Incremental Semantics-Aided Meshing from LiDAR-Inertial Odometry and RGB Direct Label Transfer

Muhammad Affan[1], Ville Lehtola[1], George Vosselman[1]

[1]Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, 7522 NB Enschede, Netherlands
(m.affan, v.v.lehtola, george.vosselman)@utwente.nl

Keywords: Meshing, Semantics, LiDAR, Sensor fusion, 3D reconstruction, Scan2USD

Abstract

Geometric high-fidelity mesh reconstruction from LiDAR-inertial scans remains challenging in large, complex indoor environments – such as cultural buildings – where point cloud sparsity, geometric drift, and fixed fusion parameters produce holes, over-smoothing, and spurious surfaces at structural boundaries. We propose a modular, incremental RGB+LiDAR pipeline that generates incremental semantics-aided high-quality meshes from indoor scans through scan frame-based direct label transfer. A vision foundation model labels each incoming RGB frame; labels are incrementally projected and fused onto a LiDAR-inertial odometry map; and an incremental semantics-aware Truncated Signed Distance Function (TSDF) fusion step produces the final mesh via marching cubes. This frame-level fusion strategy preserves the geometric fidelity of LiDAR while leveraging rich visual semantics to resolve geometric ambiguities at reconstruction boundaries caused by LiDAR point-cloud sparsity and geometric drift. We demonstrate that semantic guidance improves geometric reconstruction quality; quantitative evaluation is therefore performed using geometric metrics on the Oxford Spires dataset, while results from the NTU VIRAL dataset are analyzed qualitatively. The proposed method outperforms state-of-the-art geometric baselines ImMesh and Voxblox, demonstrating the benefit of semantics-aided fusion for geometric mesh quality. The resulting semantically labelled meshes are of value when reconstructing Universal Scene Description (USD) assets, offering a path from indoor LiDAR scanning to XR and digital modeling.

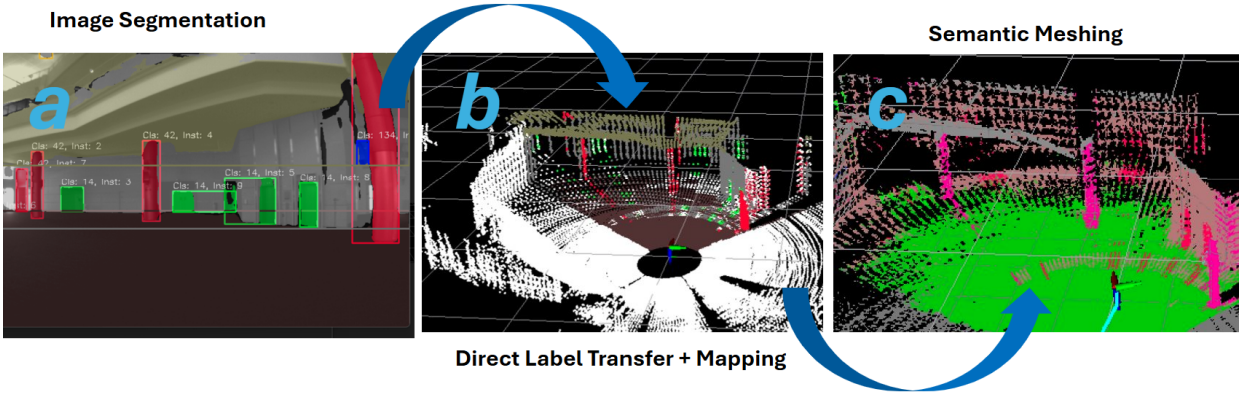


Figure 1. Overview of the proposed pipeline: (a) direct label transfer from VFM (OneFormer) segmented RGB images onto (b) LiDAR scan frames registered via LiDAR-inertial odometry, followed by (c) semantics-aided mesh reconstruction.

1. Introduction

Three-dimensional reconstruction of real-world scenes is a vital enabling technology for digital twins (Boje et al., 2020), architecture/engineering / construction workflows (Tang et al., 2010), immersive XR content for the preservation of cultural heritage (Remondino, 2011, McAvoy et al., 2023), and robotics simulation (Oleynikova et al., 2017). Across these domains, the common need is to convert sensor data into geometrically accurate, semantically-geometrically consistent scene models that can be analyzed, edited, and reused in downstream pipelines. Universal Scene Description (USD) has emerged as a widely adopted scene format for such interoperable 3D content (Pixar Animation Studios, 2023, Halacheva et al., 2025), and we refer to the conversion from raw scans to USD as Scan2USD.

Achieving high geometric fidelity in mesh reconstruction is challenging, particularly at boundaries between structurally different surfaces, where LiDAR sparsity, geometric drift, and fixed fusion parameters lead to artifacts such as holes, over-smoothing, and spurious surfaces. A promising direction is to leverage semantic scene understanding to guide the geometric reconstruction process itself: if the fusion pipeline knows that a voxel belongs to a thin railing rather than a broad wall, it can adapt its truncation distance and integration weight accordingly. Existing approaches that couple semantics with reconstruction are predominantly RGB-D-based (Rosinol et al., 2021, Hughes et al., 2024) and assume close-range, well-textured input, while LiDAR-based meshing systems (Lin et al., 2023, Zhu et al., 2025) operate without semantic information entirely.

In this work, we show that semantic guidance — transferred from RGB onto sparse LiDAR-inertial odometry maps — improves geometric mesh reconstruction. Our contributions are:

https://doi.org/10.5194/isprs-archives-XLIX-B2-2026-335-2026

- A modular, incremental pipeline that transfers Vision Foundation Model (VFM)-based panoptic labels from RGB images to sparse LiDAR-inertial odometry point clouds through ego-motion-compensated projection. Multi-stage boundary filtering improves RGB–LiDAR correspondence under sparse sampling and cross-modal temporal misalignment, in contrast to RGB-D systems that typically rely on dense, time-synchronous 2D–3D associations. The modular design makes the pipeline compatible with other segmentation backbones and readily extensible to multi-camera configurations.
- A label-aware TSDF integration scheme — combining Dirichlet-inspired per-voxel semantic fusion, class-conditioned dynamic truncation, and temporal label consolidation — that adapts fusion to class-specific geometric characteristics, improving mesh quality over geometry-only baselines.
- Evaluation on the Oxford Spires and NTU VIRAL datasets demonstrating improvements over ImMesh and Voxblox, with per-voxel uncertainty metrics that separately characterize semantic and geometric inconsistencies.

The remainder of this paper is organized as follows. Section 2 reviews related work on volumetric reconstruction, LiDAR-based meshing, and semantic 3D mapping. Section 3 describes the proposed method, including the label transfer mechanism, label-aware TSDF fusion, and mesh extraction. Section 4 presents the uncertainty analysis framework. The results are reported in Section 5, and Section 6 concludes the paper. Figure 1 illustrates the three stages of the proposed pipeline.

2. Related Work

2.1 Volumetric 3D Reconstruction

Volumetric approaches based on Truncated Signed Distance Fields (TSDFs) have become a standard framework for dense 3D reconstruction. KinectFusion (Newcombe et al., 2011) demonstrated real-time volumetric fusion from RGB-D streams, and subsequent systems such as Kintinuous (Whelan et al., 2015) and InfiniTAM (Kähler et al., 2015) extended this to larger workspaces. Scalability was further addressed through voxel hashing (Nießner et al., 2013), which allocates memory only for occupied regions, enabling reconstruction of large scenes without a fixed grid. BundleFusion (Dai et al., 2017) added global bundle adjustment for drift-free reconstruction, while Voxblox (Oleynikova et al., 2017) introduced a modular TSDF and ESDF integration pipeline with online mesh extraction that has influenced many later systems (Reijgwart et al., 2020, Grinvald et al., 2019). More recently, adaptive-resolution TSDF schemes (Wang et al., 2023) and neural implicit representations (Zhu et al., 2025) have been explored as alternatives to uniform voxel grids. A common limitation of these systems is that they treat geometry and semantics independently, if at all — fusion parameters such as truncation distance and integration weight are fixed globally, regardless of the surface type being reconstructed.

2.2 LiDAR-Based Meshing

While the volumetric methods above predominantly target RGB-D input, a separate line of work has pursued mesh reconstruction directly from LiDAR scans. ImMesh (Lin et al., 2023) maintains a sparse voxel map and performs incremental per-voxel triangulation via a pull–commit–push scheme, claimed to be the first system to reconstruct triangle meshes of large-scale scenes online on a standard CPU. Mesh-LOAM (Zhu et al., 2025) extends the LOAM paradigm by maintaining a voxel grid updated per scan and applying local marching cubes to generate mesh faces, enabling joint pose estimation and mesh mapping. Both systems demonstrate real-time capability and produce geometrically detailed meshes. However, they are purely geometric: fusion parameters are class-agnostic, and no semantic information is carried on the output mesh. ImMesh and Voxblox serve as geometry-only baselines in our evaluation.

2.3 Semantic 3D Reconstruction

A growing body of work integrates 2D semantic predictions into the 3D reconstruction loop. SemanticFusion (McCormac et al., 2017) fuses CNN-based class predictions into an ElasticFusion surfel map, providing online per-surfel semantics. Kimera (Rosinol et al., 2021) builds a metric–semantic SLAM stack with TSDF meshing on a factor-graph backend, while its successor Hydra (Hughes et al., 2024) advances this to a layered volumetric scene graph encoding geometry, semantics, and instances in real time. At the instance level, PanopticFusion (Nakajima et al., 2019) couples Mask R-CNN with a TSDF backend to maintain per-instance volumes from RGB-D streams; MaskFusion (Runz et al., 2018) leverages instance masks with surfel fusion for real-time object tracking; and Fusion++ (McCormac et al., 2018) builds per-object TSDF submaps from 2D instance masks. A unifying characteristic of these systems is their reliance on RGB-D input: the co-located depth map provides a dense, per-pixel association between the 2D segmentation and 3D geometry. This tight coupling simplifies label transfer but restricts applicability to close-range scenarios (typically < 10 m) with well-textured surfaces. In large indoor environments — museums, stations, cultural buildings — distances routinely exceed this range, lighting varies widely, and surfaces may be textureless or specular. LiDAR sensors are better suited to these conditions, yet to our knowledge no existing system transfers VFM-derived semantic labels onto a LiDAR-inertial odometry map and exploits them to improve the geometric TSDF fusion process. Our work addresses this gap.

3. Methods

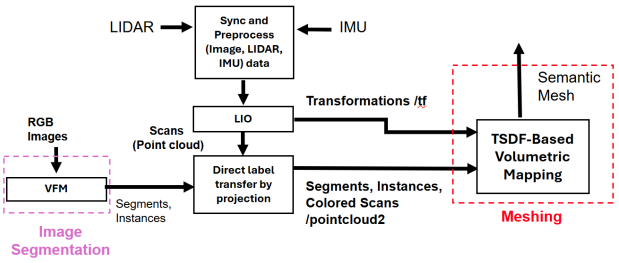


Figure 2. The flowchart of the proposed method.

Our system (Figure 2) incrementally produces a semantically labelled triangle mesh from synchronized LiDAR, RGB, and IMU streams acquired from a calibrated multi-sensor platform, exportable as a USD asset (Section 3.6) for interoperable use in XR and digital modeling pipelines. The pipeline operates at LiDAR rate in three stages: (1) panoptic segmentation of RGB frames using OneFormer (Jain et al., 2023) (dotted pink box in Figure 2); (2) tightly coupled LiDAR-inertial odometry

via FAST-LIO2, with direct label transfer from image masks onto deskewed point clouds (Section 3.2); and (3) semantics-aware TSDF fusion and marching cubes meshing, where truncation and weighting adapt per voxel based on fused labels (Sections 3.3–3.5, dotted red box in Figure 2). The following subsections detail each stage.

3.1 LiDAR-Inertial Odometry

For LiDAR-inertial odometry (LIO), we adopt FAST-LIO2 (Xu et al., 2022) as the localization backbone. Unlike classical feature-based LiDAR odometry (e.g., LOAM), which extracts planar/edge features and can be bottlenecked by feature selection, FAST-LIO2 performs tightly coupled fusion of raw LiDAR and IMU measurements and exploits an incremental k-d tree (ikd-Tree) for fast nearest-neighbor operations. This allows registration of tens of thousands of points per scan in real time on CPU, providing low-drift pose estimates and a reliable sparse map onto which our volumetric fusion and semantic meshing are built.

We use LIO instead of full SLAM with loop closure to avoid discontinuous pose corrections that would invalidate the incrementally built TSDF and require costly re-integration. For the indoor trajectories evaluated (several hundred meters), tightly coupled LIO drift remains within the voxel resolution tolerance.

3.2 Direct Label Transfer

The direct label transfer module projects 2D panoptic segmentation onto 3D LiDAR points using OneFormer (Jain et al., 2023). For each scan, the IMU state is forward-propagated to the camera timestamp t_C, yielding a dynamic LiDAR-to-camera transform

$$\mathbf{T}_{CL} = \big(\mathbf{T}_{WI}(t_C) \cdot \mathbf{T}_{IL} \cdot \mathbf{T}_{LC}\big)^{-1} \cdot \mathbf{T}_{WI}(t_L) \cdot \mathbf{T}_{IL}, \quad (1)$$

where $\mathbf{T}_{WI}(t)$ is the world-to-IMU pose from the LIO filter, t_L the LiDAR timestamp, and $\mathbf{T}_{IL}$, $\mathbf{T}_{LC}$ are the static IMU-to-LiDAR and LiDAR-to-camera extrinsics. Each deskewed point is projected via a pinhole-plus-distortion model with per-point motion correction from the instantaneous camera velocity. To suppress label bleed, morphological mask erosion, boundary distance rejection, and depth-discontinuity checks filter unreliable projections; each surviving point receives the label of the highest-scoring mask at its projected location, while ambiguous and out-of-field-of-view points enter the TSDF as geometry-only observations.

3.3 Label-Aware TSDF Fusion

After direct label transfer, each LiDAR scan arrives at the TSDF integrator as a mixture of semantically labelled and unlabelled points, together with the LIO-estimated sensor pose. These are integrated into a voxel grid using a modified TSDF scheme that maintains per-voxel semantic state alongside the signed distance and weight fields. The key modifications over standard TSDF integration (e.g., Voxblox) are described below.

Per-voxel semantic fusion. Each voxel v maintains a small label histogram $H_v(\ell)$ and a current label ℓ_v. During integration, all labelled points that map to v in the current pass are bundled and a provisional label $\tilde{\ell}_v$ is chosen by majority vote among their non-zero labels. We then add a *single* confidence increment to the histogram entry of $\tilde{\ell}_v$:

$$H_v(\tilde{\ell}_v) \leftarrow H_v(\tilde{\ell}_v) + \eta(r),$$

where $r = \|\mathbf{p}_L\|_2$ is the range of a representative LiDAR return in the sensor frame. The confidence term $\eta(r)$ follows a log-normal decay model:

$$\eta(r) = \mathrm{LogNormal}\big(\ln(\max\{0,\, r - r_0\});\, \mu, \sigma\big), \quad (2)$$

with short-range offset r_0 and log-normal parameters (μ, σ). This encodes the observation that label reliability degrades with range due to increasing projection uncertainty and decreasing LiDAR angular resolution. After the update, the voxel label is set to $\ell_v \leftarrow \arg\max_\ell H_v(\ell)$. To avoid semantic erosion, an existing non-zero label is never overwritten by background.

Greedy segment–label assignment. Before voxel updates, incoming points are grouped into segments $s \in \mathcal{S}$ and each segment is cast into the current map, tallying its overlap with existing map labels using voxels within a narrow TSDF band. We greedily assign s to the label that maximizes this overlap, enforcing per-frame uniqueness; segments with no plausible match receive a new map label. This reduces label fragmentation under sparse coverage.

Temporal label consolidation. Across frames, we maintain pairwise label co-occurrence counts $\Pi(\ell_a, \ell_b)$. When Π exceeds a threshold, the older label is merged into the newer one and per-voxel histograms containing the old label are rewritten in place. This yields stable, long-lived identities as evidence accumulates.

3.4 Dynamic Label-Aware Truncation

Standard TSDF integration applies a fixed truncation distance μ_0 to all voxels, compromising between thin structures and large planar surfaces. Semantic labels provide a prior on expected geometry: *railing* voxels use tighter truncation to preserve fine detail, while *wall* or *floor* voxels use wider truncation for completeness and noise averaging.

We define a dynamic truncation distance that adapts to both the semantic label and the LiDAR range:

$$\mu(\hat{\ell}, r) = \mathrm{clip}\Big(\mu_0\, f_{\mathrm{label}}(\hat{\ell})\, f_{\mathrm{range}}(r),\ \mu_{\min},\ \mu_{\max}\Big), \quad (3)$$

where $\hat{\ell}$ is the majority-voted voxel label, $f_{\mathrm{label}}(\cdot)$ is a class-specific multiplier derived from geometric priors on expected surface thickness (e.g., $f_{\mathrm{label}}(\text{railing}) = 0.5$, $f_{\mathrm{label}}(\text{wall}) = 1.2$), $f_{\mathrm{range}}(r)$ is a bucketed range multiplier compensating for increased range noise, and $\mathrm{clip}(\cdot, \mu_{\min}, \mu_{\max})$ enforces stability limits. Class multipliers are set empirically.

The TSDF update incorporates range-dependent weighting and sparsity compensation. Given sensor origin $\mathbf{o}$, world-frame hit point $\mathbf{p}_G$, and voxel centre $\mathbf{x}$, the signed distance is $\phi = \phi(\mathbf{o}, \mathbf{p}_G, \mathbf{x})$ (positive in front of the surface). The base weight encodes sensor confidence:

$$w_{\mathrm{base}}(\mathbf{p}_L) = \begin{cases} 1, & \text{(constant weighting)}, \\ \dfrac{1}{z^2}, & \text{otherwise, with } z = \big|(\mathbf{p}_L)_z\big|, \end{cases}$$

where the inverse-square mode down-weights distant points that carry higher range noise. Using this base weight $w = w_{\mathrm{base}}(\mathbf{p}_L)$ and a dynamic truncation band $\mu = \mu(\hat{\ell}, r)$, the

TSDF voxel with current value/weight (ϕ_v, w_v) is updated as:

$$\tilde{w} \leftarrow \begin{cases} w\dfrac{\mu+\phi}{\mu-\varepsilon}, & \phi < -\varepsilon \text{ (behind-surface dropoff)}, \\ w, & \text{otherwise}, \end{cases} \tag{4a}$$

$$\tilde{w} \leftarrow \begin{cases} \tilde{w}\,\gamma, & |\phi| < \mu \text{ (sparsity compensation)}, \\ \tilde{w}, & \text{otherwise}, \end{cases} \tag{4b}$$

$$w_{\text{new}} \leftarrow \min\big(w_{\max},\, w_v + \tilde{w}\big), \tag{4c}$$

$$\phi_{\text{new}} \leftarrow \frac{\tilde{w}\,\phi + w_v\,\phi_v}{w_{\text{new}}}, \tag{4d}$$

$$\phi_v \leftarrow \text{clip}(\phi_{\text{new}}, -\mu, \mu), \quad w_v \leftarrow w_{\text{new}}. \tag{4e}$$

Here $\varepsilon \approx \Delta$ (equal to the voxel size, 0.1 m) attenuates weight behind the surface to suppress back-face bleed, $\gamma = 1.5$ is a sparsity compensation factor that counteracts the inherently lower point density of LiDAR returns, and $w_{\max} = 255$ clamps the weight to prevent unbounded accumulation. All three values were set empirically.

3.5 Mesh Extraction and Label Assignment

After each integration step, the mesh is extracted via marching cubes on the TSDF zero-level set. Each triangle inherits the semantic label $\ell_v = \arg\max_\ell H_v(\ell)$ and instance identifier from the nearest voxel centre.

3.6 USD Export

The semantically labelled mesh can be exported as a USD asset by grouping triangles by instance identifier into `UsdGeom.Mesh` primitives with class labels as metadata, establishing a direct path from LiDAR-inertial scanning to USD-compatible engines such as Unreal Engine or RealityKit.

4. Uncertainty Analysis

Standard metrics such as accuracy, completeness, and F1 (Section 5) quantify global reconstruction quality but do not decompose error sources. In our multi-modal pipeline, inconsistencies originate from two distinct sources: geometric (LIO drift, range noise, TSDF discretisation) and semantic (segmentation errors, label flickering, projection misalignment). Since our central claim is that semantic guidance improves geometric reconstruction, we maintain per-voxel uncertainty scores that separately characterise geometric and semantic evidence. This decomposition reveals whether inconsistencies are attributable to the geometric backend, the semantic frontend, or their interaction.

4.1 Notation and Inputs

We maintain per-TSDF-voxel semantic and geometric *uncertainty scores* $s_{\text{sem}}, s_{\text{geom}} \in [0,1]$, where higher values indicate greater uncertainty about reconstruction quality at that voxel. At each integration step, we compute per-voxel evidence cues $\{e_{\text{sem}}, e_{\text{geom}}\}$ from local observations; throughout this section, all $e \in [0,1]$ are oriented such that $e = 0$ denotes full agreement and $e \to 1$ denotes maximal disagreement within the respective modality. These cues are fused into the persistent scores via an exponential moving average (EMA).

Let $\phi(\mathbf{x})$ denote the TSDF grid, and let v index a voxel with center $\mathbf{x}_v$ and voxel size Δ. From the label voxel, we have the winning label $\ell(\mathbf{v})$ and a small histogram of label confidences. We denote by $\mathcal{N}_6(v)$ the 6-neighborhood of v in the voxel grid.

4.2 Semantic Uncertainty Evidence

Dirichlet-inspired Per-voxel Consensus We use the voxel-level label histogram $H_v(\ell) \geq 0$ as unnormalized evidence analogous to a Dirichlet distribution's concentration parameters over semantic labels (excluding background). Define

$$S_v \;=\; \sum_{\ell \neq 0} H_v(\ell), \qquad M_v \;=\; \max_{\ell \neq 0} H_v(\ell).$$

The *label consensus* over each v is the dominance of the most frequent non-background label:

$$\text{consensus}(v) \;=\; \begin{cases} 1, & \text{if } S_v = 0, \\ \dfrac{M_v}{S_v}, & \text{otherwise.} \end{cases}$$

Thus, $\text{consensus}(v) \in [0,1]$, where 1 implies evidence supporting a single non-background label, and lower values reflect mixed evidence or diffuse support, computed as:

$$e_{\text{conc}} \;=\; (1.0 - \mathit{consensus}(v));$$

In the above formulation, e_{conc} is zero under full label consensus and increases toward one as the distribution becomes more mixed, providing a simple Dirichlet-entropy–like impurity measure without explicit histogram normalization.

Neighbor Disagreement. We test for any nonzero 6-neighbor with a different label:

$$e_{\text{neigh}} \;=\; \begin{cases} 1, & \exists\, u \in \mathcal{N}_6(v) :\; \ell(u) \neq 0,\; \ell(u) \neq \ell(v), \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Semantic fusion. With nonnegative weights $w_{\text{conc}}, w_{\text{neigh}}$, we form

$$e_{\text{sem}} \;=\; \frac{w_{\text{conc}}\, e_{\text{conc}} + w_{\text{neigh}}\, e_{\text{neigh}}}{w_{\text{conc}} + w_{\text{neigh}}}. \tag{6}$$

4.3 Geometric Uncertainty Evidence

We combine three cues: TSDF gradient excess, curvature from points, and normal-change across neighbors.

TSDF gradient (excess over 1) The Voxblox integrator maintains a TSDF $\phi(\mathbf{x})$ in *world units* (meters), which ideally approximates a signed distance function near surfaces. For a voxel v with center $\mathbf{x}_v$ and grid spacing Δ, we approximate the spatial derivatives of ϕ by central differences:

$$\partial_x \phi(\mathbf{x}_v) \approx \frac{\phi(\mathbf{x}_v + \Delta\,\mathbf{e}_x) - \phi(\mathbf{x}_v - \Delta\,\mathbf{e}_x)}{2\Delta}, \tag{7a}$$

$$\partial_y \phi(\mathbf{x}_v) \approx \frac{\phi(\mathbf{x}_v + \Delta\,\mathbf{e}_y) - \phi(\mathbf{x}_v - \Delta\,\mathbf{e}_y)}{2\Delta}, \tag{7b}$$

$$\partial_z \phi(\mathbf{x}_v) \approx \frac{\phi(\mathbf{x}_v + \Delta\,\mathbf{e}_z) - \phi(\mathbf{x}_v - \Delta\,\mathbf{e}_z)}{2\Delta}, \tag{7c}$$

where $\mathbf{e}_x, \mathbf{e}_y, \mathbf{e}_z$ are the canonical unit vectors in $\mathbb{R}^3$. The gradient magnitude in world units is then

$$\|\nabla\phi(\mathbf{x}_v)\| \;\approx\; \sqrt{\big(\partial_x\phi(\mathbf{x}_v)\big)^2 + \big(\partial_y\phi(\mathbf{x}_v)\big)^2 + \big(\partial_z\phi(\mathbf{x}_v)\big)^2}.$$

For a smooth Euclidean signed distance field one expects $\|\nabla\phi(\mathbf{x})\| \approx 1$ near the zero level set. We therefore measure *excess* gradient magnitude above this ideal value:

$$e_{\text{grad}}(v) \;=\; \text{clip}_{[0,1]}\left(\frac{\|\nabla\phi(\mathbf{x}_v)\| - 1}{\tau_{\text{grad}}}\right), \tag{8}$$

with scale $\tau_{\text{grad}} > 0$ controlling sensitivity. Large values of e_{grad} indicate strong deviations from a distance-like TSDF (e.g. due to fusion inconsistencies, discretization, or noisy observations).

Curvature from raw points From a voxel-neighborhood point set $\{\mathbf{x}_i\}$ we compute the covariance $\mathbf{C}$ in the global (world) frame and use the $\lambda_{\min}/\text{tr}(\mathbf{C})$ ratio, where $\lambda_{\min}(\mathbf{C})$ is the smallest eigenvalue of $\mathbf{C}$ and $\text{tr}(\mathbf{C})$ is its trace:

$$e_{\text{curv}} \;=\; \frac{\lambda_{\min}(\mathbf{C})}{\text{tr}(\mathbf{C})} \;\in\; [0,1]. \tag{9}$$

Normal-change (TSDF) Let $\mathbf{n}(u)$ be the normalized TSDF gradient at voxel u, i.e. $\mathbf{n}(u) = \nabla\phi(\mathbf{x}_u)/\|\nabla\phi(\mathbf{x}_u)\|$ whenever the gradient is nonzero. For a center voxel v we average the angle (mapped to $[0,1]$) between $\mathbf{n}(v)$ and the normals of its 6-neighbors:

$$\theta(u) = \arccos\big(|\mathbf{n}(v)^\top \mathbf{n}(u)|\big), \tag{10}$$

$$e_{\text{nchg}}(v) = \frac{1}{|\mathcal{S}|}\sum_{u\in\mathcal{S}} \min\Big(1, \frac{\theta(u)}{\pi/2}\Big), \quad \mathcal{S} = \mathcal{N}_6(v). \tag{11}$$

Surface-band gating Let $d = |\phi(v)|$ be the signed-distance magnitude at voxel v, and let $b = \kappa\,\Delta$ denote the surface-band width, where Δ is the voxel size and $\kappa > 0$ is a *global* bandwidth parameter (typically $\kappa = 2$–4). We softly gate two geometric evidence terms, e_{curv} and e_{nchg}. Define

$$g(d) \;=\; \text{clip}_{[\alpha,\,1]}\Big(1 - \tfrac{d}{b}\Big),$$

and update

$$e_{\text{curv}} \leftarrow g(d)\, e_{\text{curv}}, \qquad e_{\text{nchg}} \leftarrow g(d)\, e_{\text{nchg}},$$

where $\alpha \in [0,1]$ is a *global* attenuation floor that sets the minimum off-surface weighting. Thus, κ controls how far from the zero level set geometric evidence contributes, while α limits how much that evidence is down-weighted away from the surface. If soft gating is disabled, geometric cues are set to zero whenever $d > b$.

Geometric fusion All geometric cues are fused in a way that the gradient cue is masked to contribute only in geometrically interesting regions,

$$e_{\text{grad}} \leftarrow \max\big(e_{\text{curv}},\, e_{\text{nchg}}\big)\cdot e_{\text{grad}}, \tag{12}$$

then fuse with nonnegative weights $w_{\text{curv}}, w_{\text{nchg}}, w_{\text{grad}}$:

$$e_{\text{geom}} \;=\; \frac{w_{\text{curv}}\, e_{\text{curv}} + w_{\text{nchg}}\, e_{\text{nchg}} + w_{\text{grad}}\, e_{\text{grad}}}{w_{\text{curv}} + w_{\text{nchg}} + w_{\text{grad}}}. \tag{13}$$

4.3.1 EMA Fusion Over Time & Thresholding

The persistent uncertainty scores are updated via an exponential moving average (EMA) with smoothing factor $\beta \in (0,1]$, in which each past observation's influence decays by $(1-\beta)$ per integration step:

$$s_{\text{sem}} \leftarrow (1-\beta)\, s_{\text{sem}} + \beta\, e_{\text{sem}}, \tag{14}$$

$$s_{\text{geom}} \leftarrow (1-\beta)\, s_{\text{geom}} + \beta\, e_{\text{geom}}. \tag{15}$$

Binary flags are obtained by thresholding:

$$\text{is\_sem} = [s_{\text{sem}} \geq \tau_{\text{sem}}], \quad \text{is\_geom} = [s_{\text{geom}} \geq \tau_{\text{geom}}], \tag{16}$$

where $[\cdot]$ denotes the Iverson bracket (1 if the condition holds, 0 otherwise) and τ_{sem}, τ_{geom} are decision thresholds. These flags partition voxels into four diagnostic categories—semantic uncertainty only, geometric uncertainty only, both, or neither—thereby enabling reconstruction errors to be decomposed by source (Section 5.3).

5. Results

5.1 Qualitative Results

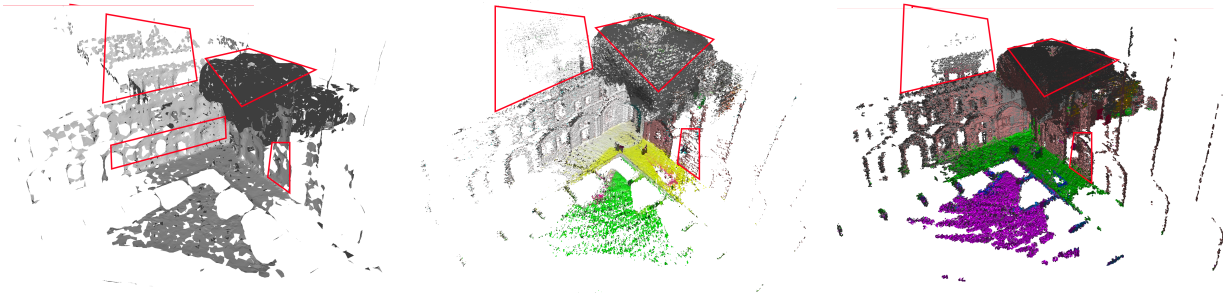
5.1.1 Oxford Spires Scene. In Figure 3c, our reconstruction exhibits robust adaptability to both outdoor and indoor spaces. The generated mesh recovers both dense ground structures and tall architectural features, such as spires and arches, with consistent color integration balancing smoothness and completeness. Voxblox (Figure 3b) again suffers from faded texture and incomplete facade geometry, while ImMesh (Figure 3a) reconstructs coarse walls but omits many thin high-frequency structures.

5.1.2 NTU VIRAL Scene. In Figure 4c, our reconstruction exhibits significantly higher surface completeness and fidelity in annotated areas without over-smoothing. Voxblox (Figure 4b) fails to capture fine details due to static TSDF truncation and limited fusion updates, resulting in blurry and incomplete surfaces. ImMesh (Figure 4a) shows sharp surfaces but lacks geometric continuity in occluded and sparse regions—noticeable on the floor, back wall, and annotated regions.

5.2 Quantitative Results

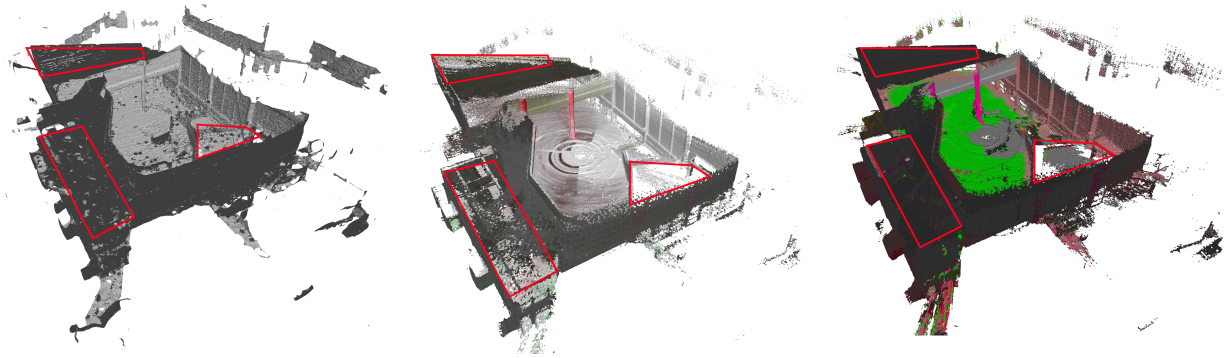
We evaluate on the Christ Church College scene (sequence 2) of the Oxford Spires dataset, selected because it contains an indoor cultural-heritage environment with terrestrial laser scanning (TLS) ground truth — aligning with our target application of large and complex indoor spaces. We adopt the accuracy, completeness, and F1 metrics commonly used in geometric reconstruction evaluation (Dai et al., 2017, Lin et al., 2023, Zhu et al., 2025). Each reconstructed mesh is sampled to 500,000 points and aligned to the ground truth via manual coarse registration followed by ICP refinement. The evaluation framework, including point sampling, outlier filtering, and metric definitions, is described below.

5.2.1 Setup. Let $\mathcal{R} = \{r_i\}_{i=1}^{N_\mathcal{R}}$ be the aligned reconstruction and $\mathcal{G} = \{g_j\}_{j=1}^{N_\mathcal{G}}$ the ground truth. We define the nearest-neighbor distances as the minimum Euclidean distance from each point in the reconstruction $\mathcal{R}$ to the closest point in the



(a) ImMesh (b) Voxblox (c) Ours

Figure 3. 3D reconstruction of Oxford Spires (Christ Church College Scene) dataset



(a) ImMesh (b) Voxblox (c) Ours

Figure 4. 3D reconstruction of NTU VIRAL (NYA01 Scene) dataset

ground truth $\mathcal{G}$, and vice versa:

$$d_{\mathcal{R}\to\mathcal{G}}(r) = \min_{g\in\mathcal{G}} \|r-g\|_2, \quad d_{\mathcal{G}\to\mathcal{R}}(g) = \min_{r\in\mathcal{R}} \|g-r\|_2. \tag{17}$$

To remove spurious correspondences we apply an outlier filter with radius $D_{\max} > 0$. For each $r \in \mathcal{R}$ we keep its distance only if it lies within this radius,

$$\tilde{d}_{\mathcal{R}\to\mathcal{G}}(r) = \begin{cases} d_{\mathcal{R}\to\mathcal{G}}(r), & d_{\mathcal{R}\to\mathcal{G}}(r) \le D_{\max}, \\ \text{discard}, & \text{otherwise}, \end{cases} \tag{18}$$

$$\tilde{d}_{\mathcal{G}\to\mathcal{R}}(g) = \begin{cases} d_{\mathcal{G}\to\mathcal{R}}(g), & d_{\mathcal{G}\to\mathcal{R}}(g) \le D_{\max}, \\ \text{discard}, & \text{otherwise}. \end{cases} \tag{19}$$

Let $\tilde{\mathcal{R}}, \tilde{\mathcal{G}}$ denote the *radius-filtered subsets* of $\mathcal{R}$ and $\mathcal{G}$, i.e., the points whose nearest-neighbor distance does not exceed $D_{\max}$.

5.2.2 Accuracy (RMSE & percentage). Accuracy reports the average geometric error of reconstruction points to the nearest ground-truth surface (lower is better).

Given a max-correspondence threshold τ, we define the inlier set $\mathcal{I}_{\mathcal{R}}(\tau) = \{\, r \in \tilde{\mathcal{R}} \mid d_{\mathcal{R}\to\mathcal{G}}(r) \le \tau \,\}$. Throughout, we set $\tau = 0.30\,\text{m}$, which is several times the expected range and registration noise of our indoor LiDAR setup (maximum range $\approx 150\,\text{m}$), so larger deviations are treated as outliers rather than measurement noise. The (inlier) RMSE accuracy is

$$\text{Acc}_\tau^{\text{RMSE}} = \sqrt{\frac{1}{|\mathcal{I}_{\mathcal{R}}(\tau)|} \sum_{r\in\mathcal{I}_{\mathcal{R}}(\tau)} d_{\mathcal{R}\to\mathcal{G}}(r)^2}\,. \tag{20}$$

Complementarily, the *percentage* (“precision at τ”) reports the fraction of retained reconstruction points within τ:

$$\text{Acc}_\tau^{\%} = \frac{|\mathcal{I}_{\mathcal{R}}(\tau)|}{|\tilde{\mathcal{R}}|}\,. \tag{21}$$

5.2.3 Completeness (RMSE & percentage). Completeness reports the average geometric miss from ground-truth points to the nearest reconstructed surface (lower is better). Using the distances $d_{\mathcal{G}\to\mathcal{R}}(g)$ from (17) and the same threshold τ, define $\mathcal{I}_{\mathcal{G}}(\tau) = \{\, g \in \tilde{\mathcal{G}} \mid d_{\mathcal{G}\to\mathcal{R}}(g) \le \tau \,\}$. The (inlier) RMSE completeness is

$$\text{Comp}_\tau^{\text{RMSE}} = \sqrt{\frac{1}{|\mathcal{I}_{\mathcal{G}}(\tau)|} \sum_{g\in\mathcal{I}_{\mathcal{G}}(\tau)} d_{\mathcal{G}\to\mathcal{R}}(g)^2}\,. \tag{22}$$

The corresponding *percentage* (“recall at τ”) is the fraction of retained ground-truth points within τ:

$$\text{Comp}_\tau^{\%} = \frac{|\mathcal{I}_{\mathcal{G}}(\tau)|}{|\tilde{\mathcal{G}}|}\,. \tag{23}$$

5.2.4 F_1 score (percentage). A thresholded balance of precision and recall reported as a single percentage (higher is better).

For threshold τ, we combine the percentage precision and recall via the harmonic mean:

$$\text{F1}_\tau = \frac{2\,\text{Acc}_\tau^{\%}\,\text{Comp}_\tau^{\%}}{\text{Acc}_\tau^{\%} + \text{Comp}_\tau^{\%}}\,. \tag{24}$$

For conciseness, Table 1 reports the most informative metric per

Method	Acc. (cm) ↓	Comp. (%) ↑	F1 (%) ↑
ImMesh	17.40	95.99%	96.31%
Voxblox	17.24	97.17%	96.98%
Ours	14.54	98.55%	98.58%

Table 1. Results from Oxford Spires dataset. Lower Acc. RMSE is better (↓); higher Completeness and F1 are better (↑).

dimension: accuracy as RMSE (capturing the magnitude of surface error) and completeness as percentage (capturing coverage of the ground truth). In Table 1, we compare the geometric reconstruction quality quantitatively across three mesh pipelines. Our method achieves the lowest accuracy error (14.54 cm) and the highest completeness (98.55%) and F1 (98.58%), improving over Voxblox (17.24 cm / 97.17% / 96.98%) and ImMesh (17.40 cm / 95.99% / 96.31%). The simultaneous reduction in accuracy error and increase in completeness indicates that our pipeline not only fits local surfaces more tightly but also covers a larger fraction of the ground-truth geometry. We attribute the improvement to semantics- and uncertainty-aware TSDF fusion, which preserves high-curvature structures while completing planar surfaces, yielding cleaner, more faithful meshes.

5.3 Uncertainty Results

Dataset	Cert. (Labelled)	Sem. Unc.	Geo. Unc.	Both
NTU-VIRAL	4515082 (83%)	776669 (14%)	122051 (2%)	33979 (1%)
Oxford Spires	1348594 (81%)	285470 (17%)	23388 (1%)	9364 (1%)

Table 2. Proportion of samples under uncertainty categories.

Table 2 and Figure 5 show a limited overlap between geometry and semantic uncertainties, indicating the presence of geometric-semantic inconsistencies. The semantic map (orange) registers class transitions even when geometry is smooth, while the geometric map (yellow) concentrates along depth discontinuities, occluding contours, and high-curvature creases.

This divergent behavior arises from different error sources. Geometric cues capture minor mapping irregularities as boundaries due to inaccuracies in point cloud registration, and sensor noise. In contrast, semantic cues inherit inaccuracies from image segmentation, and projection artifacts. Modularity poses a coupling challenge in addition to data-level discrepancies. Time de-skewing (motion-induced distortions within a single LiDAR scan), surface priors, and voxel regularity assumptions could be refuted when integrating LIO back-ends or meshing libraries. These mismatches propagate through mesh decimation and regularization, reducing overall robustness.

6. Conclusion

We explored incremental 3D mesh reconstruction by directly transferring VFM semantic segmentation labels onto LIO-registered LiDAR point clouds, followed by label-aware TSDF-based volumetric mapping. By incorporating class-conditioned truncation and confidence-weighted fusion, we improved geometric reconstruction quality in terms of both accuracy and completeness over geometry-only baselines. We further analysed the geometric-semantic inconsistencies inherent in the multi-modal TSDF framework, separately characterising their sources. The resulting mesh can be exported as a USD asset for downstream XR tooling. Limitations include odometry drift on longer trajectories, dependence on camera field-of-view coverage and calibration quality, and challenging indoor conditions such as thin structures, specular materials, and heavy clutter — which warrant further investigation.

7. Acknowledgements

This work was supported by the European Union under the Horizon Europe RIA grant agreement No. 101136006 (XTREME).

References

Boje, C., Guerriero, A., Kubicki, S., Rezgui, Y., 2020. Towards a semantic Construction Digital Twin: Directions for future research. *Automation in Construction*, 114, 103179.

Dai, A., Nießner, M., Zollhöfer, M., Izadi, S., Theobalt, C., 2017. BundleFusion: Real-Time Globally Consistent 3D Reconstruction Using On-the-Fly Surface Reintegration. *ACM Trans. Graph.*, 36(4), 1-18. https://doi.org/10.1145/3072959.3054739.

Grinvald, M., Furrer, F., Novkovic, T., Chung, J. J., Cadena, C., Siegwart, R., Nieto, J., 2019. Volumetric Instance-Aware Semantic Mapping and 3D Object Discovery. *IEEE Robotics and Automation Letters*, 4(3), 3037-3044.

Halacheva, A.-M., Miao, Y., Zaech, J.-N., Wang, X., Van Gool, L., Paudel, D. P., 2025. Articulate3d: Holistic understanding of 3d scenes as universal scene description. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Hughes, N., Chang, Y., Hu, S., Talak, R., Abdulhai, R., Strader, J., Carlone, L., 2024. Foundations of Spatial Perception for Robotics: Hierarchical Representations and Real-time Systems. *The International Journal of Robotics Research*. https://doi.org/10.1177/02783649241229725.

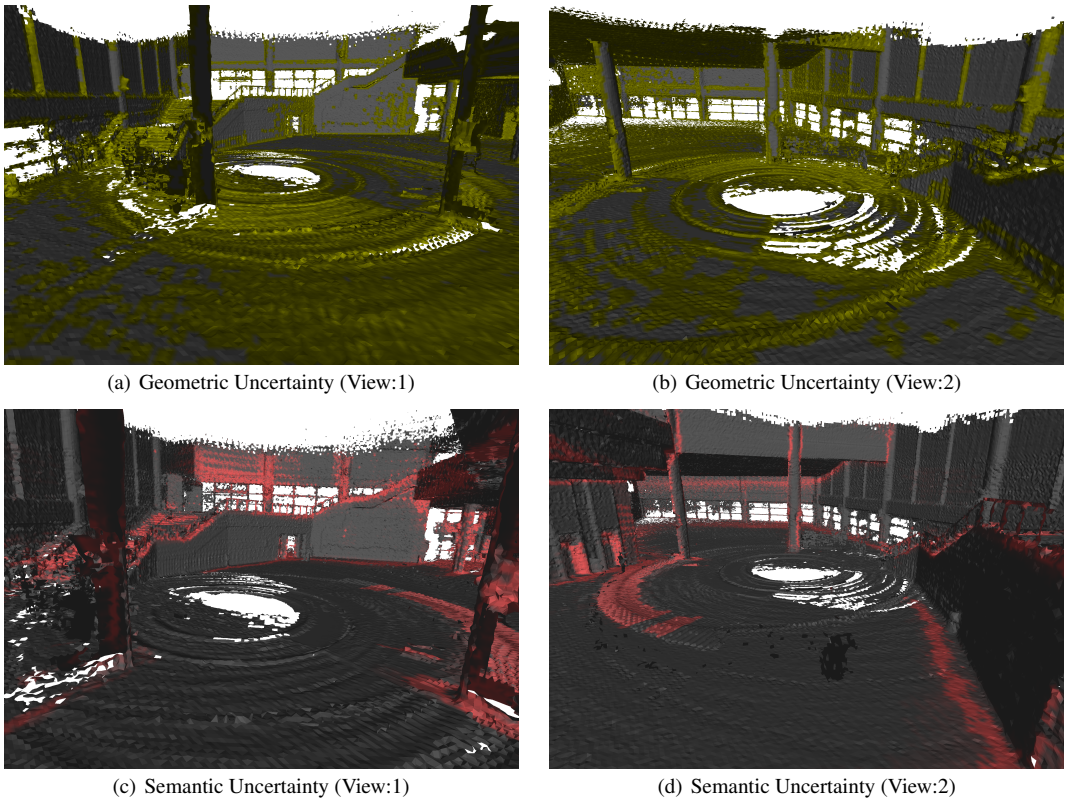
Jain, J., Li, J., Chiu, M. T., Hassani, A., Orlov, N., Shi, H., 2023. Oneformer: One transformer to rule universal image segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2989–2998.

Kähler, O., Adrian Prisacariu, V., Yuheng Ren, C., Sun, X., Torr, P., Murray, D., 2015. Very High Frame Rate Volumetric Integration of Depth Images on Mobile Devices. *IEEE Transactions on Visualization and Computer Graphics*, 21(11), 1241-1250.

Lin, J., Yuan, C., Cai, Y., Li, H., Ren, Y., Zou, Y., Hong, X., Zhang, F., 2023. ImMesh: An Immediate LiDAR Localization and Meshing Framework. *IEEE Transactions on Robotics*, 39(6), 4312-4331.

McAvoy, S., Ristevski, J., Rissolo, D., Kuester, F., 2023. OPENHERITAGE3D: BUILDING AN OPEN VISUAL ARCHIVE FOR SITE SCALE GIGA-RESOLUTION LIDAR AND PHOTOGRAMMETRY DATA. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, X-M-1-2023, 215–222. https://isprs-annals.copernicus.org/articles/X-M-1-2023/215/2023/.

McCormac, J., Clark, R., Bloesch, M., Davison, A. J., Leutenegger, S., 2018. Fusion++: Volumetric object-level SLAM. *International Conference on 3D Vision (3DV)*, 32–41.



(a) Geometric Uncertainty (View:1) (b) Geometric Uncertainty (View:2)

(c) Semantic Uncertainty (View:1) (d) Semantic Uncertainty (View:2)

Figure 5. Analyzing the boundary uncertainty in 3D reconstruction of NTU VIRAL (NYA01) dataset

McCormac, J., Handa, A., Leutenegger, S., Davison, A. J., 2017. SemanticFusion: Dense 3d semantic mapping with convolutional neural networks. *IEEE International Conference on Robotics and Automation (ICRA)*, 4628–4635.

Nakajima, Y., Sucar, E., James, S., Davison, A. J., Tateno, K., 2019. Panopticfusion: Online volumetric semantic mapping at the level of stuff and things. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 4205–4212.

Newcombe, R. A., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A. J., Kohi, P., Shotton, J., Hodges, S., Fitzgibbon, A., 2011. Kinectfusion: Real-time dense surface mapping and tracking. *2011 10th IEEE International Symposium on Mixed and Augmented Reality*, 127–136.

Nießner, M., Zollhöfer, M., Izadi, S., Stamminger, M., 2013. Real-time 3D reconstruction at scale using voxel hashing. *ACM Trans. Graph.*, 32(6), 1-11.

Oleynikova, H., Taylor, Z., Fehr, M., Siegwart, R., Nieto, J., 2017. Voxblox: Incremental 3d euclidean signed distance fields for on-board mav planning. *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1366–1373.

Pixar Animation Studios, 2023. Universal Scene Description (USD). `https://openusd.org/release/index.html`.

Reijgwart, V., Millane, A., Oleynikova, H., Siegwart, R., Cadena, C., Nieto, J., 2020. Voxgraph: Globally Consistent, Volumetric Mapping Using Signed Distance Function Submaps. *IEEE Robotics and Automation Letters*, 5(1), 227-234.

Remondino, F., 2011. Heritage Recording and 3D Modeling with Photogrammetry and 3D Scanning. *Remote Sensing*, 3(6), 1104–1138.

Rosinol, A., Abate, M., Chang, Y., Carlone, L., 2021. Kimera: an Open-Source Library for Real-Time Metric-Semantic Localization and Mapping. *IEEE Transactions on Robotics*, 37(6), 2012–2032.

Runz, M., Buffier, M., Agapito, L., 2018. MaskFusion: Real-time recognition, tracking and reconstruction of multiple moving objects. *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 10–20.

Tang, P., Huber, D., Akinci, B., Lipman, R., Lytle, A., 2010. Automatic reconstruction of as-built building information models from laser-scanned point clouds: A review of related techniques. *Automation in Construction*, 19(7), 829-843.

Wang, W., Hu, Y., Xi, W., Zou, D., Yu, W., 2023. Efficient semantic-aware tsdf mapping with adaptive resolutions. *2023 3rd International Conference on Robotics, Automation and Artificial Intelligence (RAAI)*, 39–45.

Whelan, T., Kaess, M., Johannsson, H., Fallon, M., Leonard, J. J., McDonald, J., 2015. Real-time large-scale dense RGB-D SLAM with volumetric fusion. *The International Journal of Robotics Research*, 34(4-5), 598-626. https://doi.org/10.1177/0278364914551008.

Xu, W., Cai, Y., He, D., Lin, J., Zhang, F., 2022. FAST-LIO2: Fast Direct LiDAR-Inertial Odometry. *IEEE Transactions on Robotics*, 38(4), 2053-2073.

Zhu, Y., Zheng, X., Zhu, J., 2025. Mesh-LOAM: Real-Time Mesh-Based LiDAR Odometry and Mapping. *IEEE Transactions on Intelligent Vehicles*, 10(1), 24-35.