Automatic Segmentation of 3D Gaussian Splatting in Cultural Heritage Complex

Widiatmoko A Fadilah [1], Virgile Gauthier [1], Arnadi Murtiyoso [1], Tania Landes [1], Pierre Grussenmeyer [1]

[1] Université de Strasbourg, CNRS, INSA Strasbourg, ICube Laboratory UMR 7357,
Photogrammetry and Geomatics Group, 67000, Strasbourg, France.
(widiatmoko-azis.fadilah, virgile.gauthier, arnadi.murtiyoso, tania.landes, pierre.grussenmeyer)@insa-strasbourg.fr

Keywords: Photogrammetry, Radiance Field, Gaussian Splatting, Segmentation, Urban Heritage.

Abstract

3D Gaussian Splatting (3DGS) has emerged as a promising method for photorealistic scene reconstructions, yet its application to semantic segmentation in real-world heritage documentation remains underexplored. This study proposes and evaluates an automated semantic 3DGS segmentation pipeline integrating the Segment Anything Model 3 (SAM 3) with per-class prompting for Gaussian reconstruction, applied to a nadiral UAV dataset of the Siti Inggil heritage complex in Cirebon, Indonesia. Segmentation performance of four semantic classes (ground, roofs, vegetations, and water bodies) were assessed against manually segmented 2D and 3D reference data, supplemented by geometric accuracy assessment via the M3C2 analysis. Results reveal both the promise and the inherent challenges of applying 3DGS segmentation to complex real-world heritage scenes, where acquisition geometry, surface characteristics, and foundational model limitations can be observed.

1. Introduction

Recent advancements in radiance fields methods, such as 3D Gaussian Splatting (3DGS) introduced by Kerbl et al. (2023), have achieved breakthroughs in 3D reconstruction. Several derivatives have also extended 3DGS with segmentation capabilities. 3DGS solutions demonstrated by Cen et al. (2025) and Ye et al. (2024), enable the generation of photorealistic rendering alongside scene understanding that is particularly valuable for 3D surveying applications.

Segmentation of 3D data accelerates surveying workflows by enabling efficient modelling and prediction using spatial data (Murtiyoso et al., 2022). For an urban heritage complex, 3D segmentation facilitates better scene understanding and supports informed decision-making by end-users. The use of 3DGS segmentation methods shows promise for these applications, yet its implementation on real-world data remains underexplored. Such scenes may introduce challenges absent in controlled or interior dataset.

Several approaches have been proposed to integrate segmentation into 3DGS. These spans from feature-based, point-based, and projection-based methods. Among these, projection-based methods have gained the most traction, employing foundational models such as the Segment Anything Model (SAM) by Kirillov et al. (2023), Contrastive Language-Image Pretraining (CLIP) by Radford et al. (2021), Detecting Everything in Videos (DEVA) by Cheng et al. (2023), and Self-Distillation with No labels (DINO) by Caron et al. (2021) to generate and enrich image masks that are subsequently lifted into the 3DGS model.

This study proposes a 3DGS segmentation method for heritage site reconstruction and segmentation. Due to their complex geometric and decorative features, heritage architecture serves as a challenging test case. Performance is assessed through quantitative and qualitative comparative analysis against manually segmented 2D and 3D reference data.

2. Related Works

This section introduces 3DGS as a scene representation method, followed by an overview of segmentation and its development within radiance field methods. This section will be concluded with the foundational segmentation model employed in this study.

Radiance field refers to a representation that describes the behavior and distribution of light in 3D space (Tosi et al., 2025). Among many radiance field methods, 3DGS by Kerbl et al. (2023) is one of the prominent implementations that many researchers have worked upon. It encapsulates how light interacts with the surrounding environment explicitly using 3D Gaussians as its primitives, allowing faster rendering capabilities. Each Gaussian consists of a set of geometric (i.e., position and 3D covariance matrix) and appearance parameters (i.e., opacity and color). The color in 3DGS is represented by spherical harmonics (SH), which is a set of values that describe view-dependent colors. 3DGS takes Structure from Motion (SfM) output and densifies through an optimization using densification strategies such as Adaptive Density Control (ADC). ADC allows Gaussians to be split in over-reconstructed areas, cloned in under-reconstructed areas, and pruned when the Gaussian opacity is below the threshold. Thus, 3DGS generates a correct and realistic 3D scene representation (Kerbl et al., 2023).

Segmentation in computer vision is a computational technique that partitions objects within images, video frames, or 3D scenes. It is classified into four broad types: (1) semantic segmentation, which groups objects sharing the same class label; (2) instance segmentation, which detects and delineates individual object instances; (3) panoptic segmentation, a 2D task unifying both semantic and instance segmentation; and (4) part segmentation, a 3D task that further decomposes segmented instances into constituent parts (He et al., 2024; Minaee et al., 2020; Vinodkumar et al., 2023). These capabilities support a broad range of scene understanding applications, including precise documentation in cultural heritage preservation where accurate object delineation is essential. Within 3D segmentation, He et al. (2024) organize the methods according to input modality and task goal, spanning RGB-D, point-based, image projection, voxel, and video-based approaches. Among these, image projection-

based methods have seen the widest adoption in 3DGS segmentation, leveraging foundational models to extract 2D masks that are subsequently lifted into 3D Gaussian space.

Efforts to enrich radiance fields with segmentation capabilities have been developing since the early days of Neural Radiance Fields (NeRF). Zhi et al. (2021) augmented NeRF with a segmentation renderer, while Kerr et al. (2023) introduced Language Embedded Radiance Fields (LERFs), embedding CLIP-based language grounding into NeRF to enable open-vocabulary 3D queries. Kim et al. (2024) later proposed GARField, decomposing scenes into hierarchical semantic groups using SAM-generated masks. With the introduction of 3DGS, this line of research accelerated considerably, with projection-based methods emerging as the dominant paradigm. SAM became the most widely adopted masking source, employed by OmniSeg3D (Ying et al., 2023), Gaussian Grouping (Ye et al., 2024), SAGA (Cen et al., 2025, 2024), Segment Any 4D Gaussians (Ji et al., 2024), and SA-GS (Xiong et al., 2024). Other foundational models have also been explored. LangSplat by Qin et al. (2024) leverages both SAM and DINO to embed CLIP features onto 3D Gaussians, subsequently tested on an outdoor aerial scene by Zaouali et al. (2025). Meanwhile, LabelGS by Zhang et al. (2025) employs DEVA as an alternative masking source.

Beyond projection-based methods, RT-GS2 by Jurca et al. (2024) takes a feature-based approach, extracting view-independent 3DGS features and applying a feature fusion algorithm to enhance cross-view semantic consistency without relying on external mask supervision. Jurski (2024) further explores point-based segmentation by applying PointNet++ directly to Gaussian-derived point structures. More recently, Wang et al. (2026) introduce Fusion3DGS, which performs 3D instance segmentation by jointly learning from 2D image and 3D Gaussian features, using SAM-generated masks as the source of 3D training labels without requiring manual 3D annotation. Together, these works demonstrate that while projection-based methods remain dominant, the field is actively exploring feature-based, point-based, and fusion-based directions to address their limitations.

The Segment Anything project was introduced by Meta AI Research ‘FAIR’ in 2023 (Kirillov et al., 2023). The aim of the project was to build a foundational model for segmentation of images that powers data annotation and enables zero-shot transfer via prompt. The initial version of SAM allows users to obtain segmented images by selecting a desired object. SAM 2 (Ravi et al., 2024) extended SAM capability to video through memory-based temporal tracking. SAM 3 (Carion et al., 2025) is the latest version of SAM that introduces segmentation through concept prompting. With SAM 3, a short noun, image exemplars, or a combination of those two can be used to obtain the desired segmented objects. Moreover, unlike its predecessors, SAM 3 performs exhaustive instance segmentation, which detects and segments all instances that correlate with the input prompt. Collectively, these works establish projection-based methods as the dominant paradigm and with its recent version of SAM 3, this study utilizes this technology to achieve scene segmentation in 3DGS.

3. Data and Research Methodology

3.1 Research Data

This study employs RGB Images acquired in 2025 of the Siti Inggil complex within the Kasepuhan Palace site in Cirebon, Indonesia. The dataset comprises 702 images captured at 30 m flight altitude with nadir camera orientation and standard grid flight configuration. The heritage complex presents significant challenges for segmentation, including occluded architectural features caused by dense surrounding vegetation and adjacent residential buildings.

3.2 Data Processing Pipeline

We present an automated method for semantic 3DGS segmentation following the idea in 3D point cloud segmentation proposed by Murtiyoso et al. (2022). Figure 1 illustrates the complete data processing workflow across five main stages, from image orientation to quantitative evaluation against manually segmented reference data. The semantic segmentation comprises four classes: ground (including road and path), roof (all elevated structural elements excluding terraces), vegetation (trees and shrubs), and water bodies (river). This classification scheme draws from standard photogrammetric categories and the taxonomy proposed by Pan et al., (2024), with one notable change on ‘roof’ instead of ‘building’. This substitution was caused by a technical limitation of SAM 3 in conjunction with the characteristics of the dataset. The limitation will be discussed further in section 4.1.2.

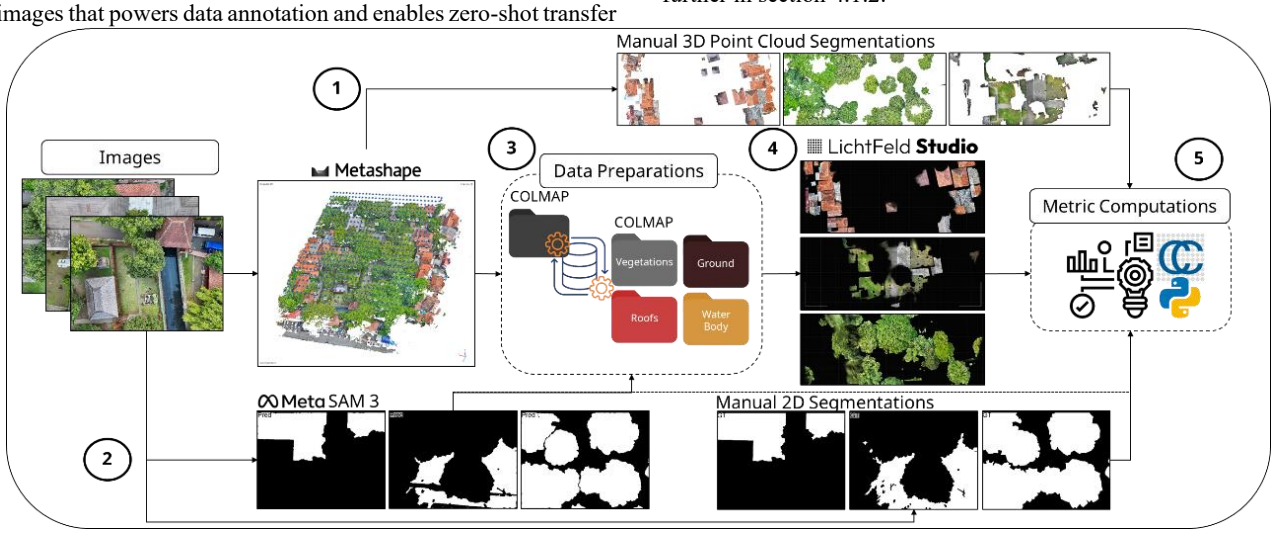

Figure 1. General workflow of proposed segmentation method using SAM 3 to generate semantic segmented 3D Gaussians. (1) image orientation and dense point cloud generation, (2) 2D semantic segmentation using SAM 3, (3) per-class COLMAP data preparation, (4) per-class 3DGS training, and (5) quantitative evaluation against manually segmented reference data.

Image orientation process was performed in Agisoft Metashape at highest-quality settings, followed by dense point cloud generation (Figure 1 step 1). Image orientation yields 8.8M tie points at a 0.6-pixel RMS reprojection error. The outputs are shown in Figure 2. The dense point cloud served as 3D reference data through manual segmentation. This process was conducted using Trimble RealWorks, using the “Auto-Classify Outdoors” tool followed by rigorous manual refinement to ensure label precision.

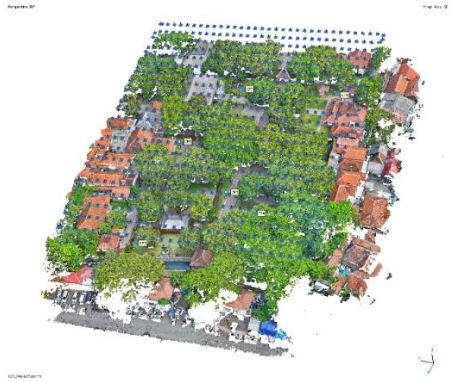

Figure 2. Image orientation and dense point cloud of Siti Inggil complex dataset.

In Figure 1 step 2, 2D segmentation was performed automatically using SAM 3, with the defined class labels as input prompts. Detected instances were merged to generate per-class semantic masks, producing 702 × N class output images. Manual 2D semantic segmentation was performed in parallel using OpenCV library following the same procedure. This manual annotation process was applied to a subset of ±100 images per class (±14% of the dataset).

Following automatic masking, per-class COLMAP workspaces were prepared (Figure 1 step 3) by shifting the workspace to a recorded coordinate space and substituting the original RGB images with RGBA masked images. Per-class 3DGS models (Figure 1 step 4) were then trained using the ADC training method in LichtFeld-Studio (LFS) (LichtFeld Studio, 2025) with the following settings: 15K iterations, 10M maximum splats, and image resize factor of 2. All preprocessing and training were conducted on a machine equipped with an Intel Xeon (3.19 GHz), 64 GB RAM, and an NVIDIA GeForce RTX 5080 (16 GB VRAM).

3.3 Evaluation

Segmentation performance (Figure 1 step 5) was evaluated using standard metrics for 3D point cloud segmentation: Intersection over Union (IoU), F1-score, precision, and recall (Betsas et al., 2025). These metrics were computed against manually segmented 2D and 3D reference data. Given the inherent presence of 3DGS floaters and positional discrepancies between Gaussian splats and reference point clouds, a voxel-based approach was adopted for 3D metric computation. With a voxel size of 0.5 m (0.5 × 0.5 × 0.5 m^3), the size balances class distribution representation and computational efficiency.

Geometric evaluation was performed using manually segmented 3D point clouds as reference against the generated 3DGS models. The evaluation was done by utilizing the M3C2 algorithm in CloudCompare. Moreover, a qualitative analysis was performed by comparing the spatial completeness of the segmented 3DGS models relative to 3D reference data.

4. Results and Discussion

This section presents the results of our proposed approach. Preliminary testing observations are reported first, followed by quantitative evaluation metrics and qualitative assessment of segmentation performance.

4.1.1 Preliminary Testing Observation: SAGA

Preliminary testing was conducted using SAGA (Cen et al., 2025), which employs native 3DGS algorithms for model generation and leverage SAM 1 for segmentation. It projects 2D masks onto the 3DGS model using depth estimation and camera poses, employing scale-aware contrastive learning to handle multi-granularity segmentation across varying physical scales. During inference, a user-selected 2D point prompt is lifted to 3D query features for interactive object segmentation.

Since SAM produces instance-level masks, individual entities are initially segmented as separate instances. Aggregating these into semantic classes requires multiple user-defined point prompts and additional post-processing. Furthermore, SAGA does not produce a standalone segmented 3D model suitable for metric analysis. Given these limitations in automation and output capability, we opted for the method described in Section 3.2, which reduces segmentation to a single text prompt input and directly produces segmented 3D models for downstream analysis.

4.1.2 2D Segmentations Metric

2D segmentation metrics were obtained by comparing predicted masks against ground truth annotation. A manual annotation strategy was adopted to ensure precise instance delineation. This annotation process was applied to 14% of the dataset, prioritizing the Siti Inggil area. It is the site’s oldest structural core and most critical zone for heritage analysis.

Class	IoU	F1	Precision	Recall
Ground	57.8%	71.5%	97.7%	58.8%
Roofs	80.4%	85.1%	96.0%	83.9%
Vegetations	82.9%	86.6%	97.1%	96.9%
Water Bodies	82.5%	87.3%	82.5%	85.1%
Mean	75.9%	82.6%	96.9%	78.3%

Table 1. 2D segmentation metrics scores on the test data with 102 out of 702 images.

Table 1 presents per-class segmentation metrics. The ‘ground’ class exhibits the weakest performance with IoU of 58%, characterized by low recall indicating under-segmentation. SAM 3 consistently fails to unify spatially disconnected ground surfaces (i.e., roads, pathways, and yards) into a single semantic class. In contrast, the ‘roofs’ and ‘vegetations’ achieve strong performance with IoU between 80-83%. The ‘vegetations’ class shows minor boundary under-segmentation, while ‘roofs’ exhibits more pronounced under-segmentation in texturally uniform flat roof areas, reflected in the precision-recall discrepancy. The class ‘Water bodies’ shows slight over-segmentation, with SAM 3 occasionally misclassifying adjacent non-water surfaces.

These results expose a fundamental limitation in SAM 3’s concept granularity. The model identifies visual patterns associated with input prompts rather than reasoning about

hierarchical semantic relationships (e.g., the ‘ground’ encompassing ‘road’ and ‘pathway’, or ‘building’ constituting ‘roof’). This pattern matching behavior, rather than semantic reasoning, explains both the ground fragmentation and the necessity of substituting the ‘building’ with ‘roof’ as the prompt. Consequently, careful choice of prompts is necessary. The choice of semantic labels and their specificity directly impact segmentation consistency. As noted by Carion et al. (2025), SAM 3 is not designed for complex expressions or reasoning-based queries. Instead, it suggests integration with Multimodal Large Language Model (MLLM) for more sophisticated prompt. For this dataset, using multiple specific prompts (“ground”, “road”, “pathway”) with subsequent aggregation may yield better result than a single general prompt.

Qualitative assessment (Figure 3) corroborates these findings. Figure 3a-b illustrates ground class segmentation behavior, where SAM 3 correctly identifies surface regions but treats semantically related elements as separate instances. Figure 3d and 3f reveal cases of total segmentation failure, where the target object is clearly present, yet SAM 3 produces no output. Figure 3h demonstrates an inverse case, where the ground truth annotation is overly strict, segmenting a partially occluded water body that is largely obscured by vegetation. The remaining classes as seen (Figure 3c, 3e, 3g) shows accurate segmentation with minor misclassification and boundary offsets. These boundary offsets are attributable to the differences between precise architectural boundaries in manual annotations and the generalized object extents predicted by SAM 3.

4.1.3 3D Segmentations Metric and Geometric Accuracy

3D segmentation metrics were obtained by comparing each semantic 3DGS output against manual segmentation of the MVS dense point cloud. The 3DGS training used only RGBA masked images to reconstruct per-class Gaussians, while the reference point cloud was generated through automatic classification followed by rigorous manual refinement.

The 3D segmentation results show markedly lower performance than 2D yet exhibit consistent class-wide behavior. Table 2 reports two distinct metric groups: 3D segmentation metrics, which measure how well reconstructed Gaussians spatially correspond to the reference point cloud, and rendering metrics (SSIM, PSNR), which measure 3DGS reconstruction fidelity from the input 2D images. These two metrics are not directly comparable, notably ‘vegetations’ achieves the highest 3D IoU yet the lowest rendering scores. Because these two evaluate fundamentally different aspects of the pipeline. Rendering metrics are more appropriately interpreted alongside the 2D segmentation results, as both operate in image space.

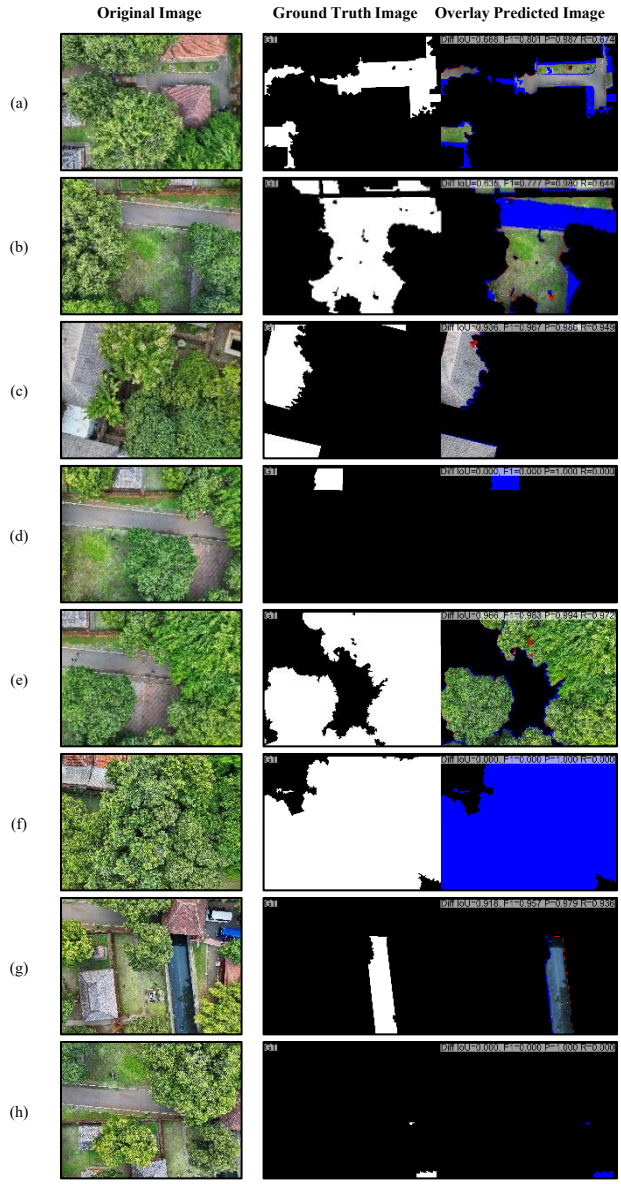

Figure 3. Visual comparison of original images, ground truth, and overlayed predicted masks. On the overlayed images, RGB-colored areas indicate true positive (TP), red indicates false positive (FP), and blue indicates false negative (FN). (a-b) Ground; (c-d) Roofs; (e-f) Vegetations; (g-h) Water Bodies.

Class	Segmentation				Rendering		Time (mins.)
	IoU	F1	Precision	Recall	SSIM ↑	PSNR ↑	
Ground	14.3%	25.0%	18.9%	37.1%	0.77	19.89	15.42
Roofs	27.7%	43.3%	29.3%	83.4%	0.85	23.43	15.07
Vegetations	44.3%	61.4%	62.5%	60.4%	0.59	19.74	15.08
Water Bodies	N/A	N/A	N/A	N/A	0.00	0.00	9.72
Mean	28.8%	43.3%	36.9%	60.3%			

Table 2. 3D segmentation, rendering quality, and training time per class. Segmentation metrics computed using voxelization with a size of 0.5 × 0.5 × 0.5 m^3. Overall accuracy (OA): 60.7%. Rendering metrics at 15K iterations.

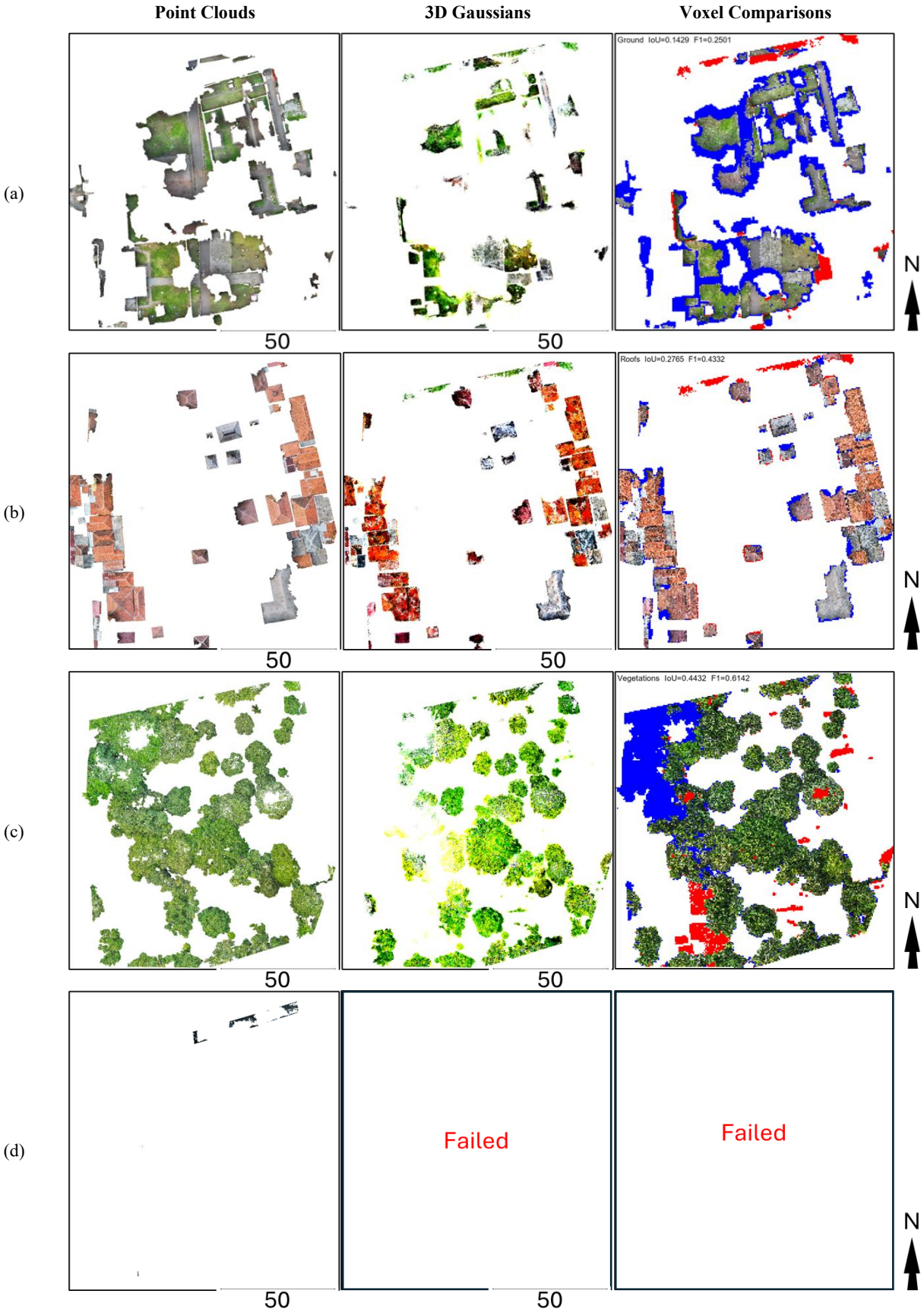

Figure 4. Visual comparison of point cloud reference, 3D Gaussian models, and voxel overlay results per class. Voxelization was adopted to compensate for positional discrepancies between MVS point clouds and 3DGS outputs. On the voxel overlay, RGB-colored areas indicate true positive (TP), red indicates false positive (FP), and blue indicates false negative (FN). (a) Ground; (b) Roofs; (c) Vegetations; (d) Water Bodies.

https://doi.org/10.5194/isprs-archives-XLIX-B2-2026-825-2026

From Table 2, ‘Vegetations’ remains the best-performing class in 3D segmentation with IoU of 44%, though its precision (63%) and recall (60%) indicate slight over segmentation, visible in Figure 4c. This is attributable to error propagations from 2D segmentation. SAM 3 miss-segmentation events such demonstrated in Figure 3f, thus it introduces inconsistent RGBA inputs, causing unstable SSIM and PSNR behavior. This behavior makes 3DGS render a geometrically correct scene while the corresponding input mask is empty (Figure 5a). Figure 5b further suggests incomplete reconstruction, potentially requiring additional training iteration. Some geometric error in the vegetation reference point cloud may also stem from the image acquisition itself, where tree canopy height may have exceeded the effective overlap coverage at 30 m altitude.

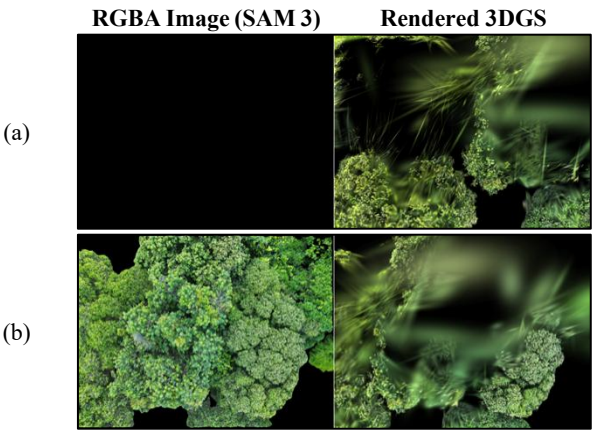

Figure 5. Comparison of SAM 3 RGBA input masks (left) and corresponding 3DGS rendered outputs (right). (a) SAM 3 fails to generate a mask yet 3DGS produces a partial reconstruction; (b) SAM 3 generates a correct mask while 3DGS fails to render a complete representation.

The ‘roofs’ class ranks second with IoU of 28% and a pronounced precision-recall asymmetry (29% vs. 83%). The high recall reflects that most true roof voxels are captured, but low precision indicates substantial false positives from two sources: dataset-edge reconstruction artefacts present across all classes, and floaters artefacts inherent to 3DGS. It is particularly impactful for roofs given the surrounding empty airspace that amplifies their relative contribution to false positive voxels. Additionally, SAM 3’s failure to detect texturally uniform flat roof surfaces propagates directly into incomplete 3D reconstruction. These behaviors are clearly visible in Figure 4b and Figure 6.

The ‘ground’ class with IoU of 14% is similarly affected by edge artefacts, depressing recall (37%) despite relatively strong 3DGS reconstruction with SSIM value of 0.77. The fragmented SAM 3 masking of ground entities further limits spatial coverage in 3D. ‘Water bodies’ class failed entirely under the ADC training method, reconstructing nearly the entire scene despite an 82% 2D IoU. This failure is attributable to the class occupying a very small spatial proportion of the scene, resulting in insufficient multi-view coverage for stable 3DGS reconstruction. Water bodies metrics are consequently excluded from further analysis.

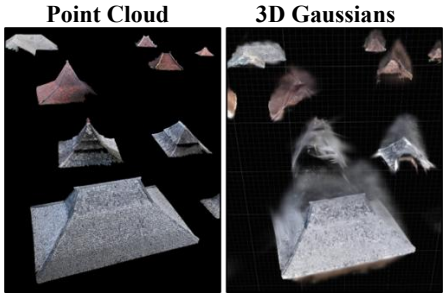

Figure 6. Perspective view of ground truth points clouds (left) and 3DGS from ‘roofs’ class (right).

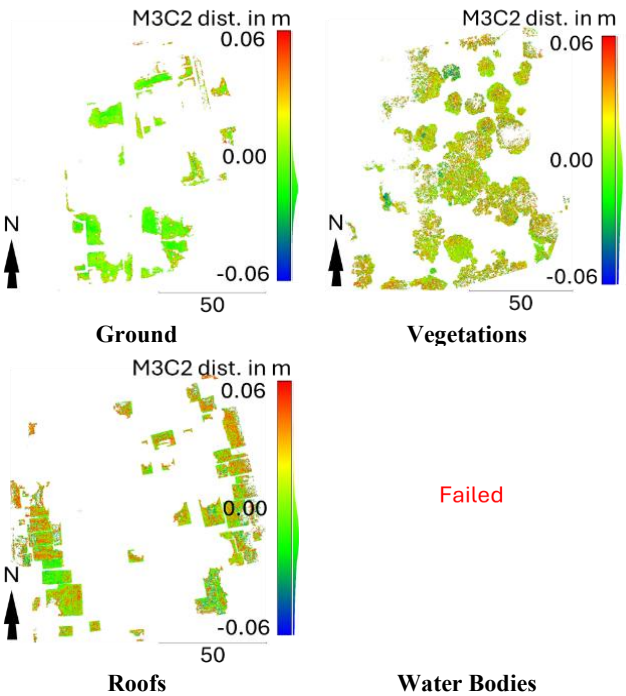

Figure 7. M3C2 signed distance per-class after applying ±6 cm outlier filtering threshold. The Water Bodies class was excluded due to anomalous reconstruction behavior.

Class	M3C2 dist. in cm		M3C2 dist. after filtered (6.0 cm) in cm	
	x̄ Original	σ Original	x̄ Clean	σ Clean
Ground	8.0	33.5	-1.5	1.4
Roofs	0.2	34.5	-1.5	2.3
Vegetations	4.5	54.0	-0.8	2.6
Water Bodies	N/A	N/A	N/A	N/A

Table 3. Geometric assessment results using M3C2 algorithm which compares each class 3D Gaussians with manual segmented point clouds.

Table 3 presents geometric assessment results. Prior to filtering, the ‘roofs’ and ‘vegetations’ show small original means (0.2 cm and 4.5 cm respectively), consistent with fewer reconstruction anomalies during masking and training. However, both classes exhibited large standard deviations (34 cm and 54 cm), reflecting floaters artefacts and geometric reconstruction errors inherent to 3DGS output (Figure 7). The ‘ground’ class shows a higher original mean value (8.0 cm) with comparable spread (33 cm). A global filtering threshold of 6 cm was applied empirically, corresponding to twice the 3× GSD acceptable error for this

dataset (3 cm), and reflecting the higher geometric uncertainty of 3DGS relative to conventional photogrammetry. This threshold isolates the central Gaussian mass of the M3C2 distribution across all valid classes.

After filtering, the ‘vegetation’ achieves the lowest mean deviation, consistent with its superior segmentation performance. In this case, 3DGS reconstructs the structural core of a complex geometry accurately, though at the cost of more floater artefacts compared to simpler geometric classes. This is corroborated by ‘vegetations’ retaining the highest filtered standard deviation despite its low mean value. The ‘roofs’ class shows increased mean deviation after filtering, reflecting Gaussian positional spread along the depth axis, which is a consequence of the nadiral acquisition geometry and texturally uniform roof surfaces discussed previously. The 'ground' class exhibits a consistently small negative offset after filtering, suggesting systematic underestimation of surface elevation attributable to the volumetric nature of Gaussian primitives optimizing photometric consistency over geometric precision on untextured planar surfaces.

4.1.4 Computational Efficiency

The proposed method processes SAM 3 semantic extraction per-image, sequentially extracting all semantic classes before advancing to the next image, totaling 1 hour 52 minutes across 702 images. Per-class COLMAP workspace generation with the process of coordinate shifting required an additional 35 minutes, with per-class 3DGS training times reported in Table 2. At approximately 9.6 seconds per image, SAM 3 extraction represents a notable scalability concern for larger UAV datasets, where large image counts may be common in heritage documentation workflows. Since the per-class COLMAP and 3DGS training steps share no inter-class dependencies, the pipeline presents a natural parallelization opportunity that could substantially reduce through multi-GPU configurations. A direct runtime comparison against SAGA is non-trivial, as its interactive prompting paradigm requires N prompts × M object interactions per scene, making automation and fair benchmarking difficult. A controlled comparison against alternative automated segmentation methods remains necessary to draw meaningful conclusions on computational efficiency and is identified as a priority for future work.

4.1.5 Experiment Limitations

Several limitations are identified in this study. These include generalizability concerns, prompt selection constraints, and the underrepresentation of the water bodies class, each of which is discussed below.

Results are derived from a single heritage site with one acquisition configuration, limiting conclusions about performance on sites with different roof typologies, vegetation density, or flight parameters. Evaluation across multiple datasets and scene types is necessary to establish the broader applicability of the proposed pipeline. Furthermore, these results reflect pipeline performance within the bounds of the reference data and should not be treated as absolute measures of segmentation accuracy.

Prompt wording in this study was constrained to the defined class labels. More systematic prompt design, including the aggregation of multiple specific prompts into a single semantic class, could mitigate SAM 3’s pattern-matching limitations and improve segmentation consistency, particularly for spatially fragmented classes such as ground.

The water bodies class was underrepresented in this dataset, precluding a reliable assessment. Preliminary testing on a dataset with greater water body coverage suggests the pipeline may yield more stable reconstruction under better-distributed spatial coverage. However, fair geometric assessment remains inherently difficult, as photogrammetric point clouds are generally unreliable on water and reflective surfaces, where alternative sensing modalities may be more appropriate.

5. Conclusions and Future Works

This study evaluated 3DGS-based semantic segmentation for urban heritage documentation using a nadiral UAV dataset of the Siti Inggil complex, Cirebon, Indonesia. An automated pipeline integrating SAM 3 concept prompting with per-class 3DGS training was proposed and assessed against manually segmented 2D and 3D reference data. In 2D, the pipeline achieved a mean IoU of 75.9%, with ‘vegetations’ and ‘roof’ performing strongest. In 3D, overall performance was substantially lower with mean IoU of 28.8%, which ‘vegetations’ remains the best-performing class.

SAM 3’s pattern-matching limitations on semantically broad classes, floater artefacts on geometrically uniform surfaces, and reconstruction instability for spatially underrepresented classes were identified as primary sources of result degradation. Geometric assessment further confirmed that while 3DGS captures structural cores accurately, floater artefacts introduce significant positional spread. Incorporating oblique imagery into the dataset could potentially reduce the effects and improve the geometric accuracy on the segmentation result.

These findings demonstrate both the potential and current limitations of 3DGS segmentation for real-world heritage applications, establishing a methodological baseline for future work. A comparison with SAGA-based segmentation is currently underway to further contextualize these results.

Acknowledgements

This paper was published within the context of the Digital Twins for Cultural Heritage (Twin4CH) Junior Professorship Chair, financed by the French National Agency for Research (*Agence nationale de la recherche* – ANR). We would like to extend our gratitude to Deni Suwardhi from ITB and Halo Robotics for their help in acquiring and sharing the Kasepuhan Palace heritage site dataset.

References

Betsas, T., Georgopoulos, A., Doulamis, A., Grussenmeyer, P., 2025. Deep Learning on 3D Semantic Segmentation: A Detailed Review. Remote Sensing 17, 298. https://doi.org/10.3390/rs17020298

Carion, N., Gustafson, L., Hu, Y.-T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., Lei, J., Ma, T., Guo, B., Kalla, A., Marks, M., Greer, J., Wang, M., Sun, P., Rädle, R., Afouras, T., Mavroudi, E., Xu, K., Wu, T.-H., Zhou, Y., Momeni, L., Hazra, R., Ding, S., Vaze, S., Porcher, F., Li, F., Li, S., Kamath, A., Cheng, H.K., Dollár, P., Ravi, N., Saenko, K., Zhang, P., Feichtenhofer, C., 2025. SAM 3: Segment Anything with Concepts. https://doi.org/10.48550/arXiv.2511.16719

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A., 2021. Emerging Properties in Self-Supervised Vision Transformers, in: Proceedings of the International Conference on Computer Vision (ICCV).

Cen, J., Fang, J., Yang, C., Xie, L., Zhang, X., Shen, W., Tian, Q., 2025. Segment Any 3D Gaussians. https://doi.org/10.48550/arXiv.2312.00860

Cen, J., Fang, J., Zhou, Z., Yang, C., Xie, L., Zhang, X., Shen, W., Tian, Q., 2024. Segment Anything in 3D with Radiance Fields. https://doi.org/10.48550/arXiv.2304.12308

Cheng, H.K., Oh, S.W., Price, B., Schwing, A., Lee, J.-Y., 2023. Tracking Anything with Decoupled Video Segmentation, in: ICCV.

He, Y., Yu, H., Liu, X., Yang, Z., Sun, W., Anwar, S., Mian, A., 2024. Deep Learning Based 3D Segmentation: A Survey. https://doi.org/10.48550/arXiv.2103.05423

Ji, S., Wu, G., Fang, J., Cen, J., Yi, T., Liu, W., Tian, Q., Wang, X., 2024. Segment Any 4D Gaussians. https://doi.org/10.48550/arXiv.2407.04504

Jurca, M.-B., Royen, R., Giosan, I., Munteanu, A., 2024. RT-GS2: Real-Time Generalizable Semantic Segmentation for 3D Gaussian Representations of Radiance Fields. https://doi.org/10.48550/arXiv.2405.18033

Jurski, K., 2024. Semantic 3D segmentation of 3D Gaussian Splats.

Kerbl, B., Kopanas, G., Leimkuehler, T., Drettakis, G., 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. ACM Trans. Graph. 42, 1–14. https://doi.org/10.1145/3592433

Kerr, J., Kim, C.M., Goldberg, K., Kanazawa, A., Tancik, M., 2023. LERF: Language Embedded Radiance Fields. https://doi.org/10.48550/arXiv.2303.09553

Kim, C.M., Wu, M., Kerr, J., Goldberg, K., Tancik, M., Kanazawa, A., 2024. GARField: Group Anything with Radiance Fields. https://doi.org/10.48550/arXiv.2401.09419

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.-Y., Dollár, P., Girshick, R., 2023. Segment Anything. https://doi.org/10.48550/arXiv.2304.02643

LichtFeld Studio, 2025. A high-performance C++ and CUDA implementation of 3D Gaussian Splatting.

Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., Terzopoulos, D., 2020. Image Segmentation Using Deep Learning: A Survey. https://doi.org/10.48550/arXiv.2001.05566

Murtiyoso, A., Pellis, E., Grussenmeyer, P., Landes, T., Masiero, A., 2022. Towards Semantic Photogrammetry: Generating Semantically Rich Point Clouds from Architectural Close-Range Photogrammetry. Sensors 22, 966. https://doi.org/10.3390/s22030966

Pan, X., Lin, Q., Ye, S., Li, L., Guo, L., Harmon, B., 2024. Deep learning based approaches from semantic point clouds to semantic BIM models for heritage digital twin. Herit Sci 12, 65. https://doi.org/10.1186/s40494-024-01179-4

Qin, M., Li, W., Zhou, J., Wang, H., Pfister, H., 2024. LangSplat: 3D Language Gaussian Splatting. https://doi.org/10.48550/arXiv.2312.16084

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I., 2021. Learning Transferable Visual Models From Natural Language Supervision. https://doi.org/10.48550/arXiv.2103.00020

Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N., Wu, C.-Y., Girshick, R., Dollár, P., Feichtenhofer, C., 2024. SAM 2: Segment Anything in Images and Videos. https://doi.org/10.48550/arXiv.2408.00714

Tosi, F., Zhang, Y., Gong, Z., Sandström, E., Mattoccia, S., Oswald, M.R., Poggi, M., 2025. How NeRFs and 3D Gaussian Splatting are Reshaping SLAM: a Survey. https://doi.org/10.48550/arXiv.2402.13255

Vinodkumar, P.K., Karabulut, D., Avots, E., Ozcinar, C., Anbarjafari, G., 2023. A Survey on Deep Learning Based Segmentation, Detection and Classification for 3D Point Clouds. Entropy 25, 635. https://doi.org/10.3390/e25040635

Wang, Z., Luo, C., Zhang, J., Li, J., Chen, Y., Zhang, G., 2026. End-to-end Fusion3DGS: label-efficient multi-modal 3D instance segmentation based on Gaussian splatting. Sci Rep 16, 3773. https://doi.org/10.1038/s41598-025-33840-8

Xiong, B., Ye, X., Tse, T.H.E., Han, K., Cui, S., Li, Z., 2024. SA-GS: Semantic-Aware Gaussian Splatting for Large Scene Reconstruction with Geometry Constrain. https://doi.org/10.48550/arXiv.2405.16923

Ye, M., Danelljan, M., Yu, F., Ke, L., 2024. Gaussian Grouping: Segment and Edit Anything in 3D Scenes. https://doi.org/10.48550/arXiv.2312.00732

Ying, H., Yin, Y., Zhang, J., Wang, F., Yu, T., Huang, R., Fang, L., 2023. OmniSeg3D: Omniversal 3D Segmentation via Hierarchical Contrastive Learning. https://doi.org/10.48550/arXiv.2311.11666

Zaouali, M.C., Charter, T., Najjaran, H., 2025. From Flight to Insight: Semantic 3D Reconstruction for Aerial Inspection via Gaussian Splatting and Language-Guided Segmentation. https://doi.org/10.48550/arXiv.2505.17402

Zhang, Y., Zheng, D., Lu, P., Zhang, H., Wang, L., xiang, L., Luo, C., Deng, K., Fu, X., Shen, L., Wang, J., 2025. LabelGS: Label-Aware 3D Gaussian Splatting for 3D Scene Segmentation. https://doi.org/10.48550/arXiv.2508.19699

Zhi, S., Laidlow, T., Leutenegger, S., Davison, A.J., 2021. In-Place Scene Labelling and Understanding with Implicit Scene Representation. https://doi.org/10.48550/arXiv.2103.15875