Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

PanoramicSplat: Scaling 3D Gaussian Splatting for Single-View Large-Scale Outdoor Reconstruction

Zijian Wu
Lishui University
Zhen Ye
yezhen@lsu.edu.cn

Lishui University

Article

Keywords: 3D gaussian splatting, Single-view reconstruction, Outdoor, Large-scale reconstruction

Posted Date: October 7th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11014476/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

PanoramicSplat: Scaling 3D Gaussian Splatting for Single-View Large-Scale Outdoor Reconstruction

Zijian Wu[1,2] and Zhen Ye[1,3*]

[1]Lishui University, Lishui, 323000, China.
[2]Zhejiang Sci-Tech University, Hangzhou, 310018, China.
[3]Lishui Key Laboratory of Low-Altitude Multimodal Sensing and Intelligent Computing, Lishui, 323000, China.

*Corresponding author(s). E-mail(s): yezhen@lsu.edu.cn;
Contributing authors: 2024110602006@mails.zstu.edu.cn;

Abstract

Although recent research on single-view 3D reconstruction has made substantial progress, accurately reconstructing complex, large-scale outdoor environments remains challenging. We present PanoramicSplat, a 3D Gaussian Splatting (3DGS) framework for monocular large-scale scene reconstruction. The framework introduces three components: panoramic supervision, which imposes cyclic multi-view consistency during training; semantic-aware depth regularization, which incorporates semantic masks and depth estimates into Gaussian optimization; and a radially weighted photometric loss combined with Gaussian pruning, which emphasizes informative image regions while suppressing redundant primitives. We also construct a dataset of large-scale outdoor scenes using Google Earth Studio, covering diverse urban landscapes and occluded structures. On the reported rendered-scene benchmark, PanoramicSplat achieves 19.28 PSNR, 0.4182 SSIM, and 0.4536 LPIPS, outperforming the evaluated retrained baselines.

Keywords: 3D gaussian splatting, Single-view reconstruction, Outdoor, Large-scale reconstruction

1 Introduction

3D reconstruction, which infers three-dimensional geometric structures from two-dimensional data, has become a cornerstone technology in computer vision and

graphics. Its applications extend across a wide range of critical domains, including autonomous driving, augmented and virtual reality (AR/VR), digital cultural heritage preservation, and smart city planning [1]. Traditional multi-view stereo (MVS) methods achieve millimeter-level precision by triangulating scene geometry from multiple calibrated viewpoints [2, 3]. However, they necessitate dense image sampling and involve significant computational overhead, making them impractical for resource-constrained scenarios. In contrast, single-view 3D reconstruction offers a compelling alternative by inferring complete 3D representations from a single monocular image, making it particularly valuable in data-scarce environments where multi-view capture is infeasible [4–8].

Recent advancements in deep learning and generative models have driven significant progress in single-view 3D reconstruction [9–11]. Diffusion-based approaches employ pre-trained 2D diffusion models to enable novel view synthesis through iterative denoising, bridging 2D generative capabilities with 3D reconstruction tasks [12–14]. While these methods produce visually plausible results, they often struggle with geometric consistency across large viewpoint changes due to their lack of explicit 3D scene understanding. Additionally, some studies employ Neural Radiance Fields (NeRF) [15, 16] to represent scenes as continuous volumetric functions using neural implicit modeling and differentiable rendering, achieving photorealistic quality but at a high computational cost. Recently, 3D Gaussian Splatting (3DGS) [17] has emerged as a computationally efficient paradigm, combining real-time rendering speeds with high-fidelity geometric reconstruction. Pioneering works have successfully extended 3DGS to monocular 3D reconstruction. GaussianAnything [18] introduces a novel point cloud-structured 3D latent representation, which, combined with a VAE and a cascaded diffusion model, enables high-quality and interactive 3D object generation conditioned on multi-modal inputs such as images and text. TGS [19] designs a hybrid triplane-Gaussian representation that synergistically integrates 3DGS for coarse geometric modeling with neural triplane encoding to simultaneously refine structural details and parameterize critical 3D Gaussian attributes, enabling photorealistic single-view reconstruction.

In spite of these breakthroughs, existing approaches predominantly focus on reconstructing individual objects or small-scale scenes [20, 21]. Therefore, their generalizability to novel view synthesis tasks in expansive outdoor scenes remains constrained by three critical scientific barriers. First, the coupling effects of multi-scale spatial attributes and heterogeneous geometric compositions (e.g., architectural clusters, transportation networks, and vegetation) in large scenes makes it challenging for conventional methods to simultaneously maintain global topological coherence and local geometric accuracy. Second, the dual limitations of texture feature degradation and parallax cue absence in distant regions induce substantial cumulative inaccuracies in monocular depth estimation, invalidating conventional geometric prior assumptions in open-world settings. Third, the scarcity of high-quality datasets for large-scale outdoor scenes hinders further advancement in this area [22–24].

To address these challenges, we propose a single-view large-scale outdoor reconstruction framework based on 3DGS. Our approach uses a UNet-based architecture [25] to predict 3D Gaussian parameters. Panoramic supervision encourages geometric

consistency across sampled front, back, left, and right viewpoints, thereby providing multi-view constraints during training. We further incorporate semantic masks and depth estimates to impose object-aware geometric regularization and preserve scene boundaries. A radially weighted photometric loss emphasizes central image regions while assigning lower weights to peripheral regions, where monocular geometric estimates are often less reliable. We also construct a dataset using Google Earth Studio, including aerial views of urban architecture, industrial complexes, and natural landscapes across 20 global cities.

To the best of our knowledge, this study is among the first to examine single-image 3D reconstruction of complex, large-scale outdoor scenes with an explicit Gaussian representation. On the reported benchmark, PanoramicSplat improves PSNR by 1.17–1.45 dB, SSIM by 0.0305–0.0318, and reduces LPIPS by 26.2–45.2% relative to the evaluated retrained baselines. These results suggest that the proposed framework is a promising approach for large-scale digital-twin construction, although its performance remains dependent on the quality of the estimated depth, semantic masks, and training data. Our contributions are:

1. We construct a large-scale outdoor dataset, providing a comprehensive benchmark for large-scale outdoor reconstruction and fostering further research in this field.
2. We propose a multi-model panoramic supervision paradigm that simultaneously enforces both object-level consistency and cross-view geometric alignment.
3. We develop a radially weighted photometric loss coupled with a Gaussian pruning mechanism that emphasizes central regions and suppresses redundant primitives, thereby significantly enhancing reconstruction fidelity.

2 Related work

2.1 3D Scene Representation

Existing 3D reconstruction techniques can be broadly categorized into implicit and explicit approaches. Traditional methods relied on explicit representations such as voxel grids, point clouds, and polygon meshes, which directly encode spatial occupancy but face scalability limitations in high-resolution scenarios.

The emergence of NeRF [15] revolutionized the field of 3D scene representation by modeling scenes as continuous volumetric functions, enabling photorealistic novel view synthesis. Subsequent works have introduced significant improvements, pushing the boundaries of performance across various aspects. For instance, Mip-NeRF [26] addresses NeRF's limitations in handling multi-scale rendering by modeling camera frustums as conical Gaussians rather than single rays. This formulation enables anti-aliasing across all viewing distances, reduces image reconstruction errors, and accelerates convergence by integrating positional encoding and optimizing volumetric sampling. Instant-NGP [27] achieves real-time training and rendering by employing a multi-resolution hash encoding for efficient spatial representation and a lightweight fully connected neural network. This design facilitates rapid convergence and high-quality rendering with a compact model size, making it highly suitable for real-time

neural rendering applications. SparseNeRF [28] further advances the field by distilling robust 3D priors from coarse depth maps, effectively compensating for the lack of constraints in sparse-view settings and bridging the gap between sparse inputs and high-fidelity 3D reconstructions.

While NeRF has demonstrated impressive novel view synthesis capabilities, its rendering speed remains a significant limitation. Recent advancements have explored alternative representations, among which 3DGS [17] stands out as the most notable, offering real-time rendering while preserving high visual fidelity. This innovative technique models scenes as a collection of continuous, differentiable 3D Gaussian primitives, where each Gaussian primitive explicitly encodes local geometric and photometric properties, coupled with a differentiable rasterization pipeline to achieve real-time rendering performance while maintaining view-consistent scene representation. Building upon this foundation, innovations such as 2DGS [29], GSDF [30], and 3DGS-Enhancer [31] have further expanded the scene representation capabilities and reconstruction performance of 3DGS. Moreover, 3DGS demonstrated remarkable potential across a diverse range of applications, including dynamic scene reconstruction, sparse-view synthesis, interactive 3D scene editing, and robotic SLAM [32–38].

2.2 Single-view 3D Reconstruction

Single-view 3D reconstruction aims to infer 3D structures from 2D images, a task inherently ill-posed due to depth ambiguity [39]. Early works like 3D-R2N2 [40] used recurrent networks to learn a mapping from images of objects to their underlying 3D shapes from a large collection of synthetic data to aggregate multi-view predictions, while PSGN [41] generated point clouds corresponding to the input image via encoder-decoder architectures. Subsequently, the introduction of NeRF inspired works such as PixelNeRF [16], SinNeRF [42], and MVSNeRF [43]. By integrating neural networks for feature extraction with implicit 3D representations, these methods have enabled more efficient 3D reconstruction from single or sparse-view inputs, improving the modeling of complex geometry and textures. Moreover, while 3DGS is still in its early stages of development, research on single-view reconstruction based on this approach has shown promising results and made significant progress. Splatter Image [44] directly predicting the 3D Gaussian parameters corresponding to each pixel through an image-to-image UNet, achieving a real-time reconstruction speed of 38 FPS. Gamba [45] presents an end-to-end trained single-view 3D reconstruction framework leveraging the linear-time Mamba network to enable fast and high-quality reconstruction. Additionally, PixelSplat [46], designed for image pair inputs, mitigates the local minima issue inherent in sparse representations by predicting a dense 3D probability distribution and sampling Gaussian means accordingly.

While existing approaches have made notable progress in single-view reconstruction of individual objects or small-scale scenes, the inherent complexity of large-scale, real-world outdoor environments and the scarcity of relevant datasets make it challenging to directly extend these methods to novel view synthesis for expansive scenes. Consequently, to overcome these limitations, we propose a novel method for single-view reconstruction of large-scale outdoor scenes based on 3DGS.

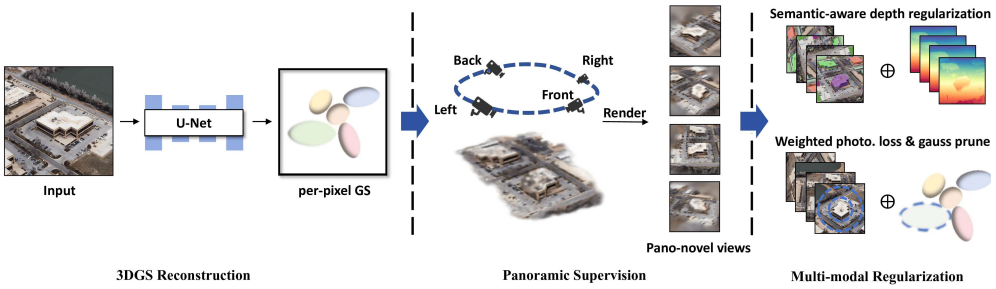


Fig. 1 Pipeline of the proposed method. A single input image is processed by a UNet to regress pixel-aligned 3D Gaussian parameters, which are then splatted and rendered from multiple panoramic viewpoints. Panoramic supervision—including semantic-aware depth and radially weighted photometric losses—guides the network toward both global geometric consistency and high-fidelity local detail.

3 Method

In this section, we discuss PanoramicSplat in detail. The full pipeline is illustrated in Fig. 1. Inspired by Splatter Image [44], we treat monocular reconstruction as a standard image-to-image task where each pixel of the input image is associated with a single Gaussian output. Formally, let i denote a pixel within an $H \times W$ image. The neural architecture f_θ, parameterized by θ, processes an input RGB image $I \in \mathbb{R}^{H \times W \times 3}$ and directly generates an output tensor $O \in \mathbb{R}^{H \times W \times K}$, where each pixel i corresponds to a K-dimensional latent vector encoding the parameters of a multivariate Gaussian distribution.

3.1 Preliminary

In our approach, the scene is represented as a collection of Gaussians $\mathcal{G} = \{(\alpha_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \mathbf{c}_i) \mid i = 1, \ldots, G\}$, where $\boldsymbol{\mu}_i \in \mathbb{R}^3$ is the center, $\alpha_i \in [0, 1]$ is the opacity, $\boldsymbol{\Sigma}_i \in \mathbb{R}^{3\times3}$ is an anisotropic covariance matrix describing the spatial extent and shape of the splat, and $\mathbf{c}_i \in \mathbb{R}^3$ is the appearance feature represented by spherical harmonics of degree 1. The Gaussians are projected onto the image plane through differentiable rasterization, yielding an explicit scene representation that can be optimized end-to-end [47, 48].

3.2 Dataset Construction

To facilitate research in large-scale outdoor single-view reconstruction and establish a standardized benchmark, we have constructed a comprehensive dataset of urban scenes and will make it publicly available. Our dataset is unique in its scale, diversity, and high-fidelity geospatial realism, and is designed to support both training and evaluation of 3D reconstruction algorithms.

We selected 20 representative metropolitan areas across five continents, including New York, Boston, Toronto, and London. Each city contributes 10–15 distinct scenes, capturing urban typologies such as dense downtown districts, waterfront promenades,

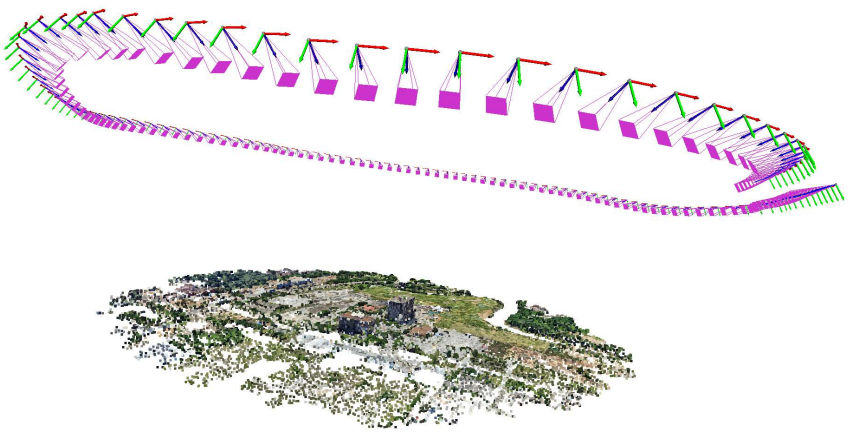


Fig. 2 Data preparation. High-resolution images are rendered using Google Earth Studio to capture diverse urban scenes from 20 global cities. For each scene, camera parameters are configured to mimic a drone's circular aerial trajectory and provide broad viewpoint coverage. Each scene consists of 70–130 images at a resolution of 1024×1024, with camera intrinsics and extrinsics estimated using COLMAP for subsequent reconstruction and evaluation.

industrial zones, and historical neighborhoods. We use Google Earth Studio's geospatial animation toolkit to simulate a drone-like circular trajectory for each scene, as illustrated in Fig. 2. Keyframe parameters, including camera latitude, longitude, yaw, pitch, and roll, were manually configured to provide uniform angular coverage (360 degrees in azimuth and ± 10 degrees in elevation). For each scene, we rendered 70–130 frames at a resolution of 1024×1024, yielding approximately 25,000 high-resolution RGB images. To obtain reference camera poses, we applied COLMAP's Structure-from-Motion and Multi-View Stereo pipelines [49, 50] and extracted the intrinsic and extrinsic parameters for each frame. We divided the 250 scenes into training, validation, and test sets at a ratio of 9:1:1.

3.3 Panoramic Supervision

To mitigate the inherent ambiguity in large-scale single-view 3D reconstruction and enhance multi-view consistency, we introduce a panoramic supervision strategy that integrates implicit multi-perspective constraints during training. This method dynamically evaluates the geometric plausibility of the predicted 3DGS model through strategic viewpoint sampling. Specifically, for each training scene, a virtual camera rig is established around the reconstruction target, aligned with the principal axes of the world coordinate system. The scene is then divided into four quadrants—front, back, left, and right—each spanning a 90° interval. For each quadrant, a candidate viewpoint is randomly sampled from the available viewpoints, yielding four selected viewpoints:

$$\mathcal{V} = \{v_f, v_b, v_l, v_r\} \tag{1}$$

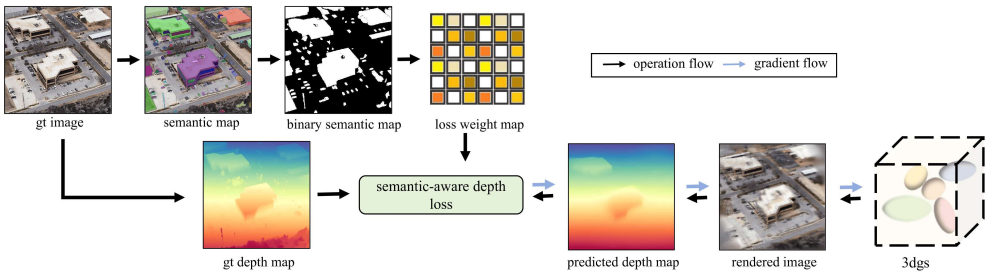


Fig. 3 Semantic-aware depth regularization mechanism. For each target image, SAM 2 generates instance masks with associated confidence scores. The masks are then fused to form semantic constraints that are incorporated into a boundary-aware depth loss. This formulation emphasizes the preservation of foreground details and object boundaries.

For a given input image I, the 3DGS model synthesizes novel views $\hat{I}_v$ corresponding to each sampled viewpoint $v \in \mathcal{V}$. The overall training loss is defined as the average panoramic differences across the four viewpoints:

$$\mathcal{L}_{reconstruction} = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \mathcal{L}_{\text{pan}}(\hat{I}_v, I_v) \tag{2}$$

where I_v denotes the ground truth image from viewpoint v and $\mathcal{L}_{\text{pan}}(\cdot)$ represents the combined multi-modal geometric loss.

This strategy provides a broad panoramic supervision signal, encouraging the 3DGS model to learn diverse viewpoint constraints and reducing overfitting to a single viewing direction. At inference time, only one input image is required; during training, the additional rendered views and their targets are obtained from the captured scene sequences.

3.4 Semantic-aware Depth Regularization

In complex large-scale scenes, the coupling effects of multi-scale spatial attributes and heterogeneous geometric compositions make it challenging for conventional methods to effectively balance global topological consistency with local geometric fidelity. To resolve this fundamental challenge, we design a dual regularization approach that incorporates semantic awareness and depth information into 3DGS optimization, thereby synergistically integrating object-level semantic constraints with geometric depth alignment. Fig. 3 illustrates the overall regularization pipeline.

For each image I_v during optimization, we use the Segment Anything Model 2 (SAM 2) [51] to establish object-level semantic constraints. SAM 2 produces K candidate instance masks with confidence scores:

$$\mathcal{M} = \{(M_k, s_k^{\text{iou}}, s_k^{\text{stab}})\}_{k=1}^{K} \tag{3}$$

where $M_k \in \{0,1\}^{H \times W}$ denotes the binary mask, $s_k^{\text{iou}} \in [0,1]$ is the predicted IoU score, and $s_k^{\text{stab}} \in [0,1]$ is the edge stability score. We then aggregate valid masks

through dual-threshold filtering and logical union:

$$M_{\text{fused}}(p) = \max_{1 \le k \le K} \left\{ M_k(p)\, \mathbb{I}\left(s_k^{\text{iou}} > \tau_{\text{iou}} \wedge s_k^{\text{stab}} > \tau_{\text{stab}}\right) \right\}, \qquad p \in \{1, \ldots, H\} \times \{1, \ldots, W\}, \tag{4}$$

where τ_{iou} and τ_{stab} are the thresholds to eliminate low-confidence predictions. Following this, semantic constraints are incorporated into the depth regularization via a boundary-aware depth loss:

$$\mathcal{L}_{\text{depth}} = \alpha \cdot \frac{1}{\max(N_{\text{fg}}, \varepsilon)} \sum_{M_{\text{fused}}=1} \Delta D^2 + \beta \cdot \frac{1}{\max(N_{\text{bg}}, \varepsilon)} \sum_{M_{\text{fused}}=0} \Delta D^2 \tag{5}$$

where α and β are the foreground and background weight coefficients, $\Delta D = D_{\text{render}} - D_{\text{gt}}$ denotes the depth error between the rendered depth map and the ground truth depth map, and

$$N_{\text{fg}} = \sum_p \mathbb{I}(M_{\text{fused}}(p) = 1), N_{\text{bg}} = \sum_p \mathbb{I}(M_{\text{fused}}(p) = 0) \tag{6}$$

represents the total number of foreground and background pixels separately. This formulation effectively balances the contributions of foreground and background regions, improving depth accuracy while preserving sharp object boundaries.

3.5 Gaussian Pruning

In our single-view 3D reconstruction framework, our network produces a dense grid of Gaussians (e.g., 512×512) for the input image. However, not all these Gaussians contribute meaningfully to the reconstruction task. To encourage sparsity, we propose incorporating a gating mechanism with L1 regularization to dynamically suppress low-importance Gaussians.

Let the network output at each pixel be decomposed into two parts: the Gaussian parameters $\tilde{G}(x, y)$ and a corresponding gating coefficient $m(x, y)$. The final output Gaussian at pixel (x, y) is defined as:

$$G(x, y) = m(x, y) \cdot \tilde{G}(x, y), \tag{7}$$

where $m(x, y) \in [0, 1]$ is obtained by applying a sigmoid activation. This coefficient effectively modulates the contribution of each Gaussian in the reconstruction process. Then, the overall loss function is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{reconstruction}} + \lambda \sum_{x,y} |m(x, y)| \tag{8}$$

where $\mathcal{L}_{\text{reconstruction}}$ is the reconstruction loss, λ is a hyperparameter balancing reconstruction accuracy and sparsity. This term drives the gating coefficients corresponding to less useful Gaussians toward zero, effectively reducing their impact during training.

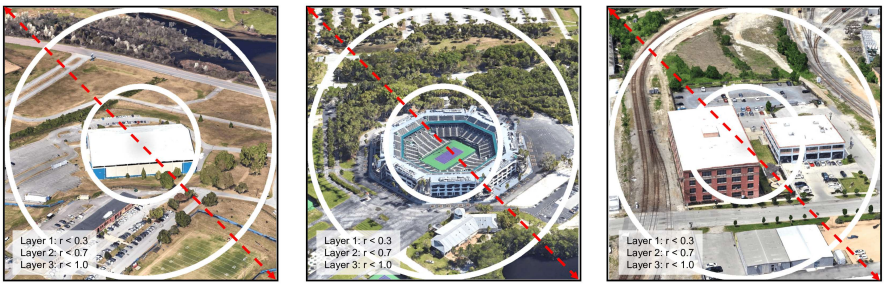


Fig. 4 This figure visualizes the design of the radially weighted photometric loss. The loss function partitions the image into three concentric zones based on their Euclidean distance from the center: an inner circle with full weight, a transitional annulus with moderate decay, and an outer region with reduced weight. This hierarchical weighting emphasizes reconstruction fidelity in the image center—where critical geometric cues reside—while relaxing constraints toward the periphery.

3.6 Model Training

The SongUNet [52] is used to predict 3DGS from the input image, with its final layer replaced by a 1×1 convolutional layer that outputs K channels. Given an $H\times W$ input, the network predicts an $H \times W$ grid of Gaussians, with each Gaussian parameterized by $\{\alpha, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{c}\}$. The rendered views of the predicted 3DGS are compared with target observations to optimize the Gaussian parameters.

Radially Weighted Photometric Loss. To address structural ambiguity in large-scale outdoor reconstruction from single-view input, we introduce a radially weighted photometric loss that emphasizes central image regions rich in critical geometric cues. As depicted in Fig. 4, our design is based on two key observations: (1) central image regions (covering 30 to 40% of FOV) predominantly contain essential structures crucial to scene understanding; (2) Peripheral areas (beyond 70% FOV) generally impose weaker geometric constraints and contribute less to the scene. To capture these spatial variations, our loss function segments pixels into three concentric zones based on their Euclidean distance from the image center: an inner circle (radius $\leq$ 30% of the diagonal length) with full weight ($\alpha = 1.0$), a transitional annulus (30%–70% of the diagonal) with moderate decay ($\beta = 0.5$), and an outer region (¿70% of the diagonal) with reduced weight ($\gamma = 0.25$). This hierarchical weighting strategy enhances gradients from geometrically informative central regions while suppressing noise from remote regions, which can be formalized as:

$$\begin{aligned} \mathcal{L}_{\text{photo}} &= \frac{\sum_{i,j} W(i,j) \|\hat{I}(i,j) - I(i,j)\|_2^2}{\max\left(\sum_{i,j} W(i,j), \varepsilon\right)} \\ \text{where} \quad W(i,j) &= \begin{cases} \alpha & \text{if } (i-c_x)^2 + (j-c_y)^2 \leq r_1^2, \\ \beta & \text{if } r_1^2 < (i-c_x)^2 + (j-c_y)^2 \leq r_2^2, \\ \gamma & \text{otherwise.} \end{cases} \end{aligned} \tag{9}$$

where i and j are pixel indices, (c_x, c_y) is the image center, $\hat{I}$ is the rendered image, and I is the target image. The normalization by the total weight keeps the loss scale approximately invariant to image resolution and the chosen radial partitions.

Loss Function. Our final loss follows the panoramic supervision framework, integrating the previously discussed semantic-aware depth loss, photometric loss, and L1 regularization, as formulated below:

$$\begin{aligned} \mathcal{L}_{\text{total}} &= \mathcal{L}_{\text{reconstruction}} + \lambda \sum_{x,y} |m(x,y)|, \\ \mathcal{L}_{\text{reconstruction}} &= \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \mathcal{L}_{\text{pan}}(\hat{I}_v, I_v) \\ &= \lambda_1 \mathcal{L}_{\text{photo}} + \lambda_2 \mathcal{L}_{\text{lpips}} + \lambda_3 \mathcal{L}_{\text{depth}}. \end{aligned} \tag{10}$$

4 Experiments

4.1 Implementation Details

We conducted our model training using a single NVIDIA A800 GPU. To improve computational efficiency and reduce memory consumption, we downsampled the input image resolution from 1024×1024 to 512×512. Depth information was extracted using Depth Anything v2 [53]. For semantic weighting, the foreground and background weight coefficients were set to 0.8 and 0.2, respectively. Furthermore, the loss function parameters λ_1, λ_2, and λ_3 were 0.8, 0.01, and 0.2, ensuring balanced optimization across various objectives.

4.2 Datasets

As mentioned previously, the training dataset was developed through Google Earth Studio simulations, facilitating the generation of diverse urban configurations spanning multiple environmental conditions and lighting variations. For rigorous evaluation of model generalizability, we implemented a systematic validation framework utilizing previously unseen real-world urban imagery acquired through drone-based aerial photography conducted at 80-meter altitude. This operational altitude was strategically maintained to ensure optimal balance between high spatial resolution (achieving ground resolution better than 5cm/pixel) and comprehensive geospatial coverage.

4.3 Baseline

We conducted a comparative analysis against representative state-of-the-art reconstruction methods using a two-stage evaluation protocol. First, we directly evaluated pretrained models from Splatter Image [44], CRM [21], InstantMesh [54], GaussianAnything [18], and Real3D [20] on our dataset. We then performed controlled retraining experiments on two representative 3DGS architectures, Splatter Image and Flash3D [55], using the same training data and experimental settings. All implementations used official codebases with only the adaptations required for dataset compatibility. We evaluated novel-view synthesis using PSNR, SSIM [56], and LPIPS.

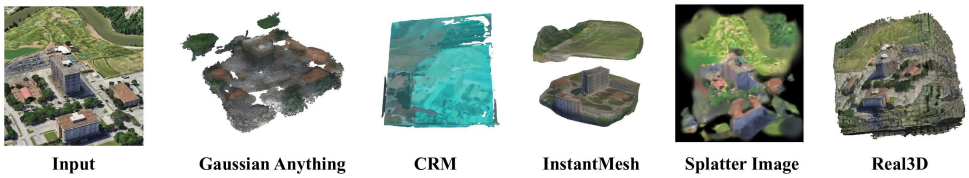


Fig. 5 Reconstructions obtained directly using pretrained models from baselines. Single-view reconstruction models designed and trained for small-scale scenes are challenging to directly apply to large-scale environments.

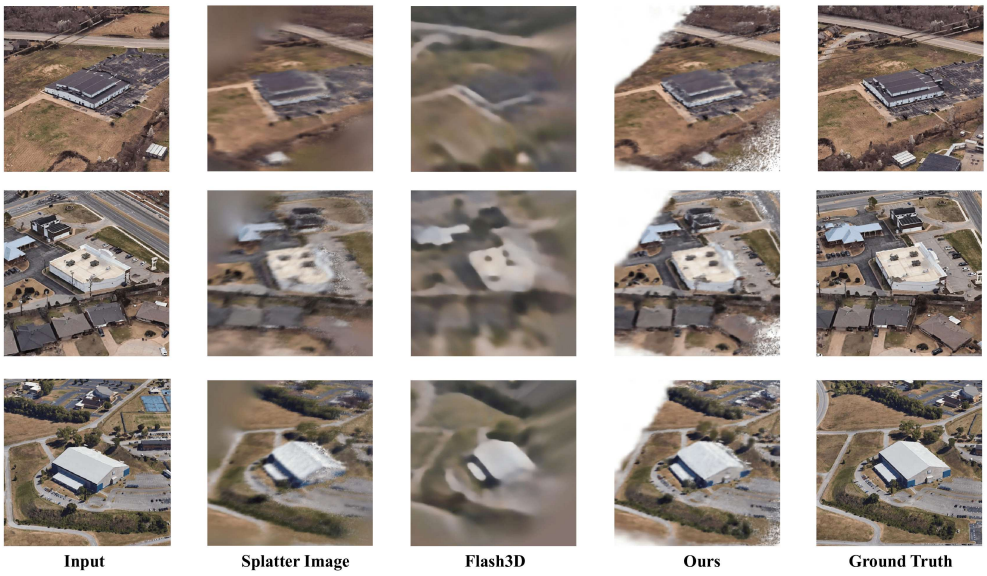


Fig. 6 Qualitative comparison of reconstructed 3D scenes produced by PanoramicSplat and representative state-of-the-art methods. The examples illustrate the recovery of geometric details and textures in complex urban environments.

4.4 Results

Qualitative Comparisons. Fig. 5 presents reconstructions obtained directly from the pretrained baselines. Models trained primarily on smaller scenes degrade when transferred to complex outdoor environments, producing inaccurate depth estimates and incomplete or distorted geometry. A controlled comparison after retraining on our dataset is shown in Fig. 6. Splatter Image and Flash3D produce more plausible reconstructions after retraining, while PanoramicSplat preserves more detailed boundaries and textures in the displayed examples.

Quantitative Comparisons. Table 1 reports PSNR, SSIM, LPIPS, and inference time for the compared methods. PanoramicSplat achieves the best results among the

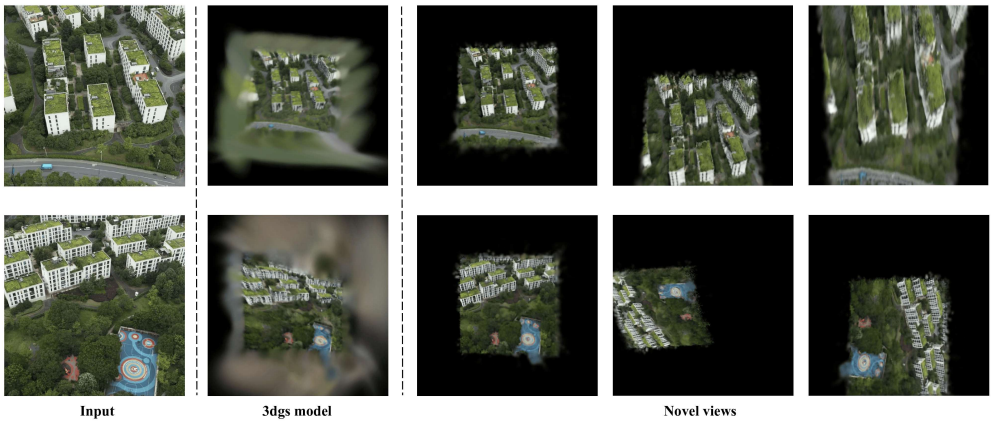


Fig. 7 Qualitative results on unseen real-world data. Our method demonstrates strong generalization capabilities by accurately reconstructing complex 3D scenes from real-world drone-captured images.

retrained methods, with 19.28 PSNR, 0.4182 SSIM, and 0.4536 LPIPS. Relative to Flash3D, it improves PSNR by 1.17 dB and reduces LPIPS by 37.6%; relative to Splatter Image, it improves PSNR by 1.45 dB and reduces LPIPS by 26.2%. The method also outperforms both zero-shot baselines: compared with Gaussian Anything, it improves PSNR by 3.66 dB and SSIM by 0.1062 while reducing LPIPS by 0.3609; compared with Real3D, the corresponding differences are 3.04 dB, 0.0837, and 0.4226. Its inference time is 1.25 s, comparable to Splatter Image and higher than Flash3D under the reported implementation.

Table 1 Quantitative results. PanoramicSplat achieves the best reconstruction metrics among the retrained methods and outperforms the reported zero-shot methods on all three quality metrics. Inference time is reported for methods for which it was available.

Category	Method	PSNR ↑	SSIM ↑	LPIPS ↓	Inference Time (s)
Retrained	Flash3D	18.11	0.3877	0.7273	0.46
	Splatter Image	17.83	0.3864	0.6149	1.25
	Ours	**19.28**	**0.4182**	**0.4536**	1.25
Zero-shot	Gaussian Anything	15.62	0.3120	0.8145	–
	Real3D	16.24	0.3345	0.8762	–

Performance on Unseen Real-world Data. To evaluate the generalization ability, we collected real-world image data using drones and directly tested our model on this dataset. As shown in Fig. 7, our method robustly handles unseen, real-world scenarios, accurately reconstructing complex 3D scenes. The reconstructions maintain high fidelity in terms of geometric structure and texture, underscoring the practical applicability of our approach in diverse and challenging environments.

4.5 Ablation Study

Table 2 Ablation study results showing the impact of each component. Removing semantic-aware depth regularization, radial photometric weighting, panoramic supervision, or Gaussian pruning degrades at least one of the reported quality metrics.

	PSNR ↑	SSIM ↑	LPIPS ↓
Full model	**11.93**	**0.342**	**0.5766**
w/o sem-depth	10.92	0.308	0.6504
w/o w. photo. loss	10.52	0.305	0.5822
w/o pano. sup.	11.90	0.285	0.5805
w/o gauss prune.	11.07	0.322	0.6518

To assess the impact of each component, we performed an ablation study by systematically removing individual modules while preserving all other settings. To improve efficiency, we used a reduced training schedule for these experiments; therefore, the absolute values in Table 2 should be compared within the ablation study rather than directly with the full-training results in Table 1. The quantitative results in Table 2 and qualitative comparisons in Fig. 8 demonstrate the contributions of the proposed components.

Semantic-aware Depth Regularization (w/o sem-depth): Removing this component caused a substantial performance degradation (-1.01 PSNR, -0.034 SSIM, +0.0738 LPIPS). As shown in Fig. 8(b), the absence of semantic guidance led to structural ambiguities in building facades and vegetation boundaries, while depth inconsistencies became evident in road surfaces and distant objects.

Radially Weighted Photometric Loss (w/o w. photo. loss): Eliminating the adaptive weighting strategy resulted in the largest PSNR decrease (-1.41) and a reduction of 0.037 in SSIM. LPIPS increased by 0.0056. The reconstruction in Fig. 8(c) exhibits over-smoothed textures in central regions while accumulating artifacts in peripheral areas.

Panoramic Supervision (w/o pano. sup.): Without multi-view consistency constraints, PSNR decreased slightly by 0.03, while SSIM decreased by 0.057 and LPIPS increased by 0.0039. The qualitative results in Fig. 8(d) reveal distorted building geometries and floating artifacts in open spaces.

Gaussian Pruning (w/o gauss prune): Disabling the sparse regularization reduced PSNR by 0.86 and SSIM by 0.020, while increasing LPIPS by 0.0752. Fig. 8(e) demonstrates how unpruned Gaussians caused texture bleeding between adjacent structures.

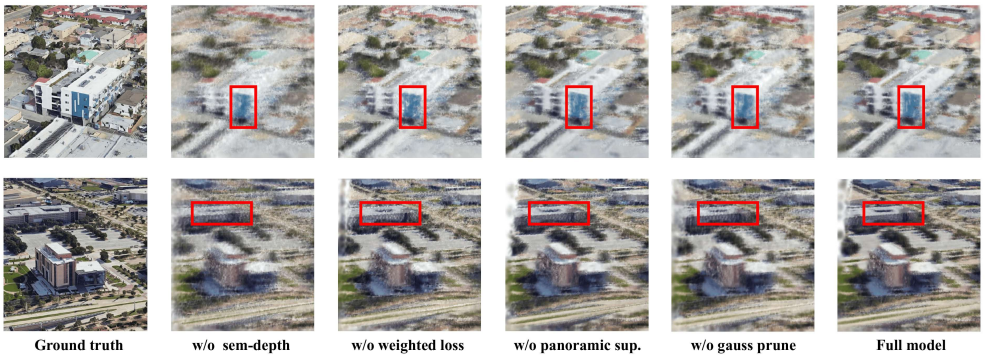


Fig. 8 Ablation experiment results. From the images, it is evident that when key components are removed, the reconstruction quality significantly degrades, resulting in missing details, structural inconsistencies, and blurred textures. In contrast, the complete model demonstrates superior performance in reconstructing fine details and maintaining global consistency, thereby validating the critical role of each module in enhancing the quality of large-scale single-view 3D reconstruction.

5 Discussion

5.1 Geometry-prior Gaussian Initialization

In our current framework, the position parameters of Gaussian primitives are initialized using a purely random strategy. While this unbiased initialization approach ensures methodological generality, it suffers from two critical limitations in large-scale urban reconstruction scenarios: 1) we observed that approximately 30% of the total training time is consumed by coarse adjustments of primitive positions during the initial optimization phase, significantly compromising computational efficiency. 2) Randomly distributed Gaussian primitives struggle to quickly align with the rigid geometric features of building facades, often causing non-physical distortions in curved areas.

Inspired by hybrid methods like MASt3R [57], which fuse multi-view stereo geometry with monocular depth estimation, we propose future research directions centered on hierarchical geometric prior integration. MASt3R is capable of directly regressing pixel-aligned point maps from raw images, which approximately represent the spatial coordinates corresponding to each pixel. These point maps can serve as an initialization basis for the Gaussian primitives. Specifically, the pixel-aligned point maps generated provide a hierarchical geometric structure that not only considers the overall shape of the buildings but also captures detailed local geometric information. This approach helps Gaussian primitives better adapt to the rigid geometric features of building facades, significantly reducing non-physical distortions, particularly in complex urban environments. Moreover, the introduction of geometric priors makes the initialization of the Gaussian primitives more precise, accelerates the model's convergence, and improves the efficiency and accuracy of training.

5.2 Towards Sparse-view Generalization

The inherent geometric constraints in our panoramic supervision paradigm establish a theoretical foundation for extending single-view reconstruction to sparse-view scenarios. The cyclic consistency mechanism, originally designed to resolve monocular ambiguities through virtual multi-perspective constraints, naturally aligns with the geometric regularization requirements in sparse-view settings. When applied to N input views, the panoramic consistency loss can be reformulated as a graph matching problem across view pairs. This formulation theoretically enables two key advantages: 1) The semantic-aware depth regularization acts as a topological prior that preserves object-level coherence across disjoint viewing frustums, mitigating the correspondence fragmentation problem in wide-baseline setups. 2) The radially weighted photometric loss induces a content-dependent attention mechanism, automatically emphasizing regions with overlapping visual coverage while downweighting occluded areas—a critical feature for sparse-view consistency.

6 Conclusion

This paper presents PanoramicSplat, a framework for single-view large-scale outdoor reconstruction using 3DGS. The method addresses depth ambiguity and geometric inconsistency by combining panoramic supervision, semantic-aware depth regularization, Gaussian pruning, and a radially weighted photometric loss. On the reported rendered-scene benchmark, PanoramicSplat achieves 19.28 PSNR, 0.4182 SSIM, and 0.4536 LPIPS, outperforming the compared retrained baselines on all three quality metrics. Its inference time is 1.25 s under the reported implementation, which is comparable to Splatter Image but slower than Flash3D. We also construct a large-scale outdoor dataset intended to support future research in this area. Further evaluation on diverse real-world scenes remains necessary to establish the method's generalization and practical utility.

Funding

This work is supported by the Zhejiang Provincial Natural Science Foundation of China (No.LTGN23F020001), the Baishanzu National Park Scientific Research Project (No.2021KFLY07), and the Public Welfare Research Plan of Lishui (No.2024GYX26).

Author contribution

Z. W. designed and completed the experiments, analyzed the results, and wrote the original draft. Z. Y. conceived the original study idea, designed the methodology, and performed the review and editing.

References

[1] Ma, Z., Liu, S.: A review of 3d reconstruction techniques in civil engineering and their applications. Advanced Engineering Informatics **37**, 163–174 (2018)

[2] Snavely, N., Seitz, S.M., Szeliski, R.: Photo tourism: exploring photo collections in 3d. In: ACM Siggraph 2006 Papers, pp. 835–846 (2006)

[3] Furukawa, Y., Ponce, J.: Accurate, dense, and robust multiview stereopsis. IEEE transactions on pattern analysis and machine intelligence **32**(8), 1362–1376 (2009)

[4] Tatarchenko, M., Richter, S.R., Ranftl, R., Li, Z., Koltun, V., Brox, T.: What do single-view 3d reconstruction networks learn? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)

[5] Fahim, G., Amin, K., Zarif, S.: Single-view 3d reconstruction: A survey of deep learning methods. Computers & Graphics **94**, 164–190 (2021)

[6] Shen, Q., Yang, X., Wang, X.: Anything-3d: Towards single-view anything reconstruction in the wild. arXiv preprint arXiv:2304.10261 (2023)

[7] Pontes, J.K., Kong, C., Sridharan, S., Lucey, S., Eriksson, A., Fookes, C.: Image2mesh: A learning framework for single image 3d reconstruction. In: Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part I 14, pp. 365–381 (2019). Springer

[8] Fu, K., Peng, J., He, Q., Zhang, H.: Single image 3d object reconstruction based on deep learning: A review. Multimedia Tools and Applications **80**(1), 463–498 (2021)

[9] Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., Zhang, W., Cui, B., Yang, M.-H.: Diffusion models: A comprehensive survey of methods and applications. ACM Computing Surveys **56**(4), 1–39 (2023)

[10] Tang, S., Chen, J., Wang, D., Tang, C., Zhang, F., Fan, Y., Chandra, V., Furukawa, Y., Ranjan, R.: Mvdiffusion++: A dense high-resolution multi-view diffusion model for single or sparse-view 3d object reconstruction. In: European Conference on Computer Vision, pp. 175–191 (2024). Springer

[11] Oussidi, A., Elhassouny, A.: Deep generative models: Survey. In: 2018 International Conference on Intelligent Systems and Computer Vision (ISCV), pp. 1–8 (2018). IEEE

[12] Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)

[13] Voleti, V., Yao, C.-H., Boss, M., Letts, A., Pankratz, D., Tochilkin, D., Laforte, C., Rombach, R., Jampani, V.: Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In: European Conference on Computer Vision, pp. 439–457 (2024). Springer

[14] Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9298–9309 (2023)

[15] Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)

[16] Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelnerf: Neural radiance fields from one or few images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4578–4587 (2021)

[17] Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4), 139–1 (2023)

[18] Lan, Y., Zhou, S., Lyu, Z., Hong, F., Yang, S., Dai, B., Pan, X., Loy, C.C.: Gaussiananything: Interactive point cloud latent diffusion for 3d generation. In: ICLR (2025)

[19] Zou, Z.-X., Yu, Z., Guo, Y.-C., Li, Y., Liang, D., Cao, Y.-P., Zhang, S.-H.: Triplane meets gaussian splatting: Fast and generalizable single-view 3d reconstruction with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10324–10335 (2024)

[20] Jiang, H., Huang, Q., Pavlakos, G.: Real3d: Scaling up large reconstruction models with real-world images. arXiv preprint arXiv:2406.08479 (2024)

[21] Wang, Z., Wang, Y., Chen, Y., Xiang, C., Chen, S., Yu, D., Li, C., Su, H., Zhu, J.: Crm: Single image to 3d textured mesh with convolutional reconstruction model. In: European Conference on Computer Vision, pp. 57–74 (2024). Springer

[22] Sitzmann, V., Zollhöfer, M., Wetzstein, G.: Scene representation networks: Continuous 3d-structure-aware neural scene representations. Advances in neural information processing systems **32** (2019)

[23] Reizenstein, J., Shapovalov, R., Henzler, P., Sbordone, L., Labatut, P., Novotny, D.: Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10901–10911 (2021)

[24] Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3d objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13142–13153 (2023)

[25] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-assisted

intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, pp. 234–241 (2015). Springer

[26] Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5855–5864 (2021)

[27] Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM transactions on graphics (TOG) **41**(4), 1–15 (2022)

[28] Wang, G., Chen, Z., Loy, C.C., Liu, Z.: Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9065–9076 (2023)

[29] Huang, B., Yu, Z., Chen, A., Geiger, A., Gao, S.: 2d gaussian splatting for geometrically accurate radiance fields. In: ACM SIGGRAPH 2024 Conference Papers, pp. 1–11 (2024)

[30] Yu, M., Lu, T., Xu, L., Jiang, L., Xiangli, Y., Dai, B.: Gsdf: 3dgs meets sdf for improved neural rendering and reconstruction. Advances in Neural Information Processing Systems **37**, 129507–129530 (2024)

[31] Liu, X., Zhou, C., Huang, S.: 3dgs-enhancer: Enhancing unbounded 3d gaussian splatting with view-consistent 2d diffusion priors. Advances in Neural Information Processing Systems **37**, 133305–133327 (2024)

[32] Sun, L.C., Bhatt, N.P., Liu, J.C., Fan, Z., Wang, Z., Humphreys, T.E., Topcu, U.: Mm3dgs slam: Multi-modal 3d gaussian splatting for slam using vision, depth, and inertial measurements. In: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 10159–10166 (2024). IEEE

[33] Wang, Y., Yi, X., Wu, Z., Zhao, N., Chen, L., Zhang, H.: View-consistent 3d editing with gaussian splatting. In: European Conference on Computer Vision, pp. 404–420 (2024). Springer

[34] Chen, Y., Xu, H., Zheng, C., Zhuang, B., Pollefeys, M., Geiger, A., Cham, T.-J., Cai, J.: Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In: European Conference on Computer Vision, pp. 370–386 (2024). Springer

[35] Zhang, B., Cheng, Y., Yang, J., Wang, C., Zhao, F., Tang, Y., Chen, D., Guo, B.: Gaussiancube: Structuring gaussian splatting using optimal transport for 3d generative modeling. arXiv e-prints, 2403 (2024)

[36] Bao, Y., Ding, T., Huo, J., Liu, Y., Li, Y., Li, W., Gao, Y., Luo, J.: 3d gaussian splatting: Survey, technologies, challenges, and opportunities. IEEE Transactions

on Circuits and Systems for Video Technology (2025)

[37] Yu, H., Gong, W., Chen, J., Ma, H.: Get3dgs: Generate 3d gaussians based on points deformation fields. IEEE Transactions on Circuits and Systems for Video Technology (2024)

[38] Guo, Z., Zhou, W., Li, L., Wang, M., Li, H.: Motion-aware 3d gaussian splatting for efficient dynamic scene reconstruction. IEEE Transactions on Circuits and Systems for Video Technology (2024)

[39] Liu, M., Xu, C., Jin, H., Chen, L., Varma T, M., Xu, Z., Su, H.: One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. Advances in Neural Information Processing Systems **36**, 22226–22246 (2023)

[40] Choy, C.B., Xu, D., Gwak, J., Chen, K., Savarese, S.: 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In: Computer vision–ECCV 2016: 14th European Conference, Amsterdam, the Netherlands, October 11-14, 2016, Proceedings, Part VIII 14, pp. 628–644 (2016). Springer

[41] Fan, H., Su, H., Guibas, L.J.: A point set generation network for 3d object reconstruction from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 605–613 (2017)

[42] Xu, D., Jiang, Y., Wang, P., Fan, Z., Shi, H., Wang, Z.: Sinnerf: Training neural radiance fields on complex scenes from a single image. In: European Conference on Computer Vision, pp. 736–753 (2022). Springer

[43] Chen, A., Xu, Z., Zhao, F., Zhang, X., Xiang, F., Yu, J., Su, H.: Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 14124–14133 (2021)

[44] Szymanowicz, S., Rupprecht, C., Vedaldi, A.: Splatter image: Ultra-fast single-view 3d reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10208–10217 (2024)

[45] Shen, Q., Wu, Z., Yi, X., Zhou, P., Zhang, H., Yan, S., Wang, X.: Gamba: Marry gaussian splatting with mamba for single view 3d reconstruction. arXiv preprint arXiv:2403.18795 (2024)

[46] Charatan, D., Li, S.L., Tagliasacchi, A., Sitzmann, V.: pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19457–19467 (2024)

[47] Meng, J., Li, H., Wu, Y., Gao, Q., Yang, S., Zhang, J., Ma, S.: Mirror-3dgs:

Incorporating mirror reflections into 3d gaussian splatting. In: 2024 IEEE International Conference on Visual Communications and Image Processing (VCIP), pp. 1–5 (2024). https://doi.org/10.1109/VCIP63160.2024.10849936

[48] Sun, X., Lee, J.C., Rho, D., Ko, J.H., Ali, U., Park, E.: F-3dgs: Factorized coordinates and representations for 3d gaussian splatting. In: Proceedings of the 32nd ACM International Conference on Multimedia, pp. 7957–7965 (2024)

[49] Schönberger, J.L., Zheng, E., Pollefeys, M., Frahm, J.-M.: Pixelwise View Selection for Unstructured Multi-View Stereo. In: European Conference on Computer Vision (ECCV) (2016)

[50] Schönberger, J.L., Frahm, J.-M.: Structure-from-Motion Revisited. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)

[51] Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)

[52] Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)

[53] Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10371–10381 (2024)

[54] Xu, J., Cheng, W., Gao, Y., Wang, X., Gao, S., Shan, Y.: Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. arXiv preprint arXiv:2404.07191 (2024)

[55] Szymanowicz, S., Insafutdinov, E., Zheng, C., Campbell, D., Henriques, J.F., Rupprecht, C., Vedaldi, A.: Flash3d: Feed-forward generalisable 3d scene reconstruction from a single image. In: 2025 International Conference on 3D Vision (3DV), pp. 670–681 (2025). https://doi.org/10.1109/3DV66043.2025.00067

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

[57] Leroy, V., Cabon, Y., Revaud, J.: Grounding image matching in 3d with mast3r. In: European Conference on Computer Vision, pp. 71–91 (2024). Springer