

# 3D Gaussian Splatting for Large-scale Surface Reconstruction from Aerial Images

Yuanzheng Wu, Jin Liu, Shunping Ji

**Abstract**—Recently, 3D Gaussian Splatting (3DGS) has demonstrated excellent ability in small-scale 3D surface reconstruction. However, extending 3DGS to large-scale scenes remains a significant challenge. To address this gap, we propose a novel 3DGS-based method for large-scale surface reconstruction using aerial multi-view stereo (MVS) images, named **Aerial Gaussian Splatting (AGS)**. First, we introduce a data chunking method tailored for large-scale aerial images, making 3DGS feasible for surface reconstruction over extensive scenes. Second, we integrate the Ray-Gaussian Intersection method into 3DGS to obtain depth and normal information. Finally, we implement multi-view geometric consistency constraints to enhance the geometric consistency across different views. Our experiments on multiple datasets demonstrate, for the first time, the 3DGS-based method can match conventional aerial MVS methods on geometric accuracy in aerial large-scale surface reconstruction, and our method also beats state-of-the-art GS-based methods both on geometry and rendering quality.

**Index Terms**—3D Gaussian splatting, 3D reconstruction, aerial images, multi-view stereo, image rendering

## I. INTRODUCTION

Large-scale surface reconstruction has long been a focal point of interest in both academic research and industrial applications, particularly in domains such as aerial surveying [1] [2] and smart city development [3] [4] [5] [6]. Recently, NeRF-based methods [7] [8] [9] have been extensively researched for image rendering, especially when applied to small-scale foreground targets. Additionally, these methods have shown potential for surface reconstruction [10] [11]. However, the high computational cost of volumetric rendering in NeRF-based approaches makes them impractical for large-scale scenes. The emergence of the 3D Gaussian Splatting (3DGS) technology [12] offers an alternative solution. In contrast to NeRF-based methods, 3DGS uses 3D Gaussian primitives instead of the implicit radiance field learned through Multi-Layer Perceptions (MLPs) to represent a scene. The training process is achieved by optimizing parameters such as positions, rotations, and scales of these Gaussian primitives. This approach markedly reduces computational requirements and enables more efficient scene rendering and reconstruction, making it a viable solution for large-scale reconstruction using aerial images captured by airplanes or unmanned aerial vehicles (UAVs).

Although 3DGS has demonstrated impressive capabilities in high-fidelity novel view synthesis and real-time rendering [13]

S. Ji, Y. Wu are with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, HB 430079, China (e-mail: ji shunping@whu.edu.cn; yuanzhengwu@whu.edu.cn).

J. Liu is with the School of Communication Engineering, Hangzhou Dianzi University, Hangzhou, ZJ 310018, China (e-mail: jinliu@hdu.edu.cn).

[14], achieving large-scale aerial surface reconstruction with geometric precision comparable to or exceeding that of mainstream conventional multi-view stereo (MVS) methods [15] [16] [17] or deep learning-based approaches [18] [19] remains a significant challenge. First, the extensive scenes and a large number of aerial images impose substantial computational demands, often leading to out-of-memory issues on modern GPUs. Second, it is challenging to determine the intersection between a Gaussian and a ray, making it difficult to obtain precise depth and normal vector information. Applying 3DGS directly to surface reconstruction tasks frequently results in low-precision surface models. Third, the original 3DGS [12] relies solely on image-related loss for optimization, which skews the Gaussian distribution toward high-fidelity image rendering at the expense of surface geometry accuracy. To address these challenges, we propose a novel framework based on 3DGS for large-scale surface reconstruction from aerial images. To the best of our knowledge, this is the first application of 3DGS methods to aerial images for achieving high-precision surface reconstruction.

The proposed large-scale surface reconstruction framework, named **Aerial Gaussian Splatting (AGS)**, builds upon 3D Gaussian Splatting (3DGS) [12] as its baseline. While 3DGS performs well for small-scale scenes, it struggles with large-scale environments due to high memory demands. To address the memory challenges, we tackle the problem by partitioning scenes for parallel training. A key issue in aerial scene partitioning is the uneven distribution of point clouds—some regions suffer from sparse views and points, while others are oversaturated with redundant views. The proposed method overcomes this by adopting the chunking method from Vast-Gaussian [20], partitioning the scene based on camera positions and expanding the boundaries of each data block. Additionally, to ensure the inclusion of more suitable viewpoints within each block, we develop a viewpoint selection and culling strategy. This approach enhances scene optimization by incorporating relevant viewpoints and discarding less useful ones, resulting in more efficient and balanced processing across blocks. We refer to the entire chunking method as **Adaptive Aerial Scene Partitioning**.

Due to the inability to precisely determine the intersection between a Gaussian and a ray, estimating accurate depth and normal vector information for each Gaussian primitive is a significant challenge, which limits the application of effective geometric constraints. To address this, the proposed method adopts the Ray-Gaussian Intersection (RGI) approach from [21] [22], which accurately retrieves both depth and normal vector information. Once this information is acquired,

we apply depth and normal consistency constraints [23] to enhance accuracy. Furthermore, to improve geometric consistency across different views, similar to MVS methods [24] [25] [26], we introduce a multi-view geometric consistency strategy. This strategy calculates the error of the rendered depth maps through projection and reprojection. By doing so, we improve the reconstruction of surface details, resulting in more accurate geometric alignment across different views.

We evaluate the geometric accuracy of our framework on the WHU-OMVS [24] and Tianjin aerial datasets. The experimental results show that the proposed method outperforms the existing 3DGS-based approaches and, in some cases, even surpasses the open-source MVS software Colmap [15] [16] and OpenMVS [17]. Furthermore, we validate the rendering quality of the proposed method on the WHU-OMVS, Mill-19 [27] and UrbanScene3D [28] datasets, where our method achieves better performance than other GS-based methods.

Our contributions are summarized as follows:

- We introduce a novel large-scale aerial surface reconstruction and rendering method based on 3DGS.
- We adapt 3DGS for large-scale scenes by employing a chunking method based on the VastGaussian and designing a viewpoint selection and culling strategy to optimize the chunking process.
- We introduce the ray-gaussian intersection strategy and multi-view geometric consistency constraints into the framework, which significantly improves geometric accuracy.
- We conduct experiments on multiple datasets, and the results demonstrate that the proposed method achieves high-quality surface reconstruction and delivers high-fidelity rendering.

## II. RELATED WORK

### A. Novel View Synthesis

Recent advances in Neural Radiance Fields (NeRF) [7] have significantly influenced the Novel View Synthesis (NVS) domain by employing neural networks to learn and render high-quality 3D representations of continuous volumetric scenes from images taken from multiple viewpoints. NeRF achieves high-fidelity scene representation by predicting density and color. However, the huge computational demands and the time required for both training and rendering present significant challenges for real-time rendering in large-scale scenarios. To address NeRF's computational demands, Plenoxels [29] uses 3D sparse grids to represent scene points, reducing computational complexity and storage requirements while enhancing speed compared to vanilla NeRF. However, the use of voxel grids in Plenoxels leads to a degradation of fine details. Efforts like Mip-NeRF [30] and Tri-MipRF [31] enhance rendering quality with multi-scale representations and anti-aliasing, achieving better visual quality without sacrificing efficiency.

More recently, 3DGS [12] has gained attention for its ability to achieve high-fidelity and real-time rendering. Subsequent work, such as Mip-Splatting [13], employs 3D smoothing filters to improve rendering quality. Scaffold-GS [14] leverages

anchor points to regulate local 3D Gaussian distributions, allowing for real-time adjustments to both the distribution and density of Gaussians.

### B. Surface Reconstruction

Image-based reconstruction has advanced considerably over the decades. Semi-global matching (SGM) [32] and patch-based methods [33] have been widely used for dense image matching. Complete surface reconstruction solutions, such as Colmap [15] [16] and OpenMVS [17], utilize dense matching techniques to generate dense point clouds or triangulated meshes. With the rapid development of deep learning, learning-based MVS methods [18] [19] have been developed to predict depth maps from multi-view images, and some of these methods are integrated into comprehensive surface reconstruction frameworks [24].

More recently, several works have attempted to apply NeRF or 3DGS-based methods to surface reconstruction, mainly for small-scale or foreground objects. NeuS [10] combines the strengths of both volumetric and surface rendering by optimizing Signed Distance Functions (SDF) and a color field to reconstruct fine surface details, but it requires substantial computational resources and inference time. Neuralangelo [11] improves surface reconstruction fidelity by combining multi-resolution 3D hash grids with neural surface rendering, but it also demands significant computational power.

The emergence of 3DGS [12] has introduced new approaches for surface reconstruction. SuGaR [34] compresses 3D Gaussian spheres into approximate 2D ellipses during training and utilizes Poisson reconstruction to extract continuous mesh from 3D point clouds, which are sampled based on the Gaussian density field. However, the absence of geometric constraints leads to dispersed point clouds and numerous holes in the meshes. 2D Gaussian Splatting [23] replaces 3D Gaussian ellipsoids with 2D disks, addressing the limitations of vanilla 3DGS in surface reconstruction. Nonetheless, 2DGS still struggles to capture the intricate details of large-scale scenes.

At present, most GS-based surface reconstruction methods predominantly focus on small-scale foreground targets, and large-scale surface reconstruction remains largely underexplored.

### C. Large Scene Reconstruction

In rendering research, the applications of the radiance field technology have extended from rendering close-range foreground objects to large-scale scenes. Block-NeRF [35] and Mega-NeRF [27] use data chunking strategies to handle large-scale scenes, while UE4-NeRF [36] integrates NeRF with Unreal Engine 4 for scalable rendering and scene editing. Switch-NeRF [37] uses a mixture of expert models for scene decomposition, enhancing large-scale scene reconstruction.

A few studies use 3DGS for large-scale scene rendering. CityGaussian [38] explores large-scale scene rendering using chunking and Level of Detail (LoD) methods to address challenges related to rendering efficiency and scalability. Similarly,

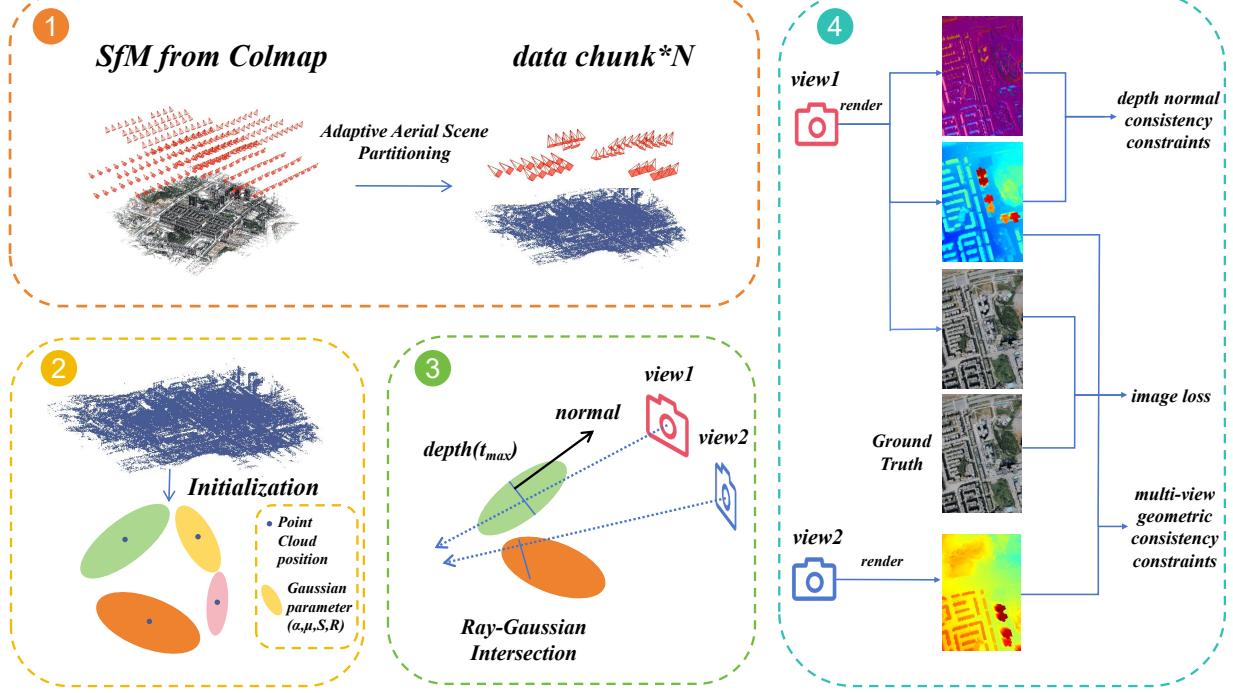


Fig. 1. The overview of AGS Framework. (1) The SfM sparse point clouds and views are divided into  $N$  data blocks. (2) The point clouds in each block are used to initialize the 3D Gaussians. (3) The ray-gaussian intersection technique is applied to obtain depth and normal vector information. (4) The depth map and normal map are utilized to compute the depth normal consistency constraints and multi-view geometric consistency constraints.

VastGaussian [20] addresses lighting effects on rendering, further enhancing the visual quality of large-scale scenes.

Most of the aforementioned NeRF and 3DGS-based studies primarily focus on image rendering, with few addressing large-scale scene surface reconstruction. In contrast, our work seeks to extend 3DGS-based methods to surface reconstruction from large-scale aerial MVS images. We borrow the idea of block chunking in the large-scale rendering method VastGaussian [20] and develop a viewpoint selection and culling strategy to bridge the huge computational resource demand and limited GPU capacity. Additionally, we introduce the ray-gaussian intersection method [21] [22] to determine the intersection point between the Gaussian and the ray, thereby enabling the acquisition of accurate depth and normal vector information. Furthermore, we apply multi-view geometric consistency constraints to ensure geometric consistency across different views, enhancing high-fidelity and high-precision large-scale scene reconstruction.

### III. PRELIMINARIES

3DGS [12] represents scenes explicitly using a large number of 3D Gaussian primitives (ellipsoids). Each Gaussian primitive is presented by four types of parameters that require optimization: position, covariance, opacity, and spherical harmonics (SH) coefficients. Using these four parameters, the  $\alpha$ -blending algorithm is employed to render a new image from these 3D Gaussians. Specifically, for a pixel  $p_i$  in the rendered image, the color of  $p_i$  can be obtained by:

$$C(p_i) = \sum_{i \in N} \alpha_i c_i \prod_{j=1}^{i-1} (1 - \alpha_j) \quad (1)$$

Where  $N$  represents the number of all Gaussians,  $c_i$  is the view-dependent color of the  $i$ -th Gaussian, derived from spherical harmonics coefficients,  $\alpha_i$  is determined by the Gaussian distribution, and the Gaussian's opacity  $\sigma_i$ , as seen in Equation (2).

$$\alpha_i = \sigma_i \exp\left(-\frac{1}{2}(p - \mu_i)^T \Sigma_i^{-1} (p - \mu_i)\right) \quad (2)$$

The parameters  $\Sigma$  is given by  $\Sigma = RSS^T R^T$ , where  $R \in \mathbb{R}^{3 \times 3}$  is the rotation matrix and  $S \in \mathbb{S}^{3 \times 1}$  is the scale matrix.

When calculating the depth map, we accumulate the blending weights  $\alpha$  at pixel  $p$ . Once the cumulative value exceeds 0.5, the depth of the current Gaussian is assigned as the depth value for that pixel. In this paper, the depth of a Gaussian is calculated by the ray-gaussian intersection method [21] [22].

### IV. METHOD

In this section, we present Aerial Gaussian Splatting (AGS), a surface reconstruction framework based on 3DGS, specifically designed for aerial MVS images. Compared to the original 3DGS [12], the proposed method incorporates several innovative and necessary modules. First, we introduce adaptive aerial scene partitioning to divide large-scale scenes effectively and ensure optimal merging at the final step. Furthermore, the ray-gaussian intersection technique [21] [22] is employed to obtain accurate depth and normal vector information. Finally, multi-view geometric consistency constraints are incorporated to improve the reconstruction quality. The framework is shown in Fig. 1.

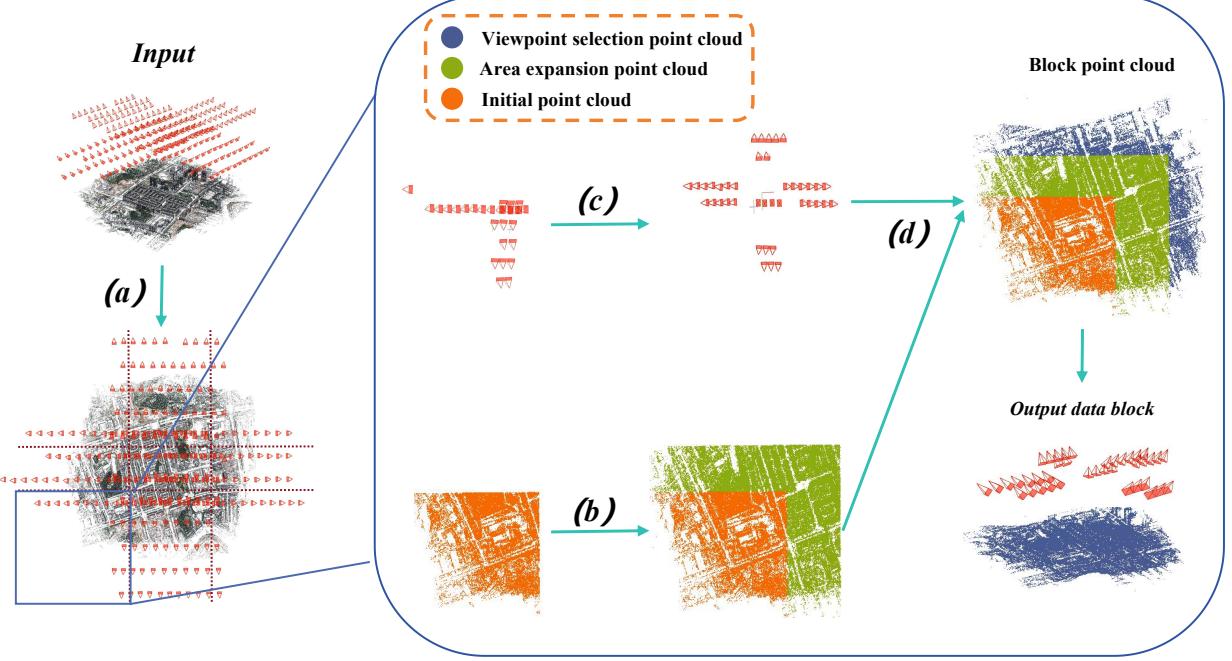


Fig. 2. Overview of Adaptive Aerial Scene Partitioning strategy. (a) The entire scene is divided into  $N$  regions based on camera positions. (b) The boundaries of each region are expanded. (c) Viewpoints (i.e., cameras) are selected and culled. (d) All points visible from the selected viewpoints within each data block are added to the block's point cloud.

### A. Adaptive Aerial Scene Partitioning

In the field of aerial photogrammetry, certain basic principles guide the partitioning of a scene into blocks [24] to alleviate the computational burden. In this paper, our method is inspired by VastGaussian [20] and involves three main steps. The first two steps are derived from VastGaussian. The process, as shown in Fig. 2, begins by dividing the scene based on camera positions and extending the block boundaries. Specifically, the scene, containing a total of  $n$  viewpoints, is divided into  $M \times N$  blocks. First, the viewpoints are horizontally divided into  $M$  blocks, with each block containing  $n/M$  viewpoints. Then, these  $M$  blocks are further divided vertically, resulting in each block containing  $n/(M \times N)$  viewpoints. After this initial partitioning, the point clouds within each region are aggregated into distinct point cloud blocks. To minimize artifacts across the scene, each region is extended by a certain proportion as in VastGaussian, ensuring that each block is adequately optimized.

However, a coarse data block partitioning method based solely on camera positions may result in insufficient optimization within each block due to suboptimal viewpoints. To address this, we develop a viewpoint selection and culling strategy, as shown in Fig. 2(c) and (d). This strategy removes erroneous viewpoints from the data block and supplements it with additional effective viewpoints. This process involves projecting all point clouds within a data chunk onto all images and calculating projection scores to determine the suitability of each viewpoint. If a point falls within the central scope (here 70%) of an image, the projection score for that image is incremented by one. For each region, the top  $N$  images with the highest scores are selected as viewpoints to optimize the

data block. Finally, to ensure sufficient points for initialization and to mitigate artifacts, the sparse point cloud generated from the SfM of all new viewpoints is incorporated into the data block.

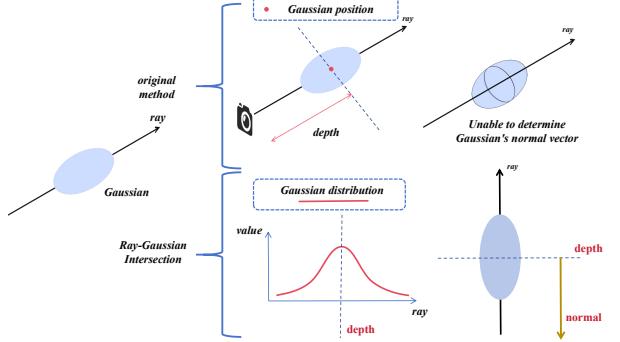


Fig. 3. Ray-Gaussian Intersection. By calculating the maximum Gaussian value along the ray, we can obtain accurate depth and normal vector information for the Gaussian.

### B. Ray-Gaussian Intersection

In vanilla 3DGS, the depth of each Gaussian primitive is assigned based on its distance from the screen, and accurate normal vector information is not provided. However, for surface reconstruction tasks, reliable depth information and normal vector data are essential for geometric constraints. To address these issues, we introduce the ray-gaussian intersection technique [21] [22], as shown in Fig. 3, to improve the accuracy of surface reconstruction. The intersection point  $t_{max}$  between a Gaussian primitive and a ray, corresponding to the

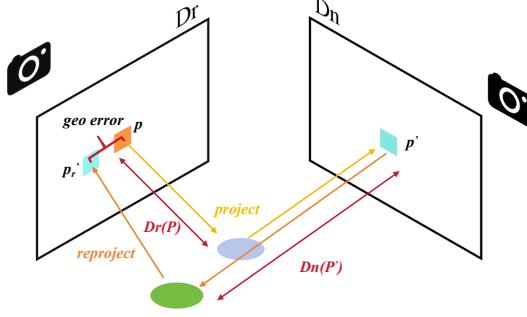


Fig. 4. Multi-view geometric consistency constraints. The multi-view geometric consistency constraints are modeled as the error between projection and reprojection of depth map across multiple views.

maximum Gaussian value along the ray, can be computed as follows [22]:

$$t_{max} = -\frac{o_g^T r_g}{r_g^T r_g} \quad (3)$$

$o_g$ ,  $r_g$  are  $o$  (the camera center) and  $r$  (the ray direction) converted into the Gaussian local coordinate system. An arbitrary point along the ray is defined as  $x = o + tr$ , where  $t$  is the depth of the ray.

Once the depth value is computed, the Gaussian's normal is derived as the normal of the intersection plane relative to the given ray direction. For image rendering, rather than projecting 3D Gaussians onto 2D screen space as done in the original 3DGS, we utilize the ray-gaussian intersection method to determine the intersection point between the Gaussian and the ray, allowing us to compute the contribution of a Gaussian to a given ray in 3D space.

After obtaining the depth and normal vector information, we apply 2DGS's depth normal consistency constraints [23]. Specifically, this involves calculating the error between the normal map and the gradient values derived from the depth map, which is used as the loss function.

$$L_n = \sum_i \omega_i (1 - n_i^T N) \quad (4)$$

Here,  $i$  represents the Gaussian index,  $\omega_i$  denotes the blending weight,  $n_i$  represents the normal vector of the Gaussian, and  $N$  is the normal vector calculated from the depth map. The normal vector in the depth map at a given point is computed as Eq.(5), in which  $p$  represents the pixel coordinates and  $\nabla$  represents the gradient calculation.

$$N(x, y) = \frac{\nabla_x p \times \nabla_y p}{|\nabla_x p \times \nabla_y p|} \quad (5)$$

### C. Multi-view geometric consistency constraints

As 3DGS-based surface construction is still in its early stages, certain beneficial empirical approaches, such as multi-view geometric consistency constraints, have yet to be fully developed. In this study, we introduce multi-view geometric consistency constraints to ensure geometric coherence across multiple views. As shown in Fig. 4, we render the depth maps

for two adjacent viewpoints  $V_r$  and  $V_n$ , denoted as  $D_r$  and  $D_n$ . Firstly, a pixel  $P$  in the reference view  $V_r$  is projected onto the adjacent view  $V_n$  through its depth value  $Dr(P)$  and the intrinsic and extrinsic parameters, yielding the projected point  $P'$  in  $V_n$ . Subsequently, the projected point  $P'$  is reprojected onto the reference view based on its rendered depth value  $Dn(P')$ , resulting in the reprojected pixel  $P'_r$ . The distance between the coordinates of  $P$  and  $P'_r$  is calculated as the geometric consistency constraints:

$$L_{geo} = \frac{1}{V} \sum_{P \in V} \|P - P'_r\| \quad (6)$$

When calculating the loss, only the non-zero values are averaged, as shown in Eq.(6), where  $V$  represents the valid pixels. To reduce the impact of occlusion, a distance threshold  $T$  (It is usually set to 1.) is applied to identify valid pixels, with distances  $\|P - P'_r\|$  exceeding  $T$  set to zero.

### D. Merging

After parameters in each data block are optimized separately, all blocks are merged to form a coherent scene. This is achieved by removing the expanded regions of each block prior to merging.

## V. EXPERIMENT

### A. Dataset

**WHU-OMVS:** This dataset covers an area of Guizhou, China, with a ground resolution of 10 cm. The images are captured using a camera rig with one nadir and four oblique viewpoints, totaling 268 images, each with a resolution of 3712×5504 pixels. The flight height is 550 meters, covering an area of 850×700 m<sup>2</sup>. Due to GPU memory limitations, we apply a 4x downsampling to the images during the training of all methods, as done in previous studies [38] [20] [37] [27] for rendering and reconstruction. The depth map is used as ground truth, and we evaluate the geometric accuracy of the proposed method based on the rendered depth map.

**Tianjin Dataset:** This dataset is captured by a camera rig with one nadir and four oblique viewpoints over Tianjin city, China, with an image size of 3840×2560 pixels and a ground resolution of 20 cm captured at a height of 200 meters. The dataset consists of 342 images and covers an area of 400×350 m<sup>2</sup>. The image overlap is 80% along the heading direction and 60% in the side direction. The ground truth data is derived from LiDAR point cloud scans. We apply a 4x downsampling operation to all images.

**Mill-19 and UrbanScene3D:** We apply the proposed method to three open-source large-scale scenes: Rubble and Building from the Mill-19 dataset [27] and Residence from the UrbanScene3D dataset [28], containing 1,678, 1,940, and 2,582 images, respectively. Following previous rendering methods [38] [20] [37] [27], we perform a 4x downsampling operation to the input image during training. Due to the absence of ground truth, we conduct a qualitative analysis of the surface reconstruction results from these datasets.



Fig. 5. Surface reconstruction results of WHU-OMVS dataset.

TABLE I  
THE QUANTITATIVE RESULTS OF SURFACE RECONSTRUCTION ON THE WHU-OMVS DATASET. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**, AND THE SECOND-BEST RESULTS ARE UNDERLINED.

method	PAG <sub>0.6m</sub> (%)	PAG <sub>0.8m</sub> (%)	PAG <sub>1.0m</sub> (%)	MAE(m)	RMSE(m)
OpenMVS	71.91	77.81	81.18	0.520	<u>1.041</u>
Colmap	71.32	<u>77.44</u>	81.36	0.623	1.287
GOF	<u>79.80</u>	<u>83.83</u>	<u>86.03</u>	<u>0.476</u>	1.155
3DGS	65.01	73.09	74.98	0.749	1.366
2DGS	56.88	65.57	69.42	0.906	1.483
proposed	<b>82.58</b>	<b>87.07</b>	<b>89.50</b>	<b>0.451</b>	<b>1.012</b>

### B. Implementation

When training the proposed method, we first perform Manhattan alignment on the target scene, aligning the y-axis to be perpendicular to the ground to facilitate the chunking process. Each block is expanded by 20%. During training, each data block is independently optimized for 50,000 iterations. The densification process begins after 500 iterations and ends at 30,000 iterations. The multi-view geometric consistency constraints and depth normal consistency constraints are introduced at 7,000 iterations. Experiments are performed on the RTX4090. Other settings remain consistent with those used in the original 3DGS method [12]. For the sparse point cloud generation, we use the SfM module in Colmap [16] [15]. For surface reconstruction, we follow the 2D Gaussian Splatting approach [23] and utilize the Truncated Signed Distance Function (TSDF) [39]. When evaluating 3DGS [12], we set the densification interval to 250 instead of the original 100 to avoid the out-of-memory problem.

### C. Results

We conduct surface reconstruction experiments on the WHU-OMVS [24], Tianjin, Mill-19 [27] and UrbanScene3D [28] dataset. Due to the limited number of 3DGS-based surface reconstruction methods capable of handling scenes with

significant depth variation, we select three for comparison: 2D Gaussian Splatting [23], 3D Gaussian Splatting [12], and Gaussian Opacity Fields [22]. Additionally, we include comparisons with the widely recognized open-source MVS software Colmap [16] [15] and OpenMVS [17].

Beyond surface reconstruction, we also evaluate the rendering quality of the proposed method in Mill-19, UrbanScene3D, and WHU-OMVS. For rendering comparisons, we select four state-of-the-art methods for large-scale rendering: Mega-NeRF [27], Switch-NeRF [37], GP-NeRF [40], and CityGaussian [38].

#### 1) Surface reconstruction:

**Results on WHU-OMVS:** Following the work [24], we use the MAE, RMSE, and PAG metrics to evaluate the rendered depth map. The specific explanations of these metrics are as follows:

**Mean Absolute Error (MAE):** MAE measures the absolute difference between the predicted values and the ground truth, which is calculated by:

$$MAE = \frac{1}{m} \sum_{i=1}^m |y_i - \hat{y}_i| \quad (7)$$

where  $y_i$  represents the ground truth,  $\hat{y}_i$  represents the estimated value and  $m$  denotes the number of valid values. In our



Fig. 6. Qualitative comparisons with different methods. The image on the left represents the original image from the viewpoint, while the three images on the right depict the error maps between the predicted depth values and the ground truth. The error maps' bands indicate the errors' magnitude, with darker colors representing larger errors.

TABLE II

THE QUANTITATIVE RESULTS OF SURFACE RECONSTRUCTION ON THE TIANJIN DATASET. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**, AND THE SECOND-BEST RESULTS ARE UNDERLINED.

method	Percentage (0.6m)↑			Percentage (0.8 m)↑			Percentage (1.0 m)↑		
	Acc.	Comp.	f-score	Acc.	Comp.	f-score	Acc.	Comp.	f-score
Colmap	<b>79.51</b>	88.79	<b>83.90</b>	<b>85.55</b>	<u>90.92</u>	<b>88.15</b>	<b>89.22</b>	<u>92.30</u>	<b>90.73</b>
GOF	77.05	86.53	81.51	83.09	89.06	85.97	86.73	90.90	88.76
3DGS	51.27	<b>93.19</b>	66.15	62.79	<b>95.49</b>	75.77	71.13	<b>96.86</b>	82.02
2DGS	69.54	81.60	75.09	79.85	85.07	82.38	86.03	86.89	86.46
proposed	79.37	85.92	<u>82.52</u>	<u>85.33</u>	89.31	<b>86.58</b>	<u>88.80</u>	89.31	<u>89.06</u>

experiments, differences larger than 10 meters are considered invalid and excluded from the calculation.

**Root Mean Square Error (RMSE):** RMSE calculates the standard deviation of the differences between the estimated values and the ground truth:

$$RMSE = \sqrt{\frac{1}{m} \sum_{i=1}^m (y_i - \hat{y}_i)^2} \quad (8)$$

Similarly, any errors exceeding 10 meters are treated as invalid and excluded from the computation.

**Percentage of Accurate Grids (PAG):** PAG measures the proportion of grids with absolute difference below a given threshold  $\alpha$  relative to the total number of grids. The evaluation is conducted using three different thresholds: 0.6m, 0.8m, and 1.0m.

$$PAG_\alpha = \left( \frac{m_\alpha}{m} \cdot 100\% \right) \quad (9)$$

$m_\alpha$  represents the valid grid, and  $m$  represents the number of all grids.

The surface reconstruction results on the WHU-OMVS dataset are presented in Table I. The experimental results demonstrate that the proposed method achieves the best reconstruction results at PAG, MAE, and RMSE metrics. For the strictest metric  $PAG_{0.6m}$ , the proposed method surpass the second-best GOF by 2.78% and the open-source software Colmap by 11.26%, highlighting its superiority in fine-grained reconstruction. In the  $PAG_{0.8m}$  and  $PAG_{1.0m}$  metrics, the proposed method outperforms other approaches by large margins, confirming that its overall reconstruction quality is substantially higher than that of the competing methods. The mesh of reconstruction results are shown in Fig. 5. The depth error distribution of different methods, shown in Fig. 6, further supports this conclusion. For example,

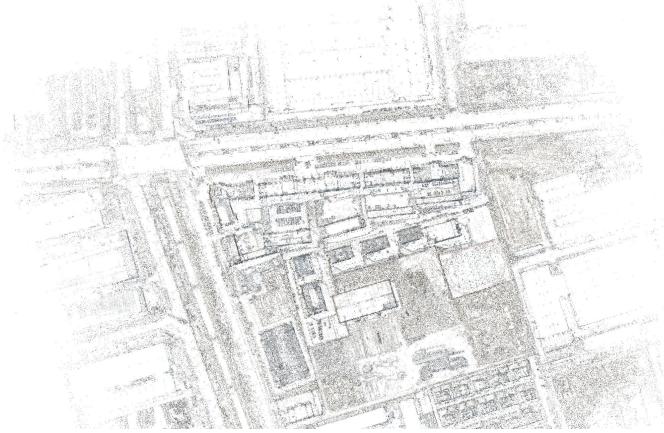


Fig. 7. Sparse point cloud of the scene.

the 2DGS results in some severe reconstruction errors in certain areas, while Colmap produces large errors along object edges and noticeable holes in some areas. In contrast, the proposed method exhibits smaller overall errors and performs exceptionally well in detailed areas.

#### Experimental Results on Tianjin:

As the Tianjin Dataset provides only ground truth point clouds instead of pixel-wise depth maps, the accuracy of the point cloud is evaluated using the following metrics. Percentage metrics [41] are employed, with thresholds set at 0.6m, 0.8m, and 1.0m. "Accuracy" represents the distance from the reconstructed point cloud to the ground truth, while "Completeness" represents the distance from the ground truth to the reconstructed point cloud. The F-score for percentage metric is defined as the harmonic mean of accuracy and completeness.

A notable characteristic of the Tianjin dataset is the prevalence of numerous weakly textured areas, which presents significant challenges for surface reconstruction. We use SfM in Colmap to generate initial point clouds. As shown in Fig. 7, certain areas—such as roads, buildings, and rooftops—exhibit notably sparse point distributions. This sparsity poses significant difficulties for 3DGS-based methods, which rely on well-distributed sparse points as the initialization for Gaussian primitives.

As shown in Table II, despite the significant challenges posed by this dataset to 3DGS-based methods, our approach achieves geometric accuracy comparable to that of Colmap, with very close metric values. Moreover, at the most strict resolution, our method outperforms other 3DGS-based methods, achieving 9.83% higher accuracy than 2DGS, 26.10% higher accuracy than 3DGS, and surpassing GOF by 2.32%. Due to the point cloud count after fusion far exceeding that of the ground truth, the completeness metrics do not effectively reflect the true quality of the reconstruction. As seen in the normal map shown in Fig. 8, the proposed method exhibits a much cleaner and smoother result than other methods. In the marked regions, GOF shows significant noise, and 2DGS lacks fine details, whereas the proposed method maintains a clean and smooth appearance, showcasing its superiority in

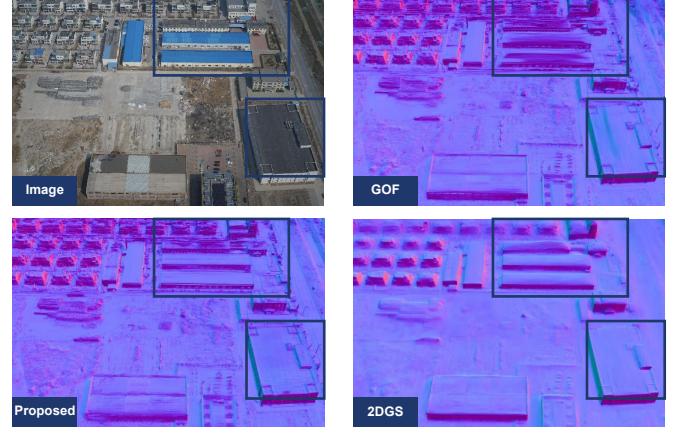


Fig. 8. Qualitative comparison of normal map with 2DGS, GOF, and the proposed method.



Fig. 9. Surface reconstruction results of Tianjin dataset.

handling weak textures and complex geometries. The mesh of reconstruction results is shown in Fig. 9.

**Results on Mill-19 and UrbanScene3D:** Due to the lack of geometric ground truth and the inability of methods like 2DGS and GOF to complete training on these extremely large scenes, we lack direct comparative methods. Therefore, we conduct a qualitative analysis of the surface reconstruction results. As shown in Fig. 10, our reconstructed scenes are generally complete with smooth surfaces. On a more detailed level, our method effectively captures fine details. For instance, in the normal map shown in Fig. 10(a), not only are the buildings reconstructed, but finer details such as windows and air conditioning units are also well represented. In Fig. 10(b), the vegetation on the ground is also successfully reconstructed. Fig. 10(c) demonstrates that even small objects on the ground, such as vehicles, can be reconstructed. The most remarkable result is seen in Fig. 10(d), where the thin streetlights along the road are clearly reconstructed, highlighting the excellent performance of our method in capturing fine details. We also apply our method to the Building and Rubble scenes, as shown in Fig. 11, the results are equally impressive.

#### 2) Novel View Synthesis:

Following the work of [37] [38] [42], we use the peak

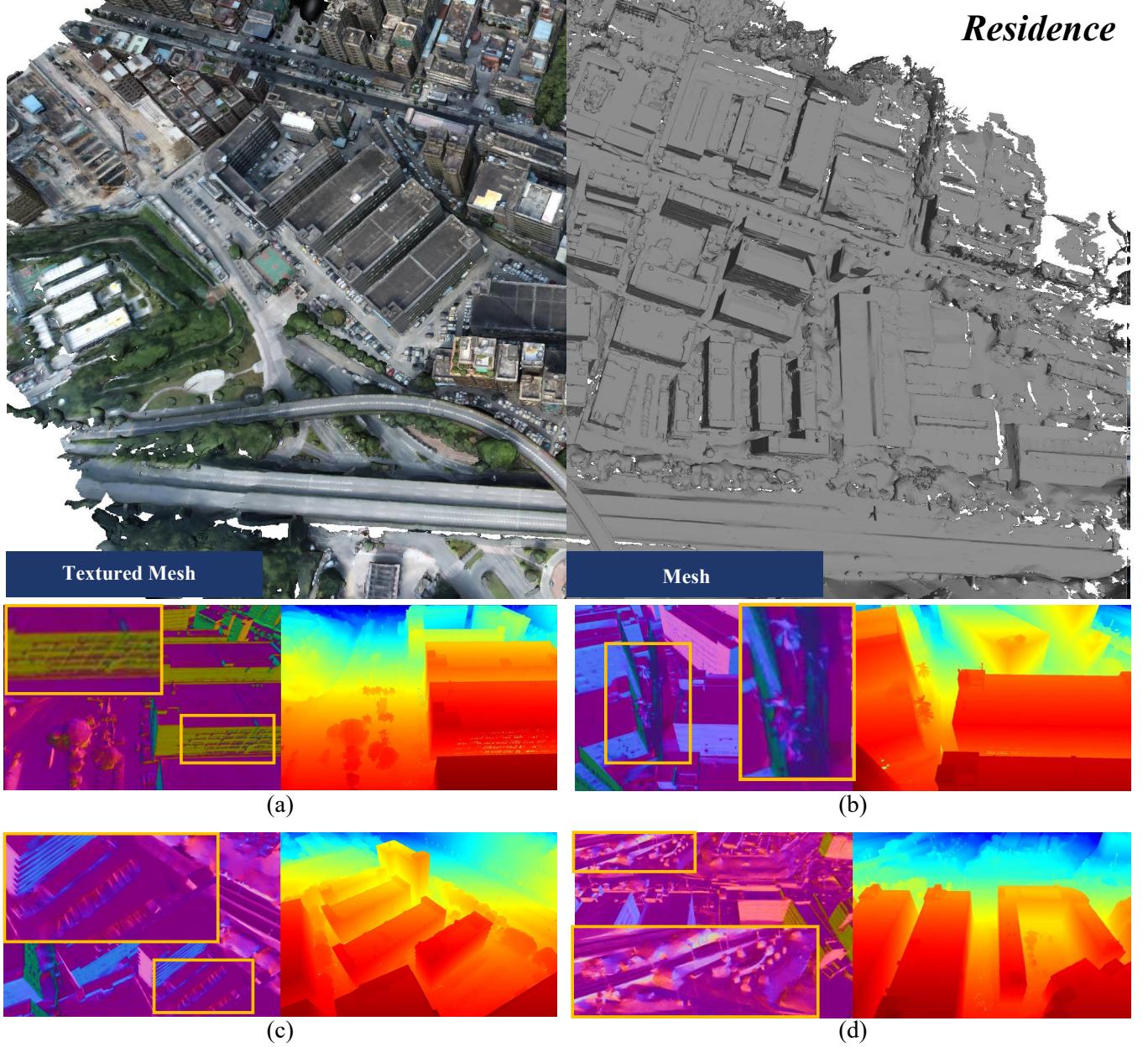


Fig. 10. Qualitative analysis of the results of the Residence surface reconstruction.

signal-to-noise ratio (PSNR), structural similarity index measure (SSIM), and the learned perceptual image patch similarity (LPIPS) metrics to evaluate the quality of the rendered image.

**Results on WHU-OMVS:** We assess the rendering quality of the proposed method on the WHU-OMVS [24] dataset, as shown in Table III. The results show that the proposed method significantly outperforms others in the LPIPS metric, with approximately a 46% improvement over 3DGS. This indicates that the proposed method achieves superior visual consistency with the ground truth images. While the PSNR score is slightly lower than that of 3DGS—primarily due to the use of lighting compensation [20] in our method, which affects image brightness—our score remains competitive. Additionally, our SSIM score is also very competitive, only 0.002 lower than the best-performing method.

TABLE III  
THE IMAGE QUALITY OF NOVEL VIEW SYNTHESIS (NVS) ON THE WHU-OMVS DATASET WAS COMPROMISED. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**, AND THE SECOND-BEST RESULTS ARE UNDERLINED.

method	SSIM↑	PSNR↑	LPIPS↓
GOF	0.925	29.07	0.666
3DGS	<b>0.940</b>	<u>31.14</u>	<u>0.514</u>
2DGS	0.927	29.89	0.668
proposed	0.938	30.46	<b>0.270</b>

**Results on Mill-19 and UrbanScene3D:** Following the work of [37] [27] and [38], we conduct experiments on the Mill-19 and UrbanScene3D datasets to further validate the rendering quality. As shown in Table IV, the proposed method

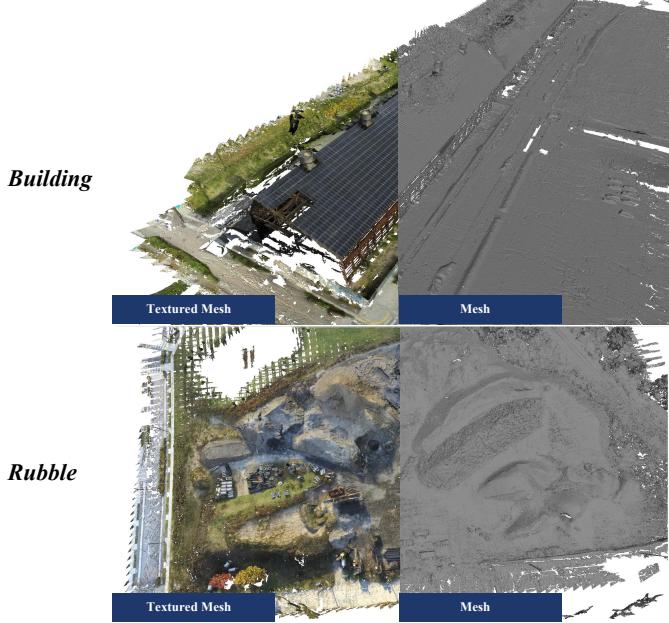


Fig. 11. Qualitative analysis of the results of the Building and Rubble surface reconstruction.

achieves the highest scores across all three metrics (PSNR, SSIM, and LPIPS) in the Building and Rubble datasets, with PSNR significantly surpassing other methods. This demonstrates the method's ability to deliver optimal perceptual quality and achieve high-fidelity rendering. In the Residence dataset, the proposed method achieves the best results in both PSNR and LPIPS. Although our SSIM score is slightly lower than that of CityGaussian, it still significantly outperforms other methods, confirming the method's effectiveness across different datasets.

#### D. Ablation

We conduct ablation experiments mainly on the WHU-OMVS dataset.

**Viewpoint Selection and Culling:** As shown in Table V, disabling the viewpoint selection and culling strategy results in a significant reduction in reconstruction accuracy for the proposed method. This decline primarily stems from the insufficient number of views. As depicted in Fig. 12, the lack of proper viewpoints leads to severe errors in localized areas of the depth error map, and certain parts of buildings in the normal map appear transparent and under-optimized. Thus, viewpoint selection and culling are crucial components of the chunking strategy, ensuring better coverage and optimization.

**Ray-Gaussian Interaction:** As shown in Table VI, the introduction of ray-gaussian interaction significantly improves reconstruction quality, evidenced by a 19.60% improvement in  $PAG_{0.6m}$ . In the original 3DGS method [12], depth maps are generated using the initialized depth from Gaussians, and normal vector information is unavailable. In contrast, our approach accurately captures both depth and normal vectors, enabling the application of geometric constraints. As a result, the ray-gaussian intersection effectively mitigates the

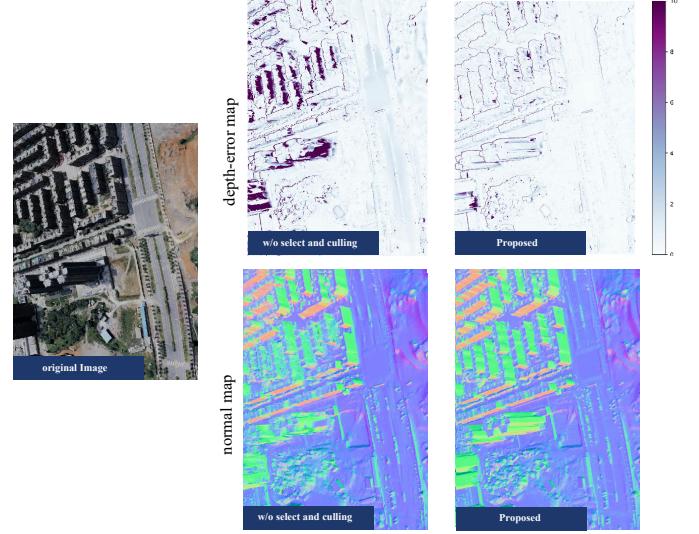


Fig. 12. Qualitative comparison with and without viewpoint selection and culling.

challenges associated with surface reconstruction, particularly those caused by the irregular distribution of Gaussian primitives.

**Multi-View Geometric Consistency Constraints:** The original 3DGS [12] applies loss within a single view, which can lead to overfitting and fails to ensure consistency across multiple viewpoints. To address this issue, we introduce multi-view geometric consistency constraints. Table VI demonstrates that the multi-view geometric consistency constraints effectively enhance reconstruction accuracy by ensuring coherence across multiple views.

## VI. DISCUSSION

**Application Prospect:** The proposed method achieves geometric accuracy comparable to conventional open-source methods like Colmap and OpenMVS. However, it should be noted that we currently can only render a 1080p image (and depth map). In [38] [27] and this work, the images are down-sampled by a factor of four. This presents a barrier to applying 3DGS-based methods to images with full resolution. Future work must explore efficient methods for rendering higher-resolution images. Nevertheless, a significant advantage of the 3DGS-based method over traditional MVS approaches lies in its ability to render high-fidelity images while reconstructing surfaces. This offers new potential for surveying applications, allowing for measurements not only from the reconstructed mesh but also from the high-fidelity rendered images. Further research is required to develop a suitable measurement and evaluation method for these rendered images.

**Gaussian Seeds:** The 3DGS-based methods rely heavily on sparse point clouds generated through SfM as the initial seeds for Gaussian primitives. Datasets like Tianjin, which contain extensive textureless regions, present significant challenges due to the absence of initialized Gaussian primitives in these areas. The densification process attempts to densify Gaussian primitives into the textureless regions. However, these

TABLE IV

QUANTITATIVE COMPARISON OF RENDERED IMAGE ON THREE DATASETS. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**, AND THE SECOND-BEST RESULTS ARE UNDERLINED.

method	Residence			Building			Rubble		
	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓
MegaNeRF	0.628	22.08	0.489	0.569	21.48	0.378	0.553	24.06	0.516
Switch-NeRF	<u>0.654</u>	<u>22.57</u>	0.457	<u>0.594</u>	<u>22.07</u>	0.332	<u>0.562</u>	24.31	0.496
GP-NeRF	0.661	<u>22.31</u>	0.448	0.566	<u>21.03</u>	0.486	0.565	24.06	0.496
CityGaussian	<b>0.813</b>	22.00	<u>0.211</u>	<u>0.778</u>	21.55	<u>0.246</u>	<u>0.813</u>	<u>25.77</u>	<u>0.228</u>
proposed	0.756	<b>22.63</b>	<b>0.182</b>	<b>0.803</b>	<b>24.31</b>	<b>0.148</b>	<b>0.827</b>	<b>27.32</b>	<b>0.143</b>

TABLE V  
ABLATION STUDY OF VIEWPOINT SELECTION AND CULLING.

method	PAG <sub>0.6m</sub> (%)	PAG <sub>0.8m</sub> (%)	PAG <sub>1.0m</sub> (%)	MAE(m)	RMSE(m)
w/o VSC	69.16	75.69	79.65	0.701	1.449
proposed	<b>82.58</b>	<b>87.07</b>	<b>89.50</b>	<b>0.451</b>	<b>1.012</b>

TABLE VI  
ABLATION EXPERIMENTS ON THE WHU-OMVS DATASET.

method	PAG <sub>0.6m</sub> (%)	PAG <sub>0.8m</sub> (%)	PAG <sub>1.0m</sub> (%)	MAE(m)	RMSE(m)
w/o RGI	62.98	71.72	77.60	0.757	1.328
w/o MVGC	81.74	86.07	88.40	0.464	1.068
proposed	<b>82.58</b>	<b>87.07</b>	<b>89.50</b>	<b>0.451</b>	<b>1.012</b>

primitives, guided solely by RGB images, do not accurately reflect the actual surface. Due to the use of  $\alpha$ -blending for image rendering, these primitives in weak texture or textureless regions will interfere with the depth estimation of surrounding areas, leading to more depth estimation errors. As a result, the proposed method performs suboptimally on the Tianjin dataset. There remains considerable room for improvement in our approach. Future work could focus on improving SfM techniques or developing specialized densification strategies tailored for weak texture or textureless regions in 3DGS-based methods.

## VII. CONCLUSION

In this paper, we present the AGS framework, the first framework to achieve large-scale high-precision surface reconstruction from aerial images using a 3DGS-based approach. The proposed method combines a data chunking strategy specifically designed for aerial images, allowing each data block to be independently trained on a GPU and merged after training. Additionally, it incorporates the ray-gaussian intersection method to impose depth normal consistency constraints and multi-view geometric consistency constraints. We validate the geometric accuracy of our approach on the WHU-OMVS and Tianjin datasets and evaluate the rendering quality on the WHU-OMVS, Mill-19, and UrbanScene3D datasets. Experimental results demonstrate that the proposed method effectively performs surface reconstruction in large-scale scenes while achieving excellent rendering quality.

## REFERENCES

- [1] Dominique Meyer, Elioth Fraijo, Eric Lo, Dominique Rissolo, and Falko Kuester. Optimizing uav systems for rapid survey and reconstruction of large scale cultural heritage sites. In *2015 Digital Heritage*, volume 1, pages 151–154. IEEE, 2015.
- [2] Surendra Pal Singh, Kamal Jain, and V Ravibabu Mandla. 3d scene reconstruction from video camera for virtual 3d city modeling. *American Journal of Engineering Research*, 3(1):140–148, 2014.
- [3] Nina Danilina, Mihail Slepnev, and Spartak Chebotarev. Smart city: Automatic reconstruction of 3d building models to support urban development and planning. In *MATEC Web of Conferences*, volume 251, page 03047. EDP Sciences, 2018.
- [4] Mehmet Buyukdemircioglu and Sultan Kocaman. Reconstruction and efficient visualization of heterogeneous 3d city models. *Remote Sensing*, 12(13):2128, 2020.
- [5] Huma H Khan, Muhammad N Malik, Raheel Zafar, Feybi A Goni, Abdoulmohammad G Chofreh, Jiří J Klemeš, and Yousef Alotaibi. Challenges for sustainable smart city development: A conceptual framework. *Sustainable Development*, 28(5):1507–1518, 2020.
- [6] Yirang Lim, Jurian Edelenbos, and Alberto Gianoli. Identifying the results of smart city development: Findings from systematic literature review. *Cities*, 95:102397, 2019.
- [7] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.
- [8] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021.
- [9] Li Ma, Xiaoyu Li, Jing Liao, Qi Zhang, Xuan Wang, Jue Wang, and Pedro V Sander. Deblur-nerf: Neural radiance fields from blurry images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12861–12870, 2022.
- [10] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021.
- [11] Zhaoshuo Li, Thomas Müller, Alex Evans, Russell H Taylor, Mathias Untereth, Ming-Yu Liu, and Chen-Hsuan Lin. Neuralangelo: High-fidelity neural surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8456–8465, 2023.
- [12] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.
- [13] Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. Mip-splatting: Alias-free 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19447–19456, 2024.
- [14] Tao Lu, Mulin Yu, Lining Xu, Yuanbo Xiangli, Limin Wang, Dahua

- Lin, and Bo Dai. Scaffold-gs: Structured 3d gaussians for view-adaptive rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20654–20664, 2024.
- [15] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [16] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016.
- [17] Dan Cernea. OpenMVS: Multi-view stereo reconstruction library. 2020.
- [18] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision (ECCV)*, pages 767–783, 2018.
- [19] Jin Liu and Shunping Ji. A novel recurrent encoder-decoder structure for large-scale multi-view stereo reconstruction from an open aerial dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6050–6059, 2020.
- [20] Jiaqi Lin, Zhihao Li, Xiao Tang, Jianzhuang Liu, Shiyong Liu, Jiayue Liu, Yangdi Lu, Xiaofei Wu, Songcen Xu, Youliang Yan, et al. Vastgaussian: Vast 3d gaussians for large scene reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5166–5175, 2024.
- [21] Leonid Keselman and Martial Hebert. Approximate differentiable rendering with algebraic surfaces. In *European Conference on Computer Vision*, pages 596–614. Springer, 2022.
- [22] Zehao Yu, Torsten Sattler, and Andreas Geiger. Gaussian opacity fields: Efficient and compact surface reconstruction in unbounded scenes. *arXiv preprint arXiv:2404.10772*, 2024.
- [23] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024.
- [24] Jin Liu, Jian Gao, Shunping Ji, Chang Zeng, Shaoyi Zhang, and JianYa Gong. Deep learning based multi-view stereo matching and 3d scene reconstruction from oblique aerial images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 204:42–60, 2023.
- [25] Qingshan Xu and Wenbing Tao. Multi-scale geometric consistency guided multi-view stereo. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5483–5492, 2019.
- [26] Haonan Dong and Jian Yao. Patchmvsnet: Patch-wise unsupervised multi-view stereo for weakly-textured surface reconstruction. *arXiv preprint arXiv:2203.02156*, 2022.
- [27] Haithem Turki, Deva Ramanan, and Mahadev Satyanarayanan. Meganerf: Scalable construction of large-scale nerfs for virtual fly-throughs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12922–12931, 2022.
- [28] Liqiang Lin, Yilin Liu, Yue Hu, Xinguang Yan, Ke Xie, and Hui Huang. Capturing, reconstructing, and simulating: the urbanscene3d dataset. In *European Conference on Computer Vision*, pages 93–109. Springer, 2022.
- [29] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinzhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5501–5510, 2022.
- [30] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5855–5864, 2021.
- [31] Wenbo Hu, Yuling Wang, Lin Ma, Bangbang Yang, Lin Gao, Xiao Liu, and Yuewen Ma. Tri-miprf: Tri-mip representation for efficient anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19774–19783, 2023.
- [32] Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on pattern analysis and machine intelligence*, 30(2):328–341, 2007.
- [33] Michael Bleyer, Christoph Rhemann, and Carsten Rother. Patchmatch stereo-stereo matching with slanted support windows. In *Bmvc*, volume 11, pages 1–11, 2011.
- [34] Antoine Guédon and Vincent Lepetit. Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5354–5363, 2024.
- [35] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P Srinivasan, Jonathan T Barron, and Henrik Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8248–8258, 2022.
- [36] Jiaming Gu, Minchao Jiang, Hongsheng Li, Xiaoyuan Lu, Guangming Zhu, Syed Afaf Ali Shah, Liang Zhang, and Mohammed Bennamoun. Ue4-nerf: Neural radiance field for real-time rendering of large-scale scene. *Advances in Neural Information Processing Systems*, 36, 2024.
- [37] Zhenxing Mi and Dan Xu. Switch-nerf: Learning scene decomposition with mixture of experts for large-scale neural radiance fields. In *The Eleventh International Conference on Learning Representations*, 2022.
- [38] Yang Liu, He Guan, Chuanchen Luo, Lue Fan, Junran Peng, and Zhaoxiang Zhang. Citygaussian: Real-time high-quality large-scale scene rendering with gaussians. *arXiv preprint arXiv:2404.01133*, 2024.
- [39] Diana Werner, Ayoub Al-Hamadi, and Philipp Werner. Truncated signed distance function: experiments on voxel size. In *Image Analysis and Recognition: 11th International Conference, ICIAR 2014, Vilamoura, Portugal, October 22–24, 2014, Proceedings, Part II 11*, pages 357–364. Springer, 2014.
- [40] Hao Li, Dingwen Zhang, Yalun Dai, Nian Liu, Lechao Cheng, Jingfeng Li, Jingdong Wang, and Junwei Han. Gp-nerf: Generalized perception nerf for context-aware 3d scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21708–21718, 2024.
- [41] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples. *ACM Transactions on Graphics*, page 1–13, Aug 2017.
- [42] Lining Xu, Yuanbo Xiangli, Sida Peng, Xingang Pan, Nanxuan Zhao, Christian Theobalt, Bo Dai, and Dahua Lin. Grid-guided neural radiance fields for large urban scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8296–8306, 2023.