The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLVIII-1/W5-2025
13th International Conference on Mobile Mapping Technology (MMT 2025)
"Mobile Mapping for Autonomous Systems and Spatial Intelligence", 20–22 June 2025, Xiamen, China

Semantic-Consistent 3D Reconstruction via Gaussian Splatting and SAM-Guided Annotation

Zhaoning Zhang, Tengfei Wang, Zipeng Xu, Quanjian Ji, Xin Wang*, Zongqian Zhan

School of Geodesy and Geomatics, Wuhan University, Wuhan 430079, China
(johnnyzhang, tf.wang, 2021302141007, 2021302141034)@whu.edu.cn, (xwang, zqzhan)@sgg.whu.edu.cn

Keywords: Semantic Segmentation, Point Cloud, Segment Anything Model, 3D Gaussian Splatting, 2D-3D projection.

Abstract

3D Gaussian Splatting (3DGS) provides a novel paradigm for multi-view semantic scene reconstruction, offering explicit and fine-grained geometric representations. However, accurate semantic expression in reconstructed scenes remains problematic as existing methods relying on multi-view 2D semantic projections inherently suffer from cross-view ambiguities in semantic boundaries and label inconsistencies. These limitations arise from occlusions, illumination variations, and object deformations, which degrade the semantic fidelity of 3D reconstructions by introducing conflicting label assignments across viewpoints. This paper proposes a geometry-verified multi-view semantic scene reconstruction framework. First, a depth-aware projection aligns 2D semantic masks with 3DGS-reconstructed point clouds, filtering semantic ambiguities in multi-view annotations via a conflict-locking mechanism. Second, a geometry-aware semantic propagation model globally diffuses semantic labels by leveraging the local geometric continuity of the point cloud. Experiments demonstrate that the proposed framework achieves significantly superior reconstruction consistency in occlusion-heavy scenes compared to conventional methods. Specifically, it outperforms multi-view voting strategies with improvements of 12.39% in Overall Accuracy (OA) and 17.03% in mean Intersection over Union (mIoU) on the BeDOI-GB dataset. Project web:https://bigbigman233.github.io/SC3D.github.io/

1. Introduction

Multi-view semantic scene reconstruction bridges the physical and digital worlds, supporting applications in cultural heritage (Gomes et al., 2014), smart agriculture (Yu et al., 2024), and autonomous driving (Ma et al., 2019). It involves two core tasks: geometric reconstruction and semantic segmentation. Extensive studies have been conducted on the geometric reconstruction of Multi-view scenes. Traditional approaches, like Multi-View Stereo (MVS) depend on camera pose and cost volume, are susceptible to illumination variations and perform poorly in textureless regions within complex scenes (Lu, 2025; Bao et al., 2025). In addition, Neural Radiance Fields (NeRF) can achieve high-fidelity novel view synthesis through implicit neural representations (Mildenhall et al., 2021). However, the issues such as low training efficiency and limited explicit control were soon solved by 3DGS (Kerbl et al., 2023). Via an explicit point cloud representation and a differentiable rasterization pipeline, 3DGS strikes an excellent balance between efficiency and accuracy in real-time rendering and dynamic scene modeling. On the other hand, the semantic segmentation of 3D reconstruction is crucial in many applications, enabling recognition of scent categories, functions, and topological relationships (Gul et al., 2019; Liu et al., 2018). However, compared to the well-explored geometric reconstruction, the integration of semantic and geometric information in 3D semantic scene reconstruction using images remains relatively underexplored.

Recent works have explored integrating vision foundation models with 3DGS (Qin et al., 2024), such as utilizing the open-vocabulary segmentation capability of the Segment Anything Model (SAM) to generate multi-view semantic masks projected onto Gaussian-splatted point clouds (Guo et al., 2024). Xiong et al. (2024b) employed semantic information to control the shape of Gaussian spheres for representing the boundaries of different types of terrain features. Furthermore, the challenge of cross-view semantic consistency emerges as a critical bottleneck: SAM-generated masks across different viewpoints suffer from boundary ambiguity or category conflicts due to occlusions, illumination variations, or object deformations. As illustrated in Fig. 1, the same building and road exhibit different semantic label masks in various images captured from different viewpoints.

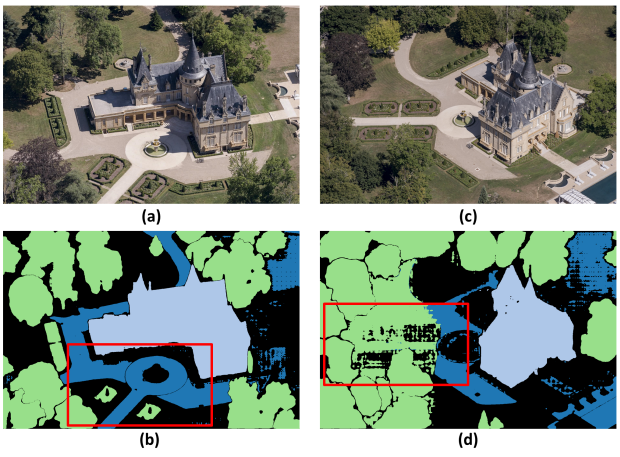

Figure 1. Inconsistent SAM-generated masks of two views from various perspectives. (a) and (c) are original images, (b) and (d) denote the corresponding semantic masks using SAM. The red box indicates the inconsistency in the labels of the road.

To address this challenge, this study proposes a multi-view semantic scene reconstruction framework based on 3DGS and mask projection fusion. Firstly, fine-grained point clouds are extracted from the 3DGS to achieve precise geometric reconstruction of the scene. Subsequently, we utilize Grounding SAM to generate segmentation masks for multiple categories of objects in multi-view images. Then, the masks with semantic

* Corresponding Author

information are projected onto the point clouds to endow the point clouds with semantics.

Specifically, to solve the semantic inconsistency caused by the direct projection mentioned above, this paper designs a geometry-semantic bidirectional verification method. First, a depth-aware voxelization filtering mechanism is constructed. This mechanism locates the first occluded surface through multi-level ray traversal and performs semantic binding exclusively at this layer, thereby eliminating deep projection noise. Second, a dynamic conflict reset mechanism is designed. It detects cross-modal annotation conflicts within spatial voxels via semantic confidence voting, automatically locks the conflicting regions, and initiates reprojection optimization. Third, a topology-driven semantic propagation model is established. This model diffuses locally verified and reliable annotations across the global scene along the point cloud surface manifolds. Finally, due to the lack of evaluation datasets for multi-view semantic reconstruction, this paper conducts multi-category semantic annotations on two existing datasets to conduct quantitative evaluations.

In summary, our contributions are as follows:

- To achieve multi-view semantic scene reconstruction, this paper proposes a novel framework based on 3DGS and SAM-generated mask projection fusion, which does not require additional semantic annotation data for supervision.
- To address the inconsistency by projection from 2D multi-view images to 3D point clouds, this paper proposes a geometry-semantic bidirectional verification mechanism.
- To tackle the difficulty in evaluating multi-view semantic reconstruction, this paper specially conducts multi-category semantic annotations on two existing datasets to quantify the reconstruction results, providing quantitative evaluation benchmarks for other relevant research.

2. Related Work

2.1 3D point cloud Reconstruction

Traditional image-based dense matching methods for 3D reconstruction (e.g., the SfM-MVS framework) have long relied on pixel-level correspondence establishment and triangulation (Pepe et al., 2022). While tools like COLMAP (Schonberger and Frahm, 2016) improve point cloud generation efficiency through SIFT feature matching (Wu et al., 2013) and photometric consistency optimization, they exhibit significant limitations in computational complexity, texture dependency, and adaptability to dynamic scenes (Fan et al., 2019; Stathopoulou et al., 2019). For instance, large-scale urban scene reconstruction is often forced into block-wise processing due to GPU memory constraints, leading to geometric discontinuity (Niu et al., 2024); weakly textured regions are prone to matching ambiguities, requiring post-processing filtering; and dynamic objects generate motion artifacts due to multi-view temporal inconsistencies (Luo et al., 2024).

In contrast, 3DGS technology achieves breakthroughs in geometry modeling without explicit matching by differentiable optimization of anisotropic Gaussian parameters. Wu et al. (2013) establishes depth-normal constraints through ray-Gaussian intersection analysis, achieving a 0.18-meter RMSE accuracy in aerial image reconstruction, significantly outperforming OpenMVS (0.32-meter RMSE) (Li et al., 2020). Meanwhile, Stuart and Pound (2025) suppresses outlier noise common in traditional MVS while preserving details through Mahalanobis distance threshold-based Gaussian probability sampling. For dynamic scenes, Cho et al. (2024) enables continuous motion trajectory reconstruction via temporal Gaussian parameter modeling, achieving over 200 times higher efficiency compared to conventional sequential image matching. Wang et al. (2025) fully exploits semantic cues for both partitioning and Gaussian kernel optimization, enabling fine-grained building surface reconstruction of large-scale urban areas without downsampling the original image resolution. These advancements indicate that 3DGS, with its explicit geometric representation and real-time optimization capabilities, is gradually replacing traditional dense matching workflows to become a new paradigm for efficient and robust point cloud generation.

2.2 Unsupervised Point Cloud Semantic Segmentation

Early research on point cloud semantic segmentation primarily relied on manually designed geometric features and statistical learning models. Vosselman et al. (2004) pioneered automated road extraction through least-squares plane fitting and elevation difference analysis, establishing the foundation for geometry-driven methods. Subsequently, Chehata et al. (2009) achieved an F1-score of 0.78 in forest scene classification by combining random forests with height-intensity joint histograms. Rutzinger et al. (2009) proposed a combined strategy of eaves line detection and plane fitting for building extraction, achieving 85% accuracy in complex roof structure recognition. Demantké et al. (2011) developed the Terrain Feature Descriptor, enabling terrain classification through elevation gradient and local roughness analysis of LiDAR point clouds.

With the rise of segmentation-driven paradigms, Golovinskiy and Funkhouser (2009) introduced graph-cut algorithms to point cloud processing, optimizing segmentation boundaries and category labels via energy minimization. Aijazi et al. (2013) proposed a voxelization-supervoxel fusion strategy for hierarchical urban element classification, leveraging surface roughness and planar fitting residuals. Serna and Marcotegui (2014) constructed a geometry-context Bayesian classifier that improved recognition accuracy by exploiting spatial relationships between adjacent objects. Papon et al. (2013) presented the Voxel Cloud Connectivity Segmentation algorithm, achieving 17× faster supervoxel generation compared to conventional methods. The integration of statistical learning and probabilistic models further advanced methodological progress. Munoz et al. (2009) first applied Conditional Random Fields (CRF) to point cloud classification, enhancing local label consistency through smoothness constraints. Weinmann et al. (2015) systematically evaluated 21 geometric features, identifying curvature variation rate as critical for anisotropic structure recognition. Hackel et al. (2016) developed density-adaptive multi-scale feature pyramids, computing 144-dimensional geometric descriptors for large-scale urban scene classification. Babahajiani et al. (2015) fused intensity reflectance and multi-echo features to significantly improve vegetation-artificial structure differentiation. Xiang et al. (2018)implemented urban road scene classification using normal-consistency-based super-segment merging and SVM classifiers.

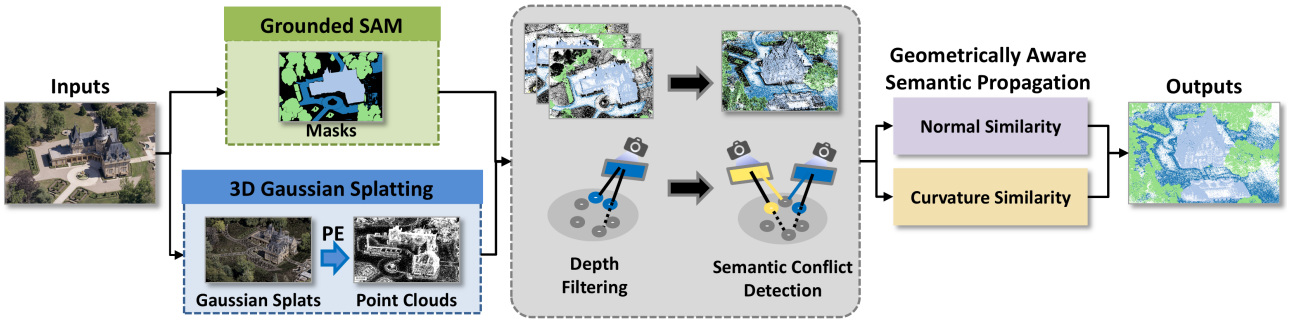

Figure 2. Workflow of our consistent semantic-aware 3D reconstruction using 3DGS and SAM.

3. Methodology

To address the critical limitations of inconsistent semantic annotations in multi-view images using SAM-based semantic masks, this study proposes a novel semantic segmentation framework based on 3D Gaussian point cloud reconstruction and mask projection fusion. Our main idea lies in constructing a spatially and semantically continuous scene representation through 3D Gaussian point cloud reconstruction, projecting 2D semantic masks into 3D space for conflict resolution and semantic fusion, and ultimately achieving 3D scene understanding that balances geometric consistency and semantic accuracy. The workflow comprises three steps: first, constructing a differentiable 3D Gaussian point cloud representation from multi-view images; second, generating multi-view semantic masks using the vision foundation model SAM and mapping them to 3D space via a differentiable projection mechanism; finally, achieving consistent 3D semantic completion through spatial conflict detection and region-growing optimization. The detailed workflow is illustrated in Fig. 2.

Section 3.1 provides a concise overview of 3D Gaussian splatting. Section 3.2 introduces the enhanced SAM-based semantic auto-extraction method. Section 3.3 elaborates on the mask projection framework and conflict resolution strategies. Section 3.4 details the geometry-continuity-based region growing algorithm for semantic completion in conflicting regions.

3.1 3D Gaussian Splatting

3DGS comprises a set of learnable 3D Gaussian splats, where each splat contains four-dimensional feature parameters: 3D coordinates μ representing spatial position, covariance matrix Σ defining geometric morphology, spherical harmonic parameters c describing color attributes, and opacity α characterizing transparency. This technology reconstructs 3D scenes from multi-view images and initializes them using Structure from Motion (SfM) point clouds (Pepe et al., 2022), thereby optimizing rendering quality and geometric structural integrity. The covariance matrix of each Gaussian point dynamically adjusts its spatial distribution through differentiable optimization, enabling the 3D representation to precisely fit complex surface geometries while maintaining efficient rendering performance.

3DGS employs point-based alpha blending to compute pixel values for 2D images. The color value C of each pixel is determined via the volumetric rendering formula along the ray path: specifically, for a ray $r(t)$ emitted from the viewpoint, its accumulated color is derived from the weighted integral of transmittance and color contributions from all Gaussian points along the path. The mathematical expression is:

$$C(r) = \sum_{i=1}^{N} \alpha_i c_i \prod_{j=1}^{i-1} (1 - \alpha_j) \tag{1}$$

where N denotes the ordered set of Gaussian points intersecting the pixel along the viewing ray, sorted by depth from front to back. α_i represents the transmittance of the i-th Gaussian point along the ray, and c_i is computed from its spherical harmonic parameters. This differentiable rendering mechanism achieves high-precision modeling of complex geometries and materials by optimizing the spatial distribution parameters of Gaussian points.

Zwicker et al. proposed a splatting method to project 3D Gaussians onto a 2D plane by computing the covariance matrix Σ' in camera coordinates. Given the world-to-camera transformation matrix W, the covariance matrix in camera coordinates is derived as:

$$\Sigma' = JW\Sigma W^\top J^\top \tag{2}$$

where J is the Jacobian matrix approximating the projective transformation. By discarding the third row and column of Σ', the 2D covariance matrix is obtained. To ensure the positive semi-definiteness of the covariance matrix, we decompose it into a learnable rotation matrix R and a scaling matrix S:

$$\Sigma = RSS^\top R^\top \tag{3}$$

where $R \in \mathrm{SO}(3)$ denotes an orthogonal rotation matrix in the 3D rotation group, and S is a diagonal matrix storing axial scaling coefficients. This parametric decomposition guarantees mathematical validity of the covariance matrix while enabling efficient optimization of Gaussian spatial distributions via gradient descent.

Gaussian splats can be regarded as specialized point clouds with additional functionalities, thereby sharing certain attributes of point clouds. An intuitive idea is to reconstruct the scene by 3DGS and extract the point cloud from it, projecting the semantic information obtained from the 2D base model onto the corresponding point cloud based on spatial relationships. In this paper, the scene reconstruction method of PGSR (Chen et al., 2024) is adopted, and the point cloud is extracted from it as the scene base of the experiment.

3.2 Semantic Mask Extraction

We employ a set of N input images $I = \{I_1, I_2, \ldots, I_N\}$, from which semantic masks are extracted using the Grounded SAM model (Ren et al., 2024). The processed mask images are categorized into distinct semantic groups, with all mask tensors

denoted as $M \in R^{H\times W\times N}$. For each pixel (x, y) in image i, its mask value $M_i(x, y)$ is associated with either a textual descriptor or a preset default value (when GroundingSAM fails to establish correspondence). The descriptor embedding vector is defined as $E(c)$, where c represents textual descriptions (e.g., "tree", "building", "road"), which are practically mapped to integer values for computational simplicity. All pixels satisfying $\forall M_i(x, y) = E(c_j)$ constitute the semantic group corresponding to descriptor c_j, with the system containing C distinct descriptor categories.

3.3 2D-3D Projection and Fusion

After obtaining the pixel-level semantic map, the method projects it onto the point cloud generated by 3DGS to build semantic components. For each pixel $\mathbf{u} = (u, v)$ in the semantic map S, the framework locates the corresponding Gaussian point $\mathbf{p} = (x, y, z)$ in 3D space. This process is achieved through camera intrinsic matrix K and world-to-camera pose matrix E, based on the pinhole camera model:

$$\tilde{\mathbf{u}} = K \cdot E \cdot \tilde{\mathbf{p}} \tag{4}$$

where $\tilde{\mathbf{u}}$ and $\tilde{\mathbf{p}}$ denote the homogeneous coordinates of $\mathbf{u}$ and $\mathbf{p}$, respectively.

To address semantic misassociation issues caused by ray penetration in discrete point cloud projection, this study proposes a depth screening mechanism. During the semantic projection phase, voxelized sampling is performed along camera ray directions: First, the point cloud is voxelized, and the key voxel containing the first occluding surface is localized through voxel region determination, thereby avoiding erroneous projections caused by deep penetration. Subsequently, pixel-level labels from 2D semantic masks are injected into corresponding voxels, achieving semantic mapping with sub-voxel precision. To enhance multi-view consistency, an inter-view semantic conflict reset mechanism is designed for semantic label determinations of the same voxel across different viewpoints. Inconsistent semantics across views trigger voxel label reset and locking, with subsequent completion via point cloud label propagation.

3.4 Point Cloud Label Propagation Method

The label propagation method for point clouds is grounded in the geometric manifold continuity hypothesis, which posits that points within a local neighborhood sharing similar geometric features have a higher probability of identical semantic labels. Given a point cloud set $P = \{\mathbf{p}_i \in R^3\}_{i=1}^N$, where a subset of points A are annotated with semantic labels $l_j \in \{0, \ldots, C-1\}$ ($j \in A \subseteq P$), the objective is to infer labels for unannotated points $U = P \setminus A$.

Local geometric structures are characterized by normal vector fields and curvature features. For any point $\mathbf{p}_i$, its normal vector $\mathbf{n}_i$ is determined by the eigenvector corresponding to the smallest eigenvalue of the neighborhood covariance matrix:

$$\Sigma = \frac{1}{k}\sum_{j\in\mathcal{N}_i}(\mathbf{p}_j - \bar{\mathbf{p}}_i)(\mathbf{p}_j - \bar{\mathbf{p}}_i)^\top \tag{5}$$

where $\mathcal{N}_i$ denotes the k-nearest neighbors of $\mathbf{p}_i$, and $\bar{\mathbf{p}}_i$ is the centroid of the neighborhood. The curvature feature κ_i measures the dispersion between the normal vector and the average neighborhood normals:

$$\kappa_i = \left\| \mathbf{n}_i - \frac{1}{|\mathcal{N}_i|}\sum_{j\in\mathcal{N}_i}\mathbf{n}_j \right\| \tag{6}$$

The joint similarity weight between an unlabeled point $\mathbf{p}_u$ and its annotated neighbor $\mathbf{p}_v \in \mathcal{N}_u$ combines normal consistency and curvature compatibility:

$$w_{uv} = \alpha \underbrace{\left(1 - \frac{1}{\pi}\arccos(\mathbf{n}_u \cdot \mathbf{n}_v)\right)}_{NormalSimilarity} + \beta \underbrace{\exp\left(-\gamma|\kappa_u - \kappa_v|\right)}_{CurvatureSimilarity} \tag{7}$$

where $\alpha + \beta = 1$ are weighting coefficients, and γ controls sensitivity to curvature differences (in this work, $\alpha = 0.6, \beta = 0.4, \gamma = 1.0$). Normal similarity is normalized via the angular cosine distance, while curvature similarity employs an exponential decay function to accentuate local geometric distinctions.

The propagation process is formulated as an energy minimization problem on a weighted graph $G = (V, E, W)$. The vertex set $V = A \cup U$ includes annotated and unannotated points, edges E are defined by radius-constrained neighborhood connections, and weights W derive from the similarity model. The objective function is:

$$\min_{\hat{L}_U}\sum_{u\in U}\sum_{v\in\mathcal{N}_u} w_{uv}\cdot\delta(l_u, l_v) \tag{8}$$

where $\delta(\cdot)$ is the Kronecker delta function. The closed-form solution via maximum a posteriori estimation yields predicted labels for unannotated points as:

$$\hat{l}_u = c \in \{0, \ldots, C-1\}\arg\max\sum_{v\in\mathcal{N}_u} w_{uv}\cdot I(l_v = c) \tag{9}$$

This method can robustly propagate semantic information to unlabeled regions while preserving geometric manifold continuity through joint feature modeling and graph diffusion theory.

4. Experiments

To comprehensively validate the effectiveness of the proposed method, two key aspects are investigated based on two outdoor scene datasets, including the analysis of geometry-semantics consistency optimization and semantic reasoning capabilities. By introducing Overall Accuracy (OA) and mean Intersection over Union (mIoU) as metrics, we quantify the resolution effectiveness of multi-view label conflicts and the coupling accuracy of 3D semantic-geometric integration, respectively. The main experimental design encompasses three core modules: Firstly, comparative experiments verify the improvement of multi-view semantic fusion mechanism on inconsistent annotations from various images. Subsequently, ablation studies analyze the individual contributions of the depth-aware filtering strategy and geometry-constrained propagation model.

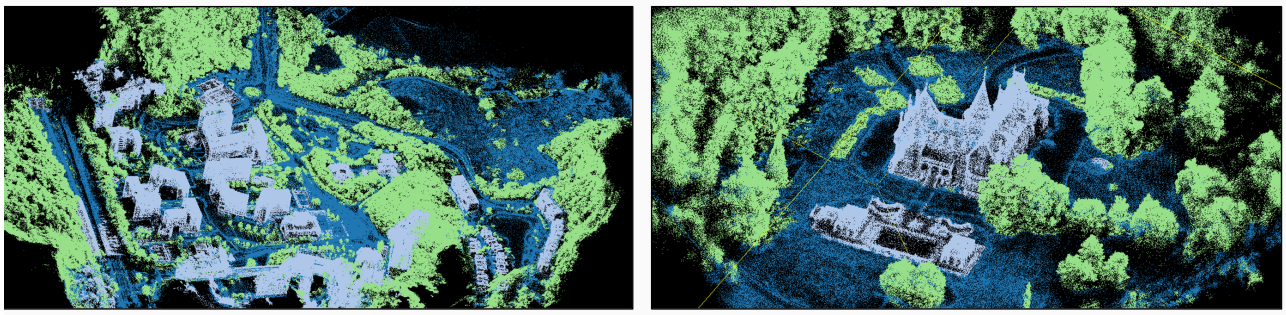

Figure 3. The overview of the sematic-consistent 3D reconstruction on the GauUScene-CUHKUPPER (left) and BeDOI-GB (right) datasets based on our proposed method.

Additionally, a cross-modal reprojection validation strategy is implemented, checking the mapping consistency between 3D semantic point clouds and 2D manually annotated masks to further verify the effectiveness of joint geometry-semantics optimization. All the tests are conducted on a machine with Intel(R) Xeon(R) Gold 6133 CPU @ 2.50GHz of 20 cores and a single RTX 4090 GPU of 24GB.

4.1 Datasets

To validate the effectiveness of the proposed method, two high-resolution outdoor drone image datasets, GauUScene-CUHKUPPER (Xiong et al., 2024a) and BeDOI-GB (Zhan et al., 2023), are utilized. The GauUScene-CUHKUPPER dataset comprises 713 urban street-view images (resolution: 5469 × 3638), capturing dynamic occlusion caused by interactions between vegetation and buildings. The BeDOI-GB dataset contains 68 images (resolution: 1840 × 1228) of densely vegetated areas, highlighting challenges in static and dynamic occlusion scenarios. Both datasets exhibit inherent complexities in generating consistent continuous masks due to overlapping vegetation and architectural structures. To standardize validation, semantic annotations are provided for three predefined categories—buildings, trees, and roads—enabling systematic quantification of geometric-semantic alignment and multi-view consistency. To tackle the difficulty in evaluating multi-view semantic reconstruction, we manually annotate masks for GauUScene-CUHKUPPER and BeDOI-GB, to ensure the semantic consistency of multi-view annotation and construct a cross-view semantic mask dataset. These annotations serve as ground truth to facilitate subsequent experimental validation. Datasets are visualized in Fig. 4 and the corresponding overview sematic-consistent 3D reconstructions are shown in Fig. 3.

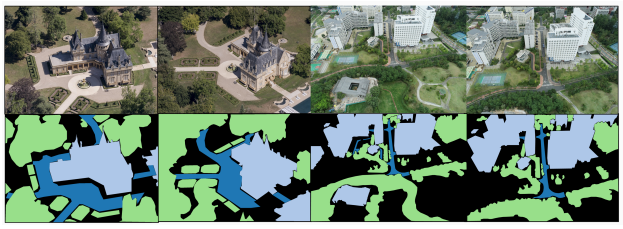

Figure 4. Sample images and semantic masks of GauUScene-CUHKUPPER and BeDOI-GB datasets.

4.2 Evaluation metrics

This paper compares various point cloud semantic segmentation methods using two evaluation metrics: Overall Accuracy (OA) and mean Intersection over Union (mIoU). We reproject semantically annotated point clouds onto manually labeled masks for metric evaluation. Note that depth filtering and voxel occlusion must still be considered during reprojection. For ease of description, we assume there are N semantic classes. p_{ij} represents the number of units where the actual semantic type is i and the predicted type is j, and vice versa for p_{ji}. p_{ii} represents the number of units with both actual and predicted semantic type i.

OA is the ratio of the number of samples correctly predicted by the segmentation algorithms to the total number of samples. Its formulation is given as:

$$OA = \frac{\sum_{i=0}^{N} p_{ii}}{\sum_{i=0}^{N} \sum_{j=0}^{N} p_{ij}} \tag{10}$$

mIOU stands out as the primary metric for evaluating segmentation methods' performance. It initially computes the intersection-over-union ratio for each category, reflecting the overlap between predicted and true regions of the models. Subsequently, it calculates the average value of the accumulated results based on the number of categories. Its formulation is given as:

$$mIoU = \frac{1}{N+1} \sum_{i=0}^{N} \frac{p_{ii}}{\sum_{j=0}^{N} p_{ij} + \sum_{i=0}^{N} p_{ji} - p_{ii}} \tag{11}$$

4.3 Performance of Different Projection Mechanisms

The effectiveness of the projection mechanism is validated on the GauUScene-CUHKUPPER and BeDOI-GB datasets, comparing three approaches: Direct Overlay, Majority Voting, and the proposed Geometry-constrained Cross-modal Fusion (GC-Fusion). The evaluation employs Overall Accuracy (OA) and mean Intersection over Union (mIoU) as core metrics, with a focus on semantic consistency in vegetation-building interaction zones and dynamic occlusion scenarios. For Direct Overlay, SAM-generated 2D semantic labels are directly projected onto the 3D point cloud using camera parameters. The Majority Voting approach resolves multi-view label conflicts through a majority voting strategy. In contrast, the proposed method leverages a multi-level voxel field to isolate credible labels at the first occlusion layer, suppresses deep penetration errors via depth-aware weighting, and completes missing semantics through 3D topological propagation. All experiments are conducted on the same point cloud to ensure geometric consistency.

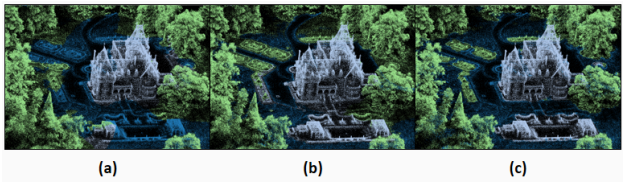

Figure 5. Qualitative results comparison on the BeDOI-GB dataset: (a) direct overlay method, (b) voting fusion method, and (c) our proposed method.

The experimental results on the BeDOI-GB dataset are presented in Fig. 5 and Tab. 1. Quantitatively, our method demonstrates significant improvements, achieving 27.1% and 17.03% higher mIoU compared to Direct Overlay and Majority Voting, respectively, along with 19.9% and 12.39% gains in OA. Furthermore, our approach outperforms all baselines across every category-specific evaluation, validating its effectiveness. The

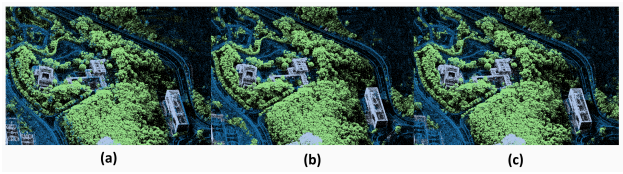

Figure 6. Qualitative results on the GauUScene-CUHKUPPER dataset: (a) direct overlay method, (b) voting fusion method, and (c) our proposed method.

results on GauUScene-CUHKUPPER are shown in Fig. 6 and Tab. 1, where our method improves by 3.34% and 2.53% on the mIOU and 3.49% and 2.21% on the OA metric, respectively. Due to the large size of the GauUScene-CUHKUPPER scenario and the complexity of the features, the accuracy of the semantic mask extracted by SAM decreases, resulting in a decline in all indicators compared with BeDOI-GB, but from the existing results, our method still demonstrates significant advantages.

Qualitatively, Direct Overlay exhibits significant semantic fragmentation on the same geographic feature. As shown in Fig. 5, vegetation point clouds on the left side of the building are split into vegetation and road semantics, while most point clouds of the front building are erroneously labeled as road, resulting in semantic discontinuity. This occurs because the projection range of current image-based methods is limited—when mask inconsistencies arise across images, only partial regions can be updated, leading to semantic cracks. Additionally, Direct Overlay is highly sensitive to projection order and lacks robustness. Majority Voting also suffers from semantic discontinuities due to viewpoint variations, though its voting mechanism avoids linear cracks, instead manifesting as intermixed categories. Our method fundamentally circumvents perspective-induced limitations: we retain conflict-prone points and resolve their labels through geometric feature consistency verification, ensuring semantically continuous and coherent results.

4.4 Geometry-Aware Semantic Propagation

This section reports the performance of semantic label propagation by comparing the conventional KNN-based voting approach (KNN-Voting) with our proposed geometry-similarity-weighted propagation method (GC-Propagation). To ensure fair comparison, all semantically labeled point clouds generated by both methods undergo back-projection onto manually annotated 2D masks for the metric computation, with

Scene(BeDOI-GB)	mIoU	OA	building	tree	road
Direct Overlay	51.74	66.72	56.13	57.25	41.85
Majority Voting	61.81	74.23	73.85	49.00	62.58
Ours	78.84	86.62	89.26	74.44	72.83
Scene(GauUScene)	-	-	-	-	-
Direct Overlay	59.34	80.36	66.55	56.56	54.92
Majority Voting	60.15	81.64	68.35	57.81	54.28
Ours	62.68	83.85	70.80	59.89	57.35

Table 1. Quantitative results of BeDOI-GB and GauUScene-CUHKUPPER in projection mechanism validation.

particular focus on semantic consistency in dynamic occlusion boundaries and fine-grained regions. The KNN-Voting baseline operates by performing Euclidean distance-based neighborhood searches followed by direct majority voting of labels. In contrast, our GC-Propagation method constructs propagation weights through comprehensive geometric considerations—incorporating surface normal consistency, curvature similarity, and spatial continuity constraints. Both approaches utilize the same point cloud and initial semantic labels.

Scene(BeDOI-GB)	mIoU	OA	building	tree	road
KNN-Voting	77.19	85.49	87.59	73.44	70.55
GC-Propagation	78.84	86.62	89.26	74.44	72.83
Scene(GauUScene)	-	-	-	-	-
KNN-Voting	59.54	80.74	67.45	56.90	54.26
GC-Propagation	62.68	83.85	70.80	59.89	57.35

Table 2. Quantitative results of BeDOI-GB and GauUScene-CUHKUPPER in geometry-aware semantic propagation validation.

Tab. 2 presents the results of the BeDOI-GB and GauUScene-CUHKUPPER datasets. On the BeDOI-GB dataset, compared with the KNN-Voting method, GC-Propagation achieves improvements of 1.13% in OA and 1.65% in mIoU. These improvements further increase to 3.11% and 3.14%, respectively, on the GauUScene-CUHKUPPER dataset. Notably, GC-Propagation outperforms KNN-Voting across all semantic categories, which can be attributed to its incorporation of geometric structure constraints during the label propagation process.

4.5 Ablation Study on Depth-Aware Voxel Filtering

To evaluate semantic projection accuracy at dynamic occlusion boundaries, a corresponding ablation study is carried out. For the baseline method without depth filtering, 2D semantic labels are directly projected onto the 3D point cloud using camera parameters, disregarding depth penetration effects. In contrast, the proposed method employs multi-level voxel fields to identify the first occluded surface and fuses multi-view labels via voxel-wise Gaussian distribution density weighting. The evaluation fixes the point cloud and input viewpoints to ensure geometric consistency, with targeted testing conducted on dynamic occlusion regions (building-vegetation interactions) across both datasets.

The experimental results on the BeDOI-GB dataset are shown in Fig. 7 and Tab. 3. From a quantitative perspective, the Depth Filtering mechanism achieves significant improvements of 24.99% in OA and 31.2% in mIoU compared to the baseline without depth filtering, demonstrating the significance of depth-based screening. Similarly, results on the GauUScene-CUHKUPPER dataset (Fig. 8 and Tab. 3) reveal that Depth Filtering improves OA and mIoU by 10.58% and 9.89%, respectively. Notably, the IoU for the building category increases

Figure 7. Qualitative results comparison on the BeDOI-GB dataset: (a) Filtering, (b) Without Filtering.

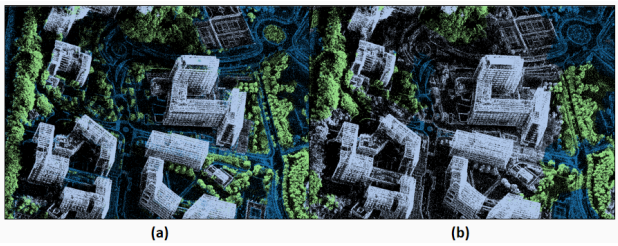

Figure 8. Qualitative results on the GauUScene-CUHKUPPER dataset: (a) Filtering, (b) Without Filtering.

by 12.98%, the highest improvement across all classes. This is attributed to the close proximity of trees and buildings in urban areas, where the absence of depth filtering causes a substantial portion of building points to be misclassified as vegetation.

Qualitative analysis shows that without depth awareness and voxel occlusion handling, the point cloud lacks volumetric constraints, allowing light rays to penetrate foreground objects and incorrectly assign semantics to occluded background elements. As highlighted by the red boxes in Fig. 7, certain viewpoints exhibit occlusion relationships between trees and buildings where portions of buildings are erroneously labeled with tree semantics. Similarly, road segments occluded by buildings in specific views receive incorrect semantic assignments. These erroneous labels rapidly accumulate conflicting, undetermined points, ultimately causing failure in subsequent semantic label propagation.

Scene(BeDOI-GB)	mIoU	OA	building	tree	road
Without Filtering	47.34	61.63	64.72	54.25	23.04
Filtering	78.84	86.62	89.26	74.44	72.83
Scene(GauUScene)	-	-	-	-	-
Without Filtering	52.79	73.27	57.82	53.80	46.77
Filtering	62.68	83.85	70.80	59.89	57.35

Table 3. Quantitative results of BeDOI-GB and GauUScene-CUHKUPPER in depth filtering validation.

5. Conclusion

This work proposes a novel method for generating consistent sematic-aware 3D reconstruction using 3D Gaussian Splatting (3DGS) and Segment Anything Model (SAM). In particular, addressing semantic mask inconsistencies across multiple images from various perspectives, we first project 2D semantic masks onto 3D Gaussian point clouds while introducing geometric feature continuity constraints to resolve cross-view semantic ambiguities through geometric regularization. Our key ideas include the construction of a depth screening mechanism, the design of a dynamic conflict reset protocol, and the development of a geometry-aware semantic propagation model. Experiments demonstrate that compared to multi-view voting strategies, our framework can significantly improve semantic consistency in occlusion-intensive scenarios, validating its practical robustness. Future work will focus on better 3D Gaussian representations, bridging 3D semantic understanding from static contexts to dynamic interactive scenarios, and also applying more large-scale scenes.

Acknowledgment

This work was jointly supported by the National Natural Science Foundation of China (42301507) and the Natural Science Foundation of Hubei Province, China (2022CFB727).

References

Aijazi, A. K., Checchin, P., Trassoudaine, L., 2013. Segmentation based classification of 3D urban point clouds: A supervoxel based approach with evaluation. *Remote Sensing*, 5(4), 1624–1650.

Babahajiani, P., Fan, L., Kamarainen, J., Gabbouj, M., 2015. Automated super-voxel based features classification of urban environments by integrating 3d point cloud and image content. *2015 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, IEEE, 372–377.

Bao, Y., Ding, T., Huo, J., Liu, Y., Li, Y., Li, W., Gao, Y., Luo, J., 2025. 3d gaussian splatting: Survey, technologies, challenges, and opportunities. *IEEE Transactions on Circuits and Systems for Video Technology*.

Chehata, N., Guo, L., Mallet, C., 2009. Airborne lidar feature selection for urban classification using random forests. *Laserscanning*.

Chen, D., Li, H., Ye, W., Wang, Y., Xie, W., Zhai, S., Wang, N., Liu, H., Bao, H., Zhang, G., 2024. Pgsr: Planar-based gaussian splatting for efficient and high-fidelity surface reconstruction. *IEEE Transactions on Visualization and Computer Graphics*.

Cho, W. O., Cho, I., Kim, S., Bae, J., Uh, Y., Kim, S. J., 2024. 4D Scaffold Gaussian Splatting for Memory Efficient Dynamic Scene Reconstruction. *arXiv preprint arXiv:2411.17044*.

Demantké, J., Mallet, C., David, N., Vallet, B., 2011. Dimensionality based scale selection in 3d lidar point clouds. *Laserscanning*.

Fan, B., Kong, Q., Wang, X., Wang, Z., Xiang, S., Pan, C., Fua, P., 2019. A performance evaluation of local features for image-based 3D reconstruction. *IEEE Transactions on Image Processing*, 28(10), 4774–4789.

Golovinskiy, A., Funkhouser, T., 2009. Min-cut based segmentation of point clouds. *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, IEEE, 39–46.

Gomes, L., Bellon, O. R. P., Silva, L., 2014. 3D reconstruction methods for digital preservation of cultural heritage: A survey. *Pattern Recognition Letters*, 50, 3–14.

Gul, F., Rahiman, W., Nazli Alhady, S. S., 2019. A comprehensive study for robot navigation techniques. *Cogent Engineering*, 6(1), 1632046.

Guo, J., Ma, X., Fan, Y., Liu, H., Li, Q., 2024. Semantic gaussians: Open-vocabulary scene understanding with 3d gaussian splatting. *arXiv preprint arXiv:2403.15624*.

Hackel, T., Wegner, J. D., Schindler, K., 2016. Fast semantic segmentation of 3D point clouds with strongly varying density. *ISPRS annals of the photogrammetry, remote sensing and spatial information sciences*, 3, 177–184.

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.

Li, S., Xiao, X., Guo, B., Zhang, L., 2020. A novel OpenMVS-based texture reconstruction method based on the fully automatic plane segmentation for 3D mesh models. *Remote Sensing*, 12(23), 3908.

Liu, C.-W., Wu, T.-H., Tsai, M.-H., Kang, S.-C., 2018. Image-based semantic construction reconstruction. *Automation in Construction*, 90, 67–78.

Lu, C., 2025. Design of a 3D Reconstruction System for Immovable Cultural Relics Based on Multi View Stereo Matching and Nerf Fusion.

Luo, H., Zhang, J., Liu, X., Zhang, L., Liu, J., 2024. Large-scale 3d reconstruction from multi-view imagery: A comprehensive review. *Remote Sensing*, 16(5), 773.

Ma, X., Wang, Z., Li, H., Zhang, P., Ouyang, W., Fan, X., 2019. Accurate monocular 3d object detection via color-embedded 3d reconstruction for autonomous driving. *Proceedings of the IEEE/CVF international conference on computer vision*, 6851–6860.

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.

Munoz, D., Bagnell, J. A., Vandapel, N., Hebert, M., 2009. Contextual classification with functional max-margin markov networks. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 975–982.

Niu, Y., Liu, L., Huang, F., Huang, S., Chen, S., 2024. Overview of image-based 3D reconstruction technology. *Journal of the European Optical Society-Rapid Publications*, 20(1), 18.

Papon, J., Abramov, A., Schoeler, M., Worgotter, F., 2013. Voxel cloud connectivity segmentation-supervoxels for point clouds. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2027–2034.

Pepe, M., Alfio, V. S., Costantino, D., 2022. UAV platforms and the SfM-MVS approach in the 3D surveys and modelling: A review in the cultural heritage field. *Applied Sciences*, 12(24), 12886.

Qin, M., Li, W., Zhou, J., Wang, H., Pfister, H., 2024. Langsplat: 3d language gaussian splatting. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20051–20060.

Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F. et al., 2024. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*.

Rutzinger, M., Elberink, S. O., Pu, S., Vosselman, G., 2009. Automatic extraction of vertical walls from mobile and airborne laser scanning data. *Laser scanning'09: ISPRS*, International Society for Photogrammetry and Remote Sensing (ISPRS), 7–11.

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

Serna, A., Marcotegui, B., 2014. Detection, segmentation and classification of 3D urban objects using mathematical morphology and supervised learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 93, 243–255.

Stathopoulou, E. K., Welponer, M., Remondino, F., 2019. Open-source image-based 3D reconstruction pipelines: Review, comparison and evaluation. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLII-2/W17*, 331–338.

Stuart, L. A., Pound, M. P., 2025. 3DGS-to-PC: Convert a 3D Gaussian Splatting Scene into a Dense Point Cloud or Mesh. *arXiv preprint arXiv:2501.07478*.

Vosselman, G., Gorte, B. G., Sithole, G., Rabbani, T., 2004. Recognising structure in laser scanner point clouds. *International archives of photogrammetry, remote sensing and spatial information sciences*, 46(8), 33–38.

Wang, T., Wang, X., Hou, Y., Xu, Y., Zhang, W., Zhan, Z., 2025. PG-SAG: Parallel Gaussian Splatting for Fine-Grained Large-Scale Urban Buildings Reconstruction via Semantic-Aware Grouping. *PFG(2025)*. https://doi.org/10.1007/s41064-025-00343-0.

Weinmann, M., Jutzi, B., Hinz, S., Mallet, C., 2015. Semantic point cloud interpretation based on optimal neighborhoods, relevant features and efficient classifiers. *ISPRS Journal of Photogrammetry and Remote Sensing*, 105, 286–304.

Wu, J., Cui, Z., Sheng, V. S., Zhao, P., Su, D., Gong, S., 2013. A Comparative Study of SIFT and its Variants. *Measurement science review*, 13(3), 122.

Xiang, B., Yao, J., Lu, X., Li, L., Xie, R., Li, J., 2018. Segmentation-based classification for 3D point clouds in the road environment. *International Journal of Remote Sensing*, 39(19), 6182–6212.

Xiong, B., Li, Z., Li, Z., 2024a. GauU-scene: A scene reconstruction benchmark on large scale 3D reconstruction dataset using Gaussian splatting. *arXiv preprint arXiv:2401.14032*.

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

Yu, S., Liu, X., Tan, Q., Wang, Z., Zhang, B., 2024. Sensors, systems and algorithms of 3D reconstruction for smart agriculture and precision farming: A review. *Computers and Electronics in Agriculture*, 224, 109229.

Zhan, H., Yu, Y., Xu, Y., Hou, Q., Xia, R., Wang, X., Feng, Y., Zhan, Z., Li, M., Gruber, M. et al., 2023. BeDOI: Benchmarks for Determining Overlapping Images With Photogrammetric Information. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 48, 1685–1692.