

SkySplat-OV: Generalizable Language Gaussian Splatting for Open-Vocabulary Scene Understanding from Sparse Satellite Views

Xuejun Huang¹, Yi Wan^{1,3,4*}, Lei Yu², Xinyi Liu^{1,3,4}, Zhi Zheng⁵, Bin Zhang⁶, Yingying Pei¹, Yi Liu¹, Changjun Zhu¹, Yi Liu¹, Xiangyuan Cai², Hongwei Hu², Xin Zhang⁷, and Yongjun Zhang^{1,3,4}

¹ School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China

² Ant Group, Hangzhou 310023, China

³ Technology Innovation Center for Collaborative Applications of Natural Resources Data in GBA, Ministry of Natural Resources, Guangzhou 510075, China

⁴ Hubei LuoJia Laboratory, Wuhan 430079, China

⁵ Department of Aeronautical and Aviation Engineering, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, China

⁶ China Railway Siyuan Survey and Design Group Co., LTD, Wuhan 430063, China

⁷ State Key Laboratory of Remote Sensing and Digital EarthAerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China
{huangxuejun, yi.wan}@whu.edu.cn

Abstract. Open-vocabulary scene understanding aims to recognize arbitrary semantic concepts in 3D beyond a fixed set of categories. Recent methods built upon 3D Gaussian Splatting (3DGS) enable scene understanding by distilling features from vision-language models, but most rely on per-scene optimization and dense observations, which limit scalability. More recent generalizable 3DGS methods show promise, but they perform poorly on sparse-view satellite imagery due to the unique pushbroom imaging mode, limited geometric constraints, and extreme scale variations. To address these limitations, we propose SkySplat-OV, a feed-forward framework that integrates the rational polynomial coefficient (RPC) model into a generalizable language 3DGS pipeline. A Geometry-Radiance-Semantics Fusion module, together with monocular relative height supervision, improves Gaussian parameter prediction under limited geometric constraints. Moreover, SkySplat-OV employs single-stage SAM3 distillation to handle extreme scale variations in satellite imagery for accurate segmentation. Compared to per-scene optimization methods, SkySplat-OV achieves up to 1070 times speedup over LangSplatV2 with higher accuracy. It also outperforms generalizable 3DGS baselines, achieving 12.91%–25.75% mIoU improvements on two benchmarks, and demonstrates strong cross-dataset generalization on the MVS3D test set. Additional qualitative results are available at <https://skysplat-ov.github.io/>.

Keywords: Scene Understanding · Gaussian Splatting · Sparse-View

* Corresponding author

1 Introduction

Open-vocabulary scene understanding from sparse satellite views is a fundamental yet challenging problem. This technique is critical for applications such as urban planning [23], robot navigation [10], and SLAM [27].

Most existing scene understanding methods rely on fully supervised learning, requiring dense ground-truth annotations for geometry and semantics. Moreover, they are typically restricted to predefined label sets and have difficulty generalizing to arbitrary categories [8, 16, 28]. In contrast, NeRF-based methods [19, 44] achieve scene understanding by distilling 2D vision-language features into radiance fields, without requiring ground-truth supervision. However, these methods rely on per-scene optimization and dense input views, making them computationally expensive and ill-suited for sparse-view satellite imagery.

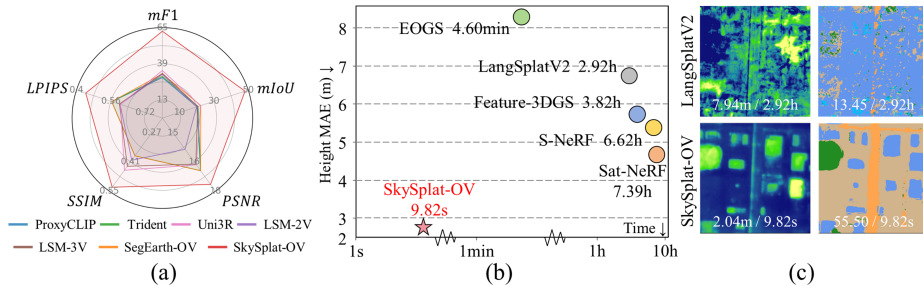


Fig. 1: Comparison with existing methods. (a) Generalization results on the DFC19 dataset [4, 22]. SkySplat-OV achieves the best performance among all competing methods. (b) Per-scene optimization results. SkySplat-OV reaches optimal performance within only 9.82 seconds. (c) DSM reconstruction and semantic segmentation results on a representative area of interest (OMA315). We report the MAE, mIoU, and reconstruction time of SkySplat-OV and Uni3R [36].

To improve efficiency, methods based on 3D Gaussian Splatting (3DGS [18]) have been proposed for open-vocabulary scene understanding. For example, Feature-3DGS [49] and LangSplatV2 [26] distill semantic features into the 3D Gaussian field, enabling promptable 3D segmentation. However, they still require per-scene optimization and dense observations, which limits their scalability.

More recently, generalizable 3DGS methods [13, 25, 36] directly predict Gaussian parameters and lift features into 3D in a single pass, improving both transferability and reconstruction speed. Despite these advances, existing methods remain unsuitable for sparse-view satellite imagery due to three challenges: (1) the pushbroom imaging mode differs from the standard pinhole imaging mode; (2) inaccurate Gaussian parameter prediction under limited geometric constraints; and (3) complex multi-stage distillation is often required to handle extreme scale variations for accurate segmentation.

To address these limitations, we propose SkySplat-OV, a generalizable framework for open-vocabulary scene understanding from sparse satellite views. It integrates the rational polynomial coefficient (RPC) model into a 3DGS pipeline, lifts language features via feature distillation into the 3D Gaussian field. Gaussian parameter predictions are then refined under limited geometric constraints via a Geometry–Radiance–Semantics Fusion module and monocular relative height supervision. Single-stage SAM3 feature distillation further handles extreme scale variations in satellite imagery, enabling accurate and view-consistent segmentation. The framework is fully self-supervised, requiring no ground-truth annotations. Extensive experiments demonstrate that SkySplat-OV significantly outperforms prior generalizable methods and achieves up to $1070\times$ speedup over state-of-the-art (SOTA) per-scene optimization methods (see Figure 1). Our main contributions are summarized as follows:

- We propose SkySplat-OV, the first self-supervised framework for scene understanding from sparse-view satellite imagery, explicitly integrating the RPC model into the language 3DGS pipeline.
- We introduce a Geometry–Radiance–Semantics Fusion (GRS-Fusion) module and incorporate monocular relative height supervision to refine Gaussian parameter predictions.
- We distill SAM3 features in a single-stage process for accurate and view-consistent segmentation, achieving $1070\times$ faster inference than per-scene optimization methods while maintaining strong cross-dataset generalization.

2 Related Works

2.1 Open-Vocabulary Semantic Segmentation

Open-vocabulary semantic segmentation (OVSS) enables segmentation of almost any category in an image through a prompt vocabulary. In natural imagery, training-based methods often require annotated datasets to adapt vision-language models (VLMs) [30, 43], while training-free methods like ProxyCLIP [21] and Trident [35] use foundation model [20, 33] to guide segmentation without extra training. However, extending OVSS methods to satellite imagery is challenging due to extreme scale variations, arbitrary object orientations, and dense layouts [24]. SegEarth-OV [24] addresses these challenges by adapting feature representations for satellite imagery, achieving strong performance. Recent SAM-based methods further introduce feature or mask guidance to improve localization, but they generally rely on complex pipelines [12, 47]. Despite the success of SAM in 2D segmentation [6, 20], its potential for open-vocabulary scene understanding from sparse satellite views remains largely unexplored. To bridge this gap, we distill SAM3 [6] features into the language embedded 3DGS field in a single-stage process, enhancing scene understanding.

2.2 Scene Reconstruction from Satellite Imagery

Scene reconstruction from satellite imagery has traditionally relied on stereo-based and multi-view stereo (MVS) reconstruction methods [14, 15, 48], which

typically require dense ground-truth supervision. To reduce this dependence, NeRF-based approaches have been introduced and adapted to satellite imagery [11, 31, 32]. However, these methods typically involve per-scene optimization and multiple input views, limiting their applicability under sparse-view settings [7].

Recently, 3DGS has emerged as an efficient scene representation, offering fast rendering and high-quality reconstruction [18]. Several works, such as SatGS [3] and EOGS [2], have adapted 3DGS to satellite imagery, but they still rely on per-scene optimization and dense observations. More recent generalizable 3DGS methods, including MVSplat [9], DepthSplat [42], HiSplat [37] and SkySplat [17], further improve efficiency and generalization by directly predicting Gaussian parameters in a single pass, enabling fast reconstruction from sparse views. However, these methods focus mainly on appearance and lack semantic understanding. In contrast, SkySplat-OV augments each 3D Gaussian with language embeddings, enabling open-vocabulary scene understanding.

2.3 Open-Vocabulary Scene Understanding

Open-vocabulary scene understanding aims to recognize arbitrary concepts in 3D beyond a fixed, annotated label space. In computer vision, early NeRF-based methods embed 2D CLIP [34] or DINO [33] features into radiance fields [19, 44], enabling flexible open-vocabulary queries but requiring per-scene optimization. To improve efficiency, 3DGS-based methods, such as Feature-3DGS [49] and LangSplatV2 [26], directly encode semantic or instance-level features within the Gaussians, supporting promptable segmentation and semantic editing while still relying on per-scene optimization.

More recently, generalizable methods lift vision-language features into 3D in a single pass. For example, LSM [13] builds upon the DUST3R [39] for spatial lifting, while Uni3R [36] leverages VGGT [38] to produce consistent 3D representations. SemanticSplat [25] further achieves accurate segmentation via a complex multi-stage distillation process. However, these generalizable methods are designed for standard pinhole camera models and do not directly support RPC-based satellite imagery. By integrating the RPC model into the language 3DGS pipeline, we enable accurate and view-consistent scene understanding.

3 Method

Given N sparse-view satellite images $\{\mathbf{I}_i\}_{i=1}^N$, where $\mathbf{I}_i \in \mathbb{R}^{H \times W \times 3}$ with height H and width W , along with their corresponding RPCs, our goal is to learn a generalizable model capable of accurate open-vocabulary scene understanding without ground-truth supervision.

To this end, we propose a generalizable framework, SkySplat-OV, which avoids per-scene optimization. SkySplat-OV first extracts deep features from sparse views and constructs a cost volume via RPC warping for regularization and height regression, achieving geometry reconstruction (Sec. 3.1). It then employs a GRS-Fusion module to accurately predict Gaussian parameters (Sec. 3.2).

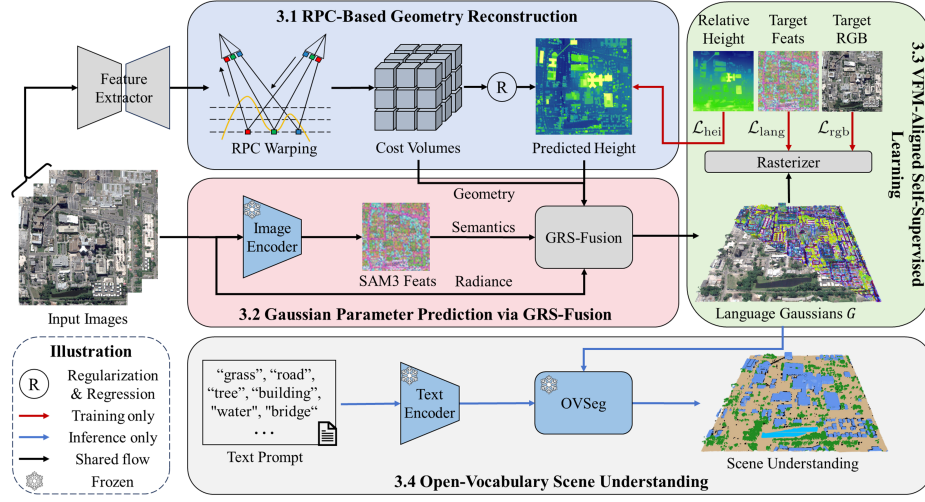


Fig. 2: Overview of the proposed SkySplat-OV. The Image Encoder, Text Encoder, and OVSeg are adopted from the frozen SAM3 [6].

Additionally, monocular relative height and language feature extracted from pre-trained 2D vision foundation models (VFMs) are incorporated as auxiliary supervision, combined with radiance supervision for self-supervised learning (Sec. 3.3). Finally, the distilled language features are fed into the Feature Pyramid Network (FPN) and decoder of SAM3 to obtain open-vocabulary segmentation results (Sec. 3.4). An overview of the SkySplat-OV framework is illustrated in Figure 2.

3.1 RPC-Based Geometry Reconstruction

To enable open-vocabulary scene understanding, SkySplat-OV first infers the scene geometry from sparse-view images.

Multi-View Feature Extraction. To extract deep features from sparse-view images, each input image is first encoded by a lightweight CNN to produce initial features $\{\mathbf{F}_i^{cnn}\}_{i=1}^N$. Features are then aggregated using the multi-view Transformer, yielding the final feature maps $\{\mathbf{F}_i\}_{i=1}^N$.

RPC-Based Cost Volume Construction and Height Regression. Next, cost volumes are constructed to encode cross-view geometric consistency. Specifically, for each reference feature map pixel coordinates (u_i, v_i) and candidate height h_m , the corresponding coordinates $(u_j^{h_m}, v_j^{h_m})$ on the source feature map $\mathbf{F}_j$ are obtained by RPC warping:

$$(u_j^{h_m}, v_j^{h_m}) = \text{RPC}_{\text{src}}(\text{RPC}_{\text{ref}}^{-1}(u_i, v_i, h_m)) \quad (1)$$

The source feature map is then sampled at these locations to obtain warped features:

$$\mathbf{F}_{j \rightarrow i}^{h_m} = \text{Inter} \left(\mathbf{F}_j, \left(u_j^{h_m}, v_j^{h_m} \right) \right) \in \mathbb{R}^{\frac{H}{s} \times \frac{W}{s} \times C} \quad (2)$$

The warped features are aggregated with the reference feature to construct the initial cost volumes. These volumes are then regularized along the height dimension to obtain cost volumes $\{\mathbf{C}_i\}_{i=1}^N$, which are subsequently normalized into probability volumes $\{\mathbf{P}_i\}_{i=1}^N$. Finally, the initial height for each pixel $\{H_i\}_{i=1}^N$, where $H_i \in \mathbb{R}^{H \times W}$, is obtained by computing the weighted sum of all candidate heights using the probability volume.

3.2 Gaussian Parameter Prediction via GRS-Fusion

Building upon previous works [13, 49], we introduce a language-aligned semantic feature parameter $\mathbf{f}$ into conventional 3DGS, resulting in a unified representation that jointly models geometry, radiometry, and semantics. This enables consistent and coherent reconstruction of a language-embedded 3DGS field $\{\mathbf{G}_m\}_{m=1}^M$, where $M = H \times W \times N$. Each 3D Gaussian is represented as:

$$\mathbf{G}_m = \{\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m, \mathbf{C}_m, \alpha_m, \mathbf{f}_m\} \quad (3)$$

where $\boldsymbol{\mu}_m$ denotes the Gaussian center, $\boldsymbol{\Sigma}_m$ denotes the covariance matrix, $\mathbf{C}_m$ denotes the color, and α_m denotes the opacity.

Geometry-Radiance-Semantics Fusion. To achieve accurate Gaussian parameter prediction, we design a Geometry-Radiance-Semantics Fusion (GRS-Fusion) module to integrate geometric information (cost volume $\mathbf{C}_i$, predicted height H_i , and confidence maps derived from the maximum values of $\mathbf{P}_i$), radiometric information (multi-view images), and semantic feature information (language features $\mathbf{F}_i^{sam}$ from SAM3 and the concatenated image features from $\mathbf{F}_i^{cnn}$ and $\mathbf{F}_i$). These features are concatenated along the channel dimension and processed by a lightweight 2D U-Net to generate unified fusion features $\{\mathbf{F}_i^{grs}\}_{i=1}^N$, where $\mathbf{F}_i^{grs} \in \mathbb{R}^{\frac{H}{s} \times \frac{W}{s} \times C_{grs}}$, which serve as the shared representation for Gaussian parameter prediction. Our pipeline proceeds as follows:

Initialization. The predicted height maps H_i are unprojected via the RPC model to obtain the Gaussian centers $\boldsymbol{\mu}_m$.

Language-Aligned Semantic Parameter Prediction. The fusion feature $\mathbf{F}_i^{grs}$ and additional feature information are concatenated and processed by a feature head to predict the language-aligned semantic parameter. Following previous methods [13, 36], the language features $\mathbf{F}_i^{sam}$ are incorporated via a residual connection to refine the prediction, yielding the final semantic parameter $\mathbf{f}_m$.

Remaining Parameters Prediction. The fusion feature $\mathbf{F}_i^{\text{grs}}$, radiometric information, and feature information are concatenated and passed through a Gaussian parameter head to estimate the covariance Σ_m and color $\mathbf{C}_m$, where the covariance matrix is parameterized as $\Sigma_m = \mathbf{R}(\theta_m)^\top \text{diag}(\mathbf{s}_m) \mathbf{R}(\theta_m)$, with $\mathbf{s}_m$ denoting the scaling factor and $\mathbf{R}(\theta_m)$ denoting the rotation matrix parameterized by a quaternion. Meanwhile, following MVSplat [9], the fusion feature $\mathbf{F}_i^{\text{grs}}$ and geometric information are concatenated and processed by convolution layers to predict the opacity α_m .

3.3 VFM-Aligned Self-Supervised Learning

We distill 2D VFMs, including DA3 [29] and SAM3 [6], into the language-embedded 3DGS field. To achieve this, we introduce three complementary self-supervision signals: (1) RGB reconstruction, (2) language feature alignment, and (3) relative height alignment.

Specifically, the rendered image $\mathbf{I}^{\text{render}}$ and the corresponding feature $\mathbf{F}^{\text{render}}$ are obtained via standard 3DGS rendering, with the RPC model approximated as a pinhole camera [45]:

$$\begin{cases} \mathbf{I}^{\text{render}} = \sum_{m=1}^M \mathbf{C}_m \alpha_m \prod_{n=1}^{m-1} (1 - \alpha_n) \\ \mathbf{F}^{\text{render}} = \sum_{m=1}^M \mathbf{f}_m \alpha_m \prod_{n=1}^{m-1} (1 - \alpha_n) \end{cases} \quad (4)$$

For RGB reconstruction, appearance consistency is enforced by supervising the rendered RGB image with the ground-truth image using MSE and LPIPS losses:

$$\mathcal{L}_{\text{rgb}} = \text{MSE}(\mathbf{I}^{\text{gt}}, \mathbf{I}^{\text{render}}) + \text{LPIPS}(\mathbf{I}^{\text{gt}}, \mathbf{I}^{\text{render}}) \quad (5)$$

For language feature alignment, the rendered feature map $\mathbf{F}^{\text{render}}$ is passed through two dedicated MLP heads to produce predicted language features $\hat{\mathbf{F}}^{\text{lang}}$ aligned with SAM3. MSE loss and cosine similarity loss are applied to align the predicted features with the ground-truth features $\mathbf{F}^{\text{lang}}$:

$$\mathcal{L}_{\text{lang}} = \text{MSE}(\mathbf{F}^{\text{lang}}, \hat{\mathbf{F}}^{\text{lang}}) + 1 - \cos(\mathbf{F}^{\text{lang}}, \hat{\mathbf{F}}^{\text{lang}}) \quad (6)$$

To further improve geometric accuracy, the reconstructed height maps H_i are supervised using the relative height maps hei from DA3 via a Pearson correlation loss [50]:

$$\mathcal{L}_{\text{hei}} = \frac{\text{Cov}(hei, H_i)}{\sqrt{\text{Var}(hei) \text{Var}(H_i)}} \quad (7)$$

The overall self-supervised objective is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{rgb}} + \mathcal{L}_{\text{lang}} + \mathcal{L}_{\text{hei}} \quad (8)$$

This self-supervision strategy aligns appearance, language features, and geometry within the unified 3D scene representation, enabling more coherent language 3DGS.

3.4 Open-Vocabulary Scene Understanding

After obtaining distilled features $\hat{\mathbf{F}}^{\text{lang}}$, SkySplat-OV performs open-vocabulary scene understanding by leveraging the SAM3. Specifically, the aligned features $\hat{\mathbf{F}}^{\text{lang}}$ are used in place of the image encoder’s output and fed into SAM3’s FPN for subsequent prediction.:

$$\mathbf{x}_i^{(m)} = \text{Conv}_m(\hat{\mathbf{F}}_i^{\text{lang}}), \quad m = 1, \dots, L \quad (9)$$

$$\mathbf{p}_i^{(m)} = \text{PE}(\mathbf{x}_i^{(m)}), \quad m = 1, \dots, L \quad (10)$$

where each $\mathbf{x}_i^{(m)}$ is the FPN feature at scale m , and $\mathbf{p}_i^{(m)}$ is its positional encoding. A text query is encoded into $\hat{\mathbf{F}}^{\text{text}}$ by the SAM3 text encoder and used by the cross-modal head to produce the segmentation mask $\mathbf{M}_i$ and probability map $\mathbf{P}_i$.

$$(\mathbf{M}_i, \mathbf{P}_i) = \text{OVSeg}(\mathbf{x}_i, \hat{\mathbf{F}}^{\text{text}}) \quad (11)$$

where OVSeg follows the original SAM3 detector design and is used to perform open-vocabulary segmentation. To achieve object-space (digital surface model, DSM-level) segmentation and reconstruction, the per-class probability maps $\mathbf{P}_i$ and height maps H_i are first projected using the RPC model. For each grid cell, the maximum height and semantic probability are selected, and an argmax over classes is applied to produce the final 3D semantic reconstruction.

4 EXPERIMENTS

4.1 Experimental Setup

Datasets. We evaluate SkySplat-OV on three datasets: DFC19 [4,22], SpaceNet4 [41], and MVS3D [5]. DFC19 has a 0.3 m ground sampling distance (GSD) and includes 11,648 training scenes and 1,472 test scenes. Following MVSplat [9], images are cropped to 256×256 , and each scene uses three reference and two target views, with five semantic categories: ground, foliage, building, water, and elevated roadway. SpaceNet4 has a 0.5 m GSD and contains 16,896 training scenes and 2,112 test scenes. Images are cropped to 256×256 in the same manner, and each scene uses two reference and two target views for testing, with three semantic categories: ground, foliage, and building. For cross-dataset evaluation, three areas of interest (AOIs) from MVS3D with a 0.3 m GSD are used. They provide four semantic categories (ground, foliage, building, and water) and are resized to 768×768 with adjusted RPC models. For all three datasets, RPC models are refined via bundle adjustment.

Table 1: Quantitative comparison with generalizable methods on the DFC19 and SpaceNet4 datasets. For 2D open-vocabulary baselines (ProxyCLIP, Trident, and SegEarth-OV), novel-view images are first rendered using HiSplat and then segmented in 2D. (**Bold** indicates best, underline indicates second best.)

Methods	DFC19					SpaceNet4				
	mF1↑	mIoU↑	PSNR↑	SSIM↑	LPIPS↓	mF1↑	mIoU↑	PSNR↑	SSIM↑	LPIPS↓
ProxyCLIP	29.12	20.07	<u>16.88</u>	0.381	<u>0.564</u>	53.07	37.64	16.46	<u>0.321</u>	<u>0.488</u>
Trident	29.93	20.79	<u>16.88</u>	0.381	<u>0.564</u>	47.86	34.08	16.46	<u>0.321</u>	<u>0.488</u>
SegEarth-OV	31.80	<u>22.37</u>	<u>16.88</u>	0.381	<u>0.564</u>	<u>53.66</u>	<u>38.11</u>	16.46	<u>0.321</u>	<u>0.488</u>
LSM-2V	31.42	19.95	15.71	0.402	0.603	46.76	31.22	<u>17.05</u>	0.257	0.655
LSM-3V	32.19	20.58	16.52	0.431	0.600	—	—	—	—	—
Uni3R	<u>34.18</u>	22.20	16.56	<u>0.450</u>	0.604	43.08	28.58	16.54	0.258	0.654
SkySplat-OV	62.42	48.12	17.63	0.531	0.443	67.40	51.02	18.49	0.377	0.400

Table 2: Cross-dataset generalization. Models trained on DFC19 and SpaceNet4 are directly evaluated on MVS3D. For 2D open-vocabulary baselines, reference-view predictions are projected to the object space using HiSplat. (**Bold** indicates best, underline indicates second best.)

Methods	DFC19 → MVS3D				SpaceNet4 → MVS3D			
	mF1↑	mIoU↑	MAE↓	PAG _{2.5} ↑	mF1↑	mIoU↑	MAE↓	PAG _{2.5} ↑
ProxyCLIP	<u>51.96</u>	36.06	18.01	9.16	50.36	34.64	16.72	9.87
Trident	51.94	<u>36.17</u>	18.01	9.16	<u>50.58</u>	<u>34.96</u>	16.72	9.87
SegEarth-OV	49.98	34.08	18.01	9.16	48.80	33.03	16.72	9.87
LSM-2V	41.74	30.04	<u>11.37</u>	<u>16.19</u>	38.79	27.97	<u>11.37</u>	<u>16.19</u>
LSM-3V	38.04	26.52	16.55	9.60	32.14	23.07	16.55	9.60
Uni3R	26.32	17.14	11.92	14.95	29.21	18.17	11.92	14.95
SkySplat-OV	58.79	43.22	4.28	38.04	60.17	45.26	4.37	41.48

Evaluation Metrics. We report standard image quality metrics for photometric fidelity, including pixel-level PSNR, patch-level SSIM [40], and feature-level LPIPS [46]. Semantic segmentation performance is evaluated using mean Intersection-over-Union (mIoU) and mean F1 score (mF1), following MVSR3D [16]. Geometric reconstruction quality on the eight AOIs is assessed using mean absolute error (MAE) and the percentage of accurate grids $PAG_{2.5}$, defined with a 2.5 m error threshold [15].

Implementation Details. All experiments are conducted on a server equipped with eight NVIDIA GeForce RTX 4090 GPUs (24 GB memory each), running Ubuntu 22.04. The model is trained using the AdamW optimizer with a learning rate of 2×10^{-4} . In all experiments, we sample 64 height candidates when constructing the cost volume.

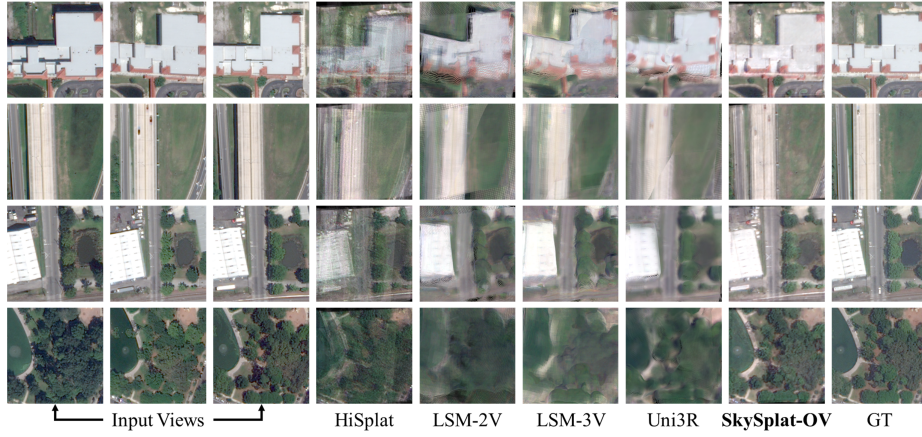


Fig. 3: Qualitative comparison of novel view synthesis on the DFC19 test set with 3 input images.

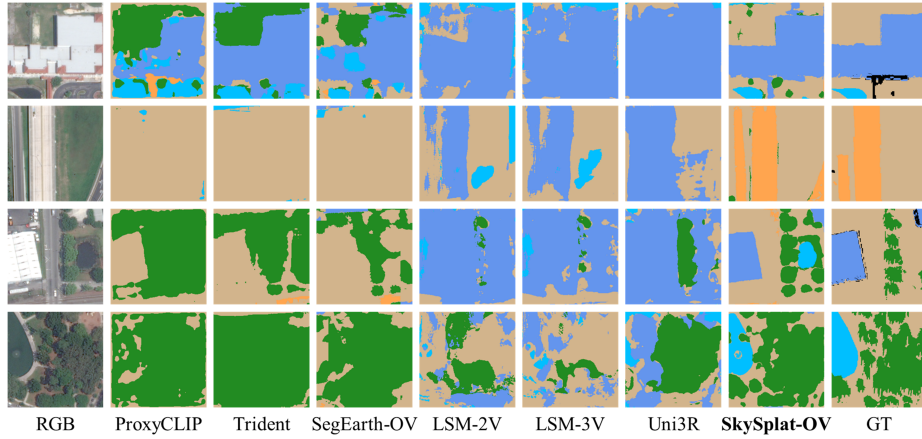


Fig. 4: Qualitative comparison of open-vocabulary novel view segmentation on the DFC19 test set with three input images.

4.2 Generalization Results

We first visualize the 3D Gaussian fields by concatenating 64 non-overlapping outputs, each reconstructed from three 256×256 images, into a large-scale scene. As shown in Figure 5, SkySplat-OV produces more coherent radiance and language fields than Uni3R, along with more accurate segmentation, highlighting the effectiveness of the RPC-based framework for scene understanding.

Evaluation of Novel View Synthesis. We compare SkySplat-OV with the SOTA method HiSplat [37] on two datasets. In addition, comparisons are made

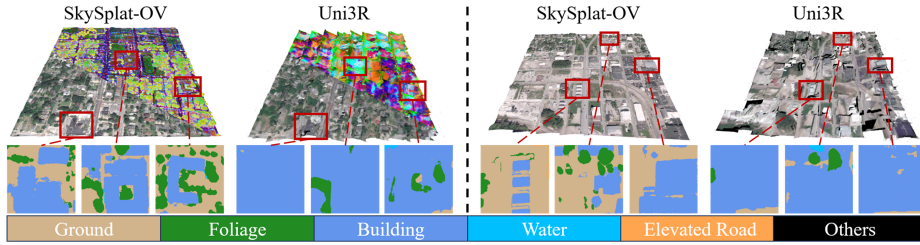


Fig. 5: Comparisons of 3D Gaussian field representations (top) and segmentation results (bottom). For each method, the 3D Gaussian visualization shows the radiance field on the left half and the language field on the right half, where PCA [1] is used for visualization. The dashed line separates JAX072 (left) and OMA315 (right).

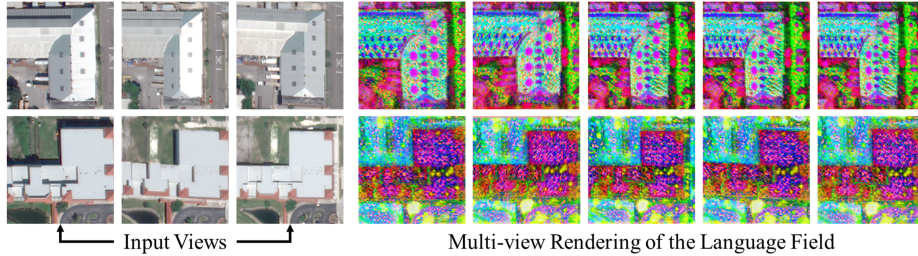


Fig. 6: Language field visualization. Multi-view rendered features show strong view consistency, demonstrating the lifting of 2D features into a 3D language field. PCA [1] is used for visualization.

with LSM [13] and Uni3R [36], both of which can perform novel view synthesis and novel view segmentation. HiSplat is used to render novel views for the open-vocabulary baselines, and its synthesis quality is reported jointly in Table 1.

As shown in Table 1, SkySplat-OV consistently outperforms all baselines across all metrics. On the DFC19 dataset, SkySplat-OV achieves a PSNR of 17.63 dB, SSIM of 0.531, and LPIPS of 0.443, compared to 16.88 dB, 0.381, and 0.564 for HiSplat, respectively. On the SpaceNet4 dataset, SkySplat-OV achieves a PSNR of 18.49 dB, SSIM of 0.377, and LPIPS of 0.400, compared to 16.46 dB, 0.321, and 0.488 for HiSplat, respectively. Figure 3 demonstrates that SkySplat-OV synthesizes sharper novel views with fewer artifacts than the baselines.

Evaluation of Open-vocabulary Novel View Segmentation. We compare SkySplat-OV with LSM and Uni3R. In addition, we include comparisons with 2D open-vocabulary SOTA methods (ProxyCLIP [21], Trident [35], SegEarth-OV [24], which perform open-vocabulary segmentation on novel views rendered by HiSplat.

Quantitative results on the DFC19 and SpaceNet4 datasets are reported in Table 1. SkySplat-OV consistently outperforms all compared methods, includ-

Table 3: Quantitative comparison with per-scene optimization methods across eight AOIs. Results marked with * are taken from prior work [17]. Each cell reports MAE (meters)/mIoU (%). (**Bold** indicates best, underline indicates second best.)

Methods	JAX004	JAX068	JAX260	OMA212	OMA315	IARPA001	IARPA002	IARPA003	Time
S-NeRF*	3.28/-	7.47/-	4.88/-	3.24/-	<u>2.98</u> /-	4.97/-	9.71/-	6.55/-	6.62h
Sat-NeRF*	<u>3.27</u> /-	<u>6.53</u> /-	5.28/-	<u>3.16</u> /-	2.99/-	4.63/-	6.65/-	<u>4.92</u> /-	7.39h
EOGS*	3.31/-	6.67/-	6.41/-	9.08/-	6.38/-	5.90/-	13.79/-	14.83/-	<u>4.60min</u>
Feature-3DGS	3.30/ <u>34.78</u>	7.77/ <u>29.32</u>	<u>3.86</u> / <u>22.01</u>	8.25/ <u>20.01</u>	7.79/ <u>22.99</u>	<u>3.33</u> / <u>13.08</u>	<u>6.09</u> / <u>19.81</u>	5.50/ <u>19.64</u>	3.82h
LangSplatV2	4.79/11.74	9.59/22.41	5.41/15.92	10.02/1.47	7.94/13.45	3.60/12.66	6.75/18.01	5.87/17.15	2.92h
SkySplat-OV	1.73/47.85	3.71/37.50	2.48/49.28	1.13/28.70	2.04/55.50	3.13/37.76	3.94/41.92	4.01/38.66	9.82s

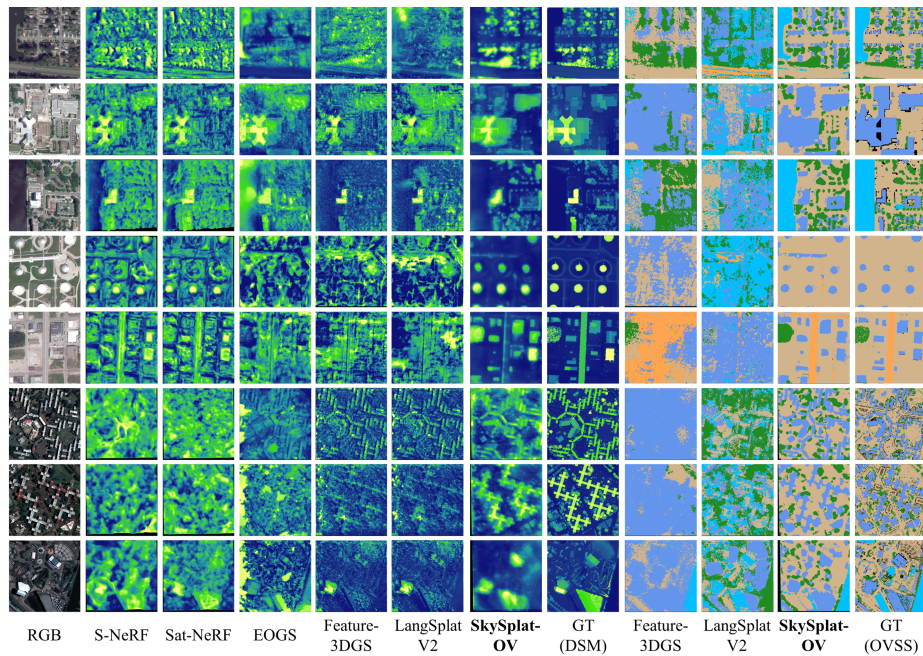


Fig. 7: Predicted DSMs and corresponding OVSS results for eight AOIs. From top to bottom: JAX004, JAX068, JAX260, OMA212, OMA315, IARPA001, IARPA002, and IARPA003.

ing the 2D open-vocabulary SOTA methods. As shown in Figure 4, SkySplat-OV generates semantic segmentation results that are both sharp and well-defined. This improvement can be attributed to distilling SAM3 features into a unified language-embedded 3DGS field, enabling more effective semantic learning and sharper boundaries. Importantly, whereas 2D open-vocabulary methods frequently exhibit inconsistencies across views, SkySplat-OV preserves high consistency across multiple perspectives. To further demonstrate this, we visualize the language field using PCA [1] in Figure 6.

Table 4: Ablation study on the DFC19 dataset. We evaluate the contributions of GRS-Fusion (Geometry-Radiance-Semantics Fusion), R.H.S. (Relative Height Supervision), and F.D. (Feature Distillation). (**Bold** indicates best.)

GRS-Fusion	R.H.S.	F.D.	mF1↑	mIoU↑	PSNR↑	SSIM↑
			–	–	15.90	0.425
		✓	50.20	37.28	15.68	0.420
✓		✓	57.31	43.12	17.64	0.496
	✓	✓	60.65	46.38	17.23	0.526
✓	✓	✓	62.42	48.12	17.63	0.531

Table 5: Ablation of Feature Distillation on the DFC19 dataset. Methods with “+ SAM3” denote applying SAM3 for segmentation on rendered novel views.

Methods	mF1↑	mIoU↑	PSNR↑	SSIM↑
HiSplat + SAM3	29.47	19.34	16.88	0.381
SkySplat + SAM3	34.03	22.44	17.53	0.546
SkySplat-OV + SAM3	43.45	31.97	17.63	0.531
SkySplat-OV	62.42	48.12	17.63	0.531

Evaluation of Cross-Dataset Generalization. To assess the generalization capability of SkySplat-OV, we conduct two cross-dataset evaluations. Specifically, we train models on DFC19 (Jacksonville and Omaha, US) or SpaceNet4 (Atlanta, US), and directly test them on MVS3D (San Fernando, Argentina). As shown in Table 2, despite the test dataset exhibiting significantly different view-point distributions and image appearances from the training datasets, SkySplat-OV achieves accurate geometric reconstruction and precise semantic segmentation. Notably, the geometric accuracy improves substantially, with MAE reduced from 18.01 m and 16.72 m for HiSplat to 4.28m and 4.37m for SkySplat-OV.

4.3 Per-Scene Optimization Results

Table 3 reports quantitative comparisons between SkySplat-OV and the SOTA per-optimization methods, including geometry-only reconstruction methods (S-NeRF [11], Sat-NeRF [31], and EOGS [2] as well as open-vocabulary scene understanding methods (LangSplatV2 [26] and Feature-3DGS [49]). Benefiting from its strong generalization capability, SkySplat-OV achieves the best overall performance while dramatically reducing the time required for scene understanding.

Specifically, SkySplat-OV reduces the average reconstruction time from several minutes (e.g., 4.60 minutes for EOGS) or even several hours (e.g., 6–8 hours for NeRF-based methods) to only 9.82 seconds. Meanwhile, the scene understanding time is shortened from several hours for LangSplatV2 (2.92 hours) and Feature-3DGS (3.82 hours) to 9.82 seconds with a single feed-forward pass. Notably, despite being nearly 1070× faster than LangSplatV2, SkySplat-OV

Table 6: Ablation of camera model under the pinhole setting. (Bold indicates best.)

Methods	DFC19			SpaceNet4		
	mF1↑	mIoU↑	LPIPS↓	mF1↑	mIoU↑	LPIPS↓
LSM (pinhole)	31.42	19.95	0.603	46.76	31.22	0.655
Uni3R (pinhole)	34.18	22.20	0.604	43.08	28.58	0.654
SkySplat-OV (pinhole)	52.75	38.45	0.530	64.78	48.13	0.551

achieves substantially higher geometric reconstruction quality across all AOIs, reducing the mean MAE averaged over eight AOIs from 4.68 m (the best-performing baseline, Sat-NeRF) to 2.77 m. In addition, SkySplat-OV significantly improves scene understanding performance, increasing the mean mIoU averaged over eight AOIs from 22.71% (Feature-3DGS) to 42.15%. Figure 7 further highlights the advantages of SkySplat-OV in preserving fine structural details and achieving more accurate scene understanding.

4.4 Ablation Study

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

Ablation of Geometry-Radiance-Semantics Fusion Module. The GRS-Fusion module is designed to refine Gaussian parameter predictions by jointly integrating geometric, radiometric, and semantic information. As shown in Table 4, the effectiveness of GRS-Fusion is consistently validated under both settings: without relative height supervision (comparing Rows 2 and 3) and with relative height supervision (comparing Rows 4 and 5). In all cases, incorporating the GRS-Fusion module leads to consistent improvements across all evaluation metrics, demonstrating its robustness and general applicability.

Ablation of Relative Height Supervision. We further investigate the impact of relative height supervision derived from DA3, which is introduced as an auxiliary signal to facilitate scene understanding. To evaluate its necessity, we compare models trained with and without relative height supervision (Rows 3 and 5 in Table 4). The results clearly indicate that relative height supervision plays a crucial role in improving scene understanding performance.

Ablation of Feature Distillation on Scene Understanding. To evaluate the effectiveness of feature distillation on scene understanding, we compare two paradigms: (1) direct SAM3 segmentation on novel views rendered by HiSplat and SkySplat-OV, and (2) SkySplat-OV, which distills SAM3 features into a

language-embedded 3DGS field. As shown in Table 5, SkySplat-OV consistently achieves the best performance, indicating that the distilled features not only enhance multi-view consistency but also improve scene understanding accuracy through joint geometric, radiometric, and semantic learning.

Ablation of Feature Distillation. To assess whether feature distillation affects the photometric fidelity, we compare novel view rendering quality between SkySplat-OV and SkySplat. As shown in Table 5, SkySplat-OV achieves a higher PSNR (17.63 vs 17.53), indicating that rendering quality is well preserved. Meanwhile, feature distillation yields large gains in open-vocabulary segmentation: SkySplat-OV improves mF1 ($34.03 \rightarrow 62.42$) and mIoU ($22.44 \rightarrow 48.12$) over SkySplat with SAM3.

Ablation of Camera Model. To assess the impact of the camera model, we evaluate all baseline methods under a unified pinhole camera setting by converting RPC cameras to pinhole models using a standard fitting procedure.

As shown in Table 6, we further evaluate a pinhole-camera variant of SkySplat-OV. Although removing explicit RPC modeling degrades performance, SkySplat-OV still consistently outperforms LSM and Uni3R on both DFC19 and SpaceNet4, indicating that the performance gap is not caused by camera model mismatch.

5 Conclusion

We propose SkySplat-OV, a self-supervised framework for open-vocabulary scene understanding from sparse-view satellite imagery. It explicitly integrates the RPC model into the language 3DGS pipeline, refines Gaussian parameters via GRS-Fusion module and relative height supervision, and employs single-stage SAM3 feature distillation, collectively improving scene understanding. SkySplat-OV achieves $1070\times$ faster inference than per-scene optimization methods while maintaining strong cross-dataset generalization.

Limitations. Open-vocabulary segmentation relies on SAM3 and struggles with visually ambiguous categories in satellite imagery. In addition, height estimation depends on MVS and degrades in highly reflective regions. Future work will explore more robust segmentation models and improved geometry estimation.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grants 42530106, 42571507), the Hubei Provincial Natural Science Foundation of China (Grant 2025AFB426), Ant Group and the Supercomputing Center of Wuhan University.

References

1. Abdi, H., Williams, L.J.: Principal component analysis. *Wiley interdisciplinary reviews: computational statistics* **2**(4), 433–459 (2010)
2. Aira, L.S., Facciolo, G., Ehret, T.: Gaussian splatting for efficient satellite image photogrammetry. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 5959–5969 (2025)
3. Bai, N., Yang, A., Chen, H., Du, C.: Satgs: Remote sensing novel view synthesis using multi-temporal satellite images with appearance-adaptive 3dgs. *Remote Sensing* **17**(9), 1609 (2025)
4. Bosch, M., Foster, K., Christie, G., Wang, S., Hager, G.D., Brown, M.: Semantic stereo for incidental satellite images. In: *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. pp. 1524–1532. IEEE (2019)
5. Bosch, M., Kurtz, Z., Hagstrom, S., Brown, M.: A multiple view stereo benchmark for satellite imagery. In: *2016 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*. pp. 1–9. IEEE (2016)
6. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719* (2025)
7. 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)
8. Chen, C., Zhao, L., He, Y., Long, Y., Chen, K., Wang, Z., Hu, Y., Sun, X.: Sem-stereo: Semantic-constrained stereo matching network for remote sensing. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 39, pp. 15758–15766 (2025)
9. 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. Springer (2024)
10. Cheng, J., Chen, H., Xue, Z., Huang, Y., Zhang, Y.: An online exploratory maximum likelihood estimation approach to adaptive kalman filtering. *IEEE/CAA Journal of Automatica Sinica* (2024)
11. Derksen, D., Izzo, D.: Shadow neural radiance fields for multi-view satellite photogrammetry. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 1152–1161 (2021)
12. Dutta, S., Vasim, A., Gole, S., Rezaatofghi, H., Banerjee, B.: Aeroseg: Harnessing sam for open-vocabulary segmentation in remote sensing images. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 2254–2264 (2025)
13. Fan, Z., Zhang, J., Cong, W., Wang, P., Li, R., Wen, K., Zhou, S., Kadambi, A., Wang, Z., Xu, D., et al.: Large spatial model: End-to-end unposed images to semantic 3d. *Advances in neural information processing systems* **37**, 40212–40229 (2024)
14. Gao, J., Liu, J., Ji, S.: 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*. pp. 6148–6157 (2021)
15. Gao, J., Liu, J., Ji, S.: 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 (2023)

16. Huang, X., Liu, X., Wan, Y., Zheng, Z., Zhang, B., Wang, Y., Guo, H., Zhang, Y.: Mvsr3d: An end-to-end framework for semantic 3d reconstruction using multi-view satellite imagery. *IEEE Transactions on Geoscience and Remote Sensing* (2025)
17. Huang, X., Liu, X., Wan, Y., Zheng, Z., Zhang, B., Xiong, M., Pei, Y., Zhang, Y.: Skysplat: Generalizable 3d gaussian splatting from multi-temporal sparse satellite images. *arXiv preprint arXiv:2508.09479* (2025)
18. 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)
19. Kerr, J., Kim, C.M., Goldberg, K., Kanazawa, A., Tancik, M.: Lerf: Language embedded radiance fields. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 19729–19739 (2023)
20. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 4015–4026 (2023)
21. Lan, M., Chen, C., Ke, Y., Wang, X., Feng, L., Zhang, W.: Proxyclick: Proxy attention improves clip for open-vocabulary segmentation. In: *European Conference on Computer Vision*. pp. 70–88. Springer (2024)
22. Le Saux, B., Yokoya, N., Hansch, R., Brown, M., Hager, G.: 2019 data fusion contest [technical committees]. *IEEE Geoscience and Remote Sensing Magazine* **7**(1), 103–105 (2019)
23. Leotta, M.J., Long, C., Jacquet, B., Zins, M., Lipsa, D., Shan, J., Xu, B., Li, Z., Zhang, X., Chang, S.F., et al.: Urban semantic 3d reconstruction from multi-view satellite imagery. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. pp. 0–0 (2019)
24. Li, K., Liu, R., Cao, X., Bai, X., Zhou, F., Meng, D., Wang, Z.: Segearth-ov: Towards training-free open-vocabulary segmentation for remote sensing images. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 10545–10556 (2025)
25. Li, Q., Sun, J., An, L., Su, Z., Zhang, H., Liu, Y.: Semanticsplat: Feed-forward 3d scene understanding with language-aware gaussian fields. *arXiv preprint arXiv:2506.09565* (2025)
26. Li, W., Zhao, Y., Qin, M., Liu, Y., Cai, Y., Gan, C.: Langsplatv2: High-dimensional 3d language gaussian splatting with 450+ fps. In: *Annual Conference on Neural Information Processing Systems* (2025)
27. Li, X., Huang, Y., Si, S., Zhang, H., Ji, Y., Zhang, Y.: Vbro: Vehicle behavior guided reliable road surface constraints for lidar self-localization in urban environments. *IEEE Transactions on Aerospace and Electronic Systems* (2025)
28. Liao, P., Zhang, X., Chen, G., Wang, T., Li, X., Yang, H., Zhou, W., He, C., Wang, Q.: S 2 net: A multitask learning network for semantic stereo of satellite image pairs. *IEEE Transactions on Geoscience and Remote Sensing* **62**, 1–13 (2023)
29. Lin, H., Chen, S., Liew, J., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647* (2025)
30. Liu, Y., Bai, S., Li, G., Wang, Y., Tang, Y.: Open-vocabulary segmentation with semantic-assisted calibration. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 3491–3500 (2024)
31. Marí, R., Facciolo, G., Ehret, T.: 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*. pp. 1311–1321 (2022)

32. 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)
33. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research Journal* pp. 1–31 (2024)
34. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: *International conference on machine learning*. pp. 8748–8763. PmLR (2021)
35. Shi, Y., Dong, M., Xu, C.: Harnessing vision foundation models for high-performance, training-free open vocabulary segmentation. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 23487–23497 (2025)
36. Sun, X., Jiang, H., Liu, L., Nam, S., Kang, G., Wang, X., Sui, W., Su, Z., Liu, W., Wang, X., et al.: Uni3r: Unified 3d reconstruction and semantic understanding via generalizable gaussian splatting from unposed multi-view images. *arXiv preprint arXiv:2508.03643* (2025)
37. Tang, S., Ye, W., Ye, P., Lin, W., Zhou, Y., Chen, T., Ouyang, W.: Hisplat: Hierarchical 3d gaussian splatting for generalizable sparse-view reconstruction. In: *The Thirteenth International Conference on Learning Representations*
38. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 5294–5306 (2025)
39. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 20697–20709 (2024)
40. 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)
41. Weir, N., Lindenbaum, D., Bastidas, A., Etten, A.V., McPherson, S., Shermeyer, J., Kumar, V., Tang, H.: Spacenet mvoi: A multi-view overhead imagery dataset. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 992–1001 (2019)
42. Xu, H., Peng, S., Wang, F., Blum, H., Barath, D., Geiger, A., Pollefeys, M.: Depth-splat: Connecting gaussian splatting and depth. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 16453–16463 (2025)
43. Xu, X., Xiong, T., Ding, Z., Tu, Z.: Masqclip for open-vocabulary universal image segmentation. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 887–898 (2023)
44. Ye, J., Wang, N., Wang, X.: Featurenerf: Learning generalizable nerfs by distilling foundation models. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 8962–8973 (2023)
45. Zhang, K., Snavely, N., Sun, J.: Leveraging vision reconstruction pipelines for satellite imagery. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. pp. 0–0 (2019)
46. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 586–595 (2018)

47. Zheng, Y., Wu, W., Li, Q., Wang, X., Zhou, X., Ren, A., Shen, J., Zhao, L., Li, G., Yang, X.: Instructsam: A training-free framework for instruction-oriented remote sensing object recognition. arXiv preprint arXiv:2505.15818 (2025)
48. Zheng, Z., Wan, Y., Zhang, Y., Hu, Z., Wei, D., Yao, Y., Zhu, C., Yang, K., Xiao, R.: Digital surface model generation from high-resolution satellite stereos based on hybrid feature fusion network. *The Photogrammetric Record* **39**(185), 36–66 (2024)
49. Zhou, S., Chang, H., Jiang, S., Fan, Z., Zhu, Z., Xu, D., Chari, P., You, S., Wang, Z., Kadambi, A.: Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 21676–21685 (2024)
50. Zhu, Z., Fan, Z., Jiang, Y., Wang, Z.: Fsgs: Real-time few-shot view synthesis using gaussian splatting. In: *European conference on computer vision*. pp. 145–163. Springer (2024)