# GS4Buildings: Prior-Guided Gaussian Splatting for 3D Building Reconstruction

Qilin Zhang, Olaf Wysocki, Boris Jutzi

Photogrammetry and Remote Sensing, TUM School of Engineering and Design, Technical University of Munich (TUM), Munich, Germany - (qilin.zhang, olaf.wysocki, boris.jutzi)@tum.de

**Keywords:** Gaussian Splatting, Semantic 3D Building Models, CityGML, LoD2, Geometric Supervision, 3D Reconstruction

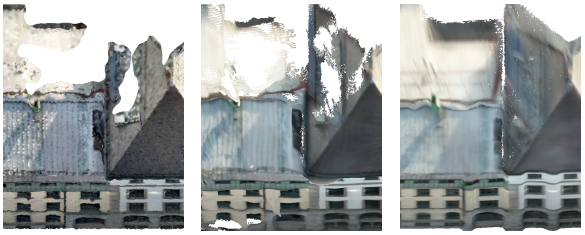
**Abstract**

Recent advances in Gaussian Splatting (GS) have demonstrated its effectiveness in photo-realistic rendering and 3D reconstruction. Among these, 2D Gaussian Splatting (2DGS) is particularly suitable for surface reconstruction due to its flattened Gaussian representation and integrated normal regularization. However, its performance often degrades in large-scale and complex urban scenes with frequent occlusions, leading to incomplete building reconstructions. We propose GS4Buildings, a novel prior-guided Gaussian Splatting method leveraging the ubiquity of semantic 3D building models for robust and scalable building surface reconstruction. Instead of relying on traditional Structure-from-Motion (SfM) pipelines, GS4Buildings initializes Gaussians directly from low-level Level of Detail (LoD)2 semantic 3D building models. Moreover, we generate prior depth and normal maps from the planar building geometry and incorporate them into the optimization process, providing strong geometric guidance for surface consistency and structural accuracy. We also introduce an optional building-focused mode that limits reconstruction to building regions, achieving a 71.8% reduction in Gaussian primitives and enabling a more efficient and compact representation. Experiments on urban datasets demonstrate that GS4Buildings improves reconstruction completeness by 20.5% and geometric accuracy by 32.8%. These results highlight the potential of semantic building model integration to advance GS-based reconstruction toward real-world urban applications such as smart cities and digital twins. Our project is available: `https://github.com/zqlin0521/GS4Buildings`.

## 1. Introduction

Novel view synthesis (NVS) and 3D reconstruction have become fundamental techniques in computer vision and photogrammetry, driven by growing demands in virtual reality, urban planning, and digital twin applications. With advances in computer graphics, Gaussian Splatting (GS) has recently emerged as a powerful approach, offering state-of-the-art performance in photo-realistic rendering and high-fidelity scene reconstruction. 3D Gaussian Splatting (3DGS) (Kerbl et al., 2023) represents scenes using anisotropic 3D Gaussian primitives, whose positions, orientations, and appearance parameters are jointly optimized from multi-view image data. Leveraging an efficient tile-based rasterization technique, 3DGS enables real-time rendering while maintaining high visual fidelity. To better support surface reconstruction tasks, 2D Gaussian Splatting (2DGS) (Huang et al., 2024) extends this framework by introducing flattened Gaussian representations and incorporating normal-based regularization.

While recent innovations have improved GS performance, existing methods still struggle with urban-scale building reconstruction. As illustrated in Figure 1, 2DGS often fails to recover complete building surfaces under occlusions or limited viewpoint coverage. Traditional Multi-View Stereo (MVS) pipelines similarly struggle in textureless or repetitive regions, where reliable feature matching is difficult. These challenges, observed in both GS and MVS approaches, limit their suitability for urban reconstruction tasks that demand high geometric completeness. Concurrently, the development of smart cities has led to the widespread availability of semantic 3D building models with more than 215 million open source building models worldwide (Wysocki et al., 2024). These models are lightweight, volumetric, and typically represented in boundary representation (B-Rep) geometry. They encode reliable geometric information such as roof structures, wall orientations,



(a) MVS (b) 2DGS (c) GS4B (Ours)

Figure 1. Unlike (a) traditional MVS and (b) vanilla 2DGS, both render incomplete building surfaces, (c) our Gaussian4Buildings (GS4B) method reconstructs complete building shape.

and building footprints, and are generally constructed under a planar surface assumption, making them well-suited as geometric priors for surface reconstruction. Although earlier studies have explored the integration of Level of Detail (LoD)2 building models with point clouds (Wysocki et al., 2023; Huang et al., 2020), the building models' potential for guiding GS-based reconstruction remains largely underexplored.

To fill this gap, we propose GS4Buildings, a prior-guided GS framework that enables robust and complete reconstruction of buildings under challenging urban conditions, including large occlusions, complex architectural geometry, and sparse camera viewpoints. Instead of relying on Structure-from-Motion (SfM) pipelines for initialization, our method samples 3D points directly from the LoD2 building model and filters them based on multi-view visibility using known camera intrinsics and extrinsics. The resulting point cloud is used to initialize the Gaussian splats in a geometry-consistent manner. In addition, depth and normal maps projected from the LoD2 geometry are integrated into the optimization process, providing surface-aware constraints that guide the Gaussians toward architecturally consistent reconstructions. We also introduce an optional

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-X-4-W6-2025-249-2025  

building-focused mode that reduces memory and computation overhead while preserving reconstruction quality. This makes GS4Buildings practical for large-scale applications with constrained resources.

In summary, the main contributions of this work are as follows:

- We present GS4Buildings, a SfM-free GS framework for building reconstruction, which directly leverages LoD2 building models and camera parameters to initialize geometry-aware Gaussian splats.
- We enhance the optimization of 2DGS by incorporating depth and normal priors derived from the LoD2 geometry, enabling more complete and accurate reconstruction of building surfaces, particularly under occlusions and limited viewpoint conditions.
- We introduce an optional building-focused reconstruction mode that concentrates computation on architecturally relevant regions, offering improved efficiency while maintaining reconstruction quality.

## 2. Related Work

This section reviews geometry-aware enhancements for GS and recent advances in building reconstruction, the main application focus of this work.

**Gaussian Splatting for 3D Reconstruction** GS has recently gained significant attention as a powerful scene representation technique for novel view synthesis and real-time rendering. While 3DGS (Kerbl et al., 2023) achieves impressive visual quality, it faces several challenges when applied to 3D reconstruction, especially surface reconstruction. The volumetric radiance representation of 3DGS is incompatible with the thin, structured nature of real-world surfaces, and its rasterization suffers from multi-view inconsistencies, leading to noisy or incomplete reconstructions (Huang et al., 2024).

To address the limitations of 3DGS, recent works have proposed incorporating geometric priors into the GS framework. One common strategy is to integrate monocular depth estimation into the optimization process. For instance, DN-Gaussian (Li et al., 2024) and CDGS (Zhang et al., 2025) use predicted depth maps to provide supervision for Gaussian placement and refinement. These methods apply depth-aware losses, such as gradient- or edge-sensitive regularization, to improve geometric consistency. Other works enhance surface modeling by adjusting the Gaussian shape and orientation to better reflect local geometry. 2DGS (Huang et al., 2024) represents each Gaussian as an elliptical disk embedded in a local tangent plane, thereby aligning the primitive with the underlying surface. DN-Splatter (Turkulainen et al., 2024) encourages Gaussians to take on a disc-like shape during optimization, with their smallest scaling axis aligned to the surface normal, guided by depth and normal priors. While these geometry-aware designs improve reconstruction quality under ideal conditions, their reliance on image-based estimations and SfM-based initialization still limits their effectiveness in complex urban environments. For urban building reconstruction in particular, directly leveraging existing 3D models of buildings as geometric priors holds strong potential for improving completeness and accuracy.

**Building Reconstruction** Traditional methods for building reconstruction primarily rely on SfM combined with MVS (Schonberger and Frahm, 2016), or on high-precision terrestrial or airborne Light Detection and Ranging (LiDAR) scanning (Haala and Kada, 2010). SfM-MVS approaches reconstruct dense point clouds or surface meshes from multi-view images and perform well in well-conditioned scenarios, but often struggle with occlusions, repetitive structures, or insufficient viewpoint coverage. LiDAR-based techniques provide accurate geometric measurements and are more robust to such challenges, yet are limited in texture detail and appearance quality, reducing their effectiveness in appearance-aware modeling tasks.

In parallel with image- and point-based reconstruction approaches, structured urban models have become an important component in large-scale 3D city modeling. Among them, the CityGML standard defining the LoD framework (Gröger et al., 2012) offers standardized, hierarchical representations of building geometry, ranging from simple building blocks to highly detailed architectural elements. Currently, low-level LoD1 represented by cuboid-like objects and LoD2 with complex roof shapes and simplified facades can be automatically reconstructed given footprints and aerial observations. This trend enabled a wide adoption of such models worldwide, totalling more than 215 million open data building models in countries such as Poland, the Netherlands, or Japan (Wysocki et al., 2024). Owing to their lightweight representation, structured geometry, and inherent planarity, LoD2 models can serve as valuable geometric priors for aligning and refining the reconstruction process.

Recent advances in urban scene representation have explored both implicit methods such as Neural Radiance Fields (NeRF) (Mildenhall et al., 2021) and explicit representations such as GS (Kerbl et al., 2023). While NeRF excels at photorealistic view synthesis, GS offers better scalability and efficiency. To extend its use in geometry-oriented tasks, recent works have focused on extracting more generalizable 3D models, such as surface meshes, from GS outputs. For example, 2DGS (Huang et al., 2024) improves mesh reconstruction quality, and Gaussian Building Mesh (Gao et al., 2024) further adapts 2DGS to structured architectural modeling using semantic masks. To the best of our knowledge, we are the first to incorporate structured geometric models, such as LoD2, into GS-based reconstruction pipelines for urban buildings, enabling both improved initialization and geometry-aware optimization.

## 3. Methodology

Our GS4B method enhances 2DGS by incorporating geometric priors derived from semantic LoD2 building models. As illustrated in Figure 2, the proposed framework begins with a geometry-consistent Gaussian initialization guided by the LoD2 mesh and camera poses (Section 3.1). We then generate dense depth and normal priors for each view using a raycasting-based approach (Section 3.2). These priors are incorporated into a prior-guided optimization scheme that extends 2DGS training process (Section 3.3). Finally, we extract a surface mesh from the optimized Gaussians using volumetric fusion.

### 3.1 Prior-Based Gaussian Initialization

With the advancement of sensor technology and lower data acquisition costs, accurate camera parameters are now accessible without relying on image-based SfM methods. Leveraging

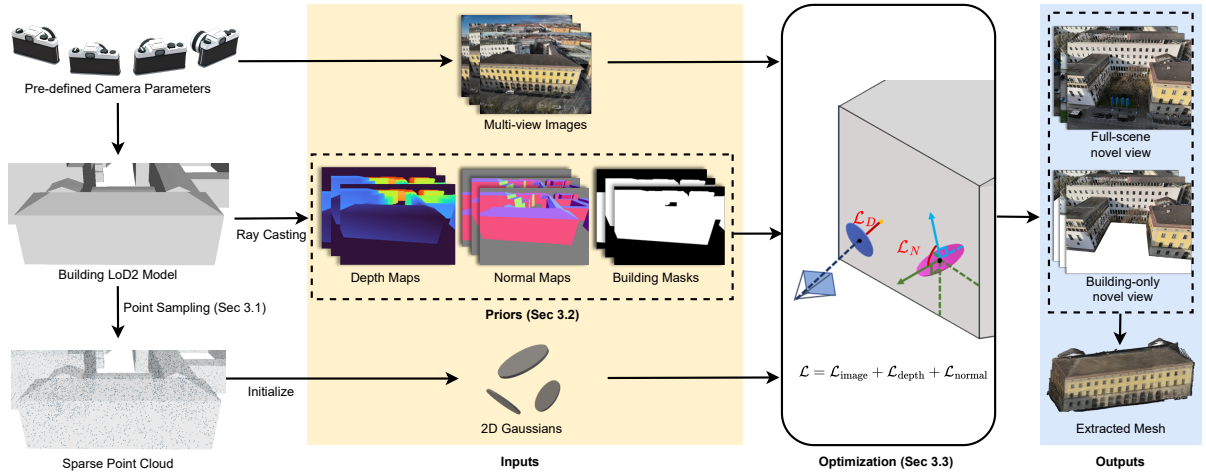



Figure 2. Overview of the proposed GS4Buildings framework. Given a LoD2 building model and camera parameters, our method samples 3D points from the mesh surface and filters them by multi-view visibility to initialize 2D Gaussians. Prior depth and normal maps projected from the LoD2 geometry are incorporated during optimization to guide the reconstruction.

this, we adopt a prior-based initialization strategy guided by the LoD2 building mesh and known camera poses. Unlike conventional GS pipelines such as 2DGS, which rely on SfM to reconstruct sparse geometry, our method directly samples a structurally reliable point cloud from the LoD2 mesh, enabling geometry-consistent initialization of 2D Gaussian primitives. We begin by uniformly sampling 3D points on the mesh surface using face-area-weighted random sampling (Dawson-Haggerty, 2019). For each sampled point $\mathbf{p}_i \in \mathbb{R}^3$ and camera $j$ with position $\mathbf{c}_j$, we define the expected depth as:

$$d_{i,j}^{\text{exp}} = \|\mathbf{p}_i - \mathbf{c}_j\|_2. \tag{1}$$

Let $d_{i,j}^{\text{int}}$ be the depth of the first ray–mesh intersection along the viewing direction from $\mathbf{c}_j$ to $\mathbf{p}_i$. The point is considered visible in view $j$ if:

$$v_{i,j} = \begin{cases} 1, & \text{if } \left|d_{i,j}^{\text{int}} - d_{i,j}^{\text{exp}}\right| < \varepsilon \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where $\varepsilon$ is a distance threshold (e.g., 5 cm). A point is retained if it is visible in at least $k$ views:

$$\sum_{j=1}^{M} v_{i,j} \geq k. \tag{3}$$

For each retained point, we store its 3D coordinates along with its 2D projections $\pi_j(\mathbf{p}_i)$ in all observing views where $v_{i,j} = 1$. The resulting data is organized in a format compatible with standard GS pipelines, enabling seamless Gaussian initialization and consistent association with image projections. Compared to SfM-based initialization, our strategy directly samples from the structured LoD2 geometry, enabling efficient, density-controllable point selection that focuses on geometrically meaningful building regions. This design also avoids the potential limitations of SfM in urban scenarios, such as failure cases or noisy reconstructions.

### 3.2 Prior Generation from Building Models

To guide the optimization of Gaussian primitives toward more accurate and complete surface reconstruction, we generate per-view depth and normal priors from the LoD2 building geometry. These priors provide reliable geometric supervision, especially in regions where image-based cues are unreliable due to occlusion or limited view coverage.

Specifically, we transform the LoD2 mesh $\mathcal{M}_{\text{raw}}$ into the global scene coordinate system defined by the camera poses, resulting in a consistent mesh representation $\mathcal{M}$. Given this aligned mesh and a set of calibrated cameras with intrinsics $\mathbf{K}_j$ and extrinsics $\mathbf{T}_j$, we adopt a raycasting-based approach (Zhou et al., 2018) to synthesize dense, view-aligned geometric priors. For each camera view $j$, the raycasting process computes:

$$\mathbf{D}_j, \mathbf{N}_j = \mathcal{R}(\mathcal{M}, \mathbf{K}_j, \mathbf{T}_j), \tag{4}$$

where $\mathbf{D}_j \in \mathbb{R}^{H \times W}$ is the depth map and $\mathbf{N}_j \in \mathbb{R}^{H \times W \times 3}$ is the corresponding normal map. During raycasting, we also record a binary visibility mask $\mathbf{M}_j \in \{0,1\}^{H \times W}$ that indicates which pixels result in valid mesh intersections; this mask is later used to exclude invalid pixels during loss computation.

Unlike image-based estimations affected by occlusions, lighting, or network uncertainty, prior-guided supervision offers consistent, occlusion-free geometric guidance derived from the building structure, while requiring less computational effort. These priors are later integrated into the optimization process.

### 3.3 2DGS with Prior-Guided Optimization

Building on the prior-guided initialization and dense priors, we extend the 2DGS optimization pipeline with structured supervision from building models.

**2DGS Formulation** 2DGS (Huang et al., 2024) represents a scene using elliptical 2D Gaussian splats, each defined by a center point $\mathbf{p}_k \in \mathbb{R}^3$, two tangent vectors $\mathbf{t}_u, \mathbf{t}_v \in \mathbb{R}^3$, and anisotropic scaling factors $(s_u, s_v)$ within the tangent plane:

$$P(u,v) = \mathbf{p}_k + s_u u \mathbf{t}_u + s_v v \mathbf{t}_v. \tag{5}$$

Each splat is rendered into screen space using an explicit ray–splat intersection strategy. Its density is modeled as a Gaussian:

$$\mathcal{G}(u,v) = \exp\left(-\frac{u^2+v^2}{2}\right), \tag{6}$$

and the final pixel color $\mathbf{c}(\mathbf{x})$ is obtained through alpha compositing:

$$\mathbf{c}(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i \mathbf{c}_i \mathcal{G}_i(\mathbf{u}(\mathbf{x})) \prod_{j=1}^{i-1} (1 - \alpha_j \mathcal{G}_j(\mathbf{u}(\mathbf{x}))). \quad (7)$$

To encourage geometric consistency, 2DGS introduces two regularization losses: a depth distortion loss $\mathcal{L}_d$ and a normal consistency loss $\mathcal{L}_n$:

$$\mathcal{L}_n = \sum_i \omega_i \left(1 - \mathbf{n}_i^\top \mathbf{N}(\mathbf{x}_i)\right), \quad (8)$$

where $\mathbf{n}_i$ is the Gaussian's surface normal and $\mathbf{N}(\mathbf{x}_i)$ is the normal estimated from the rendered depth gradient. The original 2DGS training objective is:

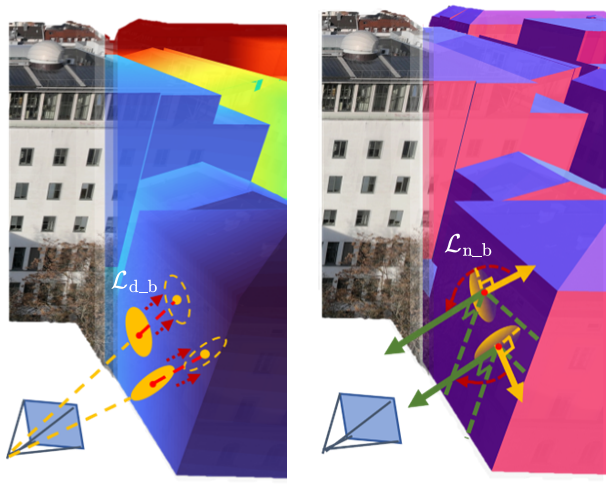
$$\mathcal{L}_{\text{2DGS}} = \mathcal{L}_c + \lambda_d \mathcal{L}_d + \lambda_n \mathcal{L}_n. \quad (9)$$

**Prior-Guided Optimization** To further enhance building-oriented reconstruction, we introduce building-aware supervision derived from LoD2-based priors $(\mathbf{D}_j, \mathbf{N}_j)$ obtained via raycasting (see Section 3.2). Given a binary mask $\mathbf{M}_j$ indicating valid mesh intersections, we define two additional losses:

$$\mathcal{L}_{\text{d\_b}} = \frac{1}{|\mathbf{M}_j|} \sum_{(x,y)\in\mathbf{M}_j} \left| \alpha \cdot \hat{\mathbf{D}}_j(x,y) - \mathbf{D}_j(x,y) \right|, \quad (10)$$

$$\mathcal{L}_{\text{n\_b}} = \frac{1}{|\mathbf{M}_j|} \sum_{(x,y)\in\mathbf{M}_j} \left(1 - \langle \hat{\mathbf{N}}_j(x,y), \mathbf{N}_j(x,y) \rangle\right), \quad (11)$$

where $\hat{\mathbf{D}}_j$ and $\hat{\mathbf{N}}_j$ are the rendered depth and normal maps from the current model, and $\alpha$ is a scale adjustment factor. Figure 3 illustrates how these losses guide 2D Gaussian splats toward accurate surface reconstruction.




(a) Depth supervision $\mathcal{L}_{\text{d\_b}}$ (b) Normal supervision $\mathcal{L}_{\text{n\_b}}$

Figure 3. Prior-guided supervision: (a) Depth priors align splats with surface geometry; (b) Normal priors ensure consistent orientation across views.

We adopt a two-phase training strategy:

- Phase 1: Emphasize $\mathcal{L}_{\text{d\_b}}$ and $\mathcal{L}_{\text{n\_b}}$ to enforce global geometric completeness and correctness using building priors.
- Phase 2: Gradually reduce prior-based loss weights while activating $\mathcal{L}_d$ and $\mathcal{L}_n$ to refine surface smoothness and alignment through local consistency.

The overall training objective becomes:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_c + \lambda_d \mathcal{L}_d + \lambda_n \mathcal{L}_n + \lambda_{\text{d\_b}} \mathcal{L}_{\text{d\_b}} + \lambda_{\text{n\_b}} \mathcal{L}_{\text{n\_b}}, \quad (12)$$

where all loss weights $\lambda$ are time-dependent and scheduled to balance prior guidance and visual fidelity throughout training.

Additionally, we support two training modes. The *building-only* mode restricts both view sampling and optimization to pixels within the building region (as defined by $\mathbf{M}_j$), effectively limiting the GS reconstruction to the building area. The *building-enhanced* mode retains full-scene training while applying additional supervision within building areas. To obtain a structured surface representation, we extract a mesh from the optimized Gaussians using TSDF fusion and Marching Cubes, following the 2DGS pipeline (Huang et al., 2024). This yields a watertight mesh suitable for downstream geometric evaluation.

In summary, our framework combines prior-consistent initialization, occlusion-free geometric supervision, flexible training modes, and TSDF-based mesh extraction to enable accurate and complete 3D reconstruction in complex urban environments.

## 4. Experiments

We evaluated our GS4B method in terms of NVS and 3D reconstruction quality, with comparisons to the original 2DGS and a conventional MVS pipeline. This section outlines the experimental setup and presents qualitative and quantitative results.

### 4.1 Dataset and Evaluation Metrics

We conducted experiments on the publicly available TUM2TWIN dataset (Wysocki and Schwab, 2025), which offers a comprehensive multi-modal capture of the central campus of the Technical University of Munich and its surrounding urban areas. For reconstruction input, we used UAV photographs and LoD2 building models, the latter serving as structured geometric priors for initialization and supervision. The UAV imagery collection (Anders et al., 2025) includes 1,179 high-resolution photographs covering more than 70 buildings. To ensure diversity and coverage, we defined nine representative subsets, each focusing on a distinct building cluster and containing approximately 10–30 images. These subsets span a variety of building forms and urban complexities under real-world conditions.

For 2D photometric evaluation, we adopted three widely used metrics: Peak Signal-to-Noise Ratio (PSNR) for pixel-wise accuracy, Structural Similarity Index (SSIM) for structural fidelity, and Learned Perceptual Image Patch Similarity (LPIPS) for perceptual quality. We further assessed 3D reconstruction quality using two types of reference point clouds from the TUM2TWIN project. The first was laser-scanned point clouds, providing fine-grained geometric accuracy and acquired synchronously with the UAV imagery (Anders et al., 2025). The second was LoD3-derived point clouds, which are structurally coherent with the LoD2 models and served as a reference for assessing structural completeness.

To evaluate geometric accuracy, we used two distance-based metrics: Chamfer Distance (CD), which measures global

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-X-4-W6-2025-249-2025 | © Author(s) 2025. CC BY 4.0 License.

similarity by averaging bidirectional nearest-neighbor distances, and Multi-Scale Model-to-Model Cloud Comparison (M3C2) (Lague et al., 2013), which captures orientation-aware surface deviations. For completeness evaluation, we employed two complementary strategies: threshold-based completeness, adapted from the Tanks and Temples benchmark (Knapitsch et al., 2017), which calculates the fraction of ground-truth points within predefined distance thresholds (e.g., 0.1 m to 0.5 m); and voxel occupancy completeness (VOC), inspired by the geometric completeness (Jäger and Jutzi, 2023), which assesses volumetric coverage by comparing voxel occupancy between the reconstruction and the reference. A voxel is considered occupied if it contains at least a predefined number of points.

### 4.2 Implementation Details

Our implementation was based on 2DGS codebase (Huang et al., 2024). We trained the model for 30,000 iterations across all experiments. Camera pose estimation and MVS-based reconstruction results were generated using Pix4Dmatic (Pix4D SA, 2024) with default parameters. For initialization and prior generation, we employed `trimesh` (Dawson-Haggerty, 2019) for surface sampling and Open3D (Zhou et al., 2018) for raycasting and mesh operations. We followed the two-stage loss scheduling strategy described in Section 3.3, with all 2DGS hyperparameters kept unchanged for fair comparison.

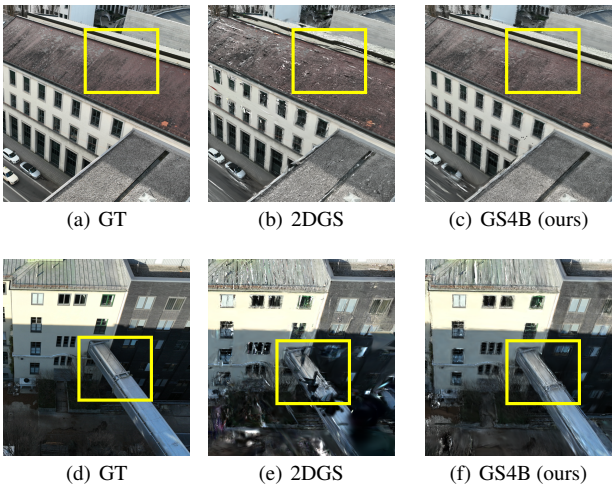
### 4.3 Results

We evaluated our method against two baselines: 2DGS (Huang et al., 2024) and the MVS pipeline from Pix4Dmatic. Evaluation was conducted from two perspectives: (i) 2D NVS quality, compared with 2DGS using perceptual and pixel-wise metrics; and (ii) 3D reconstruction quality, compared with both baselines in terms of accuracy and completeness. This dual-perspective analysis demonstrated the effectiveness of integrating building-aware priors into the GS framework.

**Novel View Synthesis** We assessed the effectiveness of our method in NVS through both qualitative and quantitative analyses. To ensure a fair comparison of scene-level reconstruction capabilities, we compared the results of our method in *building-enhanced* mode against the original 2DGS baseline. Figure 4 shows 2D comparisons of NVS results on two representative scenes, including outputs from 2DGS and our method (GS4B), alongside the ground-truth RGB images. While overall image quality was comparable between the methods, GS4B yielded more stable reconstructions in structurally challenging areas.

For quantitative evaluation, Table 1 reports the average PSNR, SSIM, and LPIPS scores across the selected scenes. Our GS4B method achieved comparable photometric performance to the 2DGS baseline, with slightly higher average PSNR (17.369 vs. 17.190) and SSIM (0.568 vs. 0.552), and a nearly identical LPIPS score (0.260). These results confirm that the integration of structural priors does not compromise image fidelity, while preserving perceptual quality across diverse urban settings.

**Training Convergence** Figure 5 illustrates the training convergence behavior on a representative scene. We visualized the composite image loss and the global normal loss over iterations. While our method initially exhibited a higher image loss due to its prior-based initialization, it quickly converged to a level comparable with the baseline 2DGS method. In terms of normal loss, our method achieved faster convergence and ultimately lower error, indicating improved surface consistency across the entire scene.



(a) GT (b) 2DGS (c) GS4B (ours)

(d) GT (e) 2DGS (f) GS4B (ours)

Figure 4. 2D visual comparison of NVS results on representative building scenes. Our method achieves better visual quality than 2DGS in challenging areas, such as textureless rooftops (b, c) and sparsely observed regions (e, f).

Table 1. 2D quantitative comparison of NVS performance across building scenes from the TUM2TWIN dataset. ↑ indicates higher is better, ↓ indicates lower is better. Best results per row are highlighted in **green**.

| Scene | 2DGS | | | GS4B (ours) | | |
|---|---|---|---|---|---|---|
| | PSNR↑ | SSIM↑ | LPIPS↓ | PSNR↑ | SSIM↑ | LPIPS↓ |
| 1 | **19.593** | **0.645** | **0.222** | 19.224 | 0.610 | 0.274 |
| 2 | 15.785 | 0.434 | 0.355 | **15.852** | **0.484** | **0.342** |
| 3 | 13.520 | 0.298 | 0.348 | **14.384** | **0.346** | **0.306** |
| 4 | **19.040** | 0.700 | 0.156 | 19.031 | **0.710** | **0.155** |
| 5 | 12.332 | 0.271 | 0.482 | **13.557** | **0.322** | **0.425** |
| 6 | **13.449** | **0.402** | **0.365** | 12.870 | 0.347 | 0.414 |
| 7 | **19.654** | **0.731** | **0.135** | 19.514 | 0.701 | 0.146 |
| 8 | **21.596** | 0.746 | 0.135 | 21.566 | **0.798** | **0.143** |
| 9 | 19.737 | 0.742 | 0.150 | **20.319** | **0.797** | **0.138** |
| Avg. | 17.190 | 0.552 | 0.261 | **17.369** | **0.568** | **0.260** |

**Geometric Accuracy and Completeness** We evaluated the 3D geometric quality of the reconstructed scenes from two perspectives: accuracy and completeness. For geometric accuracy, we compared the reconstructed meshes from MVS, 2DGS, and our method against ground-truth laser-scanned point clouds. We used two established 3D geometric metrics: Chamfer Distance (CD), which measures overall geometric similarity, and M3C2, which captures fine-grained surface deviations with local orientation awareness. Table 2 reports the results across the building scenes from the TUM2TWIN dataset, and shows that our method achieves lower errors than the baselines, indicating improved reconstruction fidelity. To assess reconstruction completeness, we used reference point clouds sampled from the LoD3 building models. Unlike laser scans that may be incomplete due to occlusion, the LoD3-derived data provides full building geometry, making it more suitable for evaluating volumetric coverage. As reported in Table 3, we evaluated completeness using two complementary metrics: threshold-based completeness and voxel occupancy completeness (VOC), as introduced in Section 4.1. Our method outperforms both baselines across most tested scenes.

In addition to quantitative comparisons, we provided a visual analysis of reconstruction quality in Figure 6. The figure sum-

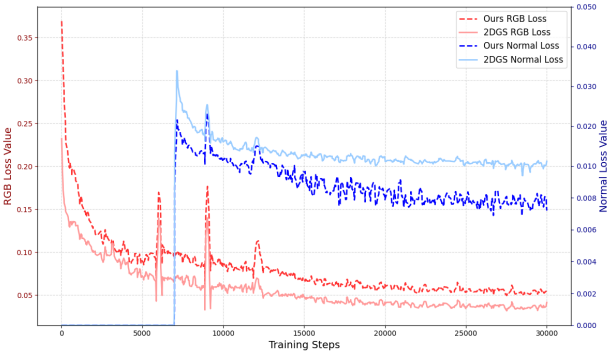



Figure 5. Training convergence of our GS4B compared to 2DGS on a representative scene. The plot shows the composite image loss and the scene-wide normal loss over 30k iterations.

Table 2. 3D quantitative comparison of reconstruction accuracy using laser-scanned point clouds as reference. Chamfer Distance (CD↓) reflects global geometric similarity, while M3C2 Distance (M3C2↓) captures fine-grained surface deviation.

| Scene | MVS | | 2DGS | | GS4B (ours) | |
|---|---|---|---|---|---|---|
| | CD↓ | M3C2↓ | CD↓ | M3C2↓ | CD↓ | M3C2↓ |
| 1 | 0.834 | **0.244** | 0.839 | 0.376 | **0.826** | 0.384 |
| 2 | **0.561** | 0.012 | 1.145 | 0.008 | 1.013 | **0.002** |
| 3 | 2.828 | 0.236 | 1.893 | 0.285 | **1.448** | **0.125** |
| 4 | **0.168** | **0.032** | 0.204 | 0.075 | 0.249 | 0.048 |
| 5 | 4.361 | 2.226 | 2.791 | 1.914 | **2.340** | **1.631** |
| 6 | 1.107 | 0.153 | **0.728** | **0.059** | 0.857 | 0.161 |
| 7 | 3.567 | 0.015 | 2.955 | 0.662 | **2.843** | **0.013** |
| 8 | **0.397** | 0.014 | 0.758 | 0.107 | 0.527 | **0.009** |
| 9 | 2.866 | 0.236 | 2.656 | 0.163 | **2.570** | **0.076** |
| Ave. | 1.854 | 0.352 | 1.552 | 0.405 | **1.408** | **0.272** |

marizes both geometric accuracy and completeness for a representative urban scene, including occluded regions. It highlights differences between MVS, 2DGS, and our method GS4B in terms of mesh reconstruction, M3C2 deviation, and completeness distribution.

**Ablation Studies** We conducted ablation studies to evaluate the contribution of three key components in our approach GS4B, considering both NVS and 3D reconstruction quality. The evaluated components included: (i) the proposed prior-based point cloud initialization, (ii) depth prior supervision via the loss term $\mathcal{L}_{\text{d\_b}}$, and (iii) normal prior supervision via $\mathcal{L}_{\text{n\_b}}$. Table 4 summarizes the quantitative results across five representative metrics: PSNR, SSIM, LPIPS for 2D synthesis quality, and Chamfer Distance (CD) and Voxel Occupancy Completeness (VOC) for geometric accuracy and completeness. Replacing prior-based initialization with an SfM-derived point cloud yielded similar image quality but required more preprocessing. Omitting either depth or normal priors resulted in noticeable degradation in 3D reconstruction accuracy and coverage.

### 4.4 Discussion

This section analyzes the performance of GS4B in both 2D and 3D evaluations and concludes with a discussion of its limitations and future prospects.

**2D View Synthesis Performance** We analyze the quality of NVS from both qualitative and quantitative perspectives, complemented by training convergence behavior. Figure 4 presents visual comparisons across two representative scenes. In the first example (Figures 4a–c), where rooftops and walls lack texture, both methods achieve similar overall photometric quality, but our results appear more coherent in these regions, likely due to geometric priors guiding the training. In contrast, the second scene (Figures 4 d–f) involves sparse viewpoints (only 11 input images), resulting in visibly degraded output from 2DGS, while our method GS4B retains structural consistency. This highlights the benefit of LoD2-derived priors in supporting synthesis under sparse view coverage. However, we also observe that in well-textured scenes with dense views, 2DGS may retain finer local details, suggesting that our regularization might slightly oversmooth certain regions. These trends are further reflected in the quantitative results in Table 1. Our method GS4B achieves slightly better overall performance across all metrics. Notably, in Scene 5, corresponding to Figures 4 d–f, we observe an 18.8% increase in SSIM, indicating improved preservation of structural details under challenging conditions.

Beyond the final rendering quality, training dynamics further highlight the behavior of our method. Figure 5 shows the convergence on a scene with sufficient viewpoint coverage, where 2DGS achieves slightly better overall photometric metrics. Our method begins with a higher image loss due to the prior-based initialization, but quickly converges to a comparable level. The modestly higher final loss is likely due to reconstruction errors in non-building areas, which are not explicitly guided by our priors. In contrast, our global normal loss declines more rapidly and reaches a lower final value, reflecting improved consistency in surface learning. These findings are consistent with the visual and quantitative trends discussed earlier.

**3D Reconstruction Analysis** We evaluate 3D reconstruction quality in terms of both geometric accuracy and completeness. As shown in Table 2, our method GS4B outperforms 2DGS in most scenes and exceeds traditional MVS in several cases. While MVS remains competitive under dense views and rich textures, it degrades in occluded or complex scenarios. Notably, GS4B reduces M3C2 error by 32.8% over 2DGS and 22.7% over MVS, indicating stronger local geometric consistency. As shown in Figure 6 (second column), GS4B achieves lower errors along façades and on the heavily occluded left wall of the building. Nevertheless, fine-scale structures such as doors and windows remain less accurately reconstructed, likely due to the coarse resolution of the LoD2 priors. The completeness results in Table 3 further highlight the strengths of our GS4B approach. Across all thresholds and in most scenes, GS4B outperforms both MVS and 2DGS in both threshold-based completeness and voxel occupancy completeness (VOC). Since many input views cover only partial façades, while the LoD3 ground truth models represent complete buildings, the absolute completeness values remain relatively low. However, our method still achieves significant improvements. For instance, in terms of VOC, GS4B shows a 63.9% increase compared to MVS.

These improvements arise from two main factors. First, the 2DGS representation, with its planar Gaussian splats, provides a more continuous approximation of surfaces compared to point-based MVS, which is more susceptible to fragmentation in low-texture or occluded regions. This explains why even the standard 2DGS generally achieves higher completeness than MVS. Second, our method benefits from integrating geometric priors derived from semantic LoD2 models. Unlike methods relying solely on images, which are constrained by occlusions and limited viewpoints, our GS4B framework leverages the volu-

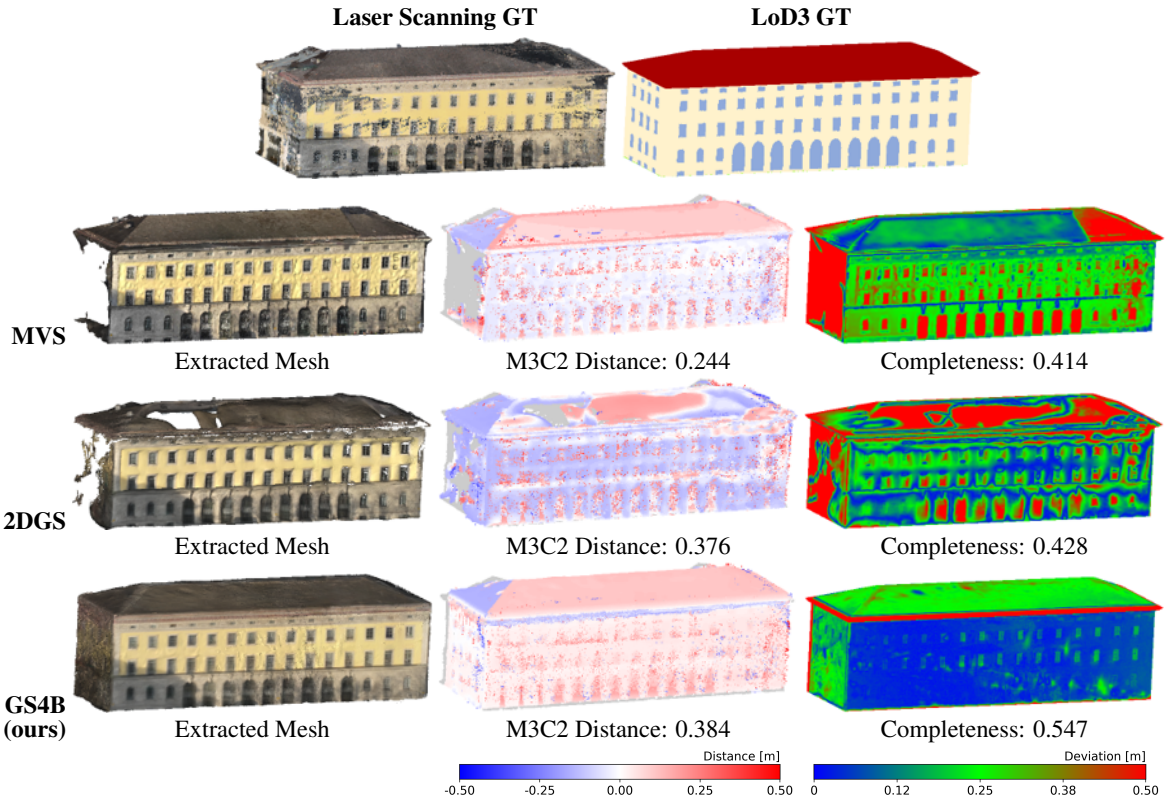



Figure 6. Visual comparison of 3D reconstruction accuracy and completeness in a representative urban scene with partial occlusion (i.e., trees, cars). The top row shows ground-truth (GT) point clouds from laser scanning (for accuracy) and LoD3 building models (for completeness). Rows 2–4 present results from MVS, 2DGS, and our method, including extracted meshes, M3C2 error maps, and completeness visualizations. For completeness, distances from ground-truth points to their nearest reconstructed points are visualized on the GT point cloud using a blue–green–red gradient. Reported values indicate the percentage of points within a 0.5 m threshold.

Table 3. Completeness comparison of 3D reconstruction results using LoD3-derived point clouds as reference. Threshold-based Completeness reports the percentage of ground truth points recovered within 0.1 m, 0.2 m, and 0.5 m. Voxel Occupancy Completeness (VOC) reflects volumetric coverage based on shared voxel occupancy. All values are in $[0, 1]$ and higher is better (↑).

| Scene | MVS | | | | 2DGS | | | | GS4B (ours) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0.1m↑ | 0.2m↑ | 0.5m↑ | VOC↑ | 0.1m↑ | 0.2m↑ | 0.5m↑ | VOC↑ | 0.1m↑ | 0.2m↑ | 0.5m↑ | VOC↑ |
| 1 | 0.024 | 0.131 | 0.414 | 0.190 | 0.098 | 0.208 | 0.428 | 0.274 | **0.265** | **0.387** | **0.547** | **0.408** |
| 2 | 0.065 | 0.140 | 0.261 | 0.154 | 0.062 | 0.137 | 0.248 | 0.184 | **0.066** | **0.141** | **0.281** | **0.188** |
| 3 | 0.012 | 0.034 | 0.096 | 0.051 | **0.046** | 0.096 | 0.215 | 0.132 | 0.038 | **0.105** | **0.273** | **0.170** |
| 4 | 0.057 | 0.237 | 0.419 | 0.242 | 0.142 | 0.223 | 0.519 | 0.345 | **0.169** | **0.322** | **0.540** | **0.396** |
| 5 | 0.000 | 0.010 | 0.015 | 0.002 | 0.001 | 0.033 | 0.082 | 0.004 | **0.006** | **0.062** | **0.143** | **0.012** |
| 6 | 0.102 | 0.208 | 0.472 | 0.281 | **0.125** | **0.253** | **0.551** | 0.314 | 0.110 | 0.241 | 0.472 | **0.319** |
| 7 | 0.002 | 0.007 | 0.021 | 0.013 | **0.008** | 0.016 | **0.045** | 0.025 | 0.004 | **0.020** | 0.042 | **0.028** |
| 8 | 0.022 | 0.146 | 0.371 | 0.107 | 0.042 | **0.199** | 0.398 | 0.167 | **0.052** | 0.137 | **0.633** | **0.245** |
| 9 | 0.078 | 0.171 | 0.334 | 0.187 | 0.076 | 0.163 | 0.374 | 0.221 | **0.079** | **0.195** | **0.394** | **0.238** |
| Avg. | 0.040 | 0.120 | 0.267 | 0.136 | 0.067 | 0.148 | 0.318 | 0.185 | **0.088** | **0.179** | **0.369** | **0.223** |

Table 4. Ablation study results on NVS and 3D reconstruction quality. Metrics include PSNR (dB), SSIM, and LPIPS for novel view synthesis (NVS); Chamfer Distance (CD) and Voxel Occupancy Completeness (VOC) for 3D geometry.

| Method | PSNR↑ | SSIM↑ | LPIPS↓ | CD↓ | VoC↑ |
|---|---|---|---|---|---|
| With SfM Init | 19.062 | **0.613** | 0.277 | 1.096 | 0.325 |
| Without Depth Prior | 18.756 | 0.601 | 0.279 | 2.425 | 0.234 |
| Without Normal Prior | 19.154 | 0.609 | 0.276 | 2.017 | 0.395 |
| **Ours (Full)** | **19.224** | 0.610 | **0.274** | **0.826** | **0.408** |

metric nature of LoD2 models to provide a complete 3D structure of buildings. This enables supervision not only in visible areas but also in regions where image cues are unreliable or entirely missing. While images contribute to realistic appearance reconstruction, especially in non-building surroundings such as vegetation and ground surfaces, the inclusion of 3D structural priors allows the geometry of occluded or unseen areas to be reconstructed more faithfully.

However, the stronger regularization introduced by our priors can, in some cases, attenuate fine-grained geometric details. As shown in Figure 6, our method produces smooth and complete surfaces but tends to miss thin structures, which MVS is able to capture more accurately when sufficient image observations are available. This trade-off reflects the inherent challenge of balancing structural completeness with local geometric precision.

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-X-4-W6-2025-249-2025 | © Author(s) 2025. CC BY 4.0 License.

**Limitations and Outlook** Our method performs well in challenging urban reconstruction scenarios, especially under occlusion and sparse viewpoints. However, in scenes with dense observations and rich textures, the strong regularization from LoD2 priors may oversmooth fine details such as eaves, windows, and doors. This underscores the need for a more adaptive use of structural priors, such as dynamically weighting their influence based on local scene characteristics. While LoD2 building models are becoming increasingly available in many countries, our method still depends on their availability and quality. Future work may explore alternative priors, such as more widely accessible LoD1 models or CAD-based representations, to improve generalizability across diverse urban environments.

## 5. Conclusions

We present GS4Buildings, a Gaussian Splatting framework guided by semantic LoD2 building models. Our method enhances both 2D view synthesis and 3D reconstruction, particularly in challenging scenarios with occlusion and sparse viewpoints. Experiments on the TUM2TWIN dataset demonstrate consistent improvements over traditional MVS and standard 2DGS, including a 32.8% reduction in surface deviation (M3C2) and a 63.9% increase in voxel occupancy completeness. These results highlight the effectiveness of integrating semantic priors into GS-based pipelines for scalable and robust urban reconstruction. Building on these results, future work could explore dynamic prior weighting and broader prior sources to enhance reconstruction fidelity across diverse scenes, including high-rise and geometrically complex buildings. Such advancements may also benefit digital city applications, including urban monitoring and digital twin updates.

## 6. Acknowledgements

The authors gratefully acknowledge the Professorship of Remote Sensing Applications at TUM for their support in the acquisition and preprocessing of the TUM2TWIN dataset. Special thanks are extended to Katharina Anders and Jiapan Wang for their dedicated contributions. The authors also thank Benjamin Busam for his insightful comments and helpful suggestions that contributed to improving this work.

## References

Anders, K., Wang, J., Wysocki, O., Huang, X., Liu, S., 2025. Uav laser scanning and photogrammetry of tum downtown campus. Zenodo. `https://doi.org/10.5281/zenodo.14899378`.

Dawson-Haggerty, M., 2019. Trimesh: A python library for triangular meshes. `https://trimsh.org/` (24 April 2025).

Gao, K., Li, L., He, H., Lu, D., Xu, L., Li, J., 2024. Gaussian Building Mesh (GBM): Extract a Building's 3D Mesh with Google Earth and Gaussian Splatting. *arXiv preprint arXiv:2501.00625*.

Gröger, G., Kolbe, T. H., Nagel, C., Häfele, K.-H., 2012. OGC City Geography Markup Language CityGML Encoding Standard.

Haala, N., Kada, M., 2010. An update on automatic 3D building reconstruction. *ISPRS Journal of Photogrammetry and Remote Sensing*, 65(6), 570–580.

Huang, B., Yu, Z., Chen, A., Geiger, A., Gao, S., 2024. 2d gaussian splatting for geometrically accurate radiance fields. *ACM SIGGRAPH 2024 conference papers*, 1–11.

Huang, H., Michelini, M., Schmitz, M., Roth, L., Mayer, H., 2020. LOD3 building reconstruction from multi-source images. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43, 427–434.

Jäger, M., Jutzi, B., 2023. 3d density-gradient based edge detection on neural radiance fields (nerfs) for geometric reconstruction. *arXiv preprint arXiv:2309.14800*.

Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G., 2023. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4), 139–1.

Knapitsch, A., Park, J., Zhou, Q.-Y., Koltun, V., 2017. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36(4), 1–13.

Lague, D., Brodu, N., Leroux, J., 2013. Accurate 3D comparison of complex topography with terrestrial laser scanner: Application to the Rangitikei canyon (NZ). *ISPRS journal of photogrammetry and remote sensing*, 82, 10–26.

Li, J., Zhang, J., Bai, X., Zheng, J., Ning, X., Zhou, J., Gu, L., 2024. Dngaussian: Optimizing sparse-view 3d gaussian radiance fields with global-local depth normalization. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 20775–20785.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., Ng, R., 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1), 99–106.

Pix4D SA, 2024. Pix4Dmatic Software, Version 1.71.0. `https://www.pix4d.com/product/pix4dmatic` (24 April 2025).

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

Turkulainen, M., Ren, X., Melekhov, I., Seiskari, O., Rahtu, E., Kannala, J., 2024. DN-Splatter: Depth and Normal Priors for Gaussian Splatting and Meshing. *arXiv preprint arXiv:2403.17822*.

Wysocki, O., Schwab, B., 2025. Tum2twin project: High-accuracy digital representations of the tum campus. `https://tum2t.win/` (24 April 2025).

Wysocki, O., Schwab, B., Beil, C., Holst, C., Kolbe, T. H., 2024. Reviewing Open Data Semantic 3D City Models to Develop Novel 3D Reconstruction Methods. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 48, 493–500.

Wysocki, O., Xia, Y., Wysocki, M., Grilli, E., Hoegner, L., Cremers, D., Stilla, U., 2023. Scan2lod3: Reconstructing semantic 3d building models at lod3 using ray casting and bayesian networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6548–6558.

Zhang, Q., Wysocki, O., Urban, S., Jutzi, B., 2025. CDGS: Confidence-Aware Depth Regularization for 3D Gaussian Splatting. *arXiv preprint arXiv:2502.14684*.

Zhou, Q.-Y., Park, J., Koltun, V., 2018. Open3D: A modern library for 3D data processing. *arXiv preprint arXiv:1801.09847*.