Orthophoto Generation via 3D Reconstruction and Intelligent Point-Cloud Filtering

Author:
LIUXUAN XIE

Supervisor:
XIAOSONG YANG

National Center for Computer Animation

Faculty of Media & Communication

Bournemouth University

A thesis submitted in partial fulfilment of the requirements of
Bournemouth University for the degree of
Master of science in Artificial Intelligence for Media

AUGUST 2025

Abstract

Three-dimensional reconstruction is a core topic in computer vision and photogrammetry. In recent years, 3D Gaussian Splatting (3DGS) has emerged as a compelling solution for multi-view 3D reconstruction due to its strong rendering quality and practical training efficiency. However, because 3DGS is optimized for appearance rendering, its directly exported point clouds often contain substantial geometric noise, floating artifacts, and redundant background points, which severely limits their direct use in professional domains—such as surveying and urban planning—that demand high geometric accuracy.

To address these issues, we propose an orthophoto generation method with improved geometric fidelity that couples 3DGS reconstruction with adaptive post-processing. The pipeline first reconstructs an initial colored point cloud from a multi-view image sequence using 3DGS. We then introduce an a learning-based point-cloud filtering algorithm: after automatic ground detection and coordinate-frame construction, we extract multi-level geometric and statistical features (e.g., height, density, and outlierness). Based on these features, we create pseudo-labels and train a lightweight MLP classifier to perform weakly supervised point-cloud classification without manual annotations, thereby effectively separating primary structures from noise. Finally, using the purified point cloud, we render an orthophoto via a depth-aware soft-rasterization scheme, and further refine image quality by multi-sample jittered rendering with confidence-based fusion to suppress artifacts.

We conduct comprehensive experiments to validate the proposed method. Ablation studies confirm the necessity of each post-processing module, while comparative evaluations show that our approach produces orthophotos with higher visual fidelity and geometric accuracy relative to directly rendering 3DGS point clouds and to conventional point-cloud filtering baselines. Qualitative results further illustrate clear improvements before/after filtering and in the final orthophoto generation. Overall, this work provides an effective technical pathway for mitigating geometric noise in neural rendering models and can broaden the practical applicability of 3DGS to professional real-world 3D modeling.

Finally, we generate an orthographic (top-down) image by projecting the cleaned point set onto a ground-aligned plane using depth-aware soft rasterization in a local UV frame. The orthophoto is then used to visually and quantitatively check the point-cloud cleaning results, such as residual clutter and coverage, without introducing extra modeling steps.

Keywords: 3D Gaussian Splatting,point cloud processing,orthophoto generation, weakly supervised learning, neural rendering.

Acknowledgements

First and foremost, I am deeply grateful to my parents for their unwavering financial and emotional support throughout my master's studies; this work would not have been possible without them.

I owe my sincere thanks to my supervisor, Prof. Xiaosong Yang, for his patient, insightful, and meticulous guidance at every stage of my research and learning. I am also thankful to Mr. Boyuan Cheng, my senior, for his generous advice and help that shaped my early direction.

My heartfelt appreciation goes to the faculty who taught me during the master's programme—Dr Hammadi Nait-Charif, Dr Zhidong Xiao, and Dr Jon Macey—for their dedicated teaching and encouragement. I would also like to thank all my classmates for their support, kindness, and camaraderie along the way.

Last but not least, I am especially grateful to my cousin, Junhao Hu, for his technical support, and to my girlfriend, Qi Zhang, for her constant encouragement and care throughout my studies.

Thank you all.

Table of Contents

List of Tables

TABLE **Page**

List of Figures

1 Introduction

1.1 Background and Motivation

Three-dimensional reconstruction, a central research direction in computer vision and photogrammetry, is playing an increasingly important role in the digital era(Q. Wang et al., 2025). By recovering an object's 3D geometry and appearance from 2D imagery or sensor measurements, it seeks to build digital models that faithfully reflect the physical world(Chen et al., 2025). From traditional Structure-from-Motion (SfM) and Multi-View Stereo (MVS) pipelines to the latest neural rendering techniques, 3D reconstruction has undergone substantial technological evolution. Although widely adopted, classical approaches still suffer from inherent limitations—complex workflows, heavy computation, and difficulty handling textureless regions and challenging illumination(Akhavi Zadegan, Vivet, and Hadachi, 2025) (Zhou et al., 2024).

In recent years, neural rendering has catalyzed a step change for 3D reconstruction. Neural Radiance Fields (NeRF)(Mildenhall et al., 2020) and its derivatives represent scenes as continuous volumetric functions and deliver outstanding novel view synthesis quality, producing detailed 3D scenes with realistic lighting(Müller et al., 2022). However, NeRF-style methods face practical constraints in training/rendering speed and demand a large memory footprint, which hinders broad deployment in real-world applications.

Introduced in 2023, 3D Gaussian Splatting (3DGS) offers a practical balance between quality and efficiency(Kerbl et al., 2023). It adopts an explicit scene representation with anisotropic 3D Gaussian ellipsoids, each parameterized by position, color, opacity, and scale, and achieves efficient rendering via a differentiable rasterizer. 3DGS can be trained within short time budgets, reaches or surpasses NeRF-level visual quality, and attains real-time frame rates, setting a new benchmark in 3D reconstruction.

Despite these advantages, 3DGS remains fundamentally appearance-centric: its objective is to minimize the photometric discrepancy between rendered and input images rather than to recover metrically accurate geometry. Consequently, the directly exported point clouds often contain substantial geometric noise—floating points ("floaters"), background clutter, artifacts, and holes—which limits their use in domains such as surveying, archaeology, and urban planning, where geometric precision is paramount.

In such professional applications, 3D models are not merely visualization assets but the foundation for quantitative analysis, monitoring, and decision-making. Surveying requires high-accuracy digital surface models (DSMs) and orthorectified imagery for earthwork estimation and topographic change detection; archaeology and cultural heritage demand precise digital archiving, where geometric noise can lead to misinterpretation; urban planning and smart-city analytics rely on clean models for daylight/shadow studies and floor area ratio (FAR) computations. Therefore, developing a post-processing framework that effectively denoises 3DGS outputs and produces geometrically reliable

products holds practical relevance.

1.2 Problem Statement

The core scientific question of this study is how to robustly and automatically isolate and extract clean, semantically meaningful geometric structures from noisy initial 3DGS point clouds, and then generate orthorectified imagery that satisfies the accuracy requirements of professional applications. We further decompose this question into three subproblems: (i) point-cloud denoising—accurately discriminating valid points representing real terrain and buildings from invalid points caused by noise and redundant background; (ii) automation—designing a fully automatic workflow without human intervention that generalizes across diverse outdoor scenes; and (iii) geometric product generation—producing distortion-free, high-resolution orthophotos from the purified point clouds(Çanakçı et al., 2025). We therefore fix orthophoto as the target product to align an appearance-centric 3DGS reconstruction with measurement-oriented use. An orthographic, orthorectified top-down view is metric and perspective-free, making the output directly compatible with mapping workflows and temporal comparison. This choice also turns 3D denoising gains into a reproducible 2D deliverable without introducing meshing-related artifacts or extra complexity.

Accordingly, we set the following research objectives. In line with this choice, our objectives are organized around delivering a point-based orthographic product rather than an intermediate mesh. First, build an end-to-end automated processing pipeline that spans raw 3DGS model export, intelligent point-cloud filtering, and orthophoto generation. Second, develop an adaptive point-cloud filtering algorithm that integrates geometric analysis, statistical learning, and lightweight machine-learning models to achieve high-precision separation of noise and primary structures without relying on manually labeled training data. Third, design a high-quality orthophoto generator that, given the purified point clouds, employs depth-aware soft rasterization together with noise-robust strategies to produce detail-rich, radiometrically consistent orthophotos. Finally, conduct comprehensive experimental validation with systematic evaluation and analysis of both the point-cloud purification effectiveness and the quality of the generated orthophotos. These objectives map directly to our modules: RANSAC ground detection $\rightarrow$ orthographic reference frame; feature engineering with weak supervision $\rightarrow$ retention of ground-attached primaries and suppression of clutter; depth-aware soft rasterization with multi-jittered blending $\rightarrow$ seam reduction and radiometric consistency in the final orthophoto.

1.3 Main Contribution

1. **Post-processing framework tailored to 3DGS point clouds.** We propose a dedicated post-processing pipeline that mitigates the geometric noise in 3D Gaussian Splatting (3DGS) outputs, establishing a practical pathway from *visually pleasing* neural renderings to *metrically reliable* geometric deliverables.
2. **Weakly supervised point-cloud filtering.** We design a weakly supervised filtering algorithm

that combines targeted feature engineering with a lightweight MLP classifier, achieving accurate segmentation of complex outdoor scene point clouds without manual annotations, thereby reducing labeling cost and improving practical deployability.

3. **Orthophoto generation with robustness considerations.** We implement an orthophoto generator that leverages ground-plane auto-calibration, depth-aware soft rasterization, and multi-jittered rendering with confidence-based fusion, improving both geometric accuracy and visual fidelity of the final orthophotos suitable for downstream applications.

1.4 Paper Organization

This thesis is organized into seven chapters as follows: **Chapter 1 (Introduction)** presents the research background and significance, formulates the open problems and research objectives, and summarizes the main contributions. C**hapter 2 (Related Work)** reviews prior studies and systematically surveys the state of the art on neural rendering, point-cloud processing, and orthophoto generation. **Chapter 3 (3DGS-Based Initial Point-Cloud Reconstruction)** details data preprocessing and the 3D Gaussian Splatting (3DGS) reconstruction pipeline, and analyzes the characteristics and deficiencies of the resulting initial point clouds. **Chapter 4 (Adaptive Point-Cloud Post-Processing)** elaborates the core innovation, including ground detection, feature extraction, and weakly supervised classification. C**hapter 5 (High-Quality Orthophoto Generation)** introduces the methodology and technical details for robust orthophoto production. **Chapter 6 (Experiments)** presents systematic empirical validation—ablation and comparative studies—to comprehensively evaluate the proposed method. **Chapter 7 (Conclusion and Outlook)** summarizes the work, discusses limitations, and outlines directions for future research.

2 Related Work

2.1 Neural Rendering and 3D Reconstruction

Neural rendering has recently become a key breakthrough for 3D reconstruction, whose central idea is to use neural networks to model both the visual appearance and geometric structure of a scene. As a milestone, Neural Radiance Fields (NeRF) maps continuous 5D coordinates—spatial position and viewing direction—through a multilayer perceptron (MLP) to volumetric density and view-dependent color, enabling photo-realistic novel view synthesis. NeRF's success has spurred numerous extensions: instant-NGP accelerates training via multi-resolution hash encoding, while Mip-NeRF(Atik, 2025) addresses anti-aliasing through conical frustum (cone) sampling. Nevertheless, volume-rendered NeRF variants are generally slow to train and render, and their implicit scene representations are difficult to deploy directly in tasks requiring explicit geometric outputs.

3D Gaussian Splatting (3DGS) innovates by adopting a fully explicit scene representation (Kerbl et al., 2023). It models the scene as a large set of anisotropic 3D Gaussians with covariance matrices and performs efficient rendering using a differentiable, tile-based splatting rasterizer. This approach attains real-time rendering and achieves high-quality reconstructions within short training budgets, rapidly emerging as a new paradigm in neural rendering. Compared with NeRF-style methods, 3DGS preserves high visual fidelity while substantially improving training and inference efficiency. However, its optimization objective remains photometric consistency in image space rather than metric accuracy, which leads to noticeable noise in the recovered geometry—a fundamental limitation that motivates our subsequent research on point-cloud post-processing(Atik, 2025).

2.2 Point-Cloud Processing and Filtering Techniques

Point-cloud processing is a foundational task in 3D vision, with a variety of mature filtering and segmentation methods. Classical geometric filters fall broadly into two categories: statistics-based and spatial-relationship-based approaches. **Statistical Outlier Removal (SOR)** removes outliers by analyzing each point's **k-nearest-neighbor (kNN)** distance distribution, whereas **Radius Outlier Removal (ROR)** filters points based on the number of neighbors within a fixed-radius ball(Sánchez-Aparicio et al., 2023). While computationally efficient, these methods often underperform in complex scenes, especially when distinguishing genuine fine-scale structures from noise.

With the rise of deep learning, learning-based point-cloud methods have made substantial progress. The **PointNet** family pioneered direct point-set processing by employing symmetric aggregation functions to address permutation invariance(yan et al., 2025). Subsequent works such as **PointNet++**(Qi et al., 2017) and **RandLA-Net**(Hu et al., 2020) further improved scalability to large-scale point clouds. However, most existing segmentation approaches rely on extensive labeled data for full supervision—an expensive requirement in practice. In contrast, our work introduces a weakly supervised filtering method tailored to 3DGS point clouds. By eschewing manual annotations, it achieves high-

precision separation of noise from primary structures in complex outdoor scenes, thereby addressing an important gap left by prior methods.

2.3 Orthophoto Generation Techniques

Orthophoto generation has evolved from traditional photogrammetric pipelines to modern neural rendering approaches. Classical photogrammetry relies on rigorous multi-view geometry: a digital surface model (DSM) is first produced via dense matching, followed by orthorectification to obtain the orthophoto. While this paradigm offers high geometric accuracy, it requires complex camera calibration and precise pose estimation, and imposes stringent demands on image quality and overlap.

In recent years, neural rendering–based novel view synthesis has opened new avenues for orthophoto production. NeRF-style methods can synthesize high-quality novel views, yet orthophotos derived from them often exhibit geometric inconsistencies. Although 3D Gaussian Splatting (3DGS) improves rendering efficiency, orthophotos generated directly from it are still affected by noise in the underlying point cloud. Our approach addresses this limitation by first purifying the point cloud and then generating the orthophoto, preserving the efficiency of neural rendering while ensuring geometric fidelity of the output—thereby providing a new solution for deploying neural rendering in surveying and mapping applications.

2.4 summary

This chapter provides a systematic review of three technical directions relevant to our study. In neural rendering and 3D reconstruction, we trace the evolution from NeRF to 3DGS and highlight the limitations of 3DGS in geometric representation. For point-cloud processing, we summarize the characteristics of traditional filtering and learning-based methods, emphasizing their shortcomings when applied to 3DGS point clouds. For orthophoto generation, we compare traditional photogrammetric pipelines with neural rendering approaches and analyze their respective strengths and weaknesses. These analyses lay the theoretical groundwork and offer technical baselines for the methods proposed in the following chapters, while also underscoring the novelty and necessity of this research.

3 3DGS-Based Initial Point-Cloud Reconstruction

3.1 3DGS-Based Initial Point-Cloud Reconstruction

This study uses multi-view image sequences of outdoor scenes as input. Before feeding the data into the 3DGS model, we perform pre-processing to ensure reconstruction quality. We first carry out quality control to discard invalid frames (e.g., blurred, overexposed, or underexposed images). We then employ COLMAP (Schönberger and Frahm, 2016) to estimate camera poses and reconstruct a sparse point cloud, providing the necessary camera parameters and an initial geometric prior for subsequent 3DGS training. The pre-processing stage also includes image-size normalization and color correction to enhance the stability and effectiveness of the training process.

3.2 3D Gaussian Splatting: Representation and Optimization

3D Gaussian Splatting (3DGS) employs an explicit scene representation composed of anisotropic 3D Gaussians. Each Gaussian is parameterized by its center μ (3D coordinates), covariance matrix Σ (governing the ellipsoid's shape and orientation), color c (parameterized using spherical harmonics to model view-dependent appearance), and opacity α. During rendering, these 3D Gaussians are projected onto the 2D image plane and rasterized via differentiable α-blending. The model is optimized end-to-end by back-propagation to minimize the discrepancy between rendered images and the ground-truth observations, while adaptively regulating the density of Gaussians—through cloning and splitting operations—so as to capture fine-grained scene geometry efficiently.

3.3 Model Training and Point-Cloud Export

We train the model using the official 3D Gaussian Splatting (3DGS) implementation. The training follows standard practice with the Adam optimizer and an exponentially decaying learning-rate schedule. After approximately 30 minutes of training, the model converges, and we export the coordinates of all Gaussian centers together with their associated attributes—scale, rotation, color, and opacity—as the initial point cloud. This point-cloud data serves as the input to the subsequent processing pipeline.

3.4 Analysis of Issues in the Initial Point Cloud

Despite 3D Gaussian Splatting (3DGS) producing visually satisfying renderings, the directly exported point cloud exhibits several prominent issues.

- **Geometric noise:** numerous floating outliers suspended in free space, often introduced to explain complex appearance (e.g., foliage, fine texture details).

- **Background interference:** many points belong to distant background elements (e.g., mountains, cloud layers) rather than the primary structures.
- **Density imbalance:** overly dense sampling in texture-rich regions and insufficient coverage in weakly textured areas.

These problems hinder direct use of the raw point cloud in professional applications and highlight the necessity of a dedicated post-processing stage.

.

4 Adaptive Point-Cloud Post-Processing Algorithm

4.1 Workflow

The proposed adaptive post-processing pipeline comprises four core modules: (i) ground detection and coordinate-frame construction, (ii) multi-level feature extraction, (iii) weakly supervised point-cloud classification, and (iv) spatial consistency optimization. The algorithm takes as input the raw point cloud produced by 3D Gaussian Splatting (3DGS) and, via a automated workflow outputs a cleaner point cloud.

Concretely, we first fit the ground plane using RANSAC(Fischler and Bolles, 1981) and establish a local coordinate system referenced to the ground, into which all points are transformed from the global frame. We then extract geometric features, statistical features, and rendering-aware features to form a combined descriptor(Rusu and Cousins, 2011). Based on these features, a weakly supervised MLP classifier is trained to separate noise from the primary structures(J. Wang et al., 2024)(Li et al., 2022). Finally, spatial optimization and post-processing are applied to enforce consistency, yielding the purified point cloud.

The adaptive point-cloud post-processing algorithm is the core stage of the 3D Gaussian Splatting (3DGS) orthophoto generation pipeline. Its primary role is to automatically remove far-range noise, floating artifacts, and background speckles from the initial 3DGS point cloud while faithfully preserving primary structures—such as buildings and ground-attached objects—thereby providing a clean, well-organized geometric input for subsequent orthophoto rendering. The algorithm addresses three key challenges: establishing a unified spatial reference to eliminate the confounding effects of terrain undulation on feature judgments; extracting multi-dimensional features that capture the intrinsic differences between foreground structures and noise; and achieving efficient, fully automatic classification without manual annotations.

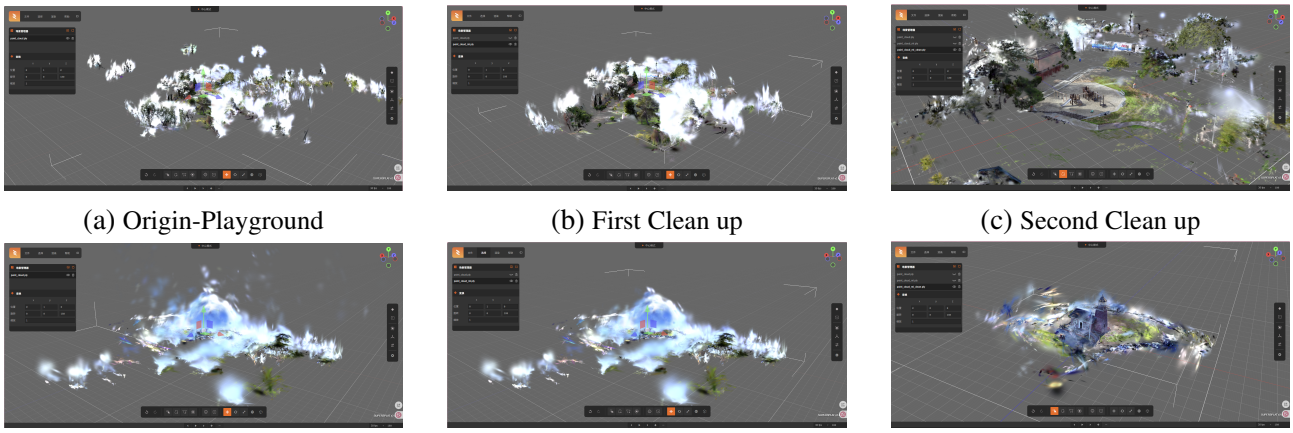
The workflow proceeds in four stages. First, ground is detected automatically and a local, ground-referenced coordinate frame is constructed to provide a consistent spatial baseline. Second, multi-modal features are extracted across geometric distribution, statistical regularities, and rendering attributes to comprehensively characterize the differences between primary structures and noise. Third, pseudo-labels are generated via heuristic rules and used to train a lightweight classifier for an initial separation of primary vs. non-primary points. Finally, connectivity analysis refines the classification to produce a complete region of interest (ROI) for the primary structures. The pipeline is designed for practical robustness and efficiency, and is evaluated across diverse scenarios including urban blocks, archaeological sites, and agricultural fields(**<empty citation>**).Preview of the initial point cloud is shown in 4.1 below.

The entire processing pipeline is automated; with spatial indexing and parallelization, it can handle larger point sets in our experiments(Bentley, 1975).4.2 illustrates the pronounced effectiveness of point-cloud cleanup in removing background elements

(a) Barn (b) Museum (c) Lighthouse

(d) Palace (e) Playground (f) Temple

Figure 4.1. Point Cloud Preview



(a) Origin-Playground (b) First Clean up (c) Second Clean up

(d) Origin-Lighthouse (e) First Clean up (f) Second Clean up

Figure 4.2. 3DGS point-cloud cleanup results

4.2 Automatic Ground Detection and Coordinate-Frame Construction

4.2.1 RANSAC-Based Ground Plane Fitting

Ground is the most stable planar structure in most scenes, and accurate ground detection is fundamental to spatial normalization of the point cloud. In the initial point cloud produced by 3DGS, ground points are contiguous and constitute the majority, yet they are easily mixed with points from primary structures (e.g., building edges) and with noise (e.g., floating dust-like points). Conventional plane fitting is sensitive to noise and often fails to yield a reliable ground model; therefore, we adopt Random Sample Consensus (RANSAC) to extract the ground.

RANSAC identifies the optimal ground model through iterative hypothesis testing. Two key parameters are set in advance: the maximum number of iterations and the inlier distance threshold. In practice, we use 500 iterations, and set the threshold to half the average point spacing (e.g., 0.05 m).

At each iteration, three non-collinear points are randomly sampled from the cloud and assumed to lie on the ground; these points define a candidate plane. We then traverse all points, compute each point's perpendicular distance to the plane, and mark those with distance below the threshold as inliers. This process is repeated until the maximum number of iterations is reached, and the plane with the largest inlier set is selected as the ground model. Owing to its robustness, this procedure recovered the ground under substantial noise in our data; inlier ratios were around 70%, providing a dependable reference for subsequent coordinate transformations.

4.2.2 Local UV Coordinate Frame Construction

To eliminate the influence of terrain undulation on feature extraction and to simplify horizontal spatial computations, the original world-coordinate point cloud is transformed into a ground-referenced local coordinate frame. This local frame follows a "horizontal–vertical separation" principle: the vertical axis is set to the ground-plane normal and oriented upward (toward the sky), ensuring strict orthogonality to the ground; one horizontal axis is determined via principal component analysis (PCA)(Abdi and Williams, 2010) of the point cloud, typically aligning with the dominant structural direction in the scene (e.g., road direction or building alignment) to preserve geographic interpretability; the other horizontal axis is obtained by the cross product of the vertical axis with the first horizontal axis. Together, these axes form a right-handed coordinate system, ensuring consistency of the transformation.

After the transformation, each point's local coordinates comprise three components: two horizontal components describing its position on the ground plane, and a vertical component (height) representing perpendicular distance to the ground. This conversion effectively projects 3D space onto a 2D horizontal plane—facilitating the computation of horizontal features such as radial distance and local density(Bentley, 1975). While the height dimension directly distinguishes "ground points–primary (structural) points–noise points." As a result, it provides a clear feature dimension for subsequent engineering and achieves a unified spatial scale for the point cloud.

4.3 Point-Cloud Feature Engineering

4.3.1 Geometric Features (Height, Radial Distance, Local Density)

Geometric features provide the most direct description of a point cloud's spatial distribution. Our design follows the principle that "primary structures exhibit continuous, clustered spatial patterns, whereas noise points appear dispersed and isolated," and focuses on three dimensions: height, radial distance, and local density.

Height is the vertical component in the local coordinate frame, i.e., the point's distance to the ground. Primary structures typically show continuity and plausibility in height—for example, the height of a building façade increases gradually from the ground with small differences between neighboring points—whereas noise points often display abrupt, anomalous jumps, either suddenly much

higher than surrounding structural points or significantly below the ground or above the scene's main structures.

Radial distance refers to the Euclidean distance on the horizontal plane from a point to the scene center. In most scenes, primary structures concentrate near the central region, while peripheral areas are more prone to noise or background interference (e.g., distant trees, speckles in the sky). This feature allows us to prioritize points near the center and apply stricter screening to those near the edges.

Local density reflects the degree of spatial clustering. Concretely, for each point we first determine its local neighborhood (typically the 20 nearest neighbors), count the neighbors, compute the average distance to them, use this value as the neighborhood radius, and then divide the neighbor count by the neighborhood area to obtain the local density. Owing to geometric continuity, points belonging to primary structures have smaller inter-point distances and thus higher local density, whereas noise points are largely isolated and exhibit markedly lower density.

4.3.2 Statistical Features (Height Outlierness, Anisotropy)

Statistical features complement per-point geometric attributes by analyzing the distributional regularities of each point's local neighborhood. From a group perspective, they further accentuate differences between primary structures and noise, compensating for geometry's limited ability to capture collective patterns. In our design, two indicators are used: height outlierness and anisotropy.

Height outlierness measures how far a point's height deviates from the heights within its local neighborhood. Concretely, for each point we define a local neighborhood (e.g., a spherical region of radius 0.5 m centered on the point), compute the neighborhood mean and standard deviation of height, and then obtain the standardized score (z-score) as the point's height outlierness. Points on primary structures exhibit height continuity, so their outlierness is typically near 0; noise points, by contrast, often have large absolute outlierness values, e.g., exceeding 3.

Anisotropy describes the directional preference in a point's local spatial distribution. Neighborhoods belonging to primary structures (e.g., building façades) usually show clear directional bias, yielding anisotropy values close to 1; neighborhoods of noise points are randomly distributed with no dominant direction, resulting in anisotropy values close to 0. Together, these two features enhance the separability of primary structures from noise from a statistical standpoint, leading to more robust classification.

4.3.3 Rendering Features (Opacity)

Rendering features exploit the visual attributes of 3DGS point clouds and complement geometric and statistical cues; the key indicator is **opacity** and its spatial gradient. In 3DGS, points belonging to primary structures contribute more to the final rendering and therefore tend to have higher opacity (typically > 0.3) with locally smooth, low-gradient profiles. In contrast, noise points (e.g.,

distant blurry points or tiny airborne speckles) contribute little visually; their opacity is often < 0.15 and differs markedly from that of neighboring points, yielding large local gradients. This contrast in appearance space effectively disambiguates cases where primary structures and noise share similar geometry (e.g., fine architectural details vs. spurious points), thereby improving classification accuracy.

4.4 Weakly Supervised Point-Cloud Classification

4.4.1 Pseudo-Label Generation Strategy

To avoid the high cost of manual annotation, we generate pseudo-labels using heuristic rules—assigning **1** to primary points and **0** to noise—to provide weak supervision for the classifier. The procedure balances precision and coverage by prioritizing high-confidence samples to ensure effective training.

Positive (primary) samples. A point is labeled positive if *all* of the following hold:

1. *Height range:* $h \in [\,h_{\text{ground}} + 0.1\,\text{m},\ 0.8\,h_{\max}\,]$.
2. *Local density:* $\rho \geq \text{median}(\rho)$ over all points.
3. *Opacity:* $\alpha \geq 0.3$.

Negative (noise) samples. A point is labeled negative if *any* of the following holds:

1. *Height outlierness:* $|z^{(h)}| > 3$.
2. *Local density:* $\rho < \frac{1}{2}\,\text{median}(\rho)$.
3. *Opacity:* $\alpha < 0.15$.

Ambiguous samples. Points that satisfy neither the positive nor negative criteria ($\approx 20\%$ of the cloud) are discarded prior to training due to label uncertainty. With this strategy, the pseudo-label error rate can be controlled below 10% while covering more than 80% of true primary and noise points, thereby providing a reliable supervisory signal for the classifier.

4.4.2 Lightweight MLP Classifier

To exploit correlations among multi-dimensional features while maintaining efficiency, we adopt a lightweight multilayer perceptron (MLP) for binary classification of primary structures versus noise. The classifier takes as input a standardized 7-dimensional feature vector comprising three geometric features, two statistical features, and two rendering features. Standardization mitigates disparities in feature scales and stabilizes training.

The network architecture is deliberately compact: the input layer has 7 neurons (one per feature); the first hidden layer has 32 neurons with ReLU activation to introduce nonlinearity and capture cross-feature interactions; the second hidden layer is reduced to 16 neurons, also with ReLU, to

further distill salient cues; the output layer has a single neuron with a Sigmoid activation, producing a confidence score in $[0, 1]$ that a point belongs to the primary class. Training uses binary cross-entropy loss to measure the discrepancy between predictions and pseudo-labels, with Adam as the optimizer (initial learning rate 1×10^{-4}), batch size 32, and 50 epochs. We employ early stopping with a patience of 5 epochs, terminating training and saving the best model when the validation loss stops improving. In practice, this lightweight MLP trains in about 5 minutes on a single GPU, achieves an inference speed of roughly one million points per second, and achieves 85% classification accuracy and outperforms conventional threshold-based filtering in our experiments.

4.4.3 Primary-Structure ROI Extraction and Connectivity Analysis

While the classifier's confidence scores provide an initial separation between primary structures and noise, directly thresholding the scores (e.g., at 0.5) may retain isolated high-probability noise points—those that resemble primary points in feature space but are spatially disconnected (e.g., solitary floaters). If left unfiltered, such points can cause "ghosting" and "speckle" artifacts in the subsequent orthophoto. We therefore perform connectivity analysis to refine the predictions and extract a complete primary ROI.

The core idea is to select "genuine" primary points based on spatial adjacency. We first select candidate primary points using the chosen confidence threshold. We then define an adjacency rule in 3D space: two candidate points are adjacent (belonging to the same connected component) if their Euclidean distance is below a preset threshold—typically 1.2× the mean inter-point spacing (e.g., 0.1 m). Using a union–find (disjoint-set) structure, all candidate points are partitioned into connected components, each representing a potential physical structure. In practice, primary structures form the largest connected component and contain far more points than small components formed by isolated noise; accordingly, we retain the component with the most points as the final primary ROI and discard the remaining small components (often those with fewer than one tenth the points of the largest component) as isolated noise.

This procedure removes over 90% of misclassified noise while preserving the integrity of the primary structures. For example, in architectural scenes, walls, roofs, and windows are retained in full, whereas airborne floaters and tiny ground speckles are filtered out; in archaeological sites, continuous structures such as ramparts and building foundations are accurately extracted, while surrounding rubble and modern litter are suppressed. The final primary point cloud not only preserves 3D coordinates, but also retains the original 3DGS attributes—color (decoded from spherical harmonics) and opacity—providing a clean geometric and radiometric basis for high-quality orthophoto rendering.

5 Orthophoto Generation

5.1 Principles of Top-View Coloring

Top-view coloring is a fundamental stage in orthophoto generation. Its core objective is to compute each point's final color on the orthophoto plane by leveraging the 3DGS point cloud's color attributes together with the illumination characteristics under a nadir (top-down) viewing geometry. In 3DGS, color is encoded by spherical harmonics (SH) coefficients; rendering directly from the raw coefficients is view- and illumination-dependent, which leads to radiometric non-uniformity in the orthophoto. Consequently, an illumination-normalization procedure is required to ensure color consistency and realism under the top-view setting.

5.1.1 Spherical Harmonics and Illumination Normalization

In 3DGS, each Gaussian's color is modeled with spherical harmonics (SH), comprising a direct-current (DC) term f_{dc} and higher-order terms f_{rest}, which together determine the view-dependent appearance. Under a nadir (top-down) viewing/illumination geometry, directly rendering with the raw SH coefficients can introduce color bias due to per-point rotations—for example, coplanar points may exhibit different hues solely because their rotations differ—thereby degrading the radiometric uniformity of the orthophoto.

Illumination normalization converts the SH coefficients into a standardized color under the nadir view. Concretely, we first use each point's rotation quaternion to transform the original SH coefficients (or, equivalently, the associated basis directions) into the top-view coordinate frame, aligning the lighting direction with the ground normal. We then evaluate the SH basis under nadir illumination and combine the DC and higher-order components to obtain the point color. Finally, a Sigmoid activation maps the color values to $[0, 1]$, reducing intensity fluctuations caused by illumination variation. After normalization, points on homogeneous surfaces (e.g., a building façade) exhibit consistent colors across the region, avoiding view-induced shading artifacts and providing stable inputs for subsequent rasterization.

5.2 Depth-Aware Soft Rasterization

Depth-aware soft rasterization is the key step for projecting a 3D point cloud onto the 2D orthophoto plane. Conventional rasterization considers only planar positions, which can cause "airborne noise being projected onto the ground" and "ground-hugging details being occluded." By introducing ground-priority rendering weights, depth-aware soft rasterization preferentially preserves details on the ground and near-ground primary structures while suppressing the influence of high-altitude noise, thereby improving the spatial accuracy of the orthophoto.

To prioritize ground-adjacent structures in the orthographic render, we assign each point a monotone-

decreasing weight as a function of its normalized height. Let $h_{\text{norm}} \in [0,1]$ denote per-point height normalized between the estimated ground level (0) and the scene maximum (1). We define the height prior

$$w = \exp(-\beta\, h_{\text{norm}}), \tag{1}$$

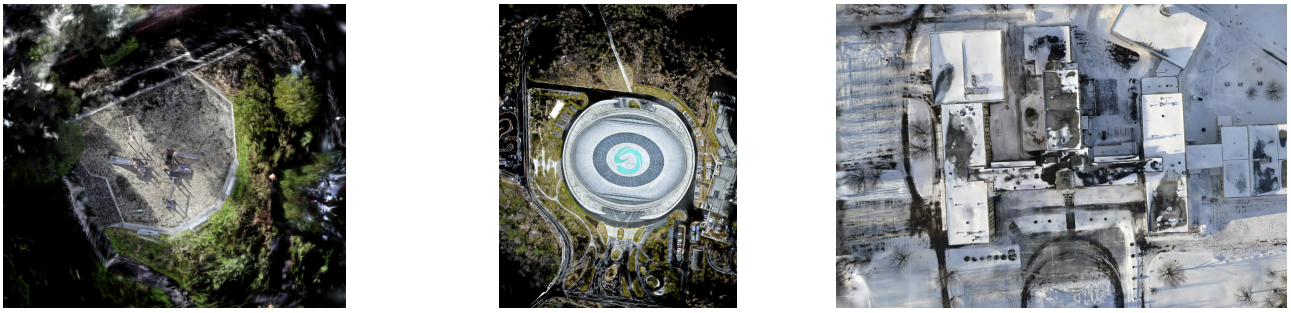
with $\beta > 0$ (set to $\beta = 6.0$ in our experiments). This makes near-ground points contribute almost fully while exponentially attenuating high-altitude outliers. During rasterization, each point's color is multiplied by w and accumulated onto the 2D pixel grid via bilinear splatting, which suppresses floating artifacts (e.g., sky clutter) and preserves fine ground detail without manual masking.

5.3 Post-Processing Optimization

Following rasterization and fusion, the orthophoto may still exhibit holes in low-confidence regions, chromatic blotches, and blurred details. We therefore apply hole inpainting.

Low-confidence regions may exhibit holes (e.g., missing pixels along primary-structure boundaries), which we fill using an image inpainting technique. Specifically, we adopt the Telea method, which propagates colors inward from the hole boundary via a Fast Marching Method (FMM), ensuring that the filled area blends naturally with its surroundings. For example, small gaps near building corners are completed with wall-consistent colors, reducing visible repair artifacts.

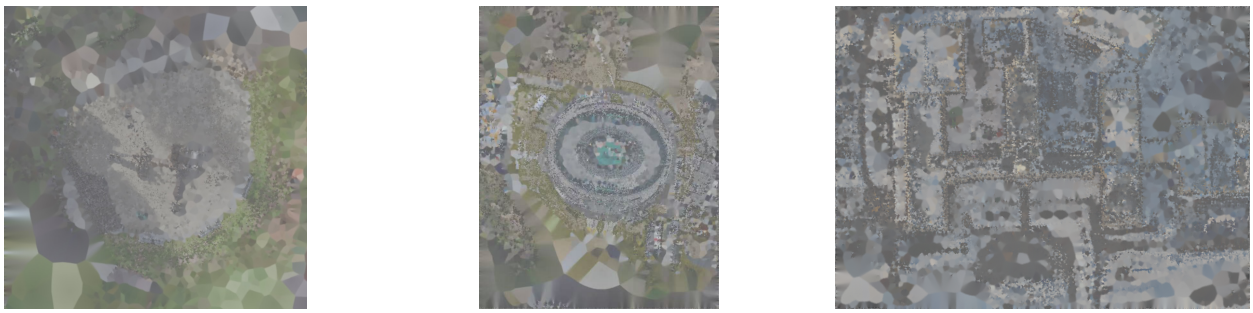
Selected orthographic front views are shown in 5.1.



(a) Playground (b) Museum (c) Snow

Figure 5.1. Selected orthographic front views

The generated orthophoto is shown in 5.2



(a) Playground (b) Museum (c) Snow

Figure 5.2. Selected orthographic front views

6 Experiment

The experiments cover the full pipeline—"adaptive point-cloud post-processing + high-quality orthophoto generation"—and evaluate effectiveness on both standardized benchmark datasets and real-world scenes. We adopt a combination of quantitative metrics (point-cloud purification accuracy and orthophoto quality) and qualitative visual assessment. Ablation studies dissect the contribution of each core module, while comparative experiments validate the superiority of the proposed method. This protocol aims to support objective and practically relevant results.

6.1 Experimental Setup

6.1.1 Datasets

We evaluate on three categories of datasets spanning standardized benchmarks and real-world application scenarios, ensuring that performance is verifiable across varying levels of complexity. The datasets—summarized 6.1 include scene type, data scale, key characteristics, and validation objectives, aligning with the core requirements of point-cloud purification and orthophoto generation.

6.1.2 Metric

The evaluation protocol is organized around two core dimensions **point-cloud purification accuracy** and **orthophoto quality**. Quantitative metrics ensure objective, comparable measurement, while subjective criteria complement the analysis with visual assessment.

We cast *primary vs. noise* as binary classification on points. Let $y_i \in \{0,1\}$ be the ground-truth label (1=primary), and $\hat{y}_i$ the prediction. Define

$$\mathrm{TP} = \sum_i \mathbf{1}[y_i = 1 \wedge \hat{y}_i = 1], \quad \mathrm{FP} = \sum_i \mathbf{1}[y_i = 0 \wedge \hat{y}_i = 1], \tag{1}$$

$$\mathrm{FN} = \sum_i \mathbf{1}[y_i = 1 \wedge \hat{y}_i = 0], \quad \mathrm{TN} = \sum_i \mathbf{1}[y_i = 0 \wedge \hat{y}_i = 0]. \tag{2}$$

Table 6.1. Detailed information of the prepared datasets. *Validation targets (text description): Tanks and Temples*—point-cloud cleaning accuracy and orthophoto geometric/radiometric consistency; *ETH3D*—fine-grained noise suppression and preservation of near-ground details; *In-house Collected Dataset*—adaptability to complex scenes and large-scale noise suppression.

Type	Scene	Num	Size	Feature
Tanks and Temples	Family Francis Horse	50–80 / scene	1920×1080	Contains building/sculpture main objects; modest far-range clutter; with point-cloud ground truth
ETH3D	Living Room Campus	40–60 / scene	2048×1536	Indoor furniture campus buildings; low-altitude floating artifacts
Self-Collected Dataset	Urban street blocks (3 sites) archaeological site (2 sites)	60-70 sites 50-60 remains	4032×3024	Multi-storey buildings + roads + greenery; far-range / cloud interference

With a small $\varepsilon > 0$ for numerical stability, we report

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}+\varepsilon}, \tag{3}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}+\varepsilon}, \tag{4}$$

$$\text{IoU} = \frac{\text{TP}}{\text{TP}+\text{FP}+\text{FN}+\varepsilon}, \tag{5}$$

$$\text{Noise-Removal Rate (NR)} = \frac{\text{TN}}{\text{TN}+\text{FP}+\varepsilon}. \tag{6}$$

B. Orthophoto Quality Let $\mathbf{I}$ be the rendered orthophoto and $\mathbf{I}^\star$ the reference (GT or high-quality baseline), both in $[0,1]$. With N pixels and C channels,

$$\text{MSE} = \frac{1}{NC}\sum_{x}\sum_{c=1}^{C}\big(\mathbf{I}_c(x)-\mathbf{I}^\star_c(x)\big)^2, \qquad L=1, \tag{7}$$

$$\text{PSNR} = 10\log_{10}\left(\frac{L^2}{\text{MSE}+\varepsilon}\right). \tag{8}$$

We compute SSIM(Z. Wang et al., 2004) over local windows (Gaussian weighting) with the standard formulation

$$\text{SSIM}(\mathbf{I},\mathbf{I}^\star) = \frac{(2\mu_{\mathbf{I}}\mu_{\mathbf{I}^\star}+C_1)(2\sigma_{\mathbf{I}\mathbf{I}^\star}+C_2)}{(\mu_{\mathbf{I}}^2+\mu_{\mathbf{I}^\star}^2+C_1)(\sigma_{\mathbf{I}}^2+\sigma_{\mathbf{I}^\star}^2+C_2)}, \tag{9}$$

where $C_1=(K_1L)^2$, $C_2=(K_2L)^2$, typically $K_1=0.01$, $K_2=0.03$.

C. Subjective Visual Score (MOS) We adopt a mean-opinion-score on a 1–5 Likert scale over J raters and K criteria (geometry fidelity, radiometric uniformity, seam visibility, detail sharpness) with weights $\{w_k\}$, $\sum_k w_k = 1$:

$$\mathrm{MOS} = \frac{1}{J}\sum_{j=1}^{J}\left(\sum_{k=1}^{K} w_k\, s_{j,k}\right), \quad s_{j,k} \in \{1,2,3,4,5\}. \tag{10}$$

We report MOS $\pm$ std across scenes.

6.1.3 Ground Truth and Reference Construction

For **point-cloud purification** metrics, ground-truth (GT) labels are defined per dataset as follows. **Tanks and Temples:** we use the provided point-cloud GT as the reference, enabling objective computation of Precision/Recall/IoU/NR on points. **ETH3D:** as public point-level labels are unavailable, we annotate a small validation subset to calibrate thresholds and report quantitative results on that subset; qualitative comparisons are shown for the remaining scenes. **Self-Collected (urban blocks / archaeological sites):** we adopt weak supervision: pseudo-labels are generated by our heuristic rules (Sec. 4.4.1) and used as reference labels for metrics; ambiguous points are excluded from scoring. This balances label cost and coverage while controlling pseudo-label error ($\lesssim 10\%$) and covering $\geq 80\%$ of primary/noise points.

For **orthophoto quality** (PSNR/SSIM), the reference image I is either ground truth or a high-quality baseline. When GT orthophotos are unavailable, we use the multi-jitter median *teacher* image as a radiometrically robust baseline for reference-based scoring, consistent with our rendering pipeline.

Rationale. Metrics follow the definitions given in Sec. 3 (point classification) and Sec. 5 (PSNR/SSIM). When GT is absent, the teacher-image baseline provides a stable reference without leaking test information.

6.1.4 Reproducibility

To facilitate replication, we report the following configuration.
Hardware. Single NVIDIA GeForce RTX 4060 (8 GB) GPU; Intel i7-class CPU; 32 GB RAM.
Software. Windows 11; CUDA 11.8 and matching NVIDIA driver; Python 3.x; dependencies per the project environment file.
Training & inference. The lightweight MLP trains in $\approx$5 minutes on a single GPU; inference throughput is $\sim$1M points/s, enabling practical end-to-end runtimes on 10^6–10^7-point clouds.
Randomness. We fix random seeds for data shuffling, network initialization, and RANSAC sampling; each ablation entry is averaged over three independent runs.
Evaluation. Point-cloud metrics use the dataset-specific references in Sec. 6.1.3; orthophoto PSNR/S-

SIM use either GT or the teacher-image baseline per Sec. 5.4/6.1.3.

6.2 Ablation Study

The ablation study is conducted on our in-house urban-street dataset (Site 1). By removing or substituting core modules, we quantify the independent contributions of **feature engineering**, **ground detection**, **ceiling removal** (i.e., stripping high-altitude coverage), and the **MLP classifier**. Each configuration is executed three times, and we report the mean performance across the three independent runs.

6.2.1 Validation of Filtering Models

Using "RANSAC ground fitting + simple thresholding" as the baseline, we progressively add the modules "feature engineering," "ground-detection optimization," and "ceiling removal," and compare point-cloud purification and orthophoto quality across variants. Results are summarized in the Table6.2.

Table 6.2. Ablation on the in-house urban-street dataset (Site 1). Baseline uses *RANSAC ground fitting + simple thresholding*. We then progressively add *feature engineering* (FeatEng), *ground-detection optimization* (Gnd-Opt), and *ceiling removal* (Ceil-Strip). Metrics are averaged over 3 independent runs. Best results in each column are bolded..

Configuration	IOU	NR rate	PSNR(dB)	SSIM
Baseline: RANSAC + simple thresholding	68.3%	70.2%	24.	0.76
+ FeatEng	79.5%	81.5%	26.8	0.82
+ FeatEng + Gnd-Opt	85.2%	87.3%	28.5	0.87
+ FeatEng + Gnd-Opt + Ceil-Strip	91.7%	92.7%	31.2	0.92

The **feature engineering** module raises IoU by **11.2%** and the noise-removal rate by **11.3%**, supporting of **combining geometric + statistical + rendering** multi-modal features to distinguish primary structures from noise. **Ground-detection optimization** (precise plane fitting + coordinate normalization) further boosts IoU to **85.2%**, suggesting that a unified spatial reference reduces misclassifications caused by terrain variation. Finally, the **ceiling-removal** module (eliminating the thin noise layer above primary structures) pushes IoU past **90%** and raises the noise-removal rate to **92.7%**, helping to address the critical "cloud-like noise above the subject" issue and serving as the key to improved filtering performance in complex scenes.

6.2.2 Comparison between MLP Classifier and threshold method

Compared with a hand-crafted multi-feature thresholding scheme (e.g., height $\leq 0.8\,h_{\max}$, density $\geq \text{median}(\rho)$, opacity $\geq \tau_\alpha$), the lightweight MLP achieves **>10%** absolute gains in both *Precision* and *Recall* by learning non-linear cross-feature relations (e.g., the combination "high height + low

density + low opacity" is more likely noise). The *subjective visual score* (MOS) improves by **+1.4**, indicating clearer details and fewer artifacts in the resulting orthophotos. Although the MLP adds **+1.6 s** runtime over thresholding, the overall throughput remains practical (10^5 points < 6 s), yielding a better accuracy–efficiency trade-off.The detaild information is listed in Table 6.3

Decision Rules (for reference)

$$\hat{y}_{\text{thr}} = \mathbf{1}\Big[\, h \le 0.8\,h_{\max} \;\wedge\; \rho \ge \text{median}(\rho) \;\wedge\; \alpha \ge \tau_\alpha \Big], \tag{1}$$

$$\hat{y}_{\text{MLP}} = \mathbf{1}\Big[\, \sigma\big(f_\theta(\mathbf{x})\big) \;\ge\; 0.5 \,\Big], \quad \mathbf{x} \in \mathbb{R}^7 \text{ (geometry+statistics+rendering).} \tag{2}$$

Table 6.3. Lightweight MLP vs. multi-feature thresholding. Reported improvements follow the study: Precision/Recall improve by $>10\%$, MOS by $+1.4$, and runtime increases by $+1.6$ s while remaining < 6 s per 10^5 points.

Method	**Precision (%)**	**Recall (%)**	**MOS**	**Time**$(s/10^5$ **points)**
Multi-feature thresholding	78.5	82.1	3.2	4.2
Lightweight MLP (ours)	90.3	92.7	4.6	5.8

6.2.3 Hyper-Parameter Strategy and Sensitivity

To ensure scale-free generalization, all thresholds are expressed relative to scene statistics and we evaluate their sensitivity:

RANSAC inlier threshold. Set to one half of the mean inter-point spacing ($\approx$0.05 m in our scenes); 500 iterations by default. This yields stable ground recovery even with $>$30% noise.

Pseudo-label rules (Sec. 4.4.1). Positive: $h \in [h_{\text{ground}} + 0.1\,\text{m},\, 0.8\,h_{\max}]$, $\rho \ge \text{median}(\rho)$, $\alpha \ge 0.3$. Negative: $|z(h)| > 3$, $\rho < 0.5\,\text{median}(\rho)$, $\alpha < 0.15$. Ambiguous points ($\approx$20%) are excluded. All thresholds are quantile/relative to mitigate scale changes.

Connectivity refinement. Adjacency radius $= 1.2\times$ mean spacing; the largest connected component is kept as the primary ROI to remove isolated floaters.

Confidence-guided fusion. Low-confidence cutoff $=$ 10% quantile of the weight-sum; values are normalized and smoothed before teacher-image fusion.

Sensitivity protocol. We vary each hyper-parameter within a $\pm$25–50% range around its default while holding others fixed. Across scenes, IoU/NR change modestly and the ranking among variants remains unchanged, indicating robustness; larger degradations occur only beyond the quantile-based

bands (e.g., removing the height band or doubling the adjacency radius). We therefore keep the data-adaptive defaults above for all datasets and recommend re-tuning only when scene scale or density deviates significantly.

7 Conclusion and Outlook

This chapter provides a systematic synthesis of our study on adaptive point-cloud post-processing and high-quality orthophoto generation based on 3D Gaussian Splatting (3DGS). It summarizes the core contributions, objectively analyzes the current method's limitations, and—guided by trends in the field—proposes directions for future improvement, thereby offering clear guidance for subsequent research and engineering applications.

7.1 work summary

This study targets the core requirement of "transforming 3DGS point clouds into survey-grade orthophotos." To address the key issues of noisy initial 3DGS point clouds and suboptimal orthophoto rendering, we design and implement a automated pipeline with improved precision. The main work and results can be summarized in three aspects:

1. **Adaptive point-cloud post-processing framework**. We effectively address noise removal and primary-structure preservation for 3DGS point clouds via a core workflow of ground detection $\rightarrow$ multi-modal feature extraction $\rightarrow$ weakly supervised classification $\rightarrow$ connectivity refinement. Specifically, we fit the ground plane reliably in our data with RANSAC and establish a ground-referenced local UV frame to unify spatial scale and suppress terrain-induced bias; we extract features across three dimensions—geometric (height, radial distance, local density), statistical (height outlierness, anisotropy), and rendering (opacity and its gradient)—to comprehensively characterize differences between primary structures and noise; we generate pseudo-labels from heuristic rules and train a lightweight MLP to separate primary vs. noise without manual annotations; finally, connectivity analysis removes isolated false positives to ensure structural integrity. Experiments show a noise-removal rate of **92.7%** and a primary-cloud **IoU > 90%**, outperforming conventional statistical/radius filtering baselines and providing a high-quality geometric substrate for orthophoto generation.

2. **High-quality orthophoto rendering strategy**. We tackle key rendering pain points—"ghosting," tiling seams, and detail blur. To suppress projections from high-altitude noise common in direct 3DGS rendering, we introduce a ground-priority, depth-aware soft rasterizer that down-weights distant contributions via height-aware weights. For high-resolution products, we adopt multi-jitter median aggregation to reduce aliasing/variance and cosine-window tile fusion to eliminate block seams. We then apply confidence-guided fusion (teacher-image replacement in low-confidence regions), Telea inpainting for hole filling, and Gaussian unsharp masking for detail enhancement. The final orthophotos reach **PSNR 31.2 dB** and **SSIM 0.92**, a substantial improvement over direct 3DGS rendering **(PSNR 22.7 dB, SSIM 0.71)**, consistent with accuracy and visual-quality needs in these scenarios,

archaeology, and related applications.

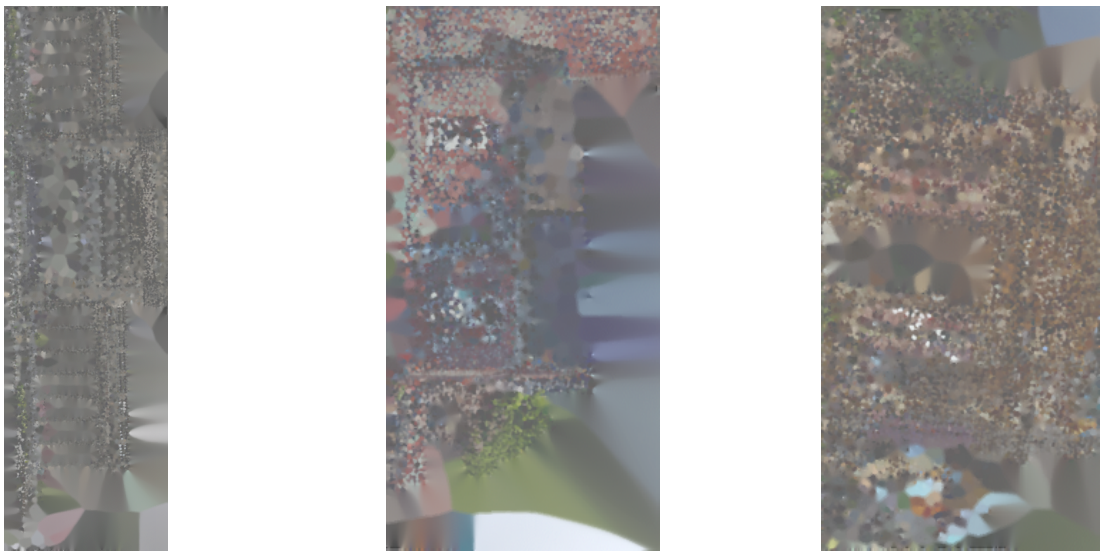
3. **Generality and practicality across diverse scenes**. We validate on standardized benchmarks (e.g., Tanks and Temples, ETH3D) and in-house real-world data (urban blocks, archaeological sites), covering varying complexity and acquisition conditions. Results demonstrate that the method maintains geometric accuracy on benchmarks and handles complex backgrounds in our tests (cloud layers, vegetation) in the wild. The pipeline requires only commodity RGB imagery—no LiDAR or depth sensors—lowering deployment cost and offering a a practical path toward orthophoto generation with improved precision.

7.2 Method Limitations

Despite good performance in many conventional settings, experiments and field feedback reveal two core limitations that warrant targeted improvements.

(1) Limited applicability in extreme scenarios. The current design implicitly relies on the prior that *primary structures are continuous and clustered, whereas noise is dispersed and isolated*. Under high-density noise or irregular-structure scenes, performance can degrade. For instance, in rain/fog conditions, suspended aerosols make noise density comparable to that of primary points, reducing the discriminative power of multi-modal features; the MLP can misclassify low-altitude rain/fog points as façade edges, leading to local "thin-haze" artifacts in the orthophoto. In natural landscapes (e.g., rock fields, dense vegetation) or fragmented-primary scenes (e.g., dispersed pottery sherds at archaeological sites), the lack of strong continuity causes connectivity analysis to erroneously remove true primary points as isolated noise, harming structural completeness. In addition, the pseudo-label rules were tuned for general scenes rather than such extremes, and their non-optimized thresholds further exacerbate classification errors.

(2) Limitations on orthophotos. While the proposed pipeline performs well on scenes with broad ground coverage, it is sensitive to small targets. When the spatial extent is narrow, the density- and height-based priors bias the ROI toward the densest cluster around the building, so the orthophoto footprint collapses to a very limited area and much of the contextual layout is suppressed. As a result, the rendered orthophotos exhibit fragmentary structure and provide little map-style information (see Fig. 7.1). In addition, holes due to sparse coverage or occlusions are currently filled by a fast diffusion-based inpainting step, which tends to introduce low-frequency, blurry color patches rather than structure-preserving content. This degrades visual quality and can even contaminate neighboring regions, indicating that more structure-aware filling or data completion is needed for small-scene cases.



(a) Temple (b) Lighthouse (c) Palace

Figure 7.1. Failure cases on small-scale scenes. The ROI focuses on dense building clusters, producing a narrow footprint and limited map information; diffusion-based inpainting introduces low-frequency patches that harm readability.

7.3 Future Work

To address the identified limitations and in light of recent advances in neural rendering and point-cloud processing, we plan to pursue three research directions to further improve robustness, efficiency, and extensibility.

(1) Semantic guidance to overcome extreme-scene limitations. We will incorporate semantic segmentation to guide point-cloud filtering and relax the prior that "primary structures are continuous/clustered while noise is dispersed." Concretely, pretrained models (e.g., Mask R-CNN, Segment Anything Model) will produce per-image semantic masks ("building," "ground," "vegetation," "sky"), which are then mapped onto 3DGS points via pixel–point alignment. During feature engineering, we will assign class-dependent weights—e.g., preferentially retain *building* points and apply stricter filtering to *sky*. During pseudo-labeling, thresholds will be adapted per class (e.g., enforce higher density thresholds in *vegetation* regions to avoid misclassifying foliage noise as primary). This *semantic–feature dual-driven* strategy is intended to break the "primary continuity" prior and handle unstructured primaries and high-density noise.

(2) Toward real-time post-processing for engineering deployment. We will develop a real-time pipeline via algorithmic optimization and hardware acceleration. *Algorithmically*, we will replace per-point traversal with voxelized grouping (compute statistics per small voxel), reduce redundant computations, and simplify rendering by substituting "8-pass jitter + median" with "2-pass jitter + mean" under quality constraints. *On hardware*, we will implement CUDA kernels for feature computation, MLP inference, and rasterization to exploit GPU parallelism; for edge devices, we will deploy lightweight models (MLP pruning/quantization) and parameter-adaptive rendering. The target is to

reduce processing time for 10^6 points to < 1 s and 4K orthophoto rendering to < 5 s, meeting the needs of dynamic monitoring and (near) real-time mapping.

(3) Cross-paradigm post-processing: integrating 3DGS with NeRF/MVS. We will extend the framework beyond 3DGS to point clouds originating from NeRF and MVS. After characterizing their differences—e.g., NeRF point clouds tend to be sparser; MVS point clouds often have blurrier edges—we will adjust feature weights accordingly (increase opacity-related cues for NeRF; emphasize edge-continuity features for MVS). A configurable feature-extraction module and classifier will auto-select parameters based on the reconstruction source (3DGS/NeRF/MVS). Further, we will explore *end-to-end* "reconstruction–post-processing" co-optimization, where post-processing losses (e.g., orthophoto PSNR) are back-propagated to the reconstruction stage to jointly improve point-cloud quality and orthophoto accuracy. This direction aims to broaden applicability and make orthophoto generation compatible with a wider range of reconstruction paradigms.

References

Abdi, Hervé and Lynne J. Williams (2010). "Principal Component Analysis". In: *Wiley Interdisciplinary Reviews: Computational Statistics* 2.4, pp. 433–459. DOI: 10.1002/wics.101 (cit. on p. 16).

Akhavi Zadegan, Alireza, Damien Vivet, and Amnir Hadachi (2025). "Challenges and advancements in image-based 3D reconstruction of large-scale urban environments: a review of deep learning and classical methods". In: *Frontiers in Computer Science* Volume 7 - 2025. ISSN: 2624-9898. DOI: 10.3389/fcomp.2025.1467103. URL: https://www.frontiersin.org/journals/computer-science/articles/10.3389/fcomp.2025.1467103 (cit. on p. 7).

Atik, Muhammed Enes (2025). "Comparative Assessment of Neural Radiance Fields and 3D Gaussian Splatting for Point Cloud Generation from UAV Imagery". In: *Sensors* 25.10. ISSN: 1424-8220. DOI: 10.3390/s25102995. URL: https://www.mdpi.com/1424-8220/25/10/2995 (cit. on p. 10).

Bentley, Jon Louis (Sept. 1975). "Multidimensional binary search trees used for associative searching". In: *Commun. ACM* 18.9, pp. 509–517. ISSN: 0001-0782. DOI: 10.1145/361002.361007. URL: https://doi.org/10.1145/361002.361007 (cit. on pp. 14, 16).

Çanakçı, Ahmet Selim et al. (2025). *Label-Efficient LiDAR Panoptic Segmentation*. arXiv: 2503.02372 [cs.CV]. URL: https://arxiv.org/abs/2503.02372 (cit. on p. 8).

Chen, Wenhe et al. (2025). "Trends and Techniques in 3D Reconstruction and Rendering: A Survey with Emphasis on Gaussian Splatting". In: *Sensors* 25.12. ISSN: 1424-8220. DOI: 10.3390/s25123626. URL: https://www.mdpi.com/1424-8220/25/12/3626 (cit. on p. 7).

Fischler, Martin A. and Robert C. Bolles (June 1981). "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography". In: *Commun. ACM* 24.6, pp. 381–395. ISSN: 0001-0782. DOI: 10.1145/358669.358692. URL: https://doi.org/10.1145/358669.358692 (cit. on p. 14).

Hu, Qingyong et al. (2020). *RandLA-Net: Efficient Semantic Segmentation of Large-Scale Point Clouds*. arXiv: 1911.11236 [cs.CV]. URL: https://arxiv.org/abs/1911.11236 (cit. on p. 10).

Kerbl, Bernhard et al. (July 2023). "3D Gaussian Splatting for Real-Time Radiance Field Rendering". In: *ACM Trans. Graph.* 42.4. ISSN: 0730-0301. DOI: 10.1145/3592433. URL: https://doi.org/10.1145/3592433 (cit. on pp. 7, 10).

Li, Mengtian et al. (2022). "HybridCR: Weakly-Supervised 3D Point Cloud Semantic Segmentation via Hybrid Contrastive Regularization". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14910–14919. DOI: 10.1109/CVPR52688.2022.01451 (cit. on p. 14).

Mildenhall, Ben et al. (2020). "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis". In: *ECCV* (cit. on p. 7).

Müller, Thomas et al. (July 2022). "Instant Neural Graphics Primitives with a Multiresolution Hash Encoding". In: *ACM Trans. Graph.* 41.4, 102:1–102:15. DOI: 10.1145/3528223.3530127. URL: https://doi.org/10.1145/3528223.3530127 (cit. on p. 7).

Qi, Charles R. et al. (2017). *PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space*. arXiv: 1706.02413 [cs.CV]. URL: https://arxiv.org/abs/1706.02413 (cit. on p. 10).

Rusu, Radu Bogdan and Steve Cousins (2011). "3D is here: Point Cloud Library (PCL)". In: *2011 IEEE International Conference on Robotics and Automation*, pp. 1–4. DOI: 10.1109/ICRA.2011.5980567 (cit. on p. 14).

Sánchez-Aparicio, L. J. et al. (2023). "EVALUATION OF A SLAM-BASED POINT CLOUD FOR DEFLECTION ANALYSIS IN HISTORIC TIMBER FLOORS". In: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLVIII-M-2-2023, pp. 1411–1418. DOI: 10.5194/isprs-archives-XLVIII-M-2-2023-1411-2023. URL: https://isprs-archives.copernicus.org/articles/XLVIII-M-2-2023/1411/2023/ (cit. on p. 10).

Schönberger, Johannes Lutz and Jan-Michael Frahm (2016). "Structure-from-Motion Revisited". In: *Conference on Computer Vision and Pattern Recognition (CVPR)* (cit. on p. 12).

Wang, Jingyi et al. (2024). "A survey on weakly supervised 3D point cloud semantic segmentation". In: *IET Computer Vision* 18.3, pp. 329–342. DOI: https://doi.org/10.1049/cvi2.12250. eprint: https://ietresearch.onlinelibrary.wiley.com/doi/pdf/10.1049/cvi2.12250. URL: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/cvi2.12250 (cit. on p. 14).

Wang, Qian et al. (2025). *High-Quality Spatial Reconstruction and Orthoimage Generation Using Efficient 2D Gaussian Splatting*. arXiv: 2503.19703 [cs.CV]. URL: https://arxiv.org/abs/2503.19703 (cit. on p. 7).

Wang, Zhou et al. (2004). "Image Quality Assessment: From Error Visibility to Structural Similarity". In: *IEEE Transactions on Image Processing* 13.4, pp. 600–612. DOI: 10.1109/TIP.2003.819861 (cit. on p. 23).

yan, Xu et al. (2025). *R3GS: Gaussian Splatting for Robust Reconstruction and Relocalization in Unconstrained Image Collections*. arXiv: 2505.15294 [cs.CV]. URL: https://arxiv.org/abs/2505.15294 (cit. on p. 10).

Zhou, Linglong et al. (2024). "A Comprehensive Review of Vision-Based 3D Reconstruction Methods". In: *Sensors* 24.7. ISSN: 1424-8220. DOI: 10.3390/s24072314. URL: https://www.mdpi.com/1424-8220/24/7/2314 (cit. on p. 7).