

# SkySplat: Generalizable 3D Gaussian Splatting from Multi-Temporal Sparse Satellite Images

Xuejun Huang<sup>1</sup>, Xinyi Liu<sup>1,2,3\*</sup>, Yi Wan<sup>1,2,3</sup>, Zhi Zheng<sup>4</sup>, Bin Zhang<sup>5</sup>,  
Mingtao Xiong<sup>1</sup>, Yingying Pei<sup>1</sup>, Yongjun Zhang<sup>1,2,3\*</sup>

<sup>1</sup>School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China

<sup>2</sup>Technology Innovation Center for Collaborative Applications of Natural Resources Data in GBA, Ministry of Natural Resources, Guangzhou 510075, China

<sup>3</sup>Hubei LuoJia Laboratory, Wuhan 430079, China

<sup>4</sup>Department of Geography and Resource Management, The Chinese University of Hong Kong, Shatin 999077 Hong Kong

<sup>5</sup>China Railway Siyuan Survey and Design Group Co., LTD, Wuhan 430063, China

{huangxuejun, liuxy0319, yi.wan, bin.zhang, xiongmingtao, painyy, zhangyj}@whu.edu.cn, zhizheng@cuhk.edu.hk

## Abstract

Three-dimensional scene reconstruction from sparse-view satellite images is a long-standing and challenging task. While 3D Gaussian Splatting (3DGS) and its variants have recently attracted attention for its high efficiency, existing methods remain unsuitable for satellite images due to incompatibility with rational polynomial coefficient (RPC) models and limited generalization capability. Recent advances in generalizable 3DGS approaches show potential, but they perform poorly on multi-temporal sparse satellite images due to limited geometric constraints, transient objects, and radiometric inconsistencies. To address these limitations, we propose SkySplat, a novel self-supervised framework that integrates the RPC model into the generalizable 3DGS pipeline, enabling more effective use of sparse geometric cues for improved reconstruction. SkySplat relies only on RGB images and radiometric-robust relative height supervision, thereby eliminating the need for ground-truth height maps. Key components include a Cross-Self Consistency Module (CSCM), which mitigates transient object interference via consistency-based masking, and a multi-view consistency aggregation strategy that refines reconstruction results. Compared to per-scene optimization methods, SkySplat achieves an 86 times speedup over EOGS with higher accuracy. It also outperforms generalizable 3DGS baselines, reducing MAE from 13.18 m to 1.80 m on the DFC19 dataset significantly, and demonstrates strong cross-dataset generalization on the MVS3D benchmark.

## Introduction

Three-dimensional scene reconstruction from sparse-view satellite images remains a fundamental challenge in photogrammetry and computer vision. This technique has a wide range of applications such as digital twins and urban planning (Zheng et al. 2024).

Recent advances in multi-view stereo (MVS) methods have demonstrated the potential of deep neural networks for accurate 3D reconstruction from satellite images (Gao, Liu, and Ji 2021, 2023). However, these methods typically rely

on fully supervised learning and often degrade on multi-temporal satellite images due to radiometric and seasonal variations. In contrast, NeRF-based methods (Mildenhall et al. 2021; Barron et al. 2021, 2022) have gained attention for 3D reconstruction without ground-truth height supervision. Variants like S-NeRF (Derksen and Izzo 2021) and Sat-NeRF (Marí, Facciolo, and Ehret 2022) have been proposed to address the challenges posed by multi-temporal satellite images. However, these methods are computationally expensive and require long training times (Charatan et al. 2024).

To improve efficiency, the original 3D Gaussian Splatting (3DGS) (Kerbl et al. 2023) has emerged as a novel 3D representation, enabling fast rendering. Several extensions (Aira, Facciolo, and Ehret 2025; Bai et al. 2025) for satellite images have been proposed to enhance reconstruction quality and training efficiency. However, they still require several minutes to reconstruct a  $256 \times 256$  m<sup>2</sup> scene through per-scene optimization and rely on a large number of input views.

More recently, data-driven generalizable 3DGS methods have been developed to improve transferability and reconstruction speed. These approaches use neural networks to directly infer per-pixel Gaussian splatting parameters for unseen scenes (Xu et al. 2025). Representative examples include MVSplat (Chen et al. 2024), TranSplat (Zhang et al. 2025), and HiSplat (Tang et al. 2024), which achieve efficient 3D reconstruction from sparse-view images in a single feed-forward pass. However, these methods are not directly applicable to multi-temporal satellite images due to several challenges: (1) the unique pushbroom imaging mode of satellites, which violates the standard pinhole camera assumption; (2) interference from transient objects, such as moving vehicles and vegetation changes; and (3) radiometric inconsistencies across multi-temporal images caused by variations in illumination and atmospheric conditions.

To address these challenges, we propose SkySplat, a novel self-supervised framework for generalizable 3D reconstruction from multi-temporal sparse satellite images. SkySplat is the first to explicitly incorporate the satellite-specific rational polynomial coefficient (RPC) model into the generalizable 3DGS pipeline, enabling accurate scene reconstruction

\*Corresponding author.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

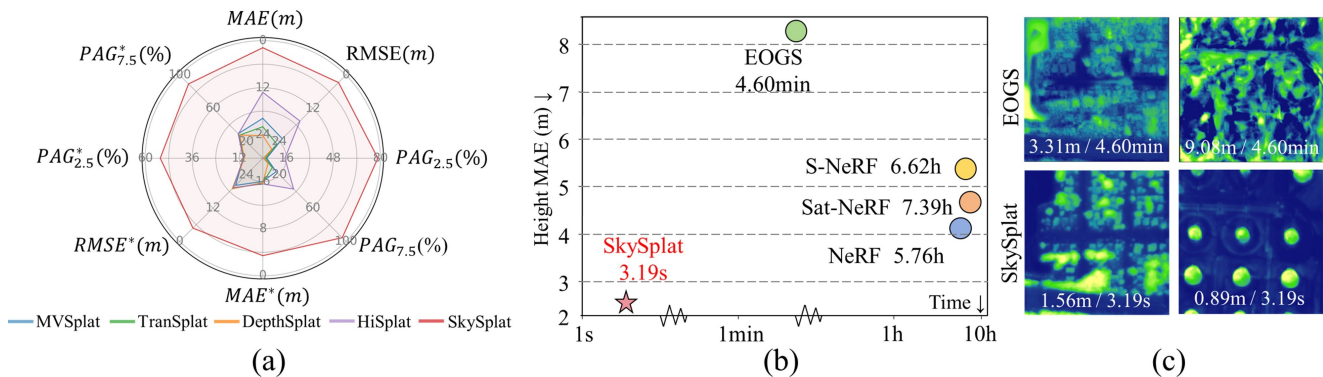


Figure 1: **Comparison with existing methods.** (a) Generalization results on the DFC19 (Bosch et al. 2019; Le Saux et al. 2019) and MVS3D (Bosch et al. 2016) datasets. Results from the MVS3D dataset are marked with \*. SkySplat achieves the best performance among all competitors. (b) Per-scene optimization results. SkySplat reaches optimal performance in just 3.19 seconds. (c) DSM reconstruction results on two areas of interest (AOIs). We report both MAE and reconstruction time.

tion without pinhole approximations. Extensive experiments show that SkySplat achieves significantly higher accuracy than previous generalizable methods while being 86× faster than SOTA per-scene optimization approaches (see Figure 1). Our main contributions can be summarized as follows:

- We propose SkySplat, the first generalizable 3DGS framework that incorporates the RPC model without approximation, achieving accurate scene reconstruction and an 86× speedup over the SOTA per-scene optimization method EOGs.
- We design a Cross-Self Consistency Module (CSCM) to minimize the impact of transient objects during training, and incorporate monocular relative heights as additional supervision to address radiometric inconsistencies in multi-temporal satellite images.
- We introduce a multi-view consistency aggregation strategy to further enhance reconstruction accuracy.

## Related Work

### NeRF and 3DGS for Satellite Images

In recent years, NeRF (Mildenhall et al. 2021) and 3DGS (Kerbl et al. 2023) have made significant progress in 3D scene reconstruction. However, the unique characteristics of satellite images—such as the RPC model and radiometric inconsistencies—pose substantial challenges for reconstruction. To address these issues, several methods (Derksen and Izzo 2021; Marí, Facciolo, and Ehret 2022, 2023) have extended NeRF to satellite domains by improving light transport models, ray sampling strategies based on the RPC model, and incorporating shadow modeling techniques. Subsequent efforts further enhanced NeRF-based scene reconstruction by introducing geometric constraints or priors (Behari et al. 2024; Liu et al. 2025b). In contrast, more recent approaches like EOGs (Aira, Facciolo, and Ehret 2025) and SatGS (Bai et al. 2025) leverage the real-time and efficient rendering capabilities of 3DGS, adapting it to multi-temporal satellite images for more efficient photogrammetry. Despite these advances, current methods still

struggle with limited generalization and often require more than ten input views to achieve satisfactory performance.

### Sparse-View Scene Reconstruction

The insufficient geometric constraints between sparse-view images poses significant challenges for scene reconstruction (Shi et al. 2024). In computer vision, existing sparse-view methods can be categorized into two groups: per-scene reconstruction methods and generalizable reconstruction methods (Zhang et al. 2025). The former typically leverage multi-view geometric constraints (Truong et al. 2023; Deng et al. 2022), or incorporate stronger supervision from pre-trained models to improve reconstruction quality (Zhu et al. 2024; Yu et al. 2022). However, these methods are often hindered by time-consuming optimization processes, leading to low efficiency. In contrast, generalizable methods perform scene reconstruction in a single feed-forward pass, demonstrating strong generalization by learning powerful priors from large-scale datasets (Yu et al. 2021; Yang, Pavone, and Wang 2023; Liu et al. 2024). However, in the context of sparse-view satellite images, existing reconstruction methods either rely on ground truth height supervision (Gao, Liu, and Ji 2021) or require per-scene optimization, which is computationally expensive (Zhang et al. 2024a).

### Generalizable 3DGS

Generalizable 3DGS models have recently gained significant attention due to their high efficiency and strong generalization capabilities. PixelSplat (Charatan et al. 2024) pioneers this approach by leveraging transformer-encoded features to predict Gaussian parameters directly. Subsequent works (Chen et al. 2024; Zhang et al. 2025; Tang et al. 2024) typically employ depth prediction networks to regress 3D Gaussians, further enhancing generalization. However, due to transient objects, radiometric inconsistencies, and the unique RPC model, directly applying them to multi-temporal satellite images results in suboptimal performance. To address this, we explicitly integrate the RPC model into the generalizable 3DGS pipeline. Our method enables ac-

curate scene reconstruction in a self-supervised manner by filtering transient objects and incorporating relative height supervision.

## Method

Given a set of  $N$  sparse-view satellite images  $\{I_i\}_{i=1}^N$  and their corresponding RPCs, we aim to learn a generalizable model that can accurately reconstruct 3D scenes without requiring ground-truth height supervision. To achieve this goal, we propose SkySplat, a feed-forward framework that avoids per-scene optimization. It first matches and fuses geometric cues from the images with the RPC model to predict height and Gaussian parameters. Then, transient masks are generated by the CSCM to minimize the impact of dynamic objects during training. Next, monocular relative heights from a pretrained depth model (Yang et al. 2024) are incorporated as additional supervision to handle radiometric inconsistencies in multi-temporal satellite images. Finally, a consistency aggregation strategy is applied to refine the scene reconstruction results. An overview of the proposed SkySplat is shown in Figure 2.

### Generalizable Gaussian Splatting with RPC

The generalizable 3DGS with RPC involves four main steps: multi-view feature extraction, RPC-guided cost volume construction, height estimation and Gaussian parameter prediction.

**Multi-View Feature Extraction.** To ensure efficiency, we follow the feature extraction strategy used in MVSPat (Chen et al. 2024). Specifically, we avoid any 3D convolutions and use a multi-view Transformer to aggregate features across different views, which produces features  $\{F_i\}_{i=1}^N$ .

**RPC-Guided Cost Volume Construction.** Next, we compute feature correlations across views using the differentiable rpc warping to construct cost volumes  $\{C_i\}_{i=1}^N$  (Gao, Liu, and Ji 2021). For each reference feature  $F_i$ , we sample  $M$  height candidates  $\{h_m\}_{m=1}^M$  within a predefined elevation range. Using inverse RPC projection, each pixel coordinates  $(u_i, v_i)$  of  $F_i$  is back-projected with height  $h_m$  to obtain corresponding 3D coordinates:

$$(\text{Lat}_i^m, \text{Lon}_i^m, \text{Hei}_i^m) = \text{RPC}_{\text{ref}}^{-1}(u_i, v_i, h_m) \quad (1)$$

These 3D points are then projected into each source view. We sample source features via interpolation at the projected positions and warp them to the reference view:

$$F_{j \rightarrow i}^{h_m} = \text{Inter}(F_j, \text{RPC}_{\text{src}}(\text{Lat}_i^m, \text{Lon}_i^m, \text{Hei}_i^m)) \quad (2)$$

where *Inter* refers to the interpolation operation. A variance-based operation (Gao, Liu, and Ji 2021) is then applied across the warped features to compute the final cost volumes  $\{C_i\}_{i=1}^N$ .

**Height Estimation.** Each regularized cost volume is then used to estimate height. We apply a *soft argmin* operation along the height dimension to produce a per-view height map, denoted as  $\{\hat{h}_i\}_{i=1}^N$ .

**Gaussian Parameter Prediction.** With the estimated heights, we compute the 3D center positions  $\mu_{3D}$  of the Gaussians by inverse RPC projection followed by coordinate correction. Then, we adopt a 2D U-Net, following MVSPat (Chen et al. 2024), to predict the remaining Gaussian parameters. These include the scaling factor  $S$ , rotation quaternion  $R$ , spherical harmonic coefficients  $C$ , and opacity  $\alpha$ :

$$G = \{\mu_{3D}, S, R, C, \alpha\} \quad (3)$$

Note that the rotation quaternion  $R$  is derived by approximating the RPC model with the pinhole camera model during training (Zhang, Snavely, and Sun 2019). Since training is conducted on  $256 \times 256$  patches, the introduced error is negligible. During inference, we use the exact  $\mu_{3D}$  obtained from the RPC projection, avoiding any approximation error.

### Transient Masking via Cross-Self Consistency

Current methods for handling transient objects in multi-temporal satellite images often rely on per-scene optimization, limiting their generalization (Bao et al. 2024). To address this issue, we propose the Cross-Self Consistency Module (CSCM), a transient masking method based on both cross-view and self-view feature similarity. As shown in Figure 3, CSCM automatically identifies uncertain regions during training, where the RGB image supervision loss is halted.

Considering the photometric stability of the features (Liu et al. 2025a), we extract feature maps  $\{feat_i\}_{i=1}^N$  using DINOv2 with FeatUp (Oquab et al. 2023; Fu et al. 2024). Following Equations (1) and (2), we build feature correspondences across views to obtain projected features  $\{feat_{j \rightarrow i}\}_{i=1}^N$ . At pixels selected by multi-view geometric consistency filters (see *Scene Reconstruction via Consistency Aggregation*), we compute the cosine similarity between  $feat_i$  and  $feat_{j \rightarrow i}$ , then convert it into a confidence map in  $[0, 1]$  (Kulhanek et al. 2024):

$$Q_{cv} = \max(2 \cdot \cos(\text{feat}_i, \text{feat}_{j \rightarrow i}) - 1, 0) \quad (4)$$

where  $Q_{cv}$  measures the features consistency across views. Similarities below 0.5 yield zero confidence. To handle invalid regions caused by geometric filtering, we introduce a self-view confidence map  $Q_{sv}$ , computed based on the similarity between the reference features  $feat_i$  and the rendered features  $feat'_i$  (Fu et al. 2025):

$$Q_{sv} = \max(2 \cdot \cos(\text{feat}_i, \text{feat}'_i) - 1, 0) \quad (5)$$

As  $Q_{sv}$  is usually lower than  $Q_{cv}$ , we calibrate its scale using the mean ratio over valid regions of  $Q_{cv}$ . The final confidence map  $Q$  is constructed by replacing the invalid regions in  $Q_{cv}$  with the calibrated  $Q_{sv}$ . Finally,  $Q$  is thresholded at  $\tau = 0.2$  to produce a binary mask  $M$ , which suppresses misleading supervision from RGB images in regions affected by transient objects. The module is applied after 35k iterations, as height estimates become more reliable. See supplementary for hyperparameter analysis.

### Radiometric-Robust Self-Supervised Learning

To mitigate suboptimal local minima caused by varying imaging conditions in images supervision (Zhang et al.

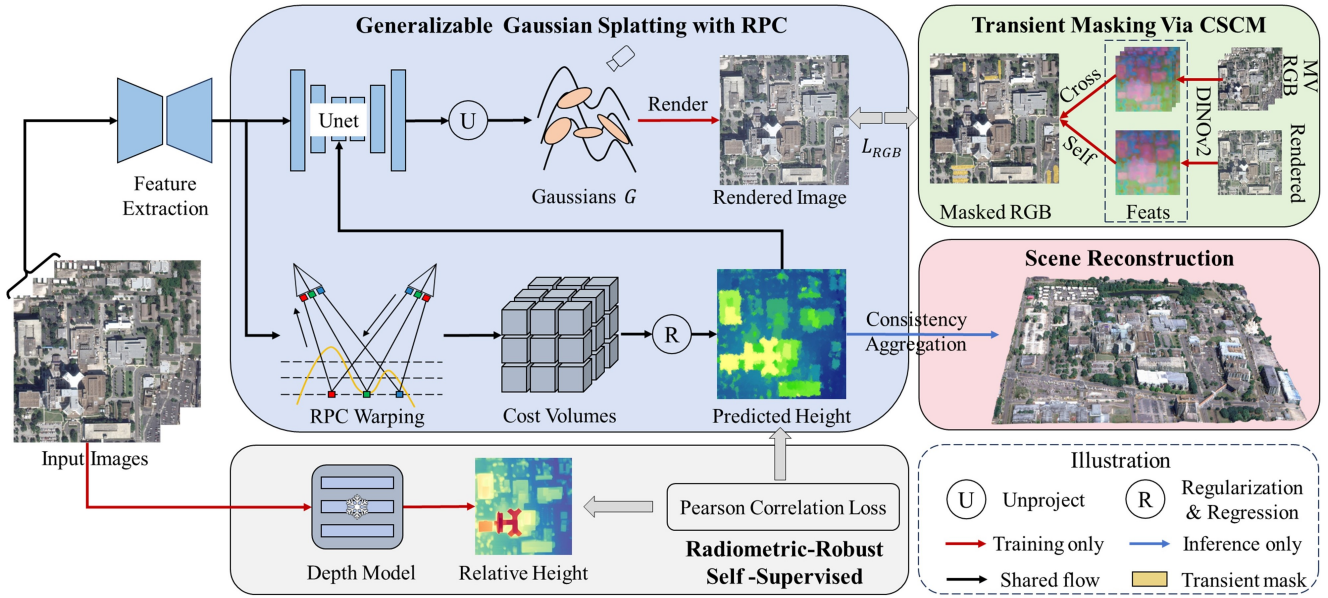


Figure 2: **Overview of the proposed SkySplat.** The depth model is from (Yang et al. 2024), Depth Anything V2 (DAMV2).

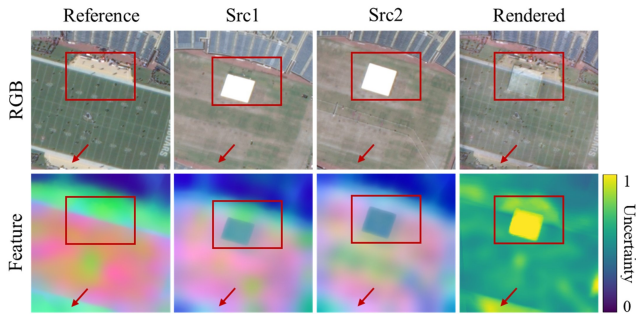


Figure 3: **Visualization of uncertainty regions generated by CSCM.** Top: Three-views input images and the rendered reference image. Bottom: DINOv2 feature visualizations (first three columns) via PCA (Abdi and Williams 2010), and uncertainty map (last column) of the reference view. Red boxes and arrows highlight transient objects, where gradient propagation is halted.

2024b), we introduce auxiliary supervision based on relative height, which is more robust to illumination changes (Liu et al. 2025b).

We first obtain relative height maps  $\{H_i\}_{i=1}^N$  from Depth Anything V2 (DAMV2) (Yang et al. 2024) for each view. To address the scale ambiguity issue, we follow previous approach (Zhu et al. 2024) and supervise the predicted absolute height maps  $\{\hat{h}_i\}_{i=1}^N$  using the Pearson correlation loss:

$$\mathcal{L}_{\text{hei}} = \frac{\text{Cov}(H, \hat{h})}{\sqrt{\text{Var}(H) \cdot \text{Var}(\hat{h})}} \quad (6)$$

where  $\text{Cov}(\cdot, \cdot)$  and  $\text{Var}(\cdot)$  denote covariance and variance, respectively. Unlike previous work (Liu et al. 2025b), SkyS-

plat does not require explicit scale alignment, as it directly captures similarity between height distributions.

In addition, we apply both LPIPS and MSE losses between rendered and ground-truth images, guided by the mask  $M$ , where 1 indicates stable (non-transient) regions:

$$\mathcal{L}_{\text{rgb}} = M \odot \text{LPIPS}(I_{\text{render}}, I_{\text{gt}}) + M \odot (I_{\text{render}} - I_{\text{gt}})^2 \quad (7)$$

where  $\odot$  denotes element-wise multiplication. The rendered image  $I_{\text{render}}$  is generated by native 3DGS rendering, where the RPC model is approximated by the pinhole camera model (Zhang, Snavely, and Sun 2019):

$$I_{\text{render}} = \sum_i c_i \cdot \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j) \quad (8)$$

As mentioned earlier, the approximation error is negligible due to small training patches, and no rendering is used during inference. The final self-supervised loss is defined as:  $\mathcal{L} = \mathcal{L}_{\text{rgb}} + \mathcal{L}_{\text{hei}}$ .

### Scene Reconstruction via Consistency Aggregation

To enhance reconstruction accuracy, we adopt a Multi-view Consistency Aggregation strategy inspired by (Liu et al. 2024; Gao, Liu, and Ji 2021), which filters out noisy Gaussian points with high reprojection errors across views.

Specifically, given a predicted height map  $\hat{h}_i$  for the reference view and  $\hat{h}_j$  for the source view, we first project a point  $p$  from the reference view to the source view using Equations (1) and (2), obtaining the projection point  $q$ . The source-view height  $\hat{h}_j(q)$  is then sampled at this location. Subsequently, the point  $q$  is reprojected back to the reference view to obtain  $p'$ , using:

$$p' = \text{RPC}_{\text{ref}} \left( \text{RPC}_{\text{src}}^{-1}(u_q, v_q, \hat{h}_j(q)) \right) \quad (9)$$

where  $(u_q, v_q)$  denote the pixel coordinates of  $q$ . Then, we get the height  $\hat{h}_i(p')$ , and compute the geometric and height reprojection errors as:

$$\delta_p = \|p - p'\|_2 \quad (10)$$

$$\delta_h = \frac{|\hat{h}_i(p) - \hat{h}_i(p')|}{|\hat{h}_i(p)|} \quad (11)$$

Only points with  $\delta_p < 3$  and  $\delta_h < 0.2$  are retained as reliable 3D points. Finally, these filtered points are orthogonally projected onto a regular 2D grid. For each grid cell, we keep the maximum height among all assigned points to generate the final DSM.

## Experiments

### Experimental Setup

**Datasets.** We train and evaluate our model on the large-scale DFC19 dataset (Bosch et al. 2019; Le Saux et al. 2019), which contains multi-temporal images from Jacksonville (JAX) and Omaha (OMA), all with a ground sampling distance (GSD) of 0.3 m. Following MVSPat (Chen et al. 2024), we select three views for each scene and crop them into  $256 \times 256$  patches, yielding 11,648 training and 1,472 test samples. For cross dataset evaluation, we use three AOIs from the MVS3D dataset (Bosch et al. 2016), which share the same GSD, consistent with EOGS (Aira, Facciolo, and Ehret 2025). To compare with per-scene optimization methods, we also evaluate on five AOIs from DFC19 dataset. RPC camera models in both datasets are refined via bundle adjustment.

**Evaluation Metrics.** We evaluate all methods using LiDAR-based DSMs with a GSD of 0.3–0.5 m. Metrics include mean absolute error (MAE), root mean square error (RMSE), and percentage of accurate grids in total ( $PAG$ ) (Gao, Liu, and Ji 2023; Huang et al. 2025). For example,  $PAG_{2.5}$  represents the ratio of grid cells with an L1 distance error below 2.5 m, and  $PAG_{7.5}$  represents the ratio for errors below 7.5 m. Additional metrics and results on novel view synthesis are presented in the Supplements (see Appendix A.2).

**Implementation Details.** We conduct all experiments on a server equipped with eight NVIDIA® GeForce RTX 4090 GPUs (24 GB VRAM each), running on Ubuntu 22.04. All models are trained for 20 epochs using AdamW with a learning rate of  $2 \times 10^{-4}$  and a batch size of 3. For each scene, we use a fixed height sampling range  $[h_{\min}, h_{\max}]$  with 64 samples, derived from publicly DEMs or LiDAR. In addition, all images of eight AOIs are resized to  $768 \times 768$  for generalizable 3DGS, and RPCs are adjusted accordingly. No post-processing (e.g., consistency aggregation) is applied during visualization for fair comparison.

It is worth noting that, to adapt the generalizable 3DGS compared in Table 1 to satellite images, the RPC model is approximated by the pinhole camera model (Zhang, Snavely, and Sun 2019). Due to the significant distance between the satellite camera and the scene, higher numerical precision is employed to prevent numerical instability and overflow issues that may arise from the large depth values.

### Generalization Results

We compare SkySplat with SOTA generalizable methods for sparse-view scene reconstruction. All models are trained on the DFC19 training set, and quantitative results on the DFC19 test set are reported in Table 1. SkySplat consistently outperforms all baselines across all metrics. Compared to HiSplat, it achieves an 11.38 m lower MAE, a 13.91 m lower RMSE, a 63.21% higher  $PAG_{2.5}$ , and a 58.49% higher  $PAG_{7.5}$ . Notably, even without the consistency aggregation (C.A.) strategy, SkySplat still delivers the best performance, as shown in the second-to-last row of Table 1.

To further evaluate generalization ability, we directly apply the model trained on the DFC19 to the MVS3D test set. SkySplat again outperforms all baselines, demonstrating strong cross-dataset generalization.

We stitch together 64 non-overlapping outputs, each generated from three  $256 \times 256$  images, to form a large-scale 3D Gaussian scene. As shown in Figure 4, SkySplat produces higher-quality 3D Gaussians and more accurate height maps, especially in challenging regions. These results highlight the superiority of our RPC-based framework in achieving accurate 3D reconstructions.

### Per-Scene Optimization Results

Table 2 reports quantitative comparisons between SkySplat and SOTA methods that rely on per-scene optimization. Benefiting from strong generalization capability, SkySplat achieves the best performance while drastically reducing reconstruction time.

Specifically, it reduces the average reconstruction time from several minutes (e.g., 4.60 min for EOGS) or even hours (e.g., 5–8 h for NeRF-based methods) to just 3.19 seconds. Despite being nearly 86× faster than EOGS, it still delivers significantly lower MAE across all AOIs. For instance, on IARPA 002 and IARPA 003, SkySplat achieves errors of 3.75 m and 3.41 m, compared to 13.79 m and 14.83 m by EOGS. Even without the C.A. strategy, our model maintains strong accuracy, outperforming all baselines in both DFC19 and MVS3D datasets.

Visual results in Figures 5 and 6 further highlight the advantages of SkySplat in preserving structural details. Even without fine-tuning, it consistently produces more accurate and coherent 3D reconstructions compared to existing methods. Additionally, the average MAE and reconstruction time of each method are visualized in Figure 1 (b).

### Ablation Study

To evaluate the effectiveness of each component, we conduct ablation studies on the DFC19 dataset. Results show that all proposed modules contribute to improved performance.

**Ablation of Cross-Self Consistency Module.** The CSCM is designed to suppress the influence of transient objects by stopping gradient propagation. It improves performance in both settings: without relative height supervision (Row 1 vs. Row 2 in Table 3) and with supervision (Row 3 vs. Row 4). In all cases, adding CSCM yields consistent improvements across metrics.

Method	DFC19 Dataset				MVS3D Dataset			
	MAE(m)↓	RMSE(m)↓	PAG <sub>2.5</sub> (%)↑	PAG <sub>7.5</sub> (%)↑	MAE(m)↓	RMSE(m)↓	PAG <sub>2.5</sub> (%)↑	PAG <sub>7.5</sub> (%)↑
pixelSplat	176.03	189.26	0.02	0.07	29.15	38.41	6.18	18.36
MVSplat	19.82	22.96	2.71	16.11	16.05	20.20	9.99	29.54
TranSplat	21.96	24.94	1.86	11.40	15.81	19.01	9.44	27.85
DepthSplat	24.21	26.76	0.89	6.33	15.81	19.01	9.44	27.85
HiSplat	13.18	16.59	15.06	37.08	15.63	19.37	9.96	29.34
SkySplat w/o C.A.	<u>2.07</u>	<u>3.07</u>	<u>75.08</u>	<u>94.91</u>	<u>3.48</u>	<u>4.80</u>	<u>50.76</u>	<u>89.16</u>
SkySplat	<b>1.80</b>	<b>2.68</b>	<b>78.27</b>	<b>95.57</b>	<b>3.42</b>	<b>4.79</b>	<b>52.32</b>	<b>89.35</b>

Table 1: **Quantitative comparison with generalizable methods on both datasets.** SkySplat achieves the best overall performance across all metrics on both datasets. The results of “SkySplat w/o C.A.” demonstrate the contribution of our core framework even without the consistency aggregation strategy. (**Bold** indicates best, underline indicates second best.)

Method	JAX004	JAX068	JAX260	OMA212	OMA315	IARPA001	IARPA002	IARPA003	Time
NeRF	3.30	6.33	3.09	1.16	3.01	4.11	6.05	6.02	5.76 h
S-NeRF	3.28	7.47	4.88	3.24	2.98	4.97	9.71	6.55	6.62 h
Sat-NeRF	<u>3.27</u>	6.53	5.28	3.16	2.99	4.63	6.65	<u>4.92</u>	7.39 h
EOGS	3.31	6.67	6.41	9.08	6.38	5.90	13.79	14.83	4.60 min
SkySplat w/o C.A.	<b>1.56</b>	4.24	2.68	<u>0.90</u>	<u>1.53</u>	<u>3.14</u>	<u>3.89</u>	<b>3.41</b>	<b>3.13 s</b>
SkySplat	<b>1.56</b>	<b>3.86</b>	<b>2.46</b>	<b>0.89</b>	<b>1.51</b>	<b>3.10</b>	<b>3.75</b>	<b>3.41</b>	<u>3.19 s</u>

Table 2: **Quantitative comparison with per-scene optimization methods across three cities.** Reported metrics include MAE on the elevation (meters) and reconstruction time for each method. (**Bold** indicates best, underline indicates second best.)

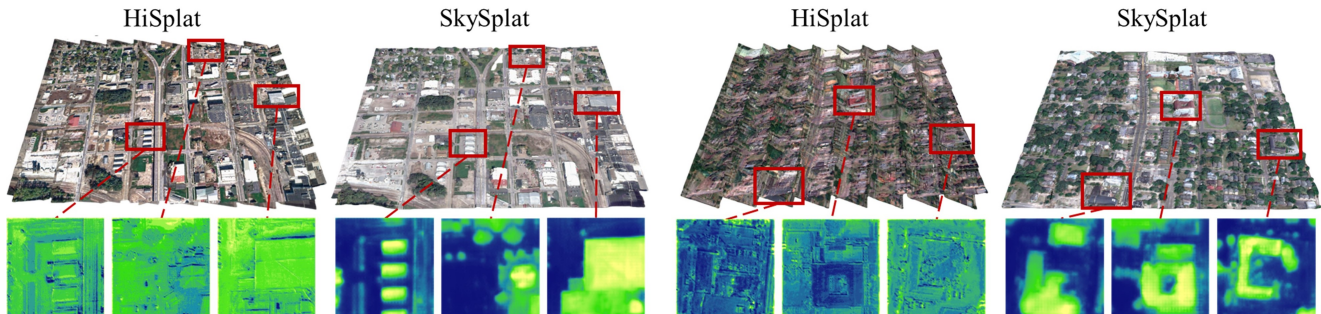


Figure 4: **Comparisons of 3D Gaussians (top) and height maps (bottom).** SkySplat generates smoother and more accurate results, highlighting its effectiveness.

CSCM	R.H.S.	C.A.	MAE (m)↓	RMSE (m)↓	PAG <sub>2.5</sub> (%)↑	PAG <sub>7.5</sub> (%)↑
			6.07	7.33	29.46	68.57
✓			5.94	7.19	29.46	70.37
	✓		2.25	3.30	72.90	94.00
✓	✓		2.07	3.07	75.08	94.91
✓	✓	✓	<b>1.80</b>	<b>2.68</b>	<b>78.27</b>	<b>95.57</b>

Table 3: **Ablation study on the DFC19 dataset.** We evaluate the contributions of CSCM (Cross-Self Consistency Module), R.H.S. (Relative Height Supervision), and C.A. (Consistency Aggregation). (**Bold** indicates best.)

**Ablation of Relative Height Supervision.** We use relative height from DAMV2 as auxiliary supervision to promote geometry learning. To assess its necessity, we compare results without and with Relative Height Supervision

(R.H.S.) (Row 2 vs. Row 4 in Table 3). The results confirm that relative height supervision is essential for better geometric reconstruction. This supervision thus acts as a crucial complementary signal to photometric cues.

Views	Method	JAX068	JAX260	OMA212	OMA315
5	S2P (MGM)*	2.35	3.29	1.50	1.79
	FVMD-ISRe*	<b>1.57</b>	2.81	1.10	1.72
3	S2P (MGM)	4.57	5.04	1.55	2.06
	SkySplat w/o C.A.	4.24	2.68	0.90	1.53
	SkySplat	3.86	<b>2.46</b>	<b>0.89</b>	<b>1.51</b>

Table 4: **Effect of the number of views on selected AOIs.** Results marked with \* are from previous work (Zhang et al. 2024a). S2P (MGM) refers to the classic stereo pipeline (De Franchis et al. 2014). The evaluation metric is MAE in meters.

Method	MFE (px)↓	MAE (m)↓	RMSE (m)↓	PAG <sub>2.5</sub> (%)↑	PAG <sub>7.5</sub> (%)↑
HiSplat (256×256)	0.27	13.18	16.59	15.06	37.08
HiSplat (512×512)	0.43	13.96	17.68	14.62	35.71
HiSplat (1024×1024)	1.23	14.04	17.77	14.49	35.48
HiSplat (2048×2048)	2.76	14.36	18.11	13.84	34.40
SkySplat (RGB sup.)	—	<b>6.07</b>	<b>7.33</b>	<b>29.46</b>	<b>68.57</b>

Table 5: **Effect of image size on the DFC19 dataset.** We compare SkySplat (trained with RGB-only supervision) with HiSplat, where all inputs are cropped to 256×256. For HiSplat, the resolutions in parentheses refer to the image sizes used when fitting the pinhole camera models before cropping.

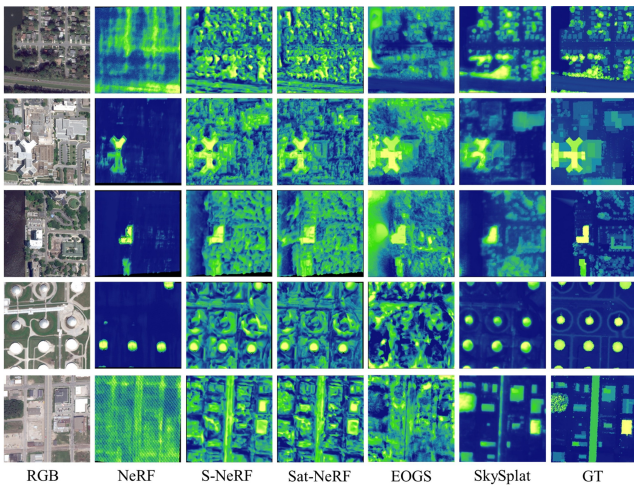


Figure 5: **Predicted DSMs of the DFC19 areas.** From top to bottom: JAX004, JAX068, JAX260, OMA212, OMA315. DSM resolution: 50 cm/pixel.

**Ablation of Consistency Aggregation.** We disable the C.A. strategy to evaluate its effect on scene refinement (Row 4 vs. Row 5 in Table 3). Removing C.A. leads to significant drops in accuracy, showing that multi-view consistency aggregation is critical for reconstruction.

**Effect of the Number of Views.** While previous methods typically require five input views, our three-view reconstruction approach achieves better results in most cases. As shown in Table 4, SkySplat outperforms these five-view baselines, indicating that high-quality scene reconstruction can be effectively accomplished with fewer views.

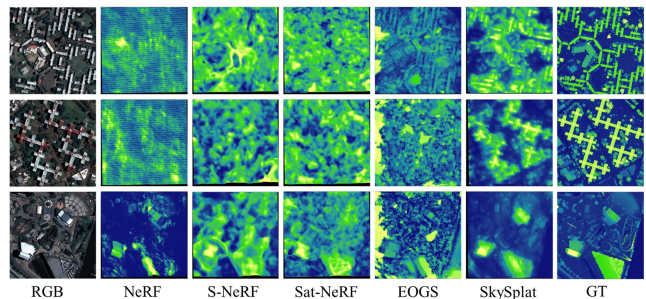


Figure 6: **Predicted DSMs of the MVS3D areas.** From top to bottom: IARPA001, IARPA002, IARPA003. DSM resolution: 30 cm/pixel.

**Effect of Image Size.** To study the effect of image size on generalizable 3DGS, we fit pinhole models at different sizes and report mean fitting error (MFE) on the test set (Table 5). Results show: (1) HiSplat’s error grows with size due to poor RPC approximation, limiting large-scale use; (2) SkySplat, even without auxiliary modules (i.e., supervised using only RGB images), showing greater precision in reconstruction.

## Conclusion

We propose SkySplat, a novel self-supervised framework for generalizable 3D reconstruction from multi-temporal sparse satellite images. Extensive experiments show that SkySplat is up to 86× faster than SOTA per-scene optimization methods, while maintaining strong cross-dataset generalization. By explicitly avoiding per-scene optimization and ground-truth height supervision, SkySplat makes significant progress toward efficient satellite-based 3D reconstruction.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grants 42530106, 42192581, 42571507, 42201474), the National Key Research and Development Program of China (Grant 2024YFB3909300), the China Railway Group Laboratory Basic Research Project (Grant L2023G014), the Major special projects of Guizhou [2022]001, and the Hubei Provincial Natural Science Foundation of China (Grant 2025AFB426).

## References

- Abdi, H.; and Williams, L. J. 2010. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4): 433–459.
- Aira, L. S.; Facciolo, G.; and Ehret, T. 2025. Gaussian Splatting for Efficient Satellite Image Photogrammetry. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 5959–5969.
- Bai, N.; Yang, A.; Chen, H.; and Du, C. 2025. SatGS: Remote Sensing Novel View Synthesis Using Multi-Temporal Satellite Images with Appearance-Adaptive 3DGS. *Remote Sensing*, 17(9): 1609.
- Bao, Y.; Liao, J.; Huo, J.; and Gao, Y. 2024. Distractor-free generalizable 3d gaussian splatting. *arXiv preprint arXiv:2411.17605*.
- Barron, J. T.; Mildenhall, B.; Tancik, M.; Hedman, P.; Martin-Brualla, R.; and Srinivasan, P. P. 2021. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, 5855–5864.
- Barron, J. T.; Mildenhall, B.; Verbin, D.; Srinivasan, P. P.; and Hedman, P. 2022. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5470–5479.
- Behari, N.; Dave, A.; Tiwary, K.; Yang, W.; and Raskar, R. 2024. SUNDIAL: 3D Satellite Understanding through Direct Ambient and Complex Lighting Decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 522–532.
- Bosch, M.; Foster, K.; Christie, G.; Wang, S.; Hager, G. D.; and Brown, M. 2019. Semantic stereo for incidental satellite images. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1524–1532. IEEE.
- Bosch, M.; Kurtz, Z.; Hagstrom, S.; and Brown, M. 2016. A multiple view stereo benchmark for satellite imagery. In *2016 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, 1–9. IEEE.
- Charatan, D.; Li, S. L.; Tagliasacchi, A.; and Sitzmann, V. 2024. 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*, 19457–19467.
- Chen, Y.; Xu, H.; Zheng, C.; Zhuang, B.; Pollefeys, M.; Geiger, A.; Cham, T.-J.; and Cai, J. 2024. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In *European Conference on Computer Vision*, 370–386. Springer.
- De Franchis, C.; Meinhardt-Llopis, E.; Michel, J.; Morel, J.-M.; and Facciolo, G. 2014. An automatic and modular stereo pipeline for pushbroom images. In *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*.
- Deng, K.; Liu, A.; Zhu, J.-Y.; and Ramanan, D. 2022. Depth-supervised nerf: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12882–12891.
- Derksen, D.; and Izzo, D. 2021. Shadow neural radiance fields for multi-view satellite photogrammetry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1152–1161.
- Fu, C.; Zhang, Y.; Yao, K.; Chen, G.; Xiong, Y.; Huang, C.; Cui, S.; and Cao, X. 2025. RobustSplat: Decoupling Den-sification and Dynamics for Transient-Free 3DGS. *arXiv preprint arXiv:2506.02751*.
- Fu, S.; Hamilton, M.; Brandt, L.; Feldman, A.; Zhang, Z.; and Freeman, W. T. 2024. Featup: A model-agnostic framework for features at any resolution. *arXiv preprint arXiv:2403.10516*.
- Gao, J.; Liu, J.; and Ji, S. 2021. Rational polynomial camera model warping for deep learning based satellite multi-view stereo matching. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6148–6157.
- Gao, J.; Liu, J.; and Ji, S. 2023. A general deep learning based framework for 3D reconstruction from multi-view stereo satellite images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 195: 446–461.
- Huang, X.; Liu, X.; Wan, Y.; Zheng, Z.; Zhang, B.; Wang, Y.; Guo, H.; and Zhang, Y. 2025. MVSR3D: An End-to-End Framework for Semantic 3D Reconstruction Using Multi-View Satellite Imagery. *IEEE Transactions on Geoscience and Remote Sensing*.
- Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3D Gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4): 139–1.
- Kulhanek, J.; Peng, S.; Kukulova, Z.; Pollefeys, M.; and Sattler, T. 2024. Wildgaussians: 3d gaussian splatting in the wild. *arXiv preprint arXiv:2407.08447*.
- Le Saux, B.; Yokoya, N.; Hansch, R.; Brown, M.; and Hager, G. 2019. 2019 data fusion contest [technical committees]. *IEEE Geoscience and Remote Sensing Magazine*, 7(1): 103–105.
- Liu, A.; Long, X.; Liu, Y.; Luo, P.; and Wang, W. 2025a. Sem-iNeRF: Camera pose refinement by inverting neural radiance fields with semantic feature consistency. *Computational Visual Media*.
- Liu, T.; Wang, G.; Hu, S.; Shen, L.; Ye, X.; Zang, Y.; Cao, Z.; Li, W.; and Liu, Z. 2024. Mvs gaussian: Fast generalizable gaussian splatting reconstruction from multi-view stereo. In *European Conference on Computer Vision*, 37–53. Springer.
- Liu, T.; Zhao, S.; Jiang, W.; and Guo, B. 2025b. Sat-DN: Implicit Surface Reconstruction from Multi-View Satellite Images with Depth and Normal Supervision. *arXiv preprint arXiv:2502.08352*.

- Marí, R.; Facciolo, G.; and Ehret, T. 2022. Sat-nerf: Learning multi-view satellite photogrammetry with transient objects and shadow modeling using rpc cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1311–1321.
- Marí, R.; Facciolo, G.; and Ehret, T. 2023. Multi-date earth observation nerf: The detail is in the shadows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2035–2045.
- Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.
- Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.
- Shi, R.; Wei, X.; Wang, C.; and Su, H. 2024. Zerorf: Fast sparse view 360deg reconstruction with zero pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21114–21124.
- Tang, S.; Ye, W.; Ye, P.; Lin, W.; Zhou, Y.; Chen, T.; and Ouyang, W. 2024. Hisplat: Hierarchical 3d gaussian splatting for generalizable sparse-view reconstruction. *arXiv preprint arXiv:2410.06245*.
- Truong, P.; Rakotosaona, M.-J.; Manhardt, F.; and Tombari, F. 2023. Sparf: Neural radiance fields from sparse and noisy poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4190–4200.
- Xu, H.; Peng, S.; Wang, F.; Blum, H.; Barath, D.; Geiger, A.; and Pollefeys, M. 2025. Depthsplat: Connecting gaussian splatting and depth. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 16453–16463.
- Yang, J.; Pavone, M.; and Wang, Y. 2023. Freenerf: Improving few-shot neural rendering with free frequency regularization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8254–8263.
- Yang, L.; Kang, B.; Huang, Z.; Zhao, Z.; Xu, X.; Feng, J.; and Zhao, H. 2024. Depth anything v2. *Advances in Neural Information Processing Systems*, 37: 21875–21911.
- Yu, A.; Ye, V.; Tancik, M.; and Kanazawa, A. 2021. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4578–4587.
- Yu, Z.; Peng, S.; Niemeyer, M.; Sattler, T.; and Geiger, A. 2022. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *Advances in neural information processing systems*, 35: 25018–25032.
- Zhang, C.; Yan, Y.; Zhao, C.; Su, N.; and Zhou, W. 2024a. Fvmd-isre: 3-d reconstruction from few-view multirate satellite images based on the implicit surface representation of neural radiance fields. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–14.
- Zhang, C.; Zou, Y.; Li, Z.; Yi, M.; and Wang, H. 2025. Transplat: Generalizable 3d gaussian splatting from sparse multi-view images with transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 9869–9877.
- Zhang, K.; Snavely, N.; and Sun, J. 2019. Leveraging vision reconstruction pipelines for satellite imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 0–0.
- Zhang, T.; Zhou, Y.; Li, Y.; and Wei, X. 2024b. Satensorf: Fast satellite tensorial radiance field for multirate satellite imagery of large size. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–15.
- Zheng, Z.; Wan, Y.; Zhang, Y.; Hu, Z.; Wei, D.; Yao, Y.; Zhu, C.; Yang, K.; and Xiao, R. 2024. Digital surface model generation from high-resolution satellite stereos based on hybrid feature fusion network. *The Photogrammetric Record*, 39(185): 36–66.
- Zhu, Z.; Fan, Z.; Jiang, Y.; and Wang, Z. 2024. Fsgs: Real-time few-shot view synthesis using gaussian splatting. In *European conference on computer vision*, 145–163. Springer.