The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLIX-B2-2026
XXV ISPRS Congress 2026 "From Imagery to Understanding", Commission II, 4–11 July 2026, Toronto, Canada

Evaluating 3D Scene Representations for Aerial Photogrammetry across Diverse Cityscapes

Shihan Chen[1], Zhaojin Li[2], Qingsong Yan[1], Haibing Liu[1], Huchen Li[1], Wubiao Huang[1], Fei Deng[1,3,*]

[1] School of Geodesy and Geomatics, Wuhan University, Wuhan 430079, China
(csh1999, yanqs_whu, liuhb_whu, lihuchen, huangwubiao)@whu.edu.cn
[2] Technology and Engineering Center for Space Utilization, University of Chinese Academy of Sciences, Beijing 100094, China - lizhaojin@csu.ac.cn
[3] Hubei Luojia Laboratory, Wuhan 430079, China – fdeng@sgg.whu.edu.cn

Keywords: 3D reconstruction, Neural Radiance Field (NeRF), 3D Gaussian Splatting (3DGS), Aerial photogrammetry.

Abstract

The proliferation of continuous Neural Radiance Field (NeRF) and 3D Gaussian Splatting (3DGS) has shifted the paradigm of 3D aerial reconstruction from relying solely on geometric stereo matching to inverse rendering optimization. However, while these emerging rendering-based frameworks excel in synthesizing photo-realistic novel views, their capability to extract accurate surfaces in complex aerial scenarios remains ambiguous compared to traditional methods. To establish a clearer understanding, this study presents a comprehensive evaluation of five representative frameworks spanning traditional Structure from Motion (SfM), purely Signed Distance Field (SDF) representations, unstructured 3D Gaussians, hybrid voxel-Gaussians, and strictly explicit sparse voxels. By systematically standardizing identical computational environments, inputs, and unified mesh-extraction pipelines on both real-world airborne LiDAR datasets and synthetic cityscapes, we assess their performance regarding visual fidelity, geometric accuracy, and resource efficiency. The experimental results reveal that while traditional MVS produces the highest overall geometric precision by strictly enforcing multi-view rigid geometry, it is prone to failures in texture-less regions. Among rendering-based representations, a fundamental trade-off exists: highly flexible, unstructured 3DGS achieve highest visual scores but degrade the underlying geometric surfaces; conversely, explicitly structured techniques demonstrate distinct superiority in regularizing topological coherence and floating artifact suppression. Furthermore, we observe that integrating structured voxels avoids the severe memory bottlenecks associated with extracting geometries from chaotic unorganized primitives. These empirical findings emphasize that for large-scale aerial photogrammetry, integrating explicit spatial structuralization into differentiable rendering pipelines is imperative for achieving scalable operations and bridging the geometric accuracy gap with traditional methods.

1. Introduction

3D reconstruction from aerial images represents a long-standing research topic in photogrammetry, remote sensing and computer vision (Colomina and Molina, 2014). Traditional photogrammetric techniques typically entail SfM (Schönberger and Frahm, 2016) and Multi-View Stereo (MVS) (Schönberger et al., 2016b) to reconstruct textured mesh models. However, these traditional pipelines fundamentally rely on robust pixel-wise feature matching and photometric consistency. Consequently, SfM often suffers from feature mismatches or camera drift in areas with repetitive patterns or severe viewpoint variations (Wu, 2013). Furthermore, subsequent MVS algorithms frequently encounter severe matching ambiguities when dealing with non-Lambertian reflective regions (e.g., glass facades), texture-less surfaces (e.g., water bodies and roofs), and topologically complex or thin structures, inevitably leading to incomplete point clouds and distorted meshes (Schönberger et al., 2016a).

On the other hand, the proliferation of downstream applications such as augmented reality, virtual reality, digital twins (Shahat et al., 2021), and autonomous driving simulation (Li et al., 2023a) has shifted the objective of 3D reconstruction from solely pursuing high geometric accuracy to achieving models with both high-fidelity rendering quality and structural precision. Consequently, the photogrammetry and remote sensing community is increasingly focusing on these dual objectives (Li et al., 2025b, Liu and Qin, 2026).

Emerging novel 3D representations, such as NeRF (Mildenhall et al., 2021) and 3DGS (Kerbl et al., 2023), offer alternative, rendering-based frameworks to traditional MVS methods. These approaches typically rely on SfM to provide accurate camera poses and initial point clouds, with a core loss function that minimizes the difference between rendered and observed images. Currently, NeRF and 3DGS are widely adopted for tasks including novel view synthesis, surface reconstruction (Wang et al., 2021), localization (Yan et al., 2024), and autonomous driving simulation (Zhou et al., 2024) due to their ability to generate high-fidelity rendered novel views and perform pixel-wise optimization. However, although these unconstrained representations excel in rendering, their lack of spatial regularization often leads to geometric degradation, such as floating artifacts in occluded or few-view regions.

To bridge this gap and embed physical geometric priors into the differentiable rendering pipeline, subsequent research has explored structured representation paradigms. By anchoring learnable features or Gaussian primitives to discrete spatial grids, recent advancements like hybrid voxel-Gaussian scaffolds (Lu et al., 2024) and explicit sparse voxels (Sun et al., 2025). These structured models present distinct advantages in regularizing geometric coherence and reducing structural ambiguities. However, the exact impact of such spatial structuralization on resolving the trade-off between rendering fidelity and geometric integrity remains under-explored in the context of large-scale aerial photogrammetry.

To address this gap, this study systematically compares these emerging representations in the context of aerial photogrammetry. We evaluate five representative methods employing distinct 3D representations in terms of visual quality, geometric accuracy, computational efficiency, and resource consumption. The experimental design encompasses challenging real-world scenarios and diverse ground features, including texture-less areas (e.g., water bodies), topologically complex and thin structures, and highly occluded regions (e.g., roads between dense high-rise buildings). Furthermore, we provide an in-depth analysis revealing the fundamental trade-offs between rendering flexibility and rigid geometric accuracy, empirically demonstrating that explicit spatial structuralization is strictly necessary to bridge the geometric gap with traditional MVS and mitigate catastrophic memory bottlenecks during mesh extraction.

2. Related Work

The standard digital photogrammetry pipeline is predominantly driven by SfM and MVS algorithms. Initially, SfM techniques extract and match robust 2D image features across overlapping aerial images. Through bundle adjustment, SfM simultaneously estimates accurate camera intrinsic and extrinsic parameters alongside a sparse 3D point cloud of the scene (Schönberger and Frahm, 2016). Following camera orientation, MVS algorithms are employed to densify the initial 3D geometry. Traditional MVS methods rely heavily on the principle of photometric consistency, searching for corresponding pixels or local patches across multiple views to estimate dense depth maps (Schönberger et al., 2016b). Foundational algorithms such as Semi-Global Matching (SGM) (Hirschmuller, 2008) and PatchMatch Stereo (Bleyer et al., 2011) have significantly advanced the field.

The introduction of NeRF (Mildenhall et al., 2021) and 3DGS (Kerbl et al., 2023) revolutionized 3D reconstruction by shifting the paradigm towards inverse rendering. By encoding complex scenes within the weights of Multilayer Perceptrons (MLPs) and employing differentiable volume rendering, NeRF achieves unprecedented novel view synthesis. By utilizing anisotropic 3D Gaussians and tile-based differentiable rasterization, 3DGS circumvents the computationally expensive ray-marching of NeRF, achieving real-time rendering and significantly faster training convergence. However, the fields of remote sensing and photogrammetry place a critical emphasis on accurate geometric reconstruction (Colomina and Molina, 2014). Therefore, the remainder of this section briefly reviews related research on NeRF- and 3DGS-based surface reconstruction methods.

Early NeRF and 3DGS models represented scenes as semi-transparent density fields or unstructured ellipsoids, making it fundamentally challenging to extract crisp, watertight geometric surfaces. To address this, recent advancements have focused on embedding geometric priors into rendering-based representations. For NeRF-based methods, a major breakthrough was substituting raw volume density with Signed Distance Fields (SDFs) (Wang et al., 2021, Yariv et al., 2021), which culminated in highly detailed surface reconstruction frameworks like Neuralangelo (Li et al., 2023b). Concurrently, to extract explicit meshes from the disordered point clouds of 3DGS, research has explored flattening Gaussians into 2D disks (Huang et al., 2024), binding them to surfaces (Guédon and Lepetit, 2024), or introducing planar constraints (Chen et al., 2024). Furthermore, expanding beyond point-based and planar priors, approaches like GeoSVR (Li et al., 2025a) propose a voxel-like representation that explicitly structures the 3D space, achieving accurate surface extraction by constraining sparse voxels while still retaining alpha-blending mechanisms for rendering.

Despite these advancements, a comprehensive evaluation of these novel representations in the context of aerial photogrammetry, which demands a strict balance between high-fidelity rendering and precise macroscopic geometry, remains an open research question. This study aims to fill this gap by systematically comparing representative methods from these emerging 3D paradigms under challenging real-world scenarios.

3. Method

To systematically evaluate various 3D representations for aerial photogrammetry, we establish a unified evaluation framework. This section describes the selected representations and the procedures for geometric reconstruction.

3.1 Selected 3D Representations

To comprehensively evaluate the performance of emerging 3D scene representations in aerial photogrammetry, we select four representative methods: Neuralangelo (NeRF-based) (Li et al., 2023b), PGSR (3DGS-based) (Chen et al., 2024), CityGS-X (hybrid voxel-Gaussian) (Gao et al., 2025), and GeoSVR (voxel-based) (Li et al., 2025a). These methods are selected not only because they represent the state-of-the-art (SOTA) for surface reconstruction within their respective paradigms, but also because they constitute a deliberate spectrum of varying spatial structuralization. They are compared against the industry-standard commercial software, iTwin Capture Modeler (Bentley Systems, 2024), which serves as a traditional photogrammetric baseline.

Neuralangelo (Li et al., 2023b) is a state-of-the-art NeRF-based approach to reconstruct with high geometric accuracy. It represents the scene as a continuous SDF encoded within a multiresolution hash encoding and a lightweight MLP. The SDF-based representation allows for the extraction of high-quality, watertight meshes via the marching cubes algorithm (Lorensen and Cline, 1987). It leverages numerical gradients to supervise the Eikonal loss, theoretically enabling the recovery of fine-grained geometric details from multi-view images. However, its global implicit nature often faces challenges in converging across large-scale aerial scenes.

Standard 3DGS (Kerbl et al., 2023) represents the scene using a collection of millions of completely unstructured, anisotropic 3D Gaussians. PGSR (Chen et al., 2024), a leading method for Gaussian-based surface extraction, introduces the first level of structuralization. It flattens Gaussians into disk-like structures and introduces geometric regularization that aligns them with local surface geometry. By rendering and supervising unbiased depth maps, PGSR bridges the gap between unstructured point clouds and coherent surfaces, making it particularly effective for reconstructing the sharp edges and flat facades common in urban environments.

We select CityGS-X (Gao et al., 2025) as the representative of the hybrid voxel-Gaussian paradigm. It utilizes a spatial partitioning strategy where Gaussians are anchored and managed within a voxel-based octree structure. This hybrid approach is designed to address the memory and optimization bottlenecks

https://doi.org/10.5194/isprs-archives-XLIX-B2-2026-367-2026 | © Author(s) 2026. CC BY 4.0 License.

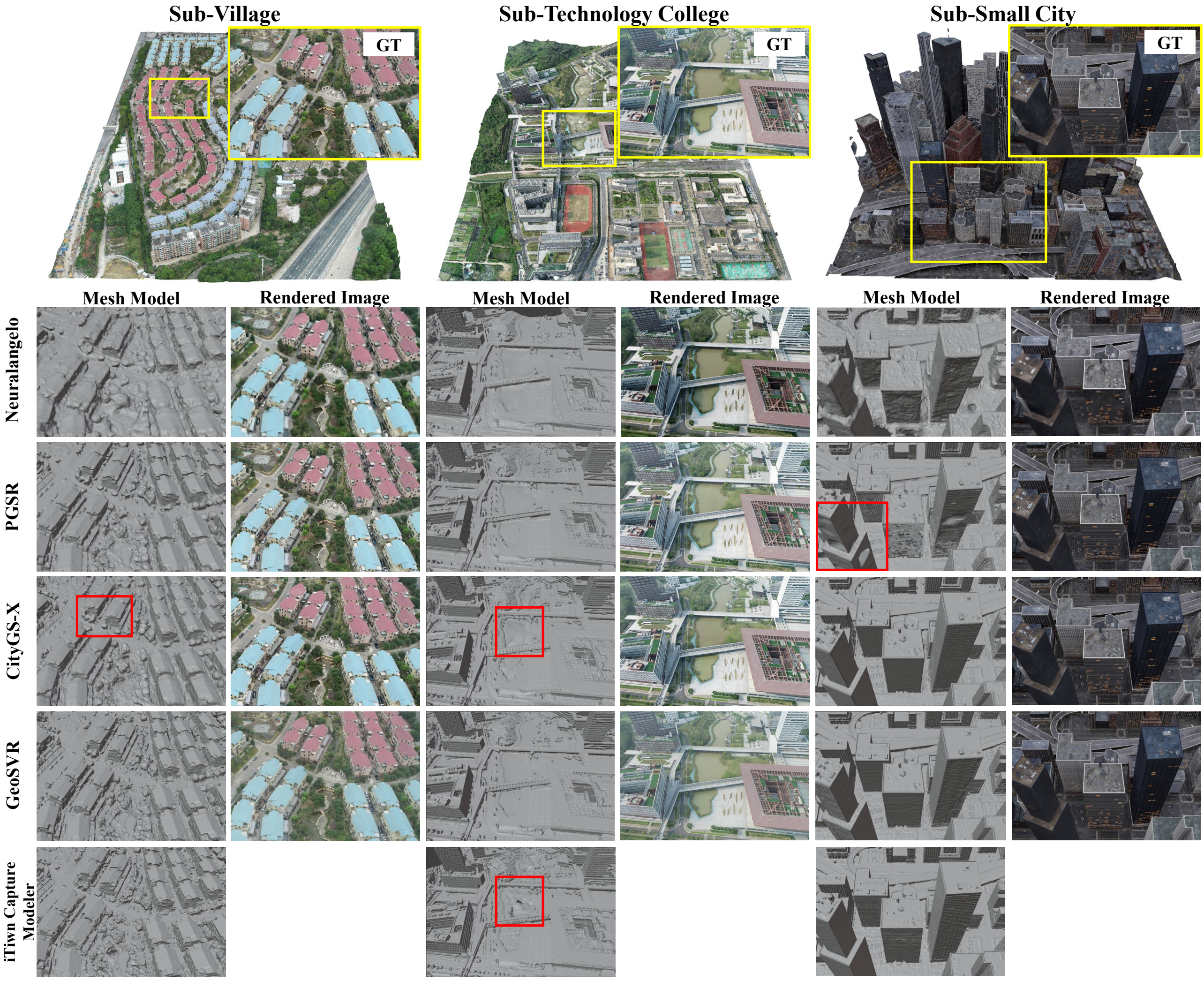


Figure 1. Comparison of reconstructed results employing different 3D representations.

of standard 3DGS when dealing with vast cityscapes. In our evaluation, it serves as a intermediate baseline to determine if a hybrid architecture can combine the flexibility of unstructured Gaussians with the regularity of strict voxel grids.

GeoSVR (Li et al., 2025a) represents the most strictly structured explicit representation in our evaluation. It employs a purely voxel-based geometry, where the scene is discretized into a sparse voxel octree. The structured representation explicitly defines the occupancy and volume density within each predefined cell. As a SOTA volumetric approach optimized for modern differentiable rendering, GeoSVR incorporates strong geometric priors to constrain the surface strictly within sparse voxels. By comparing these methods, we aim to systematically reveal how increasing degrees of spatial structuralization impact the quality of surface reconstruction in complex, occluded urban scenes.

3.2 Standardized Reconstruction Pipeline

To ensure a fair and rigorous evaluation across the diverse 3D representations, we establish a standardized reconstruction pipeline. The core principle is to isolate the inherent capabilities of the representations from algorithmic implementation biases by providing all methods with identical initial conditions and unifying the final output format.

Data Preparation and Initialization: All selected methods are fed with the exact same set of input images and intrinsics and extrinsics. We select both the real-world dataset GauU-Scene V2 (Xiong et al., 2024) and the synthetic dataset MatrixCity (Li et al., 2023a). For GauU-Scene V2, we process the multi-view images using the standard SfM pipeline via COLMAP to estimate camera poses and generate a sparse point cloud. For MatrixCity, we directly utilize the ground-truth camera poses recorded during the simulation engine's rendering process, which effectively isolates the geometric evaluation from potential errors induced by intrinsic and extrinsic inaccuracies.

Following pose estimation, the 3D scene representations are initialized. The explicit methods (PGSR, CityGS-X, and GeoSVR) are initialized using the identical sparse 3D point cloud generated by COLMAP. For MatrixCity, we fix the known intrinsics and extrinsics, and perform feature extraction, feature matching, and triangulation to simulate how sparse point clouds are realistically obtained in practical 3D reconstruction pipelines. In contrast, Neuralangelo is initialized within a normalized bounding sphere that encapsulates the entire scene. Each representation is then optimized according to its specific loss functions and regularization strategies to minimize photometric errors against the observed images.

Surface Mesh Extraction: Given that high-fidelity macroscopic geometry is a primary objective in photogrammetry, we

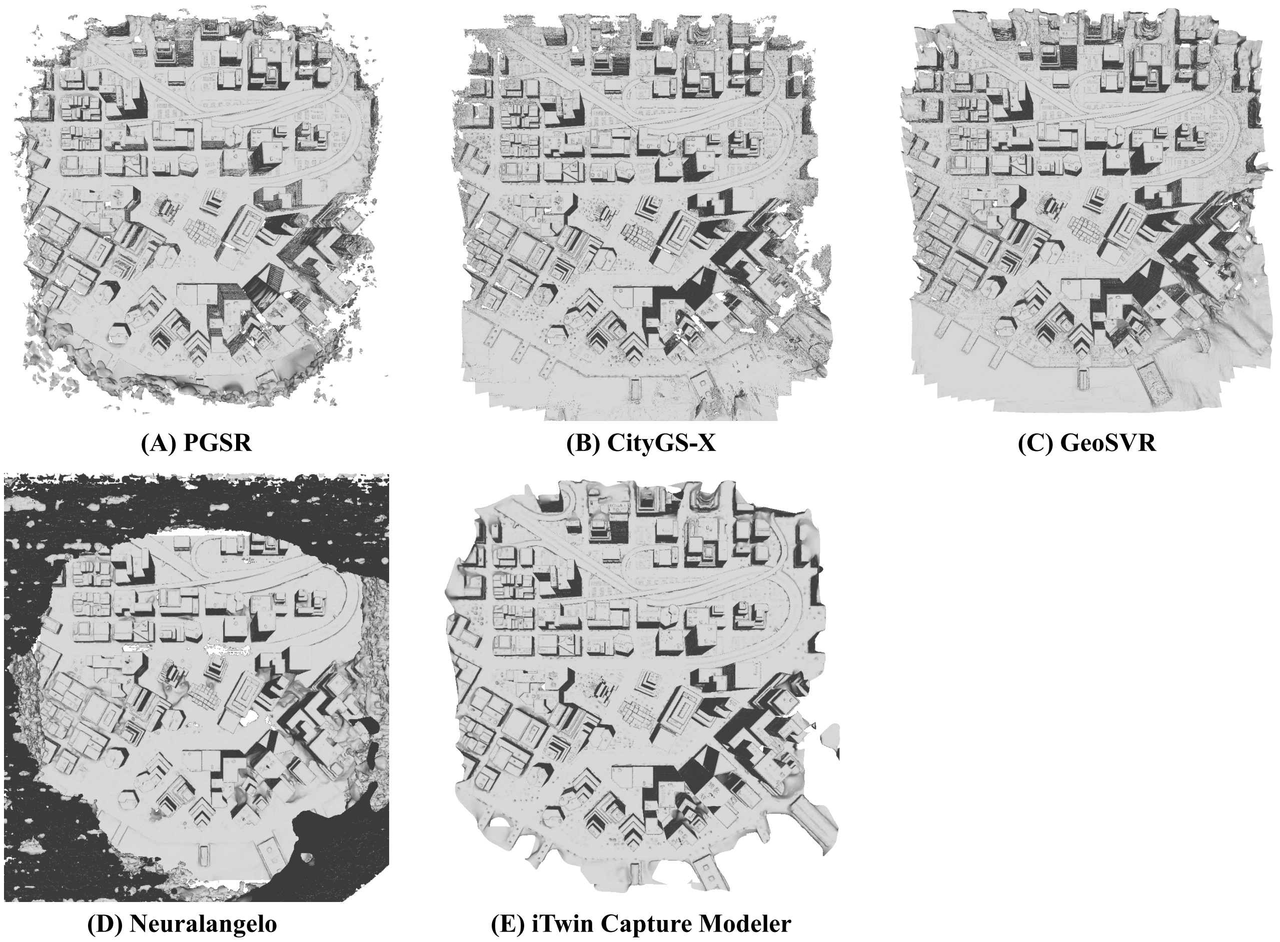


Figure 2. Overview of the reconstruction results on Sub-Small City.

extract 3D triangular meshes from the optimized models. Due to the fundamental structural differences among the representations, we employ two distinct surface extraction strategies:

- **Direct Extraction for Implicit SDFs:** For Neuralangelo, the geometric surface is intrinsically defined as the zero-level set of its learned SDF. We directly apply the marching cubes algorithm (Lorensen and Cline, 1987) to SDF at a high volumetric resolution to extract a surface mesh.

- **TSDF Fusion for Explicit Representations:** Methods based on 3D Gaussians and explicit voxels (PGSR, CityGS-X, GeoSVR) do not inherently define a globally continuous mathematical surface. To bridge this gap, we adopt a unified depth-fusion approach. First, we render multi-view dense depth maps from the optimized 3D Gaussians or voxel grids at the original camera poses. These rendered depth maps are subsequently fused into a global Truncated Signed Distance Function (TSDF) volume (Curless and Levoy, 1996). Finally, the marching cubes algorithm is applied to the TSDF volume to comprehensively reconstruct the explicit surface mesh.

This unified mesh extraction protocol ensures that despite the derived geometric models of different underlying data structures can be quantitatively compared against ground-truth data and the commercial photogrammetric baseline on equal footing.

4. Experimental Setup

4.1 Dataset

Experiments were conducted on the real-world GauU-Scene V2 (Xiong et al., 2024) and the synthetic MatrixCity (Li et al., 2023a) datasets. GauU-Scene V2 dataset is a large-scale, real-world aerial photogrammetry dataset. Crucially, It provides high-precision airborne LiDAR point clouds, which serve as the absolute ground truth for our geometric evaluation. To manage computational burden and focus on reconstruction capabilities, we selected 301 images and 600 images from ‘Village’ and ‘Technology College’ of GauU-Scene, namely ‘Sub-Village’ and ‘Sub-Technology College’. The ‘Sub-Village’ scene features low-rise buildings with vegetation, while The ‘Sub-Technology College’ includes a body of water surrounded by complex structures, testing the methods’ ability to reconstruct texture-less surfaces and thin features.

The synthetic MatrixCity dataset provides exact ground-truth camera poses directly from the simulation engine. We selected 986 images from ‘Small City’, namely ‘Sub-Small City’. It contains dense high-rise buildings whose mutual occlusion poses challenges for rendering-based methods.

4.2 Evaluation Metrics

We assess the performance of these methods across three primary dimensions: visual quality, geometric accuracy, and computational efficiency.

Methods	Visual Metrics			Geometric Metrics (m)	
	PSNR ↑	SSIM ↑	LPIPS ↓	MEA ↓	RMSE ↓
iTwin Capture Modeler	——	——	——	**0.574**	0.988
Neuralangelo	24.65	0.686	0.404	0.796	1.196
PGSR	**28.37**	**0.886**	0.174	0.733	1.076
CityGS-X	27.78	0.883	**0.150**	0.730	**0.964**
GeoSVR	26.52	0.852	0.172	0.713	0.975

Table 1. Quantitative comparison of different methods. The best results are highlighted in **bold**.

Geometric Accuracy: Accurate geometry is the core objective of aerial photogrammetry. We compute the explicit distance between the reconstructed meshes and the ground-truth point clouds (LiDAR for GauU-Scene V2, and simulation GT for MatrixCity). The geometric deviations are aggregated using Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). MAE calculates the average absolute distance between the reconstructed surface and the ground truth, reflecting the overall reconstruction precision. RMSE calculates the square root of the average squared distances. RMSE heavily penalizes large local deviations, making it particularly effective for identifying outliers, floating artifacts, or severe topological failures in the reconstructed models.

Visual Quality: We reserve 5% of images from each dataset as a test set. The visual fidelity of the rendered images from the optimized models is evaluated against the ground-truth images using Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) (Wang et al., 2004), and Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018).

Computational Efficiency: For practical applications in large-scale mapping, efficiency is critical. We record the total reconstruction time and monitor the peak GPU memory and CPU memory consumption during training.

4.3 Implementation Details

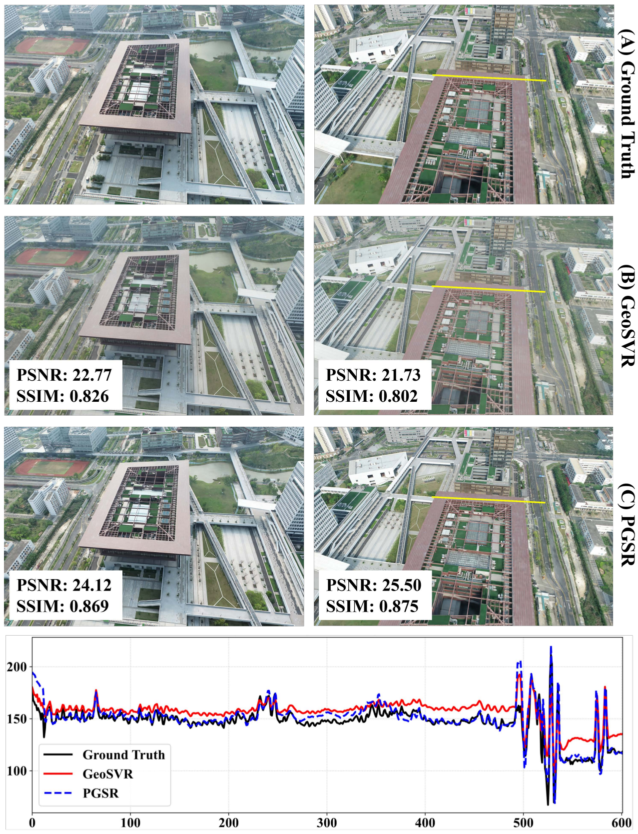
All experiments, including the neural and Gaussian-based methods, are conducted on a unified hardware platform equipped with NVIDIA RTX 4090 GPUs and 512 GB of system RAM. All images are downsampled with the long side resized to 1600 pixels following the default settings. The final reconstructed meshes are extracted at a consistent resolution to ensure comparability in geometric evaluation.

5. Results

5.1 Qualitative Evaluation

A qualitative evaluation through Figures 1, 2, and 3 intuitively reveals the distinct advantages and critical flaws of each representation. Specifically, Figure 1 compares fine geometric details, Figure 2 demonstrates the structural performance at scene boundaries and few-view areas in gaps between buildings, and Figure 3 illustrates the photometric fitting capabilities using a pixel intensity profile in regions with intense illumination.

The traditional MVS pipeline, iTwin Capture Modeler, provides a generally stable baseline but experiences evident matching failures in texture-less regions, manifesting as severe holes over textureless areas like water bodies in Figure 2(E). Conversely, Neuralangelo, representing the global implicit SDF paradigm, struggles significantly in extensive aerial scenarios. While its renderings are noticeably blurry and extracted meshes appear



(D) Intensity Profile Comparison Along the Yellow Line

Figure 3. Rendering results and intensity profile comparison. The bottom subplot shows the pixel intensity extracted along the yellow line, demonstrating that GeoSVR exhibits higher brightness, whereas PGSR closely matches the Ground Truth.

bloated and melted, as shown in Figure 1. Figure 2(D) demonstrates that the entire normalized space is filled with chaotic noise. In unobserved areas lacking multi-view supervision, such as scene boundaries, the global MLP produces entirely random geometric predictions.

Benefiting from the immense degrees of freedom inherent in unconstrained 3DGS, PGSR achieves unparalleled detail preservation. It excels at reconstructing fine structures and sharp edges, as shown in Figure 1. Furthermore, as demonstrated in Figure 3, PGSR easily handles regions with drastic color and illumination changes. It seamlessly fits these complex photometric variations by dynamically densifying and spawning new Gaussians. However, this extreme flexibility induces severe geometric degradation in less constrained areas. First, at scene boundaries where images are inherently few, as seen in Figure 2(A), the lack of multi-view constraints causes PGSR to generate massive amounts of chaotic floaters. Second, this phenomenon is exacerbated on the facades of densely packed high-rise buildings, highlighted by the red boxes in Figure

Method	Time Consumption (Min) ↓		GPU Num in Training
	Reconstruction	Mesh Extraction	
iTwin Capture Modeler	132.67		——
Neuralangelo	4539.33	6.50	4
PGSR	245.67	30.00	1
CityGS-X	327.33	21.53	1
GeoSVR	**104.00**	**8.27**	1

Table 2. Comparison of computational efficiency. The best results among rendering-based methods are highlighted in **bold**.

Method	Reconstruction Resource (GB) ↓		Mesh Extraction Resource (GB) ↓	
	RAM	GPU Memory	RAM	GPU Memory
Neuralangelo	116.99	23.22	53.09	20.75
PGSR	86.55	**19.17**	146.95	12.90
CityGS-X	**74.75**	21.00	**69.29**	16.65
GeoSVR	91.01	22.81	94.57	**7.72**

Table 3. Comparison of peak resource consumption. The best results are highlighted in **bold**.

1. This specific facade failure stems from two compounding factors: building boundaries and narrow gaps are frequently occluded in aerial multi-view images and have limited projection areas, leading to insufficient photometric supervision; moreover, for high-rise buildings, the upper facades are closer to the drone cameras, causing the 3D Gaussians to project over a larger pixel area in the images. During the pixel-wise gradient back-propagation, these Gaussians accumulate excessively large gradients, triggering abnormal densification and destroying the underlying geometric structure. Lacking spatial regularization, pure 3DGS overfits 2D training views, sacrificing foundational 3D structure.

Methods incorporating explicit spatial structures, namely CityGS-X and GeoSVR, present different trade-offs. Benefiting from rigid voxel constraints, GeoSVR suppresses floaters, yielding the cleanest geometry and best structural completeness at scene boundaries and narrow gaps, as depicted in Figure 2(C). However, CityGS-X exhibits two limitations indicating an imbalance in geometric constraints: firstly, sharp edges and fine structural details of buildings are overly smoothed out, losing crisp corners; secondly, smooth and texture-less surfaces fail to reconstruct evenly, producing erroneous holes, documented in Figure 1. Furthermore, the rigid constraints of GeoSVR significantly limit its photometric capacity. As illustrated in Figure 3, the building roof exhibits complex reflective properties. While PGSR successfully reproduces accurate color and brightness, GeoSVR suffers from overexposure and a whitish tint in these regions. This color shift is quantitatively validated by the intensity profile extracted along the yellow line. The profile demonstrates that the pixel intensity curve of GeoSVR is consistently higher than that of the Ground Truth, whereas PGSR closely aligns with the Ground Truth. The localized PSNR/SSIM metrics annotated in the figure further corroborate this performance gap. This deficiency in GeoSVR occurs because its primitive properties are bound to the vertices of voxel; unlike PGSR, it cannot dynamically spawn or adjust a massive number of local primitives to forcibly fit high-frequency illumination changes or dense grid-like textures, resulting in poorer color fidelity and blurred artifacts. Nevertheless, it must be noted that such an excessive proliferation of Gaussian primitives, while maximizing photometric scores, acts as a double-edged sword, inevitably coming at the severe cost of underlying geometric accuracy as discussed in last paragraph.

5.2 Quantitative Analysis

Table 1 presents the quantitative evaluation of the visual and geometric performance of the evaluated methods. Consistent with the qualitative observations, the results highlight a distinct trade-off between photometric fitting and geometric accuracy among emerging 3D representations.

In terms of geometric accuracy, the traditional photogrammetric baseline, iTwin Capture Modeler, achieves the lowest Mean Absolute Error, or MAE, of 0.574 m, demonstrating the robustness of traditional MVS in preserving the overall scene geometry. This superior accuracy stems from MVS strictly relying on rigid multi-view geometric constraints and Epipolar geometry; in areas with abundant textures and sufficient overlap where image matching succeeds, it produces the most precise geometric measurements. However, as noted in Section 5.1, this reliance causes severe failure in texture-less regions. Among the novel rendering-based methods, representations with explicit spatial structuralization perform significantly better. GeoSVR achieves an MAE of 0.713 m, while CityGS-X achieves the lowest Root Mean Square Error, or RMSE, of 0.964 m. The stricter spatial constraints inherent in voxel-based and hybrid representations effectively regularize the reconstruction, reducing severe local deviations and floaters. Conversely, the purely implicit Neuralangelo struggles to converge on large-scale aerial scenes, resulting in the highest geometric errors with an RMSE of 1.196 m.

For visual quality, methods based on 3DGS demonstrate a clear advantage. PGSR achieves the highest PSNR of 28.37 and an SSIM of 0.886, while CityGS-X attains the best LPIPS score of 0.150. The immense degrees of freedom provided by unstructured Gaussians enable these methods to flawlessly overfit complex photometric variations, such as the view-dependent reflections on building roofs discussed in Section 5.1. However, this extreme photometric flexibility often comes at the cost of underlying structural integrity. As evident in the geometric metrics, the superior 2D rendering scores of PGSR do not translate to the best 3D geometry.

This inherent conflict between geometric accuracy and rendering fidelity has been widely observed in recent studies (Fu et al., 2022, Chen et al., 2024). A fundamental cause of this phenomenon lies in the standard rendering formulation: both geometric depth rendering and radiance, or color, accumulation are governed by the same density or opacity variable, denoted as σ or α. However, the density distribution required to optimally fit complex view-dependent appearances and high-frequency textures is not necessarily aligned with the strict physical solid geometry. Treating them as strictly equivalent often leads to structural ambiguities. This precise issue was identified in the NeRF domain by NeuS (Wang et al., 2021), which proposed an unbiased formulation to extract density from an SDF to better align the volume rendering opacity with the true underlying surface. Nevertheless, gracefully decoupling and balancing appearance and geometry for emerging explicit representations in large-scale aerial scenarios remains an open challenge.

5.3 Computational Efficiency and Resource Consumption

Tables 2 and 3 detail the computational efficiency and resource requirements, which are critical for practical large-scale mapping applications. Regarding time consumption detailed in Table 2, GeoSVR demonstrates remarkable efficiency, completing the reconstruction in 104 minutes and mesh extraction in just 8.27 minutes, surpassing even the traditional iTwin Capture baseline of 132.67 minutes. This highlights the computational benefits of sparse voxel structures in ray tracing and optimization. In contrast, while 3DGS methods render quickly, their surface reconstruction variations, which involve complex regularizations, are time-consuming, with PGSR taking 245.67 minutes and CityGS-X taking 327.33 minutes. Neuralangelo, representing the implicit paradigm, is computationally prohibitive for practical production, requiring over 4500 minutes even with 4 GPUs.

Memory consumption, presented in Table 3, reveals further operational bottlenecks. During reconstruction, Gaussian-based methods, namely PGSR and CityGS-X, maintain relatively low GPU memory profiles of approximately 19 to 21 GB. However, a critical bottleneck emerges during the mesh extraction phase for unstructured Gaussians. PGSR experiences a massive surge in system RAM of 146.95 GB due to the demanding TSDF fusion process. This is because the unconstrained optimization of massive, unstructured Gaussians generates noisy and complex depth maps, which forces the subsequent volumetric fusion to process and integrate a vast number of conflicting surface hypotheses, thereby consuming excessive memory. Conversely, methods that inherently maintain a spatial structure, specifically GeoSVR and CityGS-X, exhibit excellent lightweight characteristics during mesh extraction. GeoSVR requires only 7.72 GB of VRAM, and CityGS-X consumes the lowest system memory at 69.29 GB. These structured representations produce cleaner, more coherent depth or density fields, making them fundamentally more amenable to direct surface extraction and avoiding the memory-intensive overhead of fusing unstructured predictions.

6. Conclusion

This paper presents a comprehensive evaluation of emerging 3D scene representations, spanning purely implicit NeRFs, unstructured 3D Gaussians, hybrid voxel-Gaussians, and explicit sparse voxels, against traditional baselines for aerial photogrammetry. Through rigorous experiments on both synthetic and real-world urban datasets, our findings reveal distinct characteristics among algorithmic paradigms. Traditional pipelines, governed by rigid multi-view epipolar geometry, continue to provide the highest macroscopic geometric accuracy when pixel-wise matched correctly, yet are vulnerable to failure in weakly textured or transparent regions. The emerging differentiable-rendering representations complete surfaces in such challenging regions via view-synthesis constraints, but exhibit an internal trade-off between photometric rendering fidelity and geometric structural integrity.

Within these novel representations, unstructured 3DGS methods excel at fitting complex view-dependent appearances and high-frequency textures, achieving the highest visual metrics. However, their unconstrained nature leads to severe structural degradation, generating noisy geometric outliers, especially in less supervised or highly occluded areas. Conversely, explicitly structured representations, particularly voxel-based approaches, demonstrate superior robustness in preserving global topological coherence and extracting clean, watertight geometric surfaces. Furthermore, from an operational perspective, explicit spatial structures significantly reduce computational bottlenecks during the critical mesh extraction phase, vastly improving processing speeds and lowering peak RAM consumption. In conclusion, for practical large-scale aerial photogrammetry mapping, optimizing solely for 2D novel view synthesis metrics is insufficient and often detrimental to the underlying 3D geometry. To harness the benefits of inverse rendering while bridging the precision gap with traditional MVS, imposing explicit spatial regularization is crucial. Future research should prioritize developing unified paradigms that elegantly decouple rendering opacity from solid geometry, striving to achieve photo-realistic rendering without compromising the rigid macroscopic accuracy of the foundational 3D structure.

References

Bentley Systems, 2024. iTwin Capture Modeler. Bentley Systems, Inc. Available: `https://www.bentley.com/software/itwin-capture-modeler/`. Accessed: 2024-11-01.

Bleyer, M., Rhemann, C., Rother, C., 2011. Patchmatch stereo-stereo matching with slanted support windows. *Bmvc*, 11number 2011, 1–11.

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*.

Colomina, I., Molina, P., 2014. Unmanned aerial systems for photogrammetry and remote sensing: A review. *ISPRS Journal of photogrammetry and remote sensing*, 92, 79–97.

Curless, B., Levoy, M., 1996. A volumetric method for building complex models from range images. *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, 303–312.

Fu, Q., Xu, Q., Ong, Y.-S., Tao, W., 2022. Geo-Neus: Geometry-Consistent Neural Implicit Surfaces Learning for Multi-view Reconstruction. *Advances in Neural Information Processing Systems (NeurIPS)*.

Gao, Y., Li, H., Chen, J., Zou, Z., Zhong, Z., Zhang, D., Sun, X., Han, J., 2025. Citygs-x: A scalable architecture for efficient and geometrically accurate large-scale scene reconstruction. *arXiv preprint arXiv:2503.23044*.

Guédon, A., Lepetit, V., 2024. Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5354–5363.

Hirschmuller, H., 2008. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on pattern analysis and machine intelligence*, 30(2), 328–341.

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.

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, J., Zhang, J., Zhang, Y., Bai, X., Zheng, J., Yu, X., Gu, L., 2025a. Geosvr: Taming sparse voxels for geometrically accurate surface reconstruction. *arXiv preprint arXiv:2509.18090*.

Li, Y., Jiang, L., Xu, L., Xiangli, Y., Wang, Z., Lin, D., Dai, B., 2023a. Matrixcity: A large-scale city dataset for city-scale neural rendering and beyond. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3205–3215.

Li, Z., Müller, T., Evans, A., Taylor, R. H., Unberath, M., Liu, M.-Y., Lin, C.-H., 2023b. Neuralangelo: High-fidelity neural surface reconstruction. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8456–8465.

Li, Z., Yao, S., Wu, T., Yue, Y., Zhao, W., Qin, R., García-Fernández, Á. F., Levers, A., Ralph, J., Zhu, X., 2025b. ULSR-GS: Urban large-scale surface reconstruction Gaussian Splatting with multi-view geometric consistency. *ISPRS Journal of Photogrammetry and Remote Sensing*, 230, 861-880.

Liu, H., Qin, R., 2026. AeroDGS: Physically Consistent Dynamic Gaussian Splatting for Single-Sequence Aerial 4D Reconstruction. *arXiv preprint arXiv:2602.22376*.

Lorensen, W. E., Cline, H. E., 1987. Marching cubes: A high resolution 3D surface construction algorithm. *ACM siggraph computer graphics*, 21(4), 163–169.

Lu, T., Yu, M., Xu, L., Xiangli, Y., Wang, L., Lin, D., Dai, B., 2024. Scaffold-gs: Structured 3d gaussians for view-adaptive rendering. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20654–20664.

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.

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

Schönberger, J. L., Zheng, E., Frahm, J.-M., Pollefeys, M., 2016a. Pixelwise view selection for unstructured multi-view stereo. *European conference on computer vision*, Springer, 501–518.

Schönberger, J. L., Zheng, E., Pollefeys, M., Frahm, J.-M., 2016b. Pixelwise view selection for unstructured multi-view stereo. *European Conference on Computer Vision (ECCV)*.

Shahat, E., Hyun, C. T., Yeom, C., 2021. City digital twin potentials: A review and research agenda. *Sustainability*, 13(6), 3386.

Sun, C., Choe, J., Loop, C., Ma, W., Wang, Y. F., 2025. Sparse voxels rasterization: Real-time high-fidelity radiance field rendering. *CVPR*.

Wang, P., Liu, L., Liu, Y., Theobalt, C., Komura, T., Wang, W., 2021. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*.

Wang, Z., Bovik, A. C., Sheikh, H. R., Simoncelli, E. P., 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4), 600–612.

Wu, C., 2013. Towards linear-time incremental structure from motion. *2013 International Conference on 3D Vision-3DV 2013*, IEEE, 127–134.

Xiong, B., Zheng, N., Liu, J., Li, Z., 2024. Gauu-scene v2: Assessing the reliability of image-based metrics with expansive lidar image dataset using 3dgs and nerf. *arXiv preprint arXiv:2404.04880*.

Yan, C., Qu, D., Xu, D., Zhao, B., Wang, Z., Wang, D., Li, X., 2024. Gs-slam: Dense visual slam with 3d gaussian splatting. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 19595–19604.

Yariv, L., Gu, J., Kasten, Y., Lipman, Y., 2021. Volume rendering of neural implicit surfaces. *Advances in neural information processing systems*, 34, 4805–4815.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., Wang, O., 2018. The unreasonable effectiveness of deep features as a perceptual metric. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhou, X., Lin, Z., Shan, X., Wang, Y., Sun, D., Yang, M.-H., 2024. Drivinggaussian: Composite gaussian splatting for surrounding dynamic autonomous driving scenes. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 21634–21643.