this issue through either semantic segmentation or depth-assisted geometry, such as DynaSLAM, MaskFusion, and StaticFusion (Bescos et al., 2018, Runz et al., 2018, Scona et al., 2018). These methods detect/segment moving objects and exclude them from tracking/mapping. While effective, these methods typically rely on pretrained detectors/segmenters or depth sensors, and are not tailored to crane-mounted, downward-looking construction capture. Geometry-based approaches filter outliers via reprojection error, depth cues, or motion consistency (Saputra et al., 2018, Ai et al., 2021), yet operate at the point level and may yield spatially fragmented selections that are not directly usable as masks. By intersecting reprojection-error and depth-based outliers and clustering them (e.g. DBSCAN), we produce spatially coherent prompts that drive a foundation segmentation model (SAM/SAM2) to generate per-frame occluder masks. This geometry-to-segmentation pathway avoids domain-specific training and plugs naturally into mask-constrained BA in standard SfM pipelines.

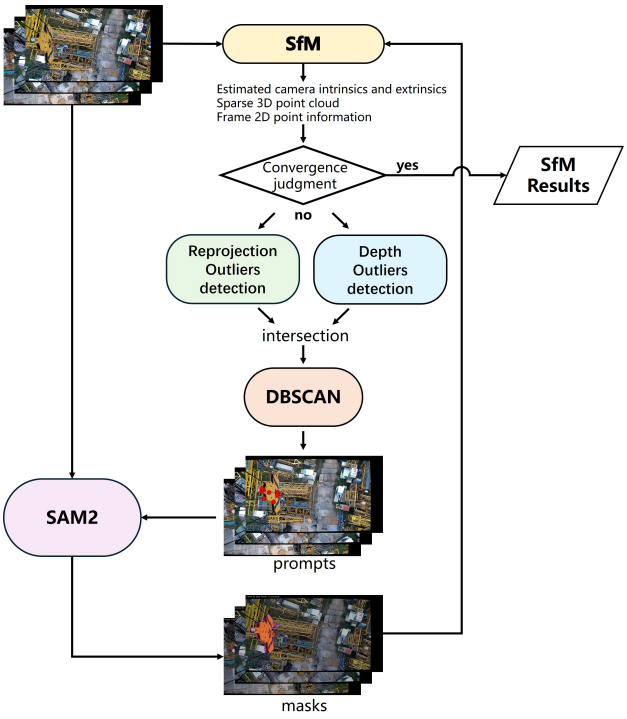


Figure 1. Framework of our method.

3. Methodology

In this section, we introduce our proposed framework (Figure 1) for occlusion-robust SfM in downward-looking crane-mounted imagery. Unlike standard SfM pipelines that treat all feature correspondences equally, our method explicitly accounts for dynamic occluders such as crane hooks and lifting accessories, which otherwise generate inconsistent tracks and degrade reconstruction. By integrating geometric outlier analysis, prompt-guided segmentation, and mask-constrained optimization, the framework ensures that pose estimation and 3D reconstruction are driven primarily by static background structures.

We first formulate the SfM problem in the presence of dynamic occlusions (Section 3.1) and describe the initial reconstruction that provides the basis for subsequent analysis (Section 3.2). We then detail the three main components of our approach: geometric outlier detection through reprojection and depth consistency (Section 3.3), segmentation guided by geometric prompts using a foundation segmentation model (Section 3.4), and mask-constrained SfM optimization (Section 3.5). Finally, we introduce an iterative refinement strategy (Section 3.6) that guarantees convergence and enhances robustness across sequences.

3.1 Problem Formulation

Let $\mathcal{I} = \{I_1, I_2, \ldots, I_N\}$ denote a sequence of N downward-looking images captured by a crane-mounted camera. For each image I_i, we denote the camera intrinsics by $\mathbf{K}_i$, and the extrinsic parameters by $(\mathbf{R}_i, \mathbf{t}_i) \in SO(3) \times \mathbb{R}^3$. A 3D point $X_j \in \mathbb{R}^3$ is observed in image I_i at location $\mathbf{x}_i^j \in \mathbb{R}^2$, and its predicted projection is

$$\mathbf{x}_{i,\mathrm{proj}}^j = \pi(\mathbf{K}_i, \mathbf{R}_i, \mathbf{t}_i, X_j), \tag{1}$$

where $\pi(\cdot)$ denotes the perspective projection.

In classical Structure-from-Motion (SfM), the goal is to jointly estimate camera parameters $\{\mathbf{R}_i, \mathbf{t}_i\}_{i=1}^N$ and 3D structure $\{X_j\}$

by minimizing the reprojection error:

$$\{\mathbf{R}_i, \mathbf{t}_i, X_j\} = \arg \min_{\{\mathbf{R}_i, \mathbf{t}_i\}, \{X_j\}} \sum_{(i,j)\in\mathcal{V}} \left\|\mathbf{x}_i^j - \mathbf{x}_{i,\text{proj}}^j\right\|_2^2, \quad (2)$$

where $\mathcal{V}$ is the set of valid 2D–3D correspondences. Under the assumption that all observed features belong to a static scene, solving (2) yields accurate camera poses and a consistent sparse 3D reconstruction.

However, in our target setting the camera is rigidly mounted on a tower crane, and the field of view is persistently occluded by dynamic foreground objects such as hooks and lifting accessories. These elements exhibit non-rigid motion (e.g., swinging due to wind or load) and thus violate the static-scene assumption. As a consequence, features detected on the foreground produce inconsistent tracks across views and lead to erroneous triangulation. When included in (2), such outliers bias the optimization, causing drift in camera poses and degradation of the reconstructed structure.

Formally, let $\mathcal{V} = \mathcal{V}_s \cup \mathcal{V}_o$, where $\mathcal{V}_s$ are correspondences on the static background and $\mathcal{V}_o$ are those contaminated by occluders. While $\mathcal{V}_s$ conforms to the geometric model, $\mathcal{V}_o$ violates multi-view consistency. The central objective of our method is therefore to identify and suppress $\mathcal{V}_o$, such that the bundle adjustment in (2) is effectively constrained to $\mathcal{V}_s$, enabling robust estimation of poses and 3D points under persistent occlusions.

3.2 Initial SfM

Our framework begins with an initial Structure-from-Motion (SfM) reconstruction to provide the necessary geometric input for subsequent outlier analysis. This is indispensable for the pipeline: without approximate camera poses and triangulated 3D points, it is impossible to evaluate geometric consistency across views. In practice, we adopt a standard incremental SfM pipeline as implemented in COLMAP, though any off-the-shelf SfM system could serve the same purpose. In our datasets, the monitoring camera was calibrated before installation, and all images were corrected in advance so that the processed imagery is well approximated by the `SIMPLE_PINHOLE` camera model. Therefore, the camera intrinsics are treated as known and fixed during reconstruction, and no additional lens-distortion parameters are estimated in the subsequent SfM stages.

The output of this stage consists of:

- Known intrinsics and estimated extrinsics $\{\mathbf{K}_i, \mathbf{R}_i, \mathbf{t}_i\}$,
- A sparse 3D point cloud $\{X_j\}$,
- Per-observation reprojection errors $r_i^j = \|\mathbf{x}_i^j - \hat{\mathbf{x}}_i^j\|_2$,
- Per-observation depths d_i^j computed in the local camera coordinate system.

These quantities provide the foundation for identifying inconsistent correspondences caused by moving occluders. In the next subsection, we exploit reprojection-error and depth statistics derived from the initial SfM to detect geometric outliers and separate static background features from dynamic foreground observations.

3.3 Geometric Outlier Detection

After the initial SfM reconstruction, we obtain camera parameters, a sparse point cloud, and per-observation reprojection errors and depths. These quantities allow us to evaluate the geometric consistency of individual observations and detect those influenced by dynamic occluders. We adopt two complementary criteria—reprojection errors and depth consistency—and combine them into a unified set of geometric outliers.

Reprojection errors. For each observation (i, j) of 3D point X_j in image I_i, the reprojection error is defined as

$$r_i^j = \left\|\mathbf{x}_i^j - \pi(\mathbf{K}_i, \mathbf{R}_i, \mathbf{t}_i, X_j)\right\|_2, \quad (3)$$

where $\mathbf{x}_i^j$ is the detected feature location and $\pi(\cdot)$ the perspective projection. Large reprojection-errors indicate inconsistency with the multi-view geometry and often arise from occluded or moving features.

To robustly decide whether r_i^j is abnormally large, we employ adaptive reprojection error thresholds derived from the reprojection error distribution $\mathcal{E} = \{r_i^j\}$. Specifically, we support:

$$\tau_r^{\text{MAD}} = \text{median}(\mathcal{E}) + k \cdot 1.4826 \cdot \text{median}(|r - \text{median}(\mathcal{E})|), \quad (4)$$

$$\tau_r^{\text{IQR}} = Q_3(\mathcal{E}) + k \cdot \big(Q_3(\mathcal{E}) - Q_1(\mathcal{E})\big), \quad (5)$$

$$\tau_r^{\text{Z}} = \mu(\mathcal{E}) + k \cdot \sigma(\mathcal{E}), \quad (6)$$

$$\tau_r^{\text{Perc}} = \text{Percentile}_p(\mathcal{E}), \quad (7)$$

where k is a scaling factor, the constant 1.4826 scales Median Absolute Deviation (MAD) to be consistent with the standard deviation under a Gaussian distribution. Q_1 and Q_3 are the first and third quartiles, and μ and σ denote mean and standard deviation. We tested all four threshold formulations in Eqs. (4)–(7) and observed no substantial differences in the overall behavior of the pipeline. Unless otherwise stated, the reported experiments use the MAD-based threshold setting.

After the thresholds are determined, we aggregate reprojection errors per 3D point and apply a set of decision rules. Let $\mathcal{E}_j = \{r_i^j\}_{i\in\mathcal{V}(j)}$ be the reprojection-errors of point X_j across its n_j observations. A point is flagged as a reprojection outlier if any of the following conditions hold:

1. **Extreme quick check:** if $\exists r_i^j > \tau_{\text{ext}}$ (default $\tau_{\text{ext}} = 8$ px), the point is directly marked as an outlier.
2. **Minimum views requirement:** if $n_j < n_{\min}$ (default $n_{\min} = 2$), the point is ignored unless it already satisfies the extreme rule.
3. **Proportion rule:** if the fraction of reprojection errors exceeding the adaptive threshold satisfies

$$\frac{N_{\{r_i^j > \tau_r\}}}{n_j} \geq \rho, \quad (8)$$

with ρ typically set to 0.5, then the point is considered an outlier.

4. **Median rule:** if the median reprojection error of the point exceeds the adaptive threshold, i.e.

$$\text{median}(\mathcal{E}_j) > \tau_r, \quad (9)$$

then the point is marked as an outlier.

These rules ensure that both isolated extreme errors and systematic inconsistencies across multiple views can be reliably detected, while points with very few observations are treated conservatively. All points satisfying any of these conditions are labeled as reprojection outliers and collected in the set $\mathcal{O}_{\text{repr}}$.

Depth consistency. Foreground occluders often lie much closer to the camera than the static scene, or in some cases triangulate to unrealistically large depths. Let d_i^j denote the depth of X_j in the local camera coordinates of image I_i, and let $m_i = \text{median}_j(d_i^j)$ be the median depth of all points visible in I_i. We consider three types of depth inconsistency:

1. **Negative depth:** if $d_i^j < 0$, the point lies behind the camera and is directly regarded as an outlier.
2. **Near outliers:** points significantly closer than the background distribution, measured by the relative deviation
$$\Delta_i^{\text{near}}(j) = \frac{|d_i^j - m_i|}{m_i}, \tag{10}$$
and marked as outliers if $\Delta_i^{\text{near}}(j) > \tau_d^{\text{near}}$ with $\tau_d^{\text{near}} < 1$.
3. **Far outliers:** points disproportionately farther than the background distribution, measured by
$$\Delta_i^{\text{far}}(j) = \frac{d_i^j}{m_i}, \tag{11}$$
and marked as outliers if $\Delta_i^{\text{far}}(j) > \tau_d^{\text{far}}$ with $\tau_d^{\text{far}} > 1$.

All points satisfying any of these conditions are labeled as depth outliers and collected in the set $\mathcal{O}_{\text{depth}}$.

Unified outlier set. We define the reprojection outlier set $\mathcal{O}_{\text{repr}}$, the depth outlier set $\mathcal{O}_{\text{depth}}$, and their intersection $\mathcal{O}_\cap = \mathcal{O}_{\text{repr}} \cap \mathcal{O}_{\text{depth}}$. Projecting these 3D outliers back into image space yields 2D outlier regions
$$\mathcal{P}_i = \text{Proj}_i(\mathcal{O}_\cap), \tag{12}$$
which serve as high-precision prompts for segmentation. In practice, restricting to the overlap $\mathcal{O}_\cap$ avoids spurious false positives and provides stable seeds for mask generation. By default, prompt candidates are derived from the overlap set for higher precision. When $\mathcal{O}_\cap$ is too sparse to form a meaningful spatial cluster in a given image, we instead use reprojection outliers as the input for the subsequent clustering step.

Clustering densest regions. Because outliers caused by dynamic occluders are spatially coherent, we further refine prompts by density-based clustering. Let $\mathcal{P}_i$ be the 2D set of outlier projections in image I_i. We apply DBSCAN with parameters $(\varepsilon, N_{\min})$ to identify connected clusters and retain only the largest one $\mathcal{C}^\star$, discarding isolated detections. The final prompts are
$$\tilde{\mathcal{P}}_i = \{\mathbf{p} \in \mathcal{P}_i \mid \mathbf{p} \in \mathcal{C}^\star\}. \tag{13}$$
We adopt DBSCAN since it can discover arbitrarily shaped dense regions and separate them from sparse noise without requiring a predefined number of clusters, which suits the irregular geometry of moving occluders. Through this two-stage process—statistical outlier detection and spatial clustering—we obtain compact and reliable prompt regions that faithfully indicate the location of moving occluders. These prompts are subsequently exploited to guide segmentation, as detailed in Section 3.4.

Threshold settings. For reprojection-error outliers, we adopt adaptive thresholds based on the residual distribution to account for varying image scales and feature densities across scenes. This data-driven strategy improves robustness and avoids manual tuning. In contrast, the remaining thresholds (e.g., depth ratios, extreme error cutoffs, and DBSCAN parameters) are empirically set because their relative magnitudes are more stable and less sensitive to scene variation. These values were selected through preliminary experiments to balance detection precision and completeness, and remain fixed for all sequences, as summarized in Table 1.

3.4 Segmentation Guided by Geometric Prompts

The reprojection- and depth-based outlier detection provides compact sets of high-confidence points $\tilde{\mathcal{P}}_i$ that indicate the approximate location of moving occluders in each image I_i. Instead of relying on handcrafted heuristics to delineate occluder regions, we exploit these points as prompts to drive a segmentation model, thereby obtaining pixel-level masks that more accurately capture the foreground objects.

Prompt-based segmentation. We employ a foundation segmentation model $S(\cdot)$, specifically SAM2, which accepts sparse point prompts $\tilde{\mathcal{P}}_i$ and an image I_i as input, and outputs a binary mask
$$M_i = S(I_i, \tilde{\mathcal{P}}_i), \quad M_i \in \{0,1\}^{H \times W}, \tag{14}$$
where $M_i(p) = 1$ indicates that pixel p belongs to the occluder. By anchoring the segmentation with geometrically inconsistent points, we avoid drifting to irrelevant regions and ensure that the generated masks are tightly aligned with the actual dynamic foreground.

Mask refinement. Although the initial masks produced by $S(\cdot)$ capture the coarse extent of the occluders, their boundaries may be fragmented or slightly misaligned. We therefore apply simple morphological operations to improve the masks: erosion removes isolated false positives, dilation fills small gaps, and connected-component analysis ensures spatial coherence. The refined mask is denoted $\hat{M}_i$.

Output. The result of this step is a sequence of masks $\{\hat{M}_i\}_{i=1}^N$ aligned with the image sequence. These masks delineate the dynamic foreground objects frame by frame, providing the basis for excluding occluded regions in subsequent SfM optimization.

3.5 Mask-Constrained SfM

With per-frame masks $\{\hat{M}_i\}$ available, we constrain the Structure-from-Motion optimization to exclude feature observations falling inside occluder regions. Specifically, given a correspondence (i, j) of 3D point X_j observed in image I_i at pixel location $\mathbf{x}_i^j$, we define its validity as
$$v_i^j = \begin{cases} 1, & \text{if } \hat{M}_i(\mathbf{x}_i^j) = 0, \\ 0, & \text{if } \hat{M}_i(\mathbf{x}_i^j) = 1, \end{cases} \tag{15}$$
where $\hat{M}_i(\mathbf{x}_i^j)$ indicates the mask value at pixel $\mathbf{x}_i^j$. Only correspondences with $v_i^j = 1$ are retained for optimization. The

Parameter	Symbol	Type	Value / Rule
Reprojection error threshold	τ_r	Adaptive	MAD
Extreme reprojection cutoff	τ_{ext}	Empirical	8 px
Minimum view count	$n_{\min}$	Empirical	2
Outlier ratio	ρ	Empirical	0.5
Near-depth threshold	τ_d^{near}	Empirical	0.95
Far-depth threshold	τ_d^{far}	Empirical	10
DBSCAN distance	ε	Empirical	10 pixels
DBSCAN minimum points	$N_{\min}$	Empirical	5

Table 1. Summary of threshold parameters used in the proposed pipeline.

effective observation set is therefore

$$\mathcal{V}' = \{(i,j) \in \mathcal{V} \mid v_i^j = 1\}. \tag{16}$$

Compared with the standard bundle adjustment in Eq. (2), the mask-constrained formulation optimizes camera poses and 3D structure only over the filtered set $\mathcal{V}'$, which denotes the post-masking observation set intended to approximate the static valid subset of observations, ensuring that features from dynamic occluders no longer bias the estimation. In practice, this leads to more stable reconstructions, improved camera trajectory consistency, and reduced reprojection error. Importantly, because the masks are derived from geometric prompts, the constrained SfM remains focused on static background structures even under persistent foreground interference.

3.6 Iterative Refinement

In our experiments, a single iteration already produced sufficiently accurate segmentation and reconstruction, and further iterations yielded negligible improvement. Nevertheless, we retain the iterative design for completeness and robustness, as it provides a principled mechanism to re-estimate outliers and masks in more challenging scenarios where the first pass may be insufficient.

Algorithm. The proposed framework forms a closed-loop process that iteratively refines both the reconstruction and the occluder masks. Starting from an initial SfM reconstruction, reprojection errors and depth values are computed for all observations. Geometric outliers are then detected based on residual and depth inconsistency, projected to image space, and clustered to generate compact prompt points. These prompts guide the segmentation model (SAM2) to produce per-frame masks, which are refined by standard morphological cleanup operations. Subsequently, a mask-constrained SfM is re-optimized using only the unmasked feature correspondences, yielding updated camera poses and 3D structure. At each new iteration, outliers are recomputed from the updated reconstruction, reclustered in image space, and used to generate new prompts and masks. This process can be repeated until convergence, defined by stability of the reprojection error, point cloud size, and the number of registered images.

Convergence criteria. We monitor four indicators: (i) the average reprojection-error change between successive iterations falls below ϵ_r, (ii) the number of inlier 3D points stabilizes, (iii) the absolute change in the number of detected outlier observations between successive iterations falls below ϵ_o, and (iv) the number of successfully registered images remains unchanged. If all conditions are satisfied, the refinement is considered converged.

4. Experiment

4.1 Experimental Setup

Datasets. We evaluate our method on three real-world construction scenarios captured from tower cranes (Figure 2). A surveillance-grade camera (Dahua bullet-type, 2 MP resolution) was rigidly mounted on the sliding trolley of the crane jib. The camera continuously looked vertically downward, producing nadir-view sequences that cover the surrounding construction site as the crane rotated. Each image has a resolution of 1920×1080 pixels, and the crane operating heights varied between 30 m and 120 m. For each scenario, we selected a sequence of 540 consecutive frames, providing sufficient overlap for Structure-from-Motion (SfM) reconstruction.

Implementation details. We use COLMAP version 3.11 as the SfM backbone, configured with the `SIMPLE_PINHOLE` camera model. The camera intrinsics were fixed to $(f, c_x, c_y) = (1600, 960, 540)$, which were obtained from pre-installation camera calibration. All images were corrected in advance to conform to the `SIMPLE_PINHOLE` model, so no additional distortion parameters were estimated during reconstruction. All other parameters follow the default COLMAP settings. Our geometric outlier detection employed the adaptive thresholding strategies described in Section 3.3, with default parameters $k = 2.0$, $\rho = 0.5$, extreme threshold $\tau_{\text{ext}} = 8$ px, and minimum view count $n_{\min} = 2$. We evaluated all threshold variants in Eqs. (4)–(7) and found similar overall behavior; the quantitative results reported below use the MAD-based setting. Depth-based filtering used thresholds $\tau_d^{\text{near}} = 0.95$ and $\tau_d^{\text{far}} = 10$. Density-based clustering of projected outliers was performed using DBSCAN with $\varepsilon = 10$ pixels and $N_{\min} = 5$.

For segmentation, we adopted the SAM2 model in the `sam2_hiera_small` configuration, which offers a favorable trade-off between accuracy and efficiency for local inference. Inference was performed on an NVIDIA GeForce RTX 3070 GPU with 8GB memory under Ubuntu 22.04. Masks were generated from clustered geometric prompts as described in Section 3.4 and refined using standard morphological cleanup and connected-component filtering to obtain spatially cleaner mask shapes. Although our framework supports iterative refinement (Section 3.6), in practice a single iteration already yields sufficiently accurate segmentation and reconstruction in our crane datasets. Therefore, all experiments in this section are reported using only the first iteration.

4.2 Quantitative Evaluation

We evaluate three settings on each of the three sequences (540 images per scene at heights 30 m/45 m/120 m): (i) the standard COLMAP without masking, (ii) our occlusion-aware pipeline

(a) Scene 1 (30 m) (b) Scene 2 (45 m) (c) Scene 3 (120 m)

Figure 2. Example frames from the three crane-mounted datasets used in our experiments.

(a) Our geometric-prompt mask (b) Rectangular baseline mask

Figure 3. Comparison of mask strategies.

with geometric-prompt masks, and (iii) a naive baseline that removes a fixed rectangular region covering the hook's swing envelope (estimated from the union of prompt projections). All configurations register all 540 images. For illustration, Figure 3 shows examples of the masks used in setting (ii) and (iii), where our geometric-prompt masks adapt to the hook position in each frame, while the rectangular baseline removes a fixed region.

Table 2 reports the SfM statistics. Across all three sequences, our method consistently reduces the mean reprojection error by approximately 2–10% relative to the unmasked baseline. The rectangular-mask baseline achieves comparable reprojection errors. However, this comes at the cost of removing a substantially larger number of valid observations and sparse points. In contrast, our approach preserves completeness much better than the rectangular baseline (point/observation counts close to the unmasked case) while improving multi-view consistency (higher track length in Scenes 2–3). Therefore, the main advantage of the proposed masks over coarse rectangular exclusion is a more selective suppression of dynamic foregrounds with less over-masking of static background features.

The relative gain varies across scenes. In Scene 1 the error drops by 9.4%, suggesting many foreground-induced mismatches that our masks successfully eliminate. In Scene 2, despite the hook covering the largest image area, the reduction is modest (3.0%), likely because vegetation and low-texture regions already limited false correspondences. In Scene 3, the hook footprint is smaller due to a longer cable, so its influence on reconstruction is weaker; nevertheless, error still decreases by 2.0%. Overall, our method provides a better accuracy–completeness trade-off by precisely isolating dynamic occluders rather than over-masking.

4.3 Qualitative Analysis

In addition to the quantitative metrics in Section 4.2, we provide qualitative visualizations to better understand the behavior of our pipeline. Specifically, we analyze (i) the spatial distribution of geometric outliers, (ii) the effect of clustering on prompt selection, and (iii) comparisons with supervised detection and open-vocabulary detection baselines. These visualizations shed light on why our method achieves robust performance under persistent crane occlusions.

4.3.1 Geometric Outliers Distributions To better understand the role of geometric filtering, we visualize the spatial distributions of detected outliers in Scene 1. Figure 4 compares depth-based, reprojection-based and overlapping outliers. Depth outliers (blue) are concentrated around the hook but also include scattered false positives in background regions. Reprojection outliers (green) form a much denser set, covering both the hook and numerous background structures. In contrast, the intersection of both criteria (red) produces a compact cluster precisely aligned with the hook. This demonstrates that the overlap provides a more reliable and noise-resistant basis for downstream segmentation, since it removes spurious detections while preserving the true dynamic occluder.

4.3.2 Effect of DBSCAN Clustering While overlapping reprojection and depth outliers already provide high-precision prompts, in some cases their intersection is too sparse to reliably initialize segmentation. Our default strategy is therefore to use overlapping outliers when available, and to fall back to reprojection outliers when the overlap set is too sparse to form a meaningful cluster. This situation occurs for Scene 3, where only a very small number of overlapping outliers are detected due to the hook's smaller footprint when the cable is extended. To handle such cases, we apply DBSCAN clustering directly on reprojection outliers. By grouping spatially coherent detections, the clustering step effectively filters isolated noise points and recovers a compact set of prompts concentrated on the hook region.

Figure 5 illustrates this effect: compared to using raw outliers that include scattered detections across the scene, DBSCAN identifies a dense cluster tightly localized around the hook, which serves as a more reliable initialization for segmentation. This example illustrates how the clustering step helps recover coherent prompt regions even when the overlap between outlier types is limited.

4.3.3 Comparison with Learning-based Detectors We further compare our method with two representative detectors: (i) YOLOv8s trained on the construction-oriented SODA dataset (Duan et al., 2022), and (ii) Grounding DINO 1.6 pro (Liu et al., 2024), prompted with “crane hook.” As shown in Figure 6, YOLOv8s produces inaccurate detections and frequently misses the hook in many frames. We attribute this to the large variations between our tower-crane scenes and the SODA dataset, including differences in camera viewpoints and hook appearance, which limit generalization. Grounding DINO achieves more frequent detections, but often mistakes surrounding crane parts for the hook due to their structural similarity, leading to many false positives. In contrast, our geometry-guided masks

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-393-2026 | © Author(s) 2026. CC BY 4.0 License.

Scene	Method	Points (↑)	Observations (↑)	Mean Track Length (↑)	Obs./Image (↑)	Reproj. Error (↓)
Scene 1 (30 m)	w/o mask	318,640	2,878,841	**9.03**	5,331.19	0.962
	Rect. mask	283,971	2,503,904	8.82	4,636.86	**0.870**
	Ours (mask)	**321,729**	**2,877,928**	8.95	**5,329.50**	0.872
Scene 2 (45 m)	w/o mask	**321,608**	2,777,510	8.64	5,143.54	0.953
	Rect. mask	286,745	2,509,818	8.75	4,647.81	**0.924**
	Ours (mask)	308,348	**2,807,908**	**9.11**	**5,199.83**	**0.924**
Scene 3 (120 m)	w/o mask	**170,056**	1,659,413	9.76	3,072.99	1.012
	Rect. mask	143,850	1,410,469	9.81	2,611.98	**0.991**
	Ours (mask)	165,726	**1,662,520**	**10.03**	**3,078.74**	0.992

Table 2. SfM statistics on three crane-mounted sequences.

consistently capture the hook region across frames, demonstrating stronger robustness in tower-crane scenes.

5. Limitations

Our design choice of mask-based occluder suppression, rather than directly deleting geometric outlier points, is motivated by robustness considerations. Outlier detection is not perfect: a few points may be flagged outside the hook region, while not all hook-interior points are guaranteed to be detected. Simply removing such points risks discarding useful correspondences or leaving residual foreground features. In contrast, mask-based exclusion enforces spatial coherence, eliminating entire occluder regions while preserving background structures. From an implementation standpoint, applying masks also integrates naturally with existing SfM pipelines, whereas point-level deletion would require deeper modification and is less practical.

Despite the promising results, our pipeline has several limitations. It assumes a successful initial SfM reconstruction, without which outlier detection and prompt generation cannot proceed. Moreover, the current experiments focus on a single dominant occluder (the crane hook), while scenes with multiple independently moving occluders remain to be studied. Evaluation is based on internal SfM measures, including reprojection error and sparse reconstruction statistics. External geometric validation was not available in the current crane-camera setup, given the complexity of the construction-site environment. The reported gains therefore mainly indicate improved internal reconstruction consistency. In addition, although one parameter setting worked reasonably well across the three sequences used here, its transferability to different image resolutions, scene scales, and motion patterns remains to be investigated. Extending the framework to richer dynamic environments remains an important direction.

6. Conclusion

In this work, we presented an occlusion-robust SfM pipeline tailored to construction-site imagery affected by persistent dynamic foregrounds. By converting geometric inconsistencies into reliable prompts for a foundation segmenter, the pipeline isolates moving foregrounds and keeps optimization focused on static structures. This leads to improved multi-view consistency with little loss of completeness, yielding a better accuracy–completeness balance than fixed rectangular exclusion. The approach requires no site-specific training and integrates directly with standard SfM toolchains, making it practical for real scenes. Future work will extend to multi-occluder scenes and explore online adaptation for cross-site deployment and real-time efficiency.

Acknowledgements

This work was supported by the Open Fund of Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Natural Resources (Grant No. KF-2023-08-14).

References

Ai, Y.-b., Rui, T., Yang, X.-q., He, J.-l., Fu, L., Li, J.-b., Lu, M., 2021. Visual SLAM in dynamic environments based on object detection. *Defence Technology*, 17(5), 1712–1721.

Al-Tawil, B., Hempel, T., Abdelrahman, A., Al-Hamadi, A., 2024. A review of visual SLAM for robotics: Evolution, properties, and future applications. *Frontiers in Robotics and AI*, 11, 1347985.

Bescos, B., Fácil, J. M., Civera, J., Neira, J., 2018. DynaSLAM: Tracking, mapping, and inpainting in dynamic scenes. *IEEE robotics and automation letters*, 3(4), 4076–4083.

Duan, R., Deng, H., Tian, M., Deng, Y., Lin, J., 2022. SODA: A large-scale open site object detection dataset for deep learning in construction. *Automation in Construction*, 142, 104499.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. E. Simoudis, J. Han, U. M. Fayyad (eds), *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96)*, AAAI Press, 226–231.

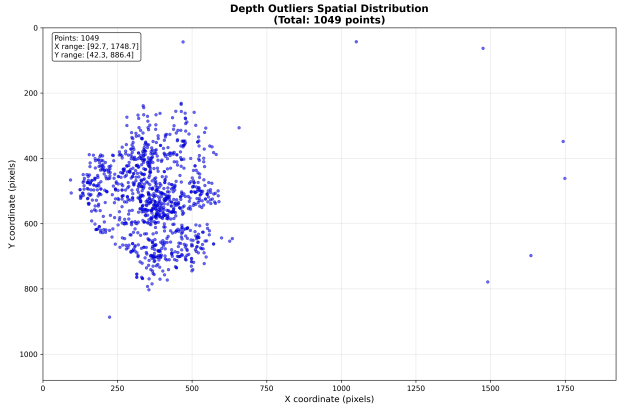
Ji, W., Li, J., Bi, Q., Liu, T., Li, W., Cheng, L., 2024. Segment Anything Is Not Always Perfect: An Investigation of SAM on Different Real-world Applications. *Machine Intelligence Research*, 21(4), 1–14.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y. et al., 2023. Segment anything. *Proceedings of the IEEE/CVF international conference on computer vision*, 4015–4026.

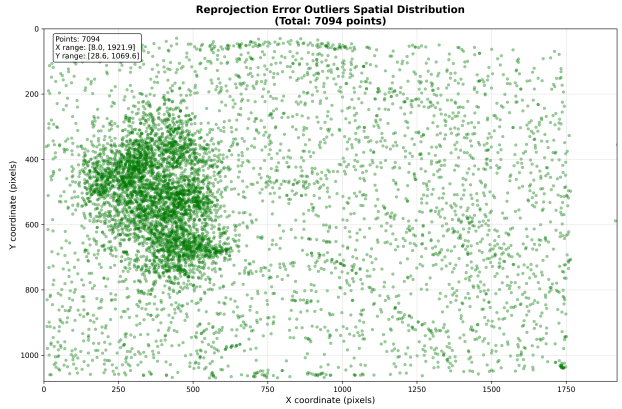
Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H. et al., 2024. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *European conference on computer vision*, Springer, 38–55.

Nietiedt, S., Helmholz, P., Luhmann, T., 2024. Occlusion handling in spatio-temporal object-based image sequence matching. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 10, 163–170.

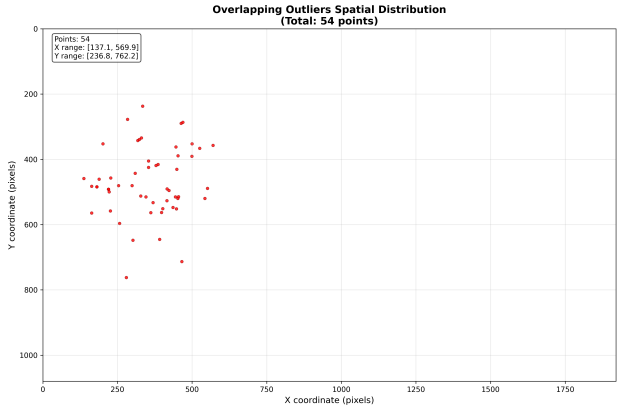
(a) Example with all outliers overlaid.



(b) Depth outliers.

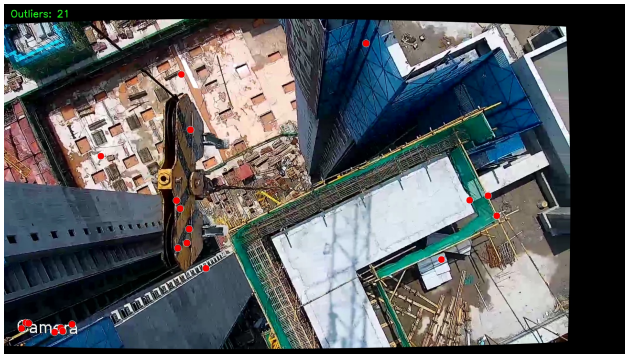


(c) Reprojection outliers.

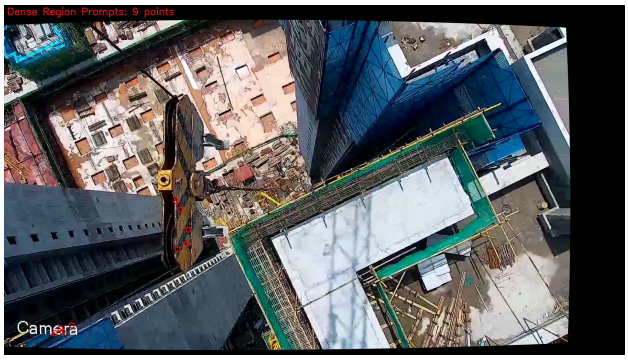


(d) Overlapping outliers.

Figure 4. Spatial distributions of different types of geometric outliers in Scene 1. The overlapping set (red) yields a compact and precise localization of the dynamic hook, while single cues (depth or reprojection) contain scattered noise.



(a) Raw reprojection outliers.



(b) Clustered prompts (DBSCAN).

Figure 5. Effect of DBSCAN clustering on reprojection outliers.

Pal, A., Lin, J. J., Hsieh, S.-H., Golparvar-Fard, M., 2023. Automated vision-based construction progress monitoring in built environment through digital twin. *Developments in the Built Environment*, 16, 100247.

Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L. et al., 2024. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*.

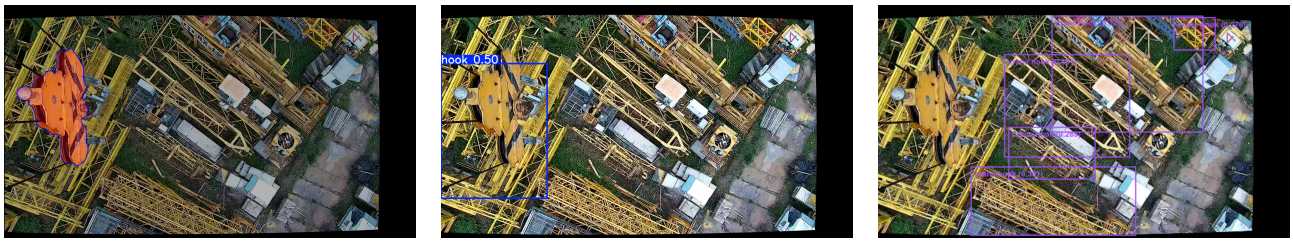
Reja, V. K., Varghese, K., Ha, Q. P., 2022. Computer vision-based construction progress monitoring. *Automation in Construction*, 138, 104245.

Runz, M., Buffier, M., Agapito, L., 2018. Maskfusion: Real-time recognition, tracking and reconstruction of multiple moving objects. *2018 IEEE international symposium on mixed and augmented reality (ISMAR)*, IEEE, 10–20.

Sami Ur Rehman, M., Shafiq, M. T., Ullah, F., 2022. Automated computer vision-based construction progress monitoring: A systematic review. *Buildings*, 12(7), 1037.

Saputra, M. R. U., Markham, A., Trigoni, N., 2018. Visual SLAM and structure from motion in dynamic environments: A survey. *ACM Computing Surveys (CSUR)*, 51(2), 1–36.

Schonberger, J. L., Frahm, J.-M., 2016. Structure-from-motion revisited. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4104–4113.



(a) Our method (b) YOLOv8s (c) Grounding DINO 1.6pro

Figure 6. Comparison of hook segmentation/detection. Learning-based detectors either produce inaccurate or missing detections, while our method robustly segments the dynamic occluder.

Scona, R., Jaimez, M., Petillot, Y. R., Fallon, M., Cremers, D., 2018. Staticfusion: Background reconstruction for dense rgb-d slam in dynamic environments. *2018 IEEE international conference on robotics and automation (ICRA)*, IEEE, 3849–3856.

Wang, Y., Zhao, Y., Petzold, L., 2024. An empirical study on the robustness of the segment anything model (sam). *Pattern Recognition*, 155, 110685.