ID-Splat: Propagating Object Identities for Segmenting 3D Aerial-view Scenes

Yijing Wang[1], Xu Tang[1*], Xiangrong Zhang[1], Jingjing Ma[1]

[1]School of Artificial Intelligence, Xidian University, Xi'an, China
yijingwang@stu.xidian.edu.cn

Abstract

High-resolution Earth Observation technologies present unprecedented opportunities for geospatial analysis, yet traditional 2D aerial-view semantic segmentation remains limited by its inability to model spatial relationships and handle object occlusions. While 3D Aerial-view Segmentation (3DAS) has emerged to address these limitations, existing methods predominantly rely on 2D discriminative models pre-trained on natural scenes. These models struggle to accurately recognize aerial-view imagery, resulting in suboptimal performance due to significant domain discrepancies. This paper introduces ID-Splat, a novel object-centric framework that directly leverages multi-view object identities without discriminative information to enhance 3D semantic understanding. ID-Splat implements a two-stage process: first, Mask-object Tracking combines SAM and Point Tracking to establish robust and consistent object identities across multi-view aerial images; second, Object Integration & Propagation assigns these identities to 3D Gaussian Splatting (3DGS) points, enabling complete 3D segmentation through semantic propagation. Experimental results on the 3D-AS dataset demonstrate that ID-Splat significantly outperforms existing methods, particularly under sparse supervision conditions. ID-Splat also achieves state-of-the-art performance while reducing the need for extensive labeled data by effectively leveraging the inherent 3D structure.

Introduction

The advancement of high-resolution Earth Observation (EO) technologies has revolutionized geospatial analysis, creating both unprecedented opportunities and significant challenges (Lu et al. 2025; Fang et al. 2024; Fassnacht et al. 2024; Abdelmajeed and Juszczak 2024). While traditional 2D aerial-view semantic segmentation is widely used in various applications, its 2D representation limits its effectiveness. Specifically, the inability to explicitly model spatial relationships, such as object occlusions, leads to content ambiguities and low practicality. To overcome this, researchers are increasingly focusing on the 3D Aerial-view Segmentation (3DAS) task, which directly performs semantic segmentation within 3D aerial-view scenes under limited labeled data. 3DAS provides benefits through its ability to analyze object structures and their surrounding environment (Cheng et al. 2024b; Fei et al. 2024). This enables the disambiguation of occluded objects and the more accurate delineation of overlapping structures, thereby facilitating a more comprehensive understanding of complex aerial-view environments.

While some progress has been made, the advancement of 3DAS remains in its initial phase. Recent efforts have primarily explored 3D open-vocabulary segmentation methods for natural scenes (Cen et al. 2025; Liu et al. 2023; Qin et al. 2023; Ye et al. 2024) based on 3D Gaussian Splatting (3DGS) (Kerbl et al. 2023). These methods first leverage 3DGS to reconstruct 3D scene geometry. Then, semantic understanding is enriched by integrating 2D pre-trained models' knowledge within reconstructed 3DGS representations. The 2D pre-trained models employed can be broadly categorized into two types: (1) discriminative models, such as CLIP (Radford et al. 2021; Cherti et al. 2023), LSeg (Li et al. 2022), and DEVA (Cheng et al. 2023), which excel at semantic classification and object recognition through learning to distinguish between different categories or objects; and (2) structure-aware models like SAM (Kirillov et al. 2023), which prioritize pixel correspondences and mask boundary delineation without necessarily understanding semantic categories.

However, directly adapting existing 3D open-vocabulary segmentation methods to 3DAS task presents significant challenges (Tang et al. 2025). These challenges primarily arise from their reliance on 2D discriminative models, such as CLIP and DEVA. These models are initially trained on natural scenes so that the domain gaps between their prior knowledge and aerial-view images are significant. Although domain-specific pre-trained models like GeoRSCLIP (Fassnacht et al. 2024) attempt to bridge this gap, they often struggle to maintain a robust understanding of viewpoint and scale variations in multi-view aerial images, as shown in Figure 1 (a). Fortunately, aerial-view images inherently possess distinct advantages: objects are static and exhibit pronounced structural features. For instance, significant structural differences exist between objects like rooftops and vehicles, enabling their clear distinction. Therefore, structure-aware methods, such as SAM (Kirillov et al. 2023) and Point Tracking (Harley et al. 2025), offer a promising avenue for more robust semantic inference. SAM can generate accurate initial object segmentation, while Point Tracking establishes

*Corresponding author

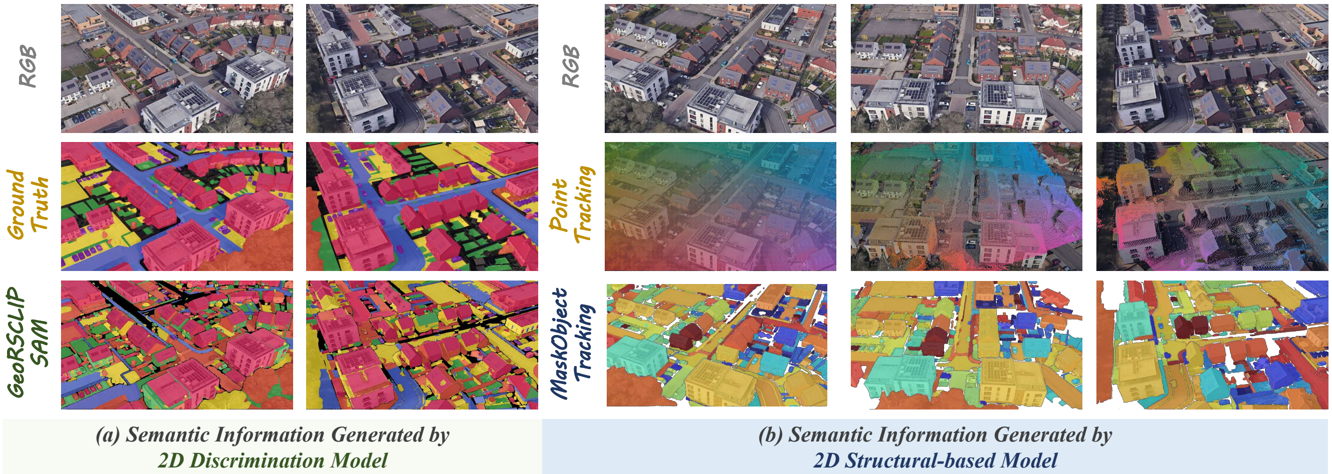

Figure 1: **Illustrates semantic information generated by different 2D models for aerial-view images.** The discriminative approach (a), utilizing GeoRSCLIP and SAM, struggles with accurate object segmentation results compared to the Ground Truth. The structural approach (b) is our proposed Mask-object Tracking, which employs Point Tracking and SAM. It demonstrates more robust segmentation and clearer object identities by effectively delineating individual objects with distinct colors.

pixel correspondences for both foreground objects and the background across multi-view aerial images, with examples of these correspondences visualized in the second row of Figure 1 (b). Their integration promises consistent and accurate multi-view object associations, providing efficient information for enhanced 3DAS performance.

Based on the above analysis, we introduce ID-Splat, an object-centric framework for the 3DAS task. The core innovation of ID-Splat is that it directly leverages multi-view object identities to perform object assignment and propagate semantic understanding within the 3DGS representation, eliminating the need for external discriminative information. This is achieved through a two-stage process. First, ***Mask-object Tracking*** combines SAM and Point Tracking to robustly establish unique and consistent object identities across multi-view aerial images, ensuring reliable object correspondences between different views. Second, ***Object Integration & Propagation*** leverages these established identities to assign object identities to 3DGS points, thereby achieving a coherent, semantically meaningful object-level representation in 3DGS representation. This object-level representation then acts as the semantic anchors, propagating labels and enabling complete object segmentation even in areas where initial labeling is sparse.

Related Work

Taking advantage of advances in 3DGS, a significant amount of recent work has explored its use for 3D segmentation. These methods leverage 3DGS for 3D representation and incorporate semantic understanding by distilling knowledge from 2D discriminative or structure-aware models (Zhou et al. 2023a; Kolides et al. 2023). However, while these methods show promise, they have primarily focused on open-vocabulary natural scene understanding (Qin et al. 2023; Gao et al. 2024) and often exhibit limited applicability to the unique challenges of 3DAS. In this section, we will first introduce existing 3DGS-based segmentation models, then review the 2D pre-trained models they commonly employ, and finally, we will discuss the challenges they face in the 3DAS task.

3DGS-based Segmentation Model: Early 3DGS-based segmentation methods face challenges related to channel inconsistency (Zhou et al. 2024) and high memory consumption (Qin et al. 2023; Gao et al. 2024). To address the issues of channel difference during feature distillation, Feature-3DGS (Zhou et al. 2024) introduced a CNN into 3DGS to guarantee the distillation of arbitrary-dimension semantic features. To mitigate the memory costs of directly embedding high-dimensional semantic features for 3DGS's semantic understanding, LangSplat (Qin et al. 2023) constructed an autoencoder to encode CLIP-distilled semantic features and further improved object boundary precision by learning hierarchical semantics from SAM. MaskField (Gao et al. 2024) eliminated the huge memory usage challenge by distilling SAM's segmented object shapes, which can avoid complex regularization and achieve efficient training speed. Furthermore, moving beyond pixel-level distillation, OpenGaussian (Wu et al. 2024) improves 3D understanding by training 3D-consistent instance features (guided by SAM) and discretizing them with a two-stage codebook. As the same, Gaussian Grouping (Ye et al. 2024) extended 3DGS to facilitate joint reconstruction and segmentation of objects in open-world 3D scenes. This was accomplished by augmenting each Gaussian with a compact identity encoding, enabling object instance grouping supervised by 2D SAM mask predictions and 3D spatial consistency. Instead of time-consuming iterative optimization as traditional methods (Ye et al. 2024; Qin et al. 2023), Occam's LGS efficiently reconstructs the semantic neural field by formulating 3DGS semantic feature assignment as a Maximum-a-

Posteriori (MAP) estimation, thereby significantly reducing computational cost compared to iterative optimization methods without sacrificing accuracy.

2D Pre-trained Models in 3DGS-based Segmentation: This section introduces two types of 2D pre-trained models commonly employed in 3DGS-based segmentation: (1) discriminative models, which facilitate semantic understanding; and (2) structure-aware models, which prioritize accurate segmentation structure. Within the realm of discriminative models (Bhat et al. 2019), CLIP models (Radford et al. 2021; Cherti et al. 2023) have demonstrated a dominant role in 3DGS-based segmentation, primarily due to their robust capacity for analyzing image-text relationships. DEVA further enhances semantic understanding across different viewpoints by enabling multi-view object tracking, thereby ensuring representational consistency. On the other hand, structure-aware models focus on refining segmentation structures. SAM (Kirillov et al. 2023) exemplifies this category through its capacity to generate accurate segmentation masks in a zero-shot fashion, a capability enabled by its pre-training on a massive and diverse dataset (Lin et al. 2014; Gupta, Dollar, and Girshick 2019; Zhou et al. 2019; Kuznetsova et al. 2020). Similarly, DINO (Oquab et al. 2023) facilitates self-supervised object discovery, which can be leveraged to refine segmentation boundaries.

Discussion: As discussed in Figure 1, existing approaches for 3DAS face primary limitations due to their reliance on discriminative models. These models are trained on natural scenes, so they struggle to generalize to the unique characteristics of aerial-view images. In contrast, structure-aware models offer a potential advantage because they focus on structural information. This information is more stable across different viewpoints in aerial-view scenes. Therefore, structure-aware approaches can provide multi-view consistent object guidance, which is crucial for improving segmentation accuracy and robustness in the 3DAS task.

Methodology

Given a dataset of M multi-view aerial images $\{\mathbf{I}_i\}_{i=1}^{M}$ ($\mathbf{I}_i \in \mathbb{R}^{C\times H\times W}$) and a sparse set of segmentation labels $\{\mathbf{L}_j\}_{j=1}^{N}$ ($\mathbf{L}_j \in \{0,1\}^{D\times H\times W}$, $N \ll M$), where H, C and W are the image color channels, height and width, while D denote the number of semantic classes. ID-Splat aims to construct an accurate 3D semantic field $\mathcal{S}$, as illustrated in Figure 2. It upon Occam's LGS (Cheng et al. 2024a), which uses 3DGS to reconstruct scene geometry and incorporates view-based semantic feature assignment for embedding semantic understanding. However, Occam's LGS relies on 2D discriminative priors, limiting its direct applicability to the 3DAS task. To overcome this, ID-Splat employs a two-stage process for enhanced 3DAS performance. First, ***Mask-object Tracking*** leverages structure-aware models to establish robust and consistent object identities across the multi-view aerial images. This provides crucial multi-view consistent object association information essential for subsequent 3D semantic field reconstruction. Second, ***Object Integration & Propagation*** leverages these established object identities to assign corresponding identities to the 3DGS points, thereby propagating semantic understanding. We will first briefly review Occam's LGS, then detail these two stages.

Preliminaries

Based on reconstructed 3DGS, Occam's LGS presents an efficient approach for 3D scene understanding. During the semantic feature assignment stage, it begins by projecting 2D pixel-space language features back to 3D space. This projection utilizes alpha blending weights and incorporates recorded 2D semantic features, followed by a weighted average calculation to determine the final semantic features of the 3DGS points. Here, 2D semantic features are obtained by CLIP and SAM as LangSplat (Qin et al. 2023). The key formula for feature aggregation in Occam's LGS is:

$$\mathbf{f}_i = \frac{\sum_{s\in\mathcal{S}_i} w_i^s \mathbf{f}_i^s}{\sum_{s\in S_i} w_i^s}, \tag{1}$$

where $\mathbf{f}_i$ is the feature of the 3DGS points i, $\mathcal{S}_i$ is the set of views where 3DGS points i is visible, w_i^s are the alpha blending weights indicating the contribution of 3DGS points i in view s, and $\mathbf{f}_i^s$ are the 2D semantic features observed at 3DGS points i's center projection in view s.

Mask-object Tracking

As discussed in Fig. 1, establishing multi-view consistent object correspondences using 2D structure-aware models is crucial for accurate 3DAS. While SAM effectively generates object proposals and Point Tracking densely tracks pixel correspondences, neither alone directly benefits 3DAS. SAM's single-view operation limits its ability to perceive objects across multiple views, and Point Tracking focuses on pixel changes without capturing object edges or structures. Therefore, effectively combining SAM and Point Tracking to overcome these limitations for the 3DAS task is a significant challenge.

To achieve multi-view consistent object IDs, we propose Mask-object Tracking, a prior information extraction stage combining SAM and Point Tracking. First, Point Tracking establishes pixel correspondences between a reference aerial-view image $\mathbf{I}_r$ and other images $\mathbf{I}_i$; let $\mathbf{c}_{r\to i}(x,y)$ be the corresponding coordinates in $\mathbf{I}_i$ for pixel (x,y) in $\mathbf{I}_r$. Next, SAM generates initial object proposals $\mathbf{S}_i = \{\mathbf{m}_{i,k}\}_{k=1}^{K_i}$ for each image $\mathbf{I}_i$, where $\mathbf{m}_{i,k}$ is the k-th object proposals with ID k. We then warp masks from $\mathbf{I}_r$ to each $\mathbf{I}_i$. For each object proposals $\mathbf{m}_{r,k}$ in $\mathbf{I}_r$, we create a warped mask $\mathbf{m}_{r,k\to i}$ in $\mathbf{I}_i$. If a pixel (x',y') in $\mathbf{I}_i$ corresponds to pixel (x,y) in $\mathbf{I}_r$, the object ID at (x',y') in $\mathbf{m}_{r\to i,k}$ is set to k; otherwise, it's without any mask object information. Finally, we refine initial SAM object proposals in each $\mathbf{I}_i$ using warped masks, considering only reliably tracked regions (have the corresponding depends in $\mathbf{c}_{r\to i}(x,y)$). If warped mask $\mathbf{m}_{r,k\to i}$ has the largest overlap with object proposals $\mathbf{m}_{i,l}$ in $\mathbf{S}_i$, SAM's object proposal is replaced by $\mathbf{m}_{i,l}$ with k, propagating the object identities and enforcing cross-view consistency. This process effectively combines the comprehensive object proposals from SAM with the dense pixel correspondences from Point Tracking, resulting in more accurate multi-view consistent object IDs.

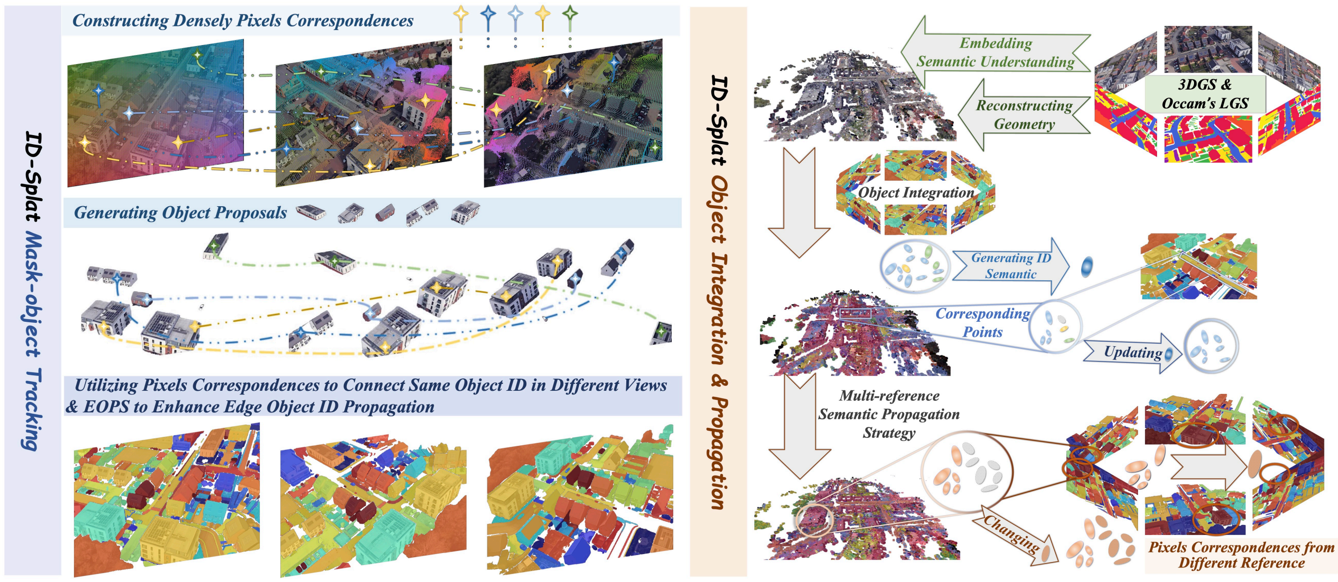

Figure 2: Overview of the ID-Splat framework. The framework consists of two main stages: Mask-object Tracking (left) and Object Integration & Propagation (right). The Mask-object Tracking stage utilizes pixel correspondences and object proposals to establish consistent object IDs across multiple views, with Edge Object Propagation Strategy (EOPS) enhancing edge object ID propagation. The Object Integration & Propagation stage first reconstructs geometry using 3DGS and Occam's LGS, performs object integration by assigning semantic information to 3DGS points, and employs a multi-reference semantic propagation strategy (MSPS) to ensure complete semantic coverage.

Considering the limitations of Point Tracking in establishing correspondences for pixels or objects partially absent in the reference aerial-view image (due to occlusion or out-of-frame scenarios), we also introduce an Edge Object Propagation Strategy (EOPS) to enhance the performance of Mask-object Tracking. Even if the mask $\mathbf{m}_{i,l}$ in the target image $\mathbf{I}_i$ only partially overlaps with corresponding coordinates and the warped mask $\mathbf{m}_{r,k\to i}$, we still replace its object ID with k from $\mathbf{m}_{r,k}$. In this way, EOPS intelligently propagates object IDs based on partial overlap between warped and SAM-generated masks, enabling robust tracking even when complete pixel correspondences are unavailable. EOPS allows us to leverage partial correspondences, enabling object ID propagation even in challenging scenarios where full correspondences are not available.

Object Integration & Propagation

When reliable multi-view consistent object identities are established, the next critical step is to achieve comprehensive 3D aerial scene understanding. Using segmentation labels $\{\mathbf{L}_j\}_{j=1}^{N}$, we can initially embed initial 3D understanding via Occam's LGS as $\mathcal{S}^{\mathcal{T}}$. However, these labeled views are limited, i.e., $(N \ll M)$. Occam's LGS can only associate the semantic understanding with limited 3DGS points corresponding to labeled views. To overcome this, the multi-view object identities play an important role in enhancing 3D aerial-view scene understanding.

A straightforward manner to propagate segmentation labels across multi-view aerial images involves assigning unique IDs to objects in the reference view and associating these IDs with their corresponding segmentation labels. These object ID-label pairs are then transferred to corresponding objects in other views based on Mask-object Tracking, enabling consistent segmentation across the scene. However, this method doesn't understand the 3D world. It's like pasting stickers onto photos – it ignores the fact that the objects are actually arranged in 3D space and that some objects might be hidden behind others or look different from different angles. To overcome the limitations of propagating 2D information, we introduce a novel approach operating directly in 3DGS space. Instead of simply "pasting" labels, we leverage multi-view consistent object IDs to link 2D observations to 3DGS points, effectively "tagging" the 3D scene. Object-labeled 3DGS points are then used to extract features and generate pseudo-labels for unlabeled views.

Specifically, object IDs are assigned from our labeled views to their corresponding 3DGS points. This creates a 3D object world where each 3DGS point is associated with object IDs. Then, to extend semantic awareness to unlabeled views, we employ feature extraction to extend semantic awareness to unlabeled views. This process identifies 3DGS points associated with known object IDs, generating pseudo-labels accordingly:

$$\mathbf{p} = \sum_{i \in \mathcal{O}} \alpha_i \cdot \mathbf{f}_i, \tag{2}$$

where $\mathbf{p}$ is the resulting pseudo-logit, $\mathcal{O}$ represents the set of 3DGS points within the unlabeled view that share the same object ID, and α_i are normalized weights inversely related to the prediction entropy $\mathbf{H}$. This strategy allows us to leverage semantic knowledge from the labeled views to inform the feature representation of objects in the unlabeled views. Furthermore, it can also effectively transfer semantic understanding across the entire scene. By weighting the contribution of each 3DGS point according to the object's entropy, we prioritize high-confidence label assignments. Therefore, the more accurate and robust the pseudo-labels, the more comprehensive the resultant 3D scene understanding. With this, pseudo 3D understanding fields are generated for all unlabeled views' 3DGS as $\mathcal{S}^P$. Finally, the 3D understanding field is represented as $\mathcal{S}^{\mathcal{E}} = \mathcal{S}^{\mathcal{T}} + 0.1 \times \mathcal{S}^P$.

Even with the expansion of object regions achieved through object ID propagation in Mask-object Tracking, some 3DGS points may still lack semantic features (as shown in Figure 2). This stems from the fundamental problem of objects being entirely absent from the initial reference aerial view, preventing their corresponding 3DGS points from being associated with any semantic label. To ensure comprehensive object ID propagation and complete semantic information transfer to all points in the scene representation, we propose a Multi-reference Semantic Propagation Strategy (MSPS). First, we construct a global semantic field, denoted as $\mathcal{S}^{\mathcal{G}}$. Then, we apply Mask-object Tracking using multiple reference frames, $\mathcal{R} = \{\mathbf{r}\}^{R \times C \times H \times W}$, to generate distinct sets of object IDs. For each reference frame $\mathbf{r}$, object assignment is performed independently on all unlabeled views for $\mathcal{S}^{\mathcal{G}}$, associating each 3DGS point i with object ID $o_{i,r}$. Following this, we aggregate object-specific features for each 3DGS point with object ID $o_{i,r}$ in $\mathcal{S}^{\mathcal{G}}$ through a weighted summation (details in Equation 2), resulting in an aggregated feature $\mathbf{f}_{i,r}$. This aggregated feature is then one-hot encoded and serves as a semantic anchor representing the object ID as $\mathbf{f}^o_{i,r}$. However, for some 3DGS points belonging to a particular object ID, no semantic label is present in $\mathcal{S}^{\mathcal{E}}$. In such cases, we assign the aggregated feature $\mathbf{f}^o_{i,r}$ to these previously unlabeled 3DGS points in $\mathcal{S}^{\mathcal{G}}$. Finally, after processing all reference frames, an argmax operation is performed on $\mathcal{S}^{\mathcal{G}}$, followed by one-hot encoding as $\mathcal{S}^{\mathcal{G}'}$. The resulting one-hot encoded label is then assigned to any previously unlabeled 3DGS points, ensuring complete and consistent semantic coverage of the entire 3D scene: $\mathcal{S} = \mathcal{S}^{\mathcal{E}} + \mathcal{S}^{\mathcal{G}'}$. In this way, MSPS bridges the 2D-3D gap by leveraging multi-view object IDs to inject 2D knowledge directly into the 3DGS representation, enabling more accurate and robust 3D aerial scene understanding.

Experiments

Due to space constraints, this section presents a selection of core experiments. Further analysis, comparisons, and visualizations are detailed in the supplementary material. The code and supplementary material can be found at https://github.com/TangXu-Group/3D-AS/tree/main/ID-Splat.

Experimental Settings

Evaluation Dataset and Metrics: We assessed the performance of our method using the multi-view aerial 3D segmentation dataset, 3D-AS (Tang et al. 2025), which encompasses three distinct scene categories: City (characterized by dense buildings, occluded roads, and significant scale variations), Country (featuring open ground, sparsely distributed buildings, and small vehicles), and Port (including ships, industrial buildings, and reflective water).$M = 70$ images and $N = 3$ views' ground truth annotations. The images have an approximate resolution of 1600×900 pixels, corresponding to a pixel resolution of approximately 0.3-0.5 meters. Ground truth annotations consist of pixel-level segmentation labels D for six or eight semantic classes. For evaluation, we used the $2\times$ down-sampled images with indices 14, 21, 35, 42, 49, 56, and 63. Images with indices 0, 7, and 28 were reserved for training. We assessed performance using mean Intersection over Union (mIoU) and mean Accuracy (mAcc) (Liu et al. 2023).

Implementation Details: ID-Splat is based on Occam's LGS (Cheng et al. 2024a) without the prune operation. For Mask-object Tracking, we used the "Default" scale setting for SAM (Kirillov et al. 2023; Qin et al. 2023), and point tracking was performed using AllTracker (Harley et al. 2025). To reduce computational demands, images were processed at half the original resolution (approximately 800×450 pixels). The reference frame for Mask-object Tracking was selected from 10 of the available images. The maximum object ID number in ID-Splat was set to 700. All results are averaged over three runs. All experiments were conducted using NVIDIA GeForce RTX 3090 GPUs (24 GB memory) with the PyTorch framework.

Comparison Experiments

To evaluate ID-Splat's effectiveness, we compared it to state-of-the-art methods in three categories as shown in Table 1: 2D open-vocabulary segmentation (LSeg (Li et al. 2022), MaskCLIP (Fu et al. 2024)), 3D open-vocabulary segmentation (LERF (Kerr et al. 2023), LangSplat (Qin et al. 2023), Feature 3DGS (Zhou et al. 2023b)), and 3DGS-based segmentation (Semantic 3DGS, Gaussian Grouping (Ye et al. 2024)). Detailed experimental settings for these comparisons are provided in the supplementary material. Observing the results reveals that while all methods demonstrate reasonable performance on the 3D-AS dataset, ID-Splat achieves the highest mean mIoU and mAcc. This superior performance directly validates the key contributions of ID-Splat: effectively leveraging multi-view object identities within a 3DGS representation to overcome the limitations of 2D discriminative priors and directly integrating structural awareness into the segmentation process. Specifically, the significant gains in mIoU demonstrate ID-Splat's ability to more accurately delineate object boundaries and segment challenging objects within complex aerial scenes, highlighting the benefits of its Mask-object Tracking and Object Integration & Propagation modules. Furthermore, the high mAcc indicates the effectiveness of ID-Splat in correctly classifying pixels within the 3D scene, showcasing its robust semantic understanding capabilities. These findings confirm that ID-Splat provides a more accurate and robust solution for 3DAS compared to existing state-of-the-art methods.

Furthermore, we provide visualization results in Figure 3, which showcases ID-Splat's superior performance in accu-

Method	City						Country						Port					
	Scene 0		Scene 1		Scene 2		Scene 0		Scene 1		Scene 2		Scene 0		Scene 1		Scene 2	
	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc
LSeg	37.5	67.0	41.1	74.9	44.5	73.2	22.5	31.4	25.4	27.2	17.6	21.5	45.2	77.0	28.2	55.8	24.0	62.0
MaskCLIP	33.1	49.5	28.8	44.9	37.8	55.0	23.8	33.6	30.2	52.0	21.5	41.6	46.9	76.0	35.6	63.4	31.8	67.2
LERF	11.7	34.5	7.2	23.3	7.5	23.8	5.5	20.2	4.0	16.3	4.1	17.6	4.5	14.7	7.6	26.5	4.6	17.7
LangSplat	11.2	26.6	8.3	20.3	1.7	4.1	5.4	12.7	3.0	11.6	2.3	9.2	5.3	10.3	4.1	9.2	2.7	5.9
Feature 3DGS	10.6	28.6	31.5	54.7	35.0	64.6	27.3	45.9	40.1	64.5	31.0	71.2	41.4	70.2	27.4	49.4	26.9	58.3
Semantic 3DGS	64.2	82.8	66.4	84.6	59.3	79.0	53.5	85.2	70.3	94.6	57.6	92.4	67.5	90.6	63.8	87.9	50.6	90.2
Gaussian Grouping	42.9	69.7	19.4	46.2	23.1	47.4	40.6	81.5	50.0	88.5	39.2	86.2	55.3	85.5	36.6	65.3	27.9	68.2
ID-Splat (Ours)	**68.9**	**85.6**	**67.6**	**86.1**	**63.6**	**80.7**	**66.9**	**91.8**	**71.9**	**95.7**	**68.0**	**94.6**	**75.2**	**93.0**	**67.7**	**90.2**	**51.9**	**91.4**

Table 1: Quantitative comparison with state-of-the-art methods across different scene types. We report mIoU and mAcc in percentages (%). The highest scores are highlighted in **bold**.

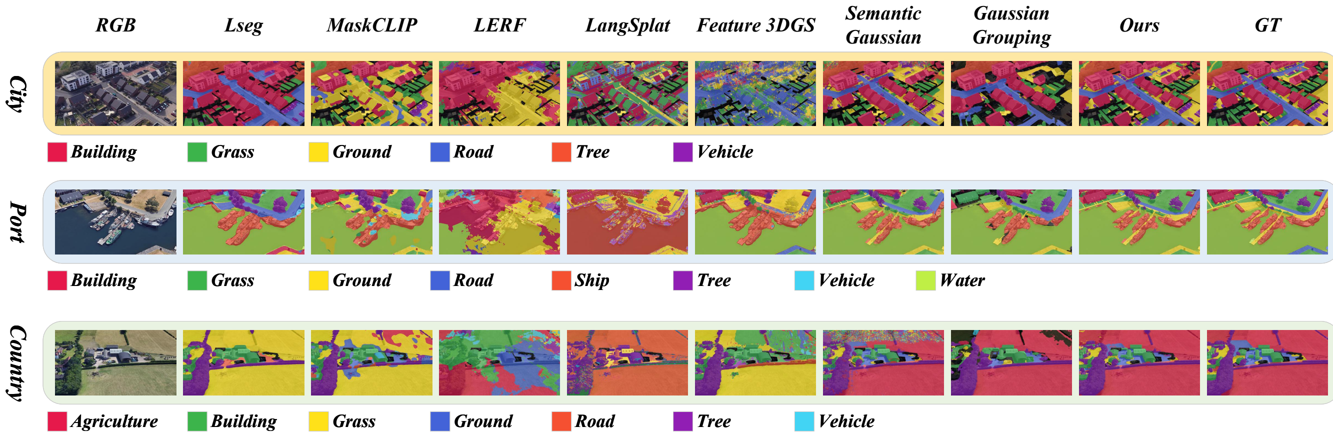

Figure 3: Visualization comparison of semantic segmentation results on aerial imagery. From left to right: RGB input, results from baseline methods, our ID-Splat, and GT. The rows show results for different scene types: City, Port, and Country. ID-Splat demonstrates improved segmentation accuracy, particularly in complex scenes and for delineating object boundaries.

rately segmenting aerial scenes. Notably, ID-Splat enhances boundary delineation, reducing both over-segmentation and misclassifications compared to the baseline methods. Specifically, in the "City" scene, ID-Splat maintains clearer distinctions between buildings and roads. In the "Port" scene, it correctly identifies ship classes, a capability often lacking in other methods such as Lseg, MaskCLIP, and Feature 3DGS. Additionally, within the "Country" scene, ID-Splat achieves a more balanced classification of agriculture, buildings, and grass, overcoming the prevalent poor segmentation of agricultural land observed in many baseline methods. This improved visual fidelity results from ID-Splat's object-centric design and multi-view consistency, leading to more robust and precise semantic understanding.

Ablation Study

Contributions of Core Components: To assess the individual contributions of ID-Splat's components, we performed a comprehensive ablation study. ID-Splat consists of two main stages: Mask-object Tracking and Object Integration & Propagation. Mask-object Tracking fuses SAM's object proposal generation with Point Tracking's pixel-level correspondences to establish initial object IDs. EOPS then refines these object IDs by leveraging partial correspondences, particularly for subtle or boundary-adjacent objects, thereby improving object ID information. Finally, Object Integration & Propagation leverages these refined object identities to enrich the 3DGS representation with contextual information. Within this stage, MSPS is also introduced to ensure complete and consistent semantic coverage of the entire 3D scene. To investigate their functions, we construct the following networks with results summarized in Table 2, as measured by mean mIoU and mAcc:

- ***Net-1***: Basis Model + GT supervision;
- ***Net-2***: Basis Model + GT supervision + Object Integration & Propagation guided by Mask-object Tracking (without EOPS and MSPS);
- ***Net-3***: Basis Model + GT supervision + Object Integration & Propagation guided by Mask-object Tracking (without MSPS);
- ***Net-4***: Basis Model + GT supervision + Object Integration & Propagation guided by Mask-object Tracking ;

Here, "Basis Model" is Occam's LGS, and the "GT supervision" denotes supervision by 3-view segmentation labels. The baseline configuration Net-1 achieved an mIoU of 65.1% and an mAcc of 88.8%, providing a foundation for comparison. Adding Object Integration & Propagation

Network	mIoU (%)	mAcc (%)
Net-1	65.1	88.8
Net-2	66.1	89.4
Net-3	66.6	89.8
Net-4	**66.8**	**89.9**

Table 2: Overall performance of different network configurations. The best results are shown in bold.

Network	mIoU (%)	mAcc (%)
Net-1 †	52.4	80.8
Net-2 †	57.6	84.8
Net-3 †	57.8	85.2
Net-4 †	**58.4**	**85.6**

† Based on Sparse GT.

Table 3: Overall performance of different network configurations with sparse (single-view) GT supervision. The best results are shown in bold.

guided by Mask-object Tracking in Net-2, without EOPS and MSPS, resulted in a slight improvement to 66.1% mIoU and 89.4% mAcc, suggesting the benefit of establishing initial object correspondences. The inclusion of EOPS in Net-3 further increased performance to 66.6% mIoU and 89.8% mAcc, highlighting the importance of refining object correspondences, especially for subtle or boundary-adjacent objects. Our full ID-Splat model, Net-4, incorporating all components, achieved the best results (66.8% mIoU and 89.9% mAcc), demonstrating that MSPS contributes to overall semantic coverage by leveraging multi-view information to ensure that the entire 3D scene is semantically labeled. These results indicate that each component contributes to the overall performance.

Contributions of Core Components under Sparse Ground Truth: To assess the effectiveness of ID-Splat under more realistic, limited supervision scenarios, we conducted an ablation study with sparse GT, using only a single-view segmentation label. Table 3 summarizes the performance of the different network configurations under this sparse GT setting, as measured by mean mIoU and mAcc. The results reveal a more pronounced benefit from our proposed components under sparse supervision compared to the 3-view supervised setting. While the absolute performance is lower due to the limited information, the relative gains achieved by ID-Splat are significantly higher. This demonstrates that our method effectively compensates for the lack of labels, improving semantic understanding by a larger margin than in the 3-view supervised case. For instance, the absolute gain of Net-4 over Net-1 is 6.0% mIoU scores compared to the 3-view supervised performance improvement over Net-1 (1.7%), highlighting ID-Splat's strength in leveraging limited information. This improved performance under sparse supervision underscores the potential for ID-Splat to reduce the need for laborious annotation in real-world 3D scene understanding applications.

Network	mIoU (%)	mAcc (%)
Default	**66.8**	**89.9**
Small	66.8	89.6
Medium	66.6	89.8
Large	66.5	89.8

Table 4: Impact of different SAM scales on 3DAS Performance. Best results are highlighted in bold.

Impact of Different SAM Scale on Performance: We also conducted an ablation study to evaluate the impact of the SAM scale on the overall performance of our ID-Splat framework. We varied the SAM scale, using "Default", "Small", "Medium," and "Large" settings, while keeping all other parameters constant. The results, presented in Table 4, demonstrate the influence of the SAM scale on the mean mIoU and mAcc of the 3D-AS dataset. As observed in Table 4, the "Default" SAM scale achieves the best mean mIoU (66.8%) and mAcc (89.9%) compared to other scales. A smaller scale ("Small") demonstrates a slightly lower mAcc, potentially indicating difficulty in accurately capturing object boundaries or merging adjacent object fragments. Conversely, larger scales ("Medium" and "Large") also perform worse. The degraded performance with "Medium" and "Large" scales suggests an over-segmentation issue. In particular, the "Large" scale shows the lowest result, likely due to over-segmentation. This is indicated by a higher rate of false positives and difficulty with proper object assignment. Since the ID-Splat method relies on 2D-3D label propagation via ID assignment, errors in object identification have a certain impact on the final result.

Conclusion

ID-Splat presents a novel approach to the 3DAS task by directly leveraging multi-view object identities within a 3DGS representation. The Mask-object Tracking effectively fuses SAM and Point Tracking to establish consistent object correspondences, mitigating the challenges associated with relying on 2D discriminative priors prone to domain shift issues. Object Integration & Propagation effectively translates these 2D object identities into the 3DGS space, propagating semantic understanding and enabling a more complete and consistent representation of the scene. Extensive experimental validation on the 3D-AS dataset demonstrates that ID-Splat outperforms existing methods, achieving a significant improvement in mIoU and mAcc. The performance gains are particularly pronounced under sparse supervision, indicating that ID-Splat can reduce the reliance on large labeled datasets. Importantly, our ablation studies highlight the contribution of each component of ID-Splat. This research paves the way for more robust and efficient 3DAS solutions, enabling accurate geospatial analysis with limited labeled data. Future work includes extending ID-Splat to handle dynamic aerial-view scenes and exploring self-supervised learning techniques to further reduce the need for labeled data.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62571387.

References

Abdelmajeed, A. Y. A.; and Juszczak, R. 2024. Challenges and limitations of remote sensing applications in northern peatlands: present and future prospects. *Remote Sensing*, 16(3): 591.

Bhat, G.; Danelljan, M.; Gool, L. V.; and Timofte, R. 2019. Learning discriminative model prediction for tracking. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6182–6191.

Cen, J.; Fang, J.; Yang, C.; Xie, L.; Zhang, X.; Shen, W.; and Tian, Q. 2025. Segment any 3D gaussians. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 1971–1979.

Cheng, H. K.; Oh, S. W.; Price, B.; Schwing, A.; and Lee, J.-Y. 2023. Tracking anything with decoupled video segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1316–1326.

Cheng, J.; Zaech, J.-N.; Van Gool, L.; and Paudel, D. P. 2024a. Occam's LGS: A Simple Approach for Language Gaussian Splatting. *arXiv preprint arXiv:2412.01807*.

Cheng, K.; Long, X.; Yang, K.; Yao, Y.; Yin, W.; Ma, Y.; Wang, W.; and Chen, X. 2024b. Gaussianpro: 3D gaussian splatting with progressive propagation. In *Forty-first International Conference on Machine Learning*.

Cherti, M.; Beaumont, R.; Wightman, R.; Wortsman, M.; Ilharco, G.; Gordon, C.; Schuhmann, C.; Schmidt, L.; and Jitsev, J. 2023. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2818–2829.

Fang, L.; Yang, Z.; Ma, T.; Yue, J.; Xie, W.; Ghamisi, P.; and Li, J. 2024. Open-world recognition in remote sensing: Concepts, challenges, and opportunities. *IEEE Geoscience and Remote Sensing Magazine*, 12(2): 8–31.

Fassnacht, F. E.; White, J. C.; Wulder, M. A.; and Næsset, E. 2024. Remote sensing in forestry: current challenges, considerations and directions. *Forestry: An International Journal of Forest Research*, 97(1): 11–37.

Fei, B.; Xu, J.; Zhang, R.; Zhou, Q.; Yang, W.; and He, Y. 2024. 3D gaussian splatting as new era: A survey. *IEEE Transactions on Visualization and Computer Graphics*.

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, Z.; Li, L.; Jiao, L.; Liu, F.; Liu, X.; Ma, W.; Guo, Y.; and Yang, S. 2024. Fast and Efficient: Mask Neural Fields for 3D Scene Segmentation. *arXiv preprint arXiv:2407.01220*.

Gupta, A.; Dollar, P.; and Girshick, R. 2019. LVIS: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5356–5364.

Harley, A. W.; You, Y.; Sun, X.; Zheng, Y.; Raghuraman, N.; Gu, Y.; Liang, S.; Chu, W.-H.; Dave, A.; Tokmakov, P.; You, S.; Ambrus, R.; Fragkiadaki, K.; and Guibas, L. J. 2025. AllTracker: Efficient Dense Point Tracking at High Resolution. In *ICCV*.

Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics*, 42(4).

Kerr, J.; Kim, C. M.; Goldberg, K.; Kanazawa, A.; and Tancik, M. 2023. LERF: Language Embedded Radiance Fields. In *International Conference on Computer Vision (ICCV)*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.

Kolides, A.; Nawaz, A.; Rathor, A.; Beeman, D.; Hashmi, M.; Fatima, S.; Berdik, D.; Al-Ayyoub, M.; and Jararweh, Y. 2023. Artificial intelligence foundation and pre-trained models: Fundamentals, applications, opportunities, and social impacts. *Simulation Modelling Practice and Theory*, 126: 102754.

Kuznetsova, A.; Rom, H.; Alldrin, N.; Uijlings, J.; Krasin, I.; Pont-Tuset, J.; Kamali, S.; Popov, S.; Malloci, M.; Kolesnikov, A.; et al. 2020. The open images dataset V4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7): 1956–1981.

Li, B.; Weinberger, K. Q.; Belongie, S.; Koltun, V.; and Ranftl, R. 2022. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Liu, K.; Zhan, F.; Zhang, J.; Xu, M.; Yu, Y.; El Saddik, A.; Theobalt, C.; Xing, E.; and Lu, S. 2023. Weakly supervised 3D open-vocabulary segmentation. *Advances in Neural Information Processing Systems*, 36: 53433–53456.

Lu, S.; Guo, J.; Zimmer-Dauphinee, J. R.; Nieusma, J. M.; Wang, X.; Wernke, S. A.; Huo, Y.; et al. 2025. Vision foundation models in remote sensing: A survey. *IEEE Geoscience and Remote Sensing Magazine*.

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*.

Qin, M.; Li, W.; Zhou, J.; Wang, H.; and Pfister, H. 2023. LangSplat: 3D Language Gaussian Splatting. *arXiv preprint arXiv:2312.16084*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.;

et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Tang, X.; Jia, J.; Wang, Y.; Ma, J.; and Zhang, X. 2025. Semantic-aware DropSplat: Adaptive Pruning of Redundant Gaussians for 3D Aerial-View Segmentation.

Wu, Y.; Meng, J.; Li, H.; Wu, C.; Shi, Y.; Cheng, X.; Zhao, C.; Feng, H.; Ding, E.; Wang, J.; et al. 2024. Opengaussian: Towards point-level 3d gaussian-based open vocabulary understanding. *Advances in Neural Information Processing Systems*, 37: 19114–19138.

Ye, M.; Danelljan, M.; Yu, F.; and Ke, L. 2024. Gaussian Grouping: Segment and edit anything in 3D scenes. In *European Conference on Computer Vision*, 162–179. Springer.

Zhou, B.; Zhao, H.; Puig, X.; Xiao, T.; Fidler, S.; Barriuso, A.; and Torralba, A. 2019. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127: 302–321.

Zhou, C.; Li, Q.; Li, C.; Yu, J.; Liu, Y.; Wang, G.; Zhang, K.; Ji, C.; Yan, Q.; He, L.; et al. 2023a. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. *arXiv preprint arXiv:2302.09419*.

Zhou, S.; Chang, H.; Jiang, S.; Fan, Z.; Zhu, Z.; Xu, D.; Chari, P.; You, S.; Wang, Z.; and Kadambi, A. 2023b. Feature 3DGS: Supercharging 3D gaussian splatting to enable distilled feature fields. *arXiv preprint arXiv:2312.03203*.

Zhou, S.; Chang, H.; Jiang, S.; Fan, Z.; Zhu, Z.; Xu, D.; Chari, P.; You, S.; Wang, Z.; and Kadambi, A. 2024. Feature 3DGS: Supercharging 3D gaussian splatting to enable distilled feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21676–21685.