Evolving Semantic Propagation for Aerial Semantic 3D Gaussian Splatting

Zihan Gao, Lingling Li*, Xu Liu, Fang Liu, Licheng Jiao, Puhua Chen, Wenping Ma, Shuyuan Yang

School of Artificial Intelligence, Xidian University, Xi'an, 710071, China
z1han_gao@163.com

Abstract

Semantic understanding of large-scale aerial scenes represents a critical challenge in 3D computer vision, hindered by the prohibitive cost of dense annotation. This paper introduces EvoPropGS, a novel approach for the semantic segmentation of 3D Gaussian Splatting models that requires only minimal supervision. Our core insight is to leverage the inherent structural repetitions within aerial environments to propagate semantic information from a sparse set of annotations across the entire 3D scene. Our approach constructs a prompt library by pairing SAM-generated mask candidates with DINOv2 feature embeddings from annotated views. For unannotated regions, we generate pseudo-labels by matching region proposals with these featured prompts via cosine similarity. We then formulate optimal prompt selection as a discrete optimization problem solved via evolutionary search, guided by our novel fitness function that evaluates both 3D consistency and 2D semantic coherence. Extensive experiments demonstrate that EvoPropGS achieves accurate segmentation with only 2 percent annotated pixels.

Introduction

A comprehensive understanding of 3D Aerial scenes represents a fundamental goal in computer vision and remote sensing, requiring the construction of detailed and accurate representations of vast environments. Recently, radiance fields (Mildenhall et al. 2021; Tewari et al. 2022; Fei et al. 2024; Tancik et al. 2022; Lin et al. 2024; Gao et al. 2024a,c) have emerged as powerful representations for modeling 3D aerial scenes, capturing how light interacts with and emanates from surfaces throughout a volume. Within this paradigm, 3D Gaussian Splatting (3DGS) (Kerbl et al. 2023) marks a pivotal advance, representing scenes as explicit 3D Gaussian primitives that can be efficiently rendered through splatting techniques. Beyond geometric fidelity, however, unlocking the full potential of the radiance fields lies in imbuing the scene with semantic meaning—a capability critical for applications ranging from urban planning and autonomous navigation to virtual reality experiences(Jiao et al. 2023; Huang et al. 2025).

While substantial progress has been made in semantically understanding natural scenes with Gaussian splatting, aerial scenes remain largely unexplored territory. This gap persists primarily due to a significant bottleneck: *the prohibitive cost of acquiring dense semantic annotations for vast aerial data.* Current approaches to this challenge fall into two categories, neither of which adequately addresses the unique characteristics of large-scale aerial scenarios.

Feature-based approaches (Kerr et al. 2023; Liu et al. 2023; Qin et al. 2023; Gao et al. 2024b; Zhou et al. 2024) that leverage vision-language models like CLIP (Radford et al. 2021) face substantial limitations in aerial applications. These methods suffer from a pronounced domain gap between CLIP's training data (He et al. 2025), predominantly consisting of ground-level imagery, and the aerial perspectives typical in aerial images. Furthermore, CLIP's relatively coarse feature representations inadequately capture the numerous small objects that characterize aerial scenes, resulting in imprecise semantic boundaries and object identification (Zhou, Loy, and Dai 2022; Kerr et al. 2023). Alternative label-based approaches (Zhi et al. 2021; Kundu et al. 2022; Siddiqui et al. 2023; Liu et al. 2025; Wang et al. 2025) that directly learn semantic labels from annotated 2D views encounter different but equally significant challenges. The immense spatial scale and repetitive object patterns in aerial environments would necessitate an impractically large number of annotated viewpoints to achieve comprehensive coverage. This requirement renders such methods prohibitively labor-intensive for large-scale aerial applications, creating an annotation bottleneck that impedes progress in the field.

Recognizing these challenges, we observe that aerial scenes present a unique opportunity despite their vast scale: *they often contain recurring structures and objects distributed across the landscape*. This characteristic repetition suggests that strategic selection of viewpoints or objects could efficiently propagate semantics across the entire scene. Our key insight leverages the inherent pattern similarities in aerial environments—buildings, roads, vegetation, and other landscape elements maintain consistent visual characteristics across different regions. By exploiting these similarities, we may dramatically reduce annotation requirements while maintaining semantic accuracy, effectively addressing the annotation bottleneck that has constrained progress in aerial scene understanding with radiance fields.

We introduce EvoPropGS, a novel approach for semantic segmentation of aerial 3D Gaussian Splatting models requir-

*Corresponding author

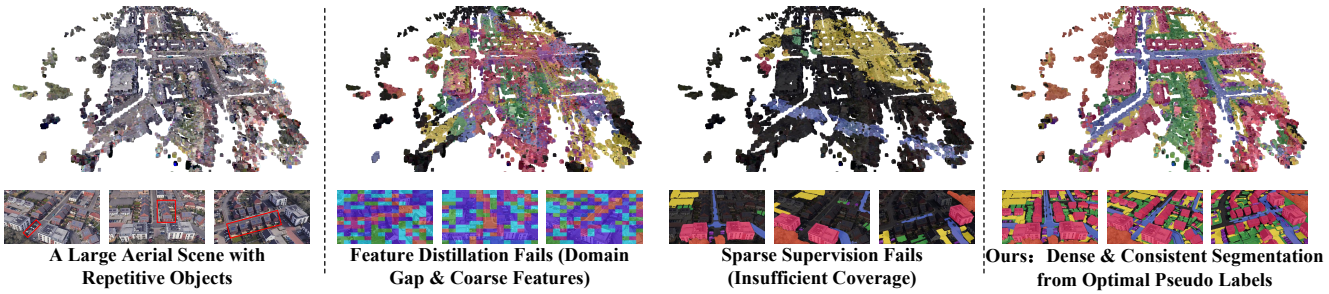

Figure 1: Given a 3D Gaussian Splatting reconstruction from aerial views (left), the goal is to achieve dense and consistent semantic segmentation. Existing methods fall short: feature distillation approaches often produce noisy and incoherent labels due to the domain gap between ground-level training data and aerial perspectives (second from left), while direct supervision from sparse annotations provides insufficient coverage for the vast and repetitive scene structures (third from left). In contrast, our proposed method leverages optimal prompt propagation to effectively utilize the sparse input, generating a dense, consistent, and accurate semantic understanding of the entire 3D scene (right). The top row of each panel shows the full 3D scene segmentation, while the bottom row displays the corresponding 2D semantic maps.

ing minimal annotation effort. Our method efficiently propagates semantic information across entire scenes through a two-phase process. First, we leverage SAM to extract a comprehensive library of high-quality mask candidates from annotated views, each associated with discriminative DINOv2 feature embeddings to form a visual prompt pool. Second, to generate high-fidelity pseudo-labels for unannotated views, we formulate prompt selection as a discrete optimization problem solved via evolutionary search. This process is guided by a novel hybrid fitness function that quantifies both 3D consistency (by back-projecting candidate pseudo-labels onto 3D Gaussians and measuring conflicts with semantic priors) and 2D semantic coherence (by evaluating intra-class similarity versus inter-class distinctiveness). Our framework effectively exploits repetitive structures in aerial scenes to overcome the annotation bottleneck, providing robust signals for pseudo-label generation. Extensive experiments demonstrate that EvoPropGS achieves accurate segmentation while only requiring 2% of pixels to be annotated. The main contributions are summarized as follows:

1. We introduce EvoPropGS, a novel framework for training semantic 3D Gaussian Splatting models in aerial scenes with minimal supervision. To the best of our knowledge, this represents the first approach addressing aerial semantic 3D Gaussian Splatting under minimal supervision, offering a practical solution for 3D aerial scene understanding with 3D Gaussian Splatting.
2. We identify and exploit the inherent structural repetition in aerial environments for semantic knowledge transfer, enabling efficient propagation from sparse annotations to similar areas. We hope this valuable property can inspire further research on training semantic Gaussians for aerial environments.
3. We propose an evolutionary prompt optimization mechanism guided by a hybrid fitness function that evaluates both 3D global consistency through probabilistic back-projection and 2D local semantic coherence through feature similarity.
4. Through extensive experiments, we validate the effectiveness and robustness of our approach across diverse aerial scenes.

Related Work

Semantic Radiance Fields

To imbue radiance fields with semantic understanding, two dominant strategies have emerged in the literature. The first approach distills knowledge from powerful, pre-trained 2D foundation models, leveraging the rich, open-vocabulary capabilities of Vision-Language Models (VLMs) such as CLIP (Radford et al. 2021) and segmentation models like the Segment Anything Model (SAM) (Kirillov et al. 2023; Ke et al. 2024). This methodology lifts 2D semantic features into 3D scene representations by establishing correspondences between visual features and 3D points.

Several significant works exemplify this distillation approach. LERF (Kerr et al. 2023) pioneered a technique for learning dense, multi-scale language fields supervised by CLIP embeddings extracted from image crop pyramids, enabling open-vocabulary queries at varying levels of detail. LEGaussians (Shi et al. 2024) introduced an innovative quantization scheme to compress high-dimensional language features, facilitating their integration into 3D Gaussian primitives. LangSplat (Qin et al. 2024) employed a scene-specific autoencoder to capture low-dimensional latent features from CLIP while utilizing SAM to establish a semantic hierarchy, resolving ambiguities and enabling direct querying across different semantic scales. MaskField (Gao et al. 2024b) decoupled shape and semantic distillation through a mask feature field with learnable queries, circumventing the need to process dense high-dimensional CLIP features during training.

As an alternative to feature distillation, a second major strategy focuses on directly learning from semantic annotations. This approach effectively fuses inconsistent or sparse 2D labels into coherent 3D representations. Semantic-NeRF (Zhi et al. 2021) extends the original NeRF architecture with

a semantic output branch trained jointly with appearance and geometry, generating complete semantic labels by exploiting multi-view consistency. Building on this foundation, Panoptic Lifting (Siddiqui et al. 2023) constructs consistent 3D panoptic fields from noisy 2D masks by solving a linear assignment problem that establishes correspondence between 2D instances and 3D surrogate identifiers. More recently, PLGS (Wang et al. 2025) adapted this concept for 3D Gaussian Splatting, introducing a structured representation specifically designed to handle noisy 2D supervision. Gaussian Grouping (Ye et al. 2024) further advances this paradigm by augmenting Gaussian primitives with learnable "Identity Encodings" supervised by DEVA (Cheng et al. 2023) masks, enabling object-instance grouping of Gaussian primitives.

While existing approaches have shown promise in natural scenes, they encounter substantial limitations in aerial environments. Feature-based methods suffer from a fundamental domain gap between their ground-level training data and aerial perspectives, resulting in inadequate representation of the fine-grained objects prevalent in aerial imagery. Concurrently, label-based approaches fail to propagate semantic information effectively across extensive aerial domains, necessitating prohibitively dense annotations to achieve adequate coverage. Our method directly addresses these constraints by identifying and exploiting the inherent structural repetitions characteristic of aerial scenes, enabling efficient semantic propagation from minimal annotations throughout the entire 3D representation.

Vision Foundation Models

Recent years have witnessed a paradigm shift towards foundation models that can adapt to diverse downstream tasks. CLIP exemplifies this trend by aligning image and text encoders through contrastive learning on 400 million image-text pairs, creating a shared embedding space that enables zero-shot classification via text prompts(Radford et al. 2021). In contrast, purely visual foundation models leverage self-supervised learning without textual guidance(Caron et al. 2021; Oquab et al. 2023; He et al. 2022). By solving pretext tasks such as patch prediction, these models develop robust representations of visual elements. Its pre-training on the diverse dataset ensures effective transfer to fine-grained tasks like segmentation and depth estimation with minimal adaptation. Complementing these semantic-focused models, the Segment Anything Model (SAM) (Kirillov et al. 2023; Ke et al. 2024) addresses geometric understanding through promptable, class-agnostic segmentation. Trained on over one billion masks, SAM learns to generate precise segmentation masks for objects specified by prompts (points, boxes), regardless of their semantic category. This approach yields impressive zero-shot performance on unseen objects, providing geometric proposals that can be effectively integrated into larger vision systems. Building upon these complementary advances, our work introduces a framework that synergizes the precise geometric proposals from SAM with the robust feature representations of DINOv2 (Oquab et al. 2023) to effectively propagate semantic labels from extremely sparse annotations across an entire 3D scene.

Methodology

Our method addresses the challenge of training semantic 3D Gaussians splatting in aerial scenes with minimal supervision. Formally, given a pre-trained 3D Gaussian Splatting (3DGS) model that reconstructs scene geometry and appearance, along with the complete set of N posed RGB images $\{I_i, P_i\}_{i=1}^{N}$ used for initial reconstruction, we require only a small subset of K views $\{I_k, M_k\}_{k=1}^{K}$ where $K \ll N$, accompanied by sparse ground-truth 2D semantic masks M_k. The objective is to leverage this sparse supervision to generate a fully semantic 3DGS model in which each of the millions of Gaussian primitives receives a coherent semantic label. This framework efficiently propagates semantic information from sparse 2D annotations to the comprehensive 3D representation, eliminating the need for exhaustive manual labeling.

Preliminaries

3DGS represents a scene as a collection of explicit anisotropic 3D Gaussians. Each Gaussian is characterized by a set of optimizable attributes: a 3D position (mean) μ, a covariance matrix Σ, an opacity α, and a color c typically represented by Spherical Harmonics (SH) coefficients. This explicit representation allows for high-quality, real-time rendering of novel views. The rendering process involves projecting these 3D Gaussians onto the 2D image plane for a given camera view. A fast, tile-based rasterizer then sorts the Gaussians by depth and blends them together to compute the final color for each pixel. The color C for a pixel is determined by alpha-compositing the N ordered Gaussians that overlap it:

$$C(v) = \sum_{i \in \mathcal{N}} c_i \alpha'_i \prod_{j=1}^{i-1} (1 - \alpha'_j), \tag{1}$$

where α'_i is the effective opacity, calculated by multiplying the learned opacity α_i with the 2D Gaussian distribution evaluated at the pixel location v.

To incorporate semantic information, the standard 3DGS model is extended by augmenting each Gaussian with an additional learnable feature vector f_i. This feature vector is rendered similarly to the color attribute. It is projected and blended onto the 2D image plane using the same differentiable rasterization and alpha-compositing process, resulting in a dense 2D feature map $F(v)$:

$$F(v) = \sum_{i \in \mathcal{N}} f_i \alpha'_i \prod_{j=1}^{i-1} (1 - \alpha'_j) \tag{2}$$

This rendered feature map can then be supervised using features extracted from 2D images, enabling the "lifting" of 2D understanding into the 3D scene representation.

Candidate Prompt Generation

The first stage of our framework is to construct a high-quality library of candidate visual prompts from the sparsely annotated reference views. Let the set of annotated views be $\{(I_k, M_k)\}_{k=1}^{K}$, where $I_k \in \mathbb{R}^{H \times W \times 3}$ is an RGB image and

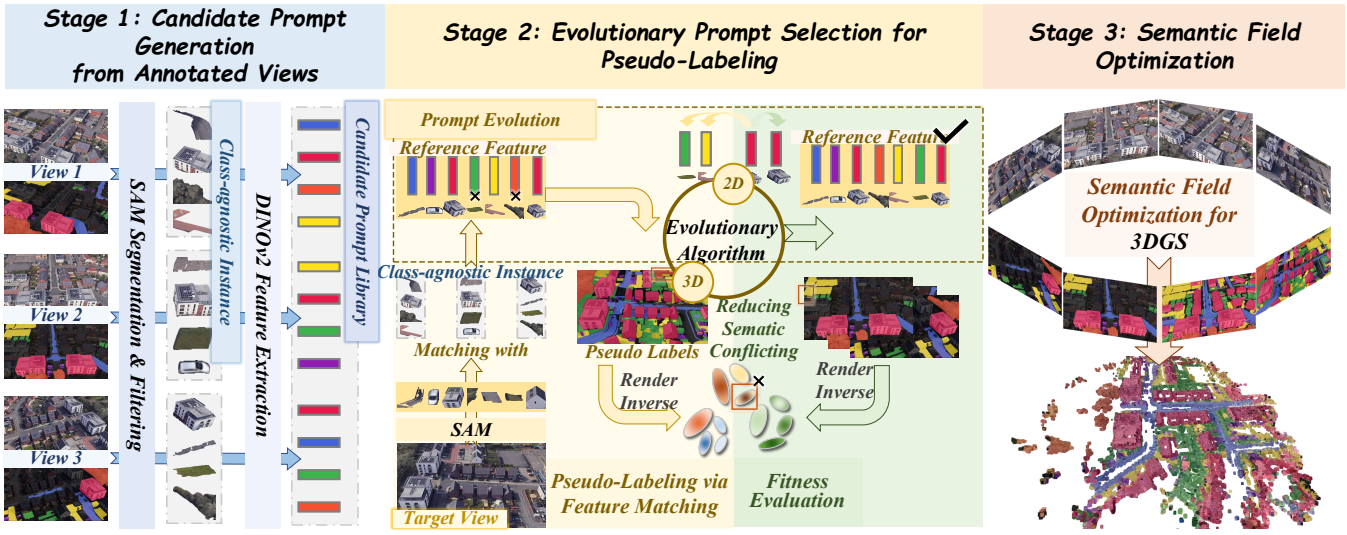

Figure 2: The EvoPropGS framework for semantic 3D Gaussian Splatting with minimal supervision. The method comprises three stages: (1) **Candidate Prompt Generation**, where SAM-generated masks are paired with DINOv2 feature embeddings to create a prompt library. (2) **Evolutionary Prompt Selection**, which uses an evolutionary search algorithm guided by a hybrid 3D-2D consistency fitness function to select the optimal prompts for pseudo-labeling unannotated views. (3) **Semantic Field Optimization**, where the pseudo-labels and ground-truth annotations are used to train semantic 3D Gaussians.

$M_k \in \{0, \ldots, C\}^{H \times W}$ is its corresponding ground-truth semantic mask with C classes.

First, for each reference image I_k, we apply the Segment Anything Model (SAM) to generate a large set of class-agnostic instance masks, denoted as $\mathcal{S}_k = \{s_j\}_{j=1}^{J_k}$, where each s_j is a binary mask. This provides a rich pool of fine-grained region proposals. Next, we filter these proposals to select a subset that are accurately annoated. For each class $c \in \{1, \ldots, C\}$ present in the ground-truth mask M_k, we define its corresponding binary ground-truth region as $M_{k,c}$. We then evaluate each SAM-generated mask $s_j \in \mathcal{S}_k$ using two criteria: the Intersection over Union (IoU) with $M_{k,c}$, and a containment ratio $\mathcal{C}$ defined as:

$$\mathcal{C}(s_j, M_{k,c}) = \frac{|s_j \cap M_{k,c}|}{|s_j|}, \tag{3}$$

where $|\cdot|$ denotes the area of a mask. This containment ratio measures the proportion of the SAM mask that falls within the ground-truth region. We retain only the masks for which $\mathcal{C}(s_j, M_{k,c}) \geq \tau_{\text{contain}}$. This filtered set then serves as candidate prompts for class c from view k.

To create a discriminative descriptor for each candidate prompt, we process the reference image I_k with a pre-trained DINOv2 model to obtain a dense feature map $F_k^{\text{DINOv2}} \in \mathbb{R}^{\frac{H}{p} \times \frac{W}{p} \times D}$, where p is the patch size and D is the feature dimension. For each selected candidate mask $s_{j,c}^*$, we perform mask-pooling on F_k^{DINOv2} to compute its corresponding feature vector $f_{j,c,k} \in \mathbb{R}^D$. This process yields a library of candidate prompts $\mathcal{P}$, storing prompt pairs $\{(s_{j,c}^*, f_{j,c,k})\}$ for each reference view k and class c, providing a reliable foundation for the subsequent optimization stage.

Pseudo-Labeling via Feature Matching

Before optimizing the selection of prompts, we first define the mechanism for generating a pseudo-semantic label for an unannotated target view I_t. This process of pseudo-labeling is guided by a set of candidate visual prompts and leverages the powerful feature representations of a DINOv2 model. The core principle is to match class-agnostic regions in the target image to the semantic concepts defined by the visual prompts from our reference views.

Given a set of candidate prompts for a semantic class c, drawn from our library $\mathcal{P}$, we form a reference feature set $\mathcal{F}_c = \{f_{j,c,k}\}$. Next, for the unannotated target image I_t, we generate a set of class-agnostic region proposals $\mathcal{S}_t = \{s_i\}_{i=1}^{N_t}$ using SAM. For each proposed mask $s_i \in \mathcal{S}_t$, we then extract its corresponding DINOv2 feature vector $f_i^t \in \mathbb{R}^D$ using the same mask-pooling procedure applied to the reference views. Finally, each target region proposal s_i is classified by comparing its feature vector f_i^t against the reference feature sets $\{\mathcal{F}_c\}_{c=1}^C$ of all semantic classes. The similarity score between a target region s_i and a class c is defined as the maximum cosine similarity between the target feature and any of the reference features for that class:

$$\text{Score}(s_i, c) = \max_{f \in \mathcal{F}_c} \left((f_i^t)^T f \right). \tag{4}$$

The target region s_i is then assigned the class label c^* that yields the highest similarity score, provided this score exceeds a predefined threshold. The final pseudo-semantic mask for the target view, M_t^{pseudo}, is constructed by composing all the classified SAM masks. This generated mask serves as the basis for the fitness evaluation in our evolutionary optimization framework.

Evolutionary Prompt Selection

While the pseudo-labeling mechanism can generate a semantic mask for a target view given any set of prompts, the quality of this mask is highly dependent on the chosen prompts. Given our candidate prompt library $\mathcal{P}$, the number of possible prompt combinations for all target views is immense, making an exhaustive search intractable. To address this, we formulate the task of finding the best set of prompts as a discrete optimization problem and employ an evolutionary algorithm to efficiently search the solution space.

Problem Formulation and Chromosome Encoding Our objective is to identify the optimal combination of candidate prompts for pseudo-labeling that maximizes both 3D consistency and segmentation quality. The potential solutions are encoded as chromosomes, which are represented as an integer vector $\mathbf{v}$. For flexibility, we allocate a fixed number of slots per class c from each reference view, with each element $v_{k,c,m}$ either indexing a specific prompt from our library or containing -1 to indicate no selection. Each index is bounded by the number of available candidate prompts for its corresponding class-view pair. The evolutionary algorithm searches the complex solution space to determine the optimal configuration vector $\mathbf{v}^*$ that produces the most coherent 3D segmentation result.

Fitness Evaluation via a Hybrid 3D-2D Consistency Metric Our chromosome fitness is evaluated through a hybrid metric that combines global 3D consistency with local 2D semantic coherence for robust pseudo-label quality assessment. Inspired by (Shen, Yang, and Wang 2024; Cheng et al. 2024), the 3D consistency component of our fitness function relies on a coarse semantic prior embedded within the 3DGS model. This prior is pre-computed by back-projecting one-hot encoded ground-truth labels from K views onto Gaussian primitives. For each Gaussian, we aggregate a semantic prior distribution $\mathbf{P}_i \in [0,1]^C$ by weighted averaging of labels from all visible views:

$$\mathbf{P}_i = \frac{\sum_{k=1}^{K} w_i^k \mathbf{l}_{p_i^k,k}}{\sum_{k=1}^{K} w_i^k + \epsilon}, \tag{5}$$

where w_i^k represents the rendering contribution of Gaussian g_i at its projected 2D center p_i^k in view k, and $\mathbf{l}_{p_i^k,k}$ is the one-hot ground-truth label. Each Gaussian's class is $c_i = \arg\max \mathbf{P}_i$. To evaluate chromosome $\mathbf{v}$, we generate its pseudo-label and determine the predicted class c_i' for each visible Gaussian $g_i \in \mathcal{V}_t$. The 3D consistency score is defined as the proportion of non-conflicting Gaussians:

$$\mathcal{F}_{\text{3D}}(\mathbf{v}) = 1 - \frac{\sum_{i \in \mathcal{V}_t} \mathbb{I}(c_i' \neq c_i)}{|\mathcal{V}_t|}, \tag{6}$$

where $\mathcal{V}_t$ is the set of Gaussians visible in view I_t and $\mathbb{I}(\cdot)$ is the indicator function.

While 3D consistency ensures global coherence, we introduce a second component to evaluate the intrinsic 2D semantic quality of the pseudo-label itself. This 2D coherence score, $\mathcal{F}_{\text{2D}}$, is based on the principle that a good segmentation should exhibit high feature similarity within classes (intra-class) and low similarity between different classes (inter-class). To compute this, we reuse the pre-extracted DINOv2 feature map from the target image I_t. We then group these feature vectors according to the generated pseudo-label M_t^{pseudo} and calculate the average intra-class similarity, $\mathcal{S}_{\text{intra}}$, and the average inter-class similarity, $\mathcal{S}_{\text{inter}}$, using cosine similarity. The 2D coherence score is then defined as:

$$\mathcal{F}_{\text{2D}}(M_t^{\text{pseudo}}) = \max(0, \mathcal{S}_{\text{intra}} - \mathcal{S}_{\text{inter}}). \tag{7}$$

This score rewards segmentations that are both internally cohesive and semantically distinct.

The final fitness combines both metrics with appropriate weights:

$$\mathcal{F}(\mathbf{v}) = \lambda_{\text{3D}} \mathcal{F}_{\text{3D}}(\mathbf{v}) + \lambda_{\text{2D}} \mathcal{F}_{\text{2D}}(M_t^{\text{pseudo}}). \tag{8}$$

This hybrid function guides the evolutionary search toward prompt combinations that yield pseudo-labels with both 3D scene consistency and strong 2D semantic coherence.

Evolutionary Search We employ a Differential Evolution algorithm (Storn 1996) to efficiently navigate the extensive search space of prompt combinations. The algorithm maintains a population of solution vectors $\{\mathbf{v}_i\}$, iteratively refining them through selection, crossover, and mutation operations guided by the fitness function. The optimal prompt combination $\mathbf{v}^*$ is then utilized to generate the final high-confidence pseudo-label for the target view, providing supervision for semantic field optimization. For computational efficiency, we optimize prompts only for a subset of target views at regular intervals, exploiting the observation that nearby views exhibit similar appearance characteristics. During testing, we simply reuse the optimal prompt from the closest optimized view, effectively balancing quality and computational cost.

Semantic Field Optimization

With a complete set of ground-truth and high-confidence pseudo-labels, we optimize the semantic features f_i of each Gaussian primitive. The model is trained by minimizing a pixel-wise cross-entropy loss between the rendered semantic predictions and the target labels across all training views. The total semantic loss $\mathcal{L}_{\text{semantic}}$ is defined as:

$$\mathcal{L}_{\text{semantic}} = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \text{CE}(\text{softmax}(\hat{S}(p)), M(p)), \tag{9}$$

where $\mathcal{P}$ is the set of all pixels across both the original annotated views and the newly generated pseudo-labeled views, $\hat{S}(p)$ is the rendered semantic logit vector for a given pixel p, $M(p)$ is the corresponding one-hot ground-truth or pseudo-label for that pixel, and CE denotes the standard cross-entropy loss function.

Experiments

To validate the approach, a series of experiments were designed. First, a comparison is made against current state-of-the-art techniques to demonstrate the method's performance in a sparse-supervision context. Detailed ablation studies

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-Gaussian	33.5	53.7	28.0	52.0	25.5	47.1	36.9	72.4	48.2	87.0	33.5	82.0	41.2	77.7	28.0	61.9	26.4	72.9
Gaussian Grouping	23.9	44.2	11.9	36.7	12.3	32.3	24.1	67.4	33.4	63.4	26.0	77.5	38.8	73.4	18.7	51.1	17.2	56.8
EvoPropGS (Ours)	**72.4**	**87.8**	**61.3**	**83.4**	**64.5**	**82.7**	**54.7**	**89.0**	**61.8**	**93.1**	**56.7**	**93.5**	**46.3**	**77.7**	**56.5**	**82.1**	**47.7**	**90.1**

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

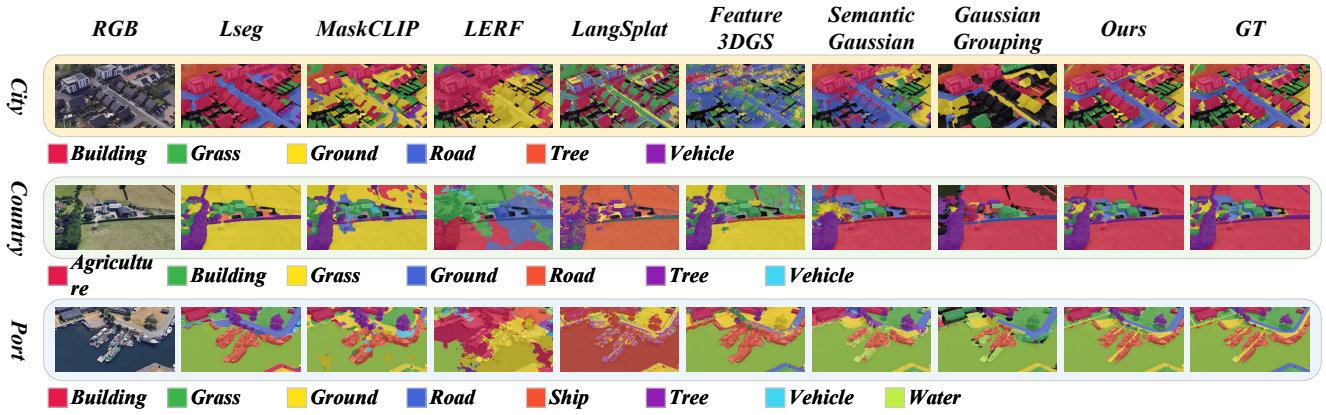

Figure 3: Qualitative comparisons of 3 different scenes. Our method successfully gives the most accurate segmentation maps.

then analyze the contribution of each core component of the framework. Finally, an analysis of different foundation models justifies the selection of DINOv2, and the model's performance is contextualized by comparing it to a model trained with dense supervision.

Datasets and matrics

To evaluate EvoPropGS, we adopt the 3D-AS dataset (**?**). The dataset contains nine diverse real-world scenes captured from Google Earth with a resolution of 1600×900 pixels. Each pixel corresponds to approximately 0.3-0.5 meters in the physical world. Unlike previous approaches that require dense annotations, we employ an extremely sparse labeling strategy where only 2% of pixels are annotated. Specifically, we select only 3-5 objects per class in 3 views for each scene. This minimal supervision setting better simulates real-world scenarios with limited annotation while creating a more challenging environment for evaluating semantic propagation methods. The dataset covers diverse environments, including urban, rural, and port areas, each presenting distinct spatial and semantic distributions that challenge model generalization capabilities.

Comparison Experiments

Our primary evaluation examines an extremely sparse setting where only several objects in 3 views provide supervision, with annotations limited to approximately 2% of pixels—creating a demanding testbed for information propagation. To ensure a comprehensive evaluation, we compare EvoPropGS against a diverse set of state-of-the-art baselines. This includes 2D vision-language segmentation models LSeg and MaskCLIP, which we use for per-view prediction, along with feature-distillation methods such as LERF (Kerr et al. 2023), LangSplat (Qin et al. 2024), and Feature 3DGS (Zhou et al. 2024), which distill knowledge from pre-trained 2D foundation models. For label-based approaches that learn directly from 2D masks, we select the Semantic-Gaussian (a 3DGS adaptation of Semantic-NeRF (Zhi et al. 2021)) and a modified version of Gaussian Grouping (Ye et al. 2024) that uses our sparse annotations instead of Grounding DINO for classification. The quantitative and qualitative results are given in Table 1 and Figure 3.

The results reveal consistent limitations in existing approaches for aerial semantic segmentation. Feature-based methods underperform due to a substantial domain gap between their training data (predominantly ground-level imagery) and aerial perspectives, coupled with reliance on coarse foundation model features inadequate for capturing fine-grained objects in aerial scenes. Label-based approaches perform adequately in annotated regions but lack mechanisms to effectively propagate semantic information across the spatial domain—a critical weakness when han-

Configuration	mIoU	mAcc
(1) Sparse-Supervision Only (No Propagation)	33.4	67.4
EvoPropGS with the following modifications:		
(2) w/ Random Selection	52.6	79.1
(3) w/ Feature Averaging	56.4	80.1
(4) w/ 2D-Fitness Only	53.1	79.2
(5) w/ 3D-Fitness Only	56.2	84.6
(6) w/ Rigid Encoding	50.9	79.0
(7) EvoPropGS (Full Model)	**58.3**	**86.6**

Table 2: Ablation study of our framework.

dling inherently sparse annotations characteristic of large-scale aerial environments. In contrast, EvoPropGS effectively leverages structural repetitions to propagate semantics from minimal annotations while maintaining precise boundaries through evolutionary prompt optimization.

Ablations

To systematically analyze our framework, we conduct a series of ablation studies. To ensure a clear and consolidated comparison, all metrics reported in this section are the average scores across all test scenes.

Analysis of Semantic Propagation Gain. To quantify the performance benefits of EvoPropGS and validate its design principles, we conduct a comprehensive ablation study. The complete EvoPropGS model is first compared against a Sparse-Supervision Only baseline lacking semantic propagation capabilities. We then systematically dissected to attribute improvements to specific components. Three key aspects are analyzed: (1) Search Mechanism: The evolutionary search is replaced with two non-optimizing strategies—Random Selection and Feature Averaging—to demonstrate the necessity of an intelligent search process; (2) Fitness Function: Variants using only 2D-Fitness or 3D-Fitness are tested to highlight the importance of their synergy; (3) Chromosome Encoding: A comparison against a Rigid Encoding scheme validates the benefit of the flexible prompt selection. All results are reported in Table 2.

Our findings confirm the critical contribution of each EvoPropGS component. The substantial improvement over the Sparse-Supervision Only baseline demonstrates the essential role of semantic propagation in aerial scene understanding. Our evolutionary search significantly outperforms both alternative selection methods, underscoring the importance of intelligent prompt optimization. Neither isolated fitness component achieves optimal results, validating our hybrid approach that balances local semantic coherence with global 3D consistency. Finally, our flexible chromosome encoding proves superior to rigid alternatives by enabling more adaptive prompt selection that accommodates diverse semantic patterns in aerial environments. These results collectively confirm that each design element contributes meaningfully.

Analysis of Foundation Model Choices. The effectiveness of EvoPropGS depends critically on the quality of foundation models used for feature representation and geometric proposal generation. To validate these choices, we benchmark our DINOv2-based implementation against two alternative powerful pre-trained vision encoders: CLIP's vision encoder and Masked Autoencoder (MAE). As shown in Table 3, DINOv2 delivers superior performance, which we attribute to its robust self-supervised pre-training methodology that produces highly discriminative and generalizable visual features suited to aerial imagery analysis.

Feature Extractor	mIoU	mAcc
EvoPropGS with different visual foundation models:		
CLIP (ViT-L/14)	33.1	53.2
MAE (ViT-L/16)	40.6	72.3
DINOv2 (ViT-L/14)	**58.3**	**86.6**

Table 3: Analysis of different pre-trained vision models.

Supervision Setting	mIoU	mAcc
EvoPropGS with different supervision levels:		
Sparse-Supervision Only	33.4	67.4
EvoPropGS (Ours, Sparse)	58.3	86.6
Gaussian Grouping (Dense Supervision)	37.2	70.9
Semantic Gaussian (Dense Supervision)	61.4	87.7

Table 4: Comparison with different levels of supervision.

Performance Context with Dense Supervision. To contextualize our method's effectiveness, we compare against models trained with Dense 3-View Supervision, utilizing complete ground-truth from the same three views in our sparse setting. Table 4 demonstrates our approach substantially closes the performance gap between sparse and dense supervision paradigms. Remarkably, EvoPropGS outperforms Gaussian Grouping despite the latter's access to dense labels—a limitation stemming from Gaussian Grouping's dependence on DEVA mask tracking, which performs poorly on small objects prevalent in aerial scenes. These results validate our approach as a cost-effective alternative for aerial semantic 3D Gaussian Splatting.

Conclusion

This paper presents EvoPropGS, a novel framework for training semantic 3D Gaussian Splatting models in aerial scenes with minimal supervision. Our approach identifies and exploits inherent structural repetitions within aerial environments to propagate semantic information effectively from sparse annotations across entire scenes. By integrating SAM's masks with DINOv2's feature embeddings, we develop an evolutionary optimization framework that identifies optimal visual prompts for robust semantic propagation. Our novel hybrid fitness function simultaneously enforces global 3D consistency and local 2D semantic coherence, ensuring coherent segmentation results throughout the scene. Extensive experimental validation demonstrates that EvoPropGS achieves high-quality semantic segmentation while requiring only 2% annotated pixels. We hope to give some insights to the annotation bottleneck that has limited progress in aerial scene understanding with Gaussian Splatting.

Acknowledgements

This work was supported in part by the Joint Funds of the National Natural Science Foundation of China (U22B2054), the National Natural Science Foundation of China (62076192, 62276199, 62431020 and 62276201), the 111 Project, the Program for Cheung Kong Scholars and Innovative Research Team in University (IRT 15R53), the Science and Technology Innovation Project from the Chinese Ministry of Education, the National Key Laboratory of Human-Machine Hybrid Augmented Intelligence, Xi'an Jiaotong University (HMHAI-202404 and HMHAI-202405), the Fundamental Research Funds for the Central Universities (YJSJ25004).

References

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9650–9660.

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. 2024. Occam's LGS: A Simple Approach for Language Gaussian Splatting. *arXiv preprint arXiv:2412.01807*.

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*.

Gao, Z.; Jiao, L.; Li, L.; Liu, X.; Liu, F.; Chen, P.; and Guo, Y. 2024a. Multiplane prior guided few-shot aerial scene rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5009–5019.

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

Gao, Z.; Li, L.; Liu, X.; Jiao, L.; Liu, F.; and Yang, S. 2024c. Uncertainty Guided Progressive Few-Shot Learning Perception for Aerial View Synthesis. *IEEE Transactions on Multimedia*.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.

He, Y.; Zhu, J.; Li, Y.; Huang, Q.; Wang, Z.; and Yang, K. 2025. Rethinking Remote Sensing CLIP: Leveraging Multimodal Large Language Models for High-Quality Vision-Language Dataset. In *International Conference on Neural Information Processing*, 417–431. Springer.

Huang, Z.; Yan, H.; Zhan, Q.; Yang, S.; Zhang, M.; Zhang, C.; Lei, Y.; Liu, Z.; Liu, Q.; and Wang, Y. 2025. A survey on remote sensing foundation models: From vision to multimodality. *arXiv preprint arXiv:2503.22081*.

Jiao, L.; Huang, Z.; Lu, X.; Liu, X.; Yang, Y.; Zhao, J.; Zhang, J.; Hou, B.; Yang, S.; Liu, F.; et al. 2023. Brain-inspired remote sensing foundation models and open problems: A comprehensive survey. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16: 10084–10120.

Ke, L.; Ye, M.; Danelljan, M.; Tai, Y.-W.; Tang, C.-K.; Yu, F.; et al. 2024. Segment anything in high quality. *Advances in Neural Information Processing Systems*, 36.

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.

Kundu, A.; Genova, K.; Yin, X.; Fathi, A.; Pantofaru, C.; Guibas, L. J.; Tagliasacchi, A.; Dellaert, F.; and Funkhouser, T. 2022. Panoptic neural fields: A semantic object-aware neural scene representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12871–12881.

Lin, J.; Li, Z.; Tang, X.; Liu, J.; Liu, S.; Liu, J.; Lu, Y.; Wu, X.; Xu, S.; Yan, Y.; et al. 2024. Vastgaussian: Vast 3d gaussians for large scene reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5166–5175.

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.

Liu, X.; Sun, X.; Xie, H.; Li, Z.; Li, R.; and Zhang, S. 2025. Multi-view Consistent 3D Panoptic Scene Understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 5613–5621.

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*.

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

Qin, M.; Li, W.; Zhou, J.; Wang, H.; and Pfister, H. 2024. Langsplat: 3D language gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20051–20060.

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.

Shen, Q.; Yang, X.; and Wang, X. 2024. Flashsplat: 2d to 3d gaussian splatting segmentation solved optimally. In *European Conference on Computer Vision*, 456–472. Springer.

Shi, J.-C.; Wang, M.; Duan, H.-B.; and Guan, S.-H. 2024. Language embedded 3d gaussians for open-vocabulary scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5333–5343.

Siddiqui, Y.; Porzi, L.; Buló, S. R.; Müller, N.; Nießner, M.; Dai, A.; and Kontschieder, P. 2023. Panoptic lifting for 3d scene understanding with neural fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9043–9052.

Storn, R. 1996. On the usage of differential evolution for function optimization. In *Proceedings of North American fuzzy information processing*, 519–523. Ieee.

Tancik, M.; Casser, V.; Yan, X.; Pradhan, S.; Mildenhall, B.; Srinivasan, P. P.; Barron, J. T.; and Kretzschmar, H. 2022. Block-nerf: Scalable large scene neural view synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8248–8258.

Tewari, A.; Thies, J.; Mildenhall, B.; Srinivasan, P.; Tretschk, E.; Yifan, W.; Lassner, C.; Sitzmann, V.; Martin-Brualla, R.; Lombardi, S.; et al. 2022. Advances in neural rendering. In *Computer Graphics Forum*, volume 41, 703–735. Wiley Online Library.

Wang, Y.; Wei, X.; Lu, M.; and Kang, G. 2025. Plgs: Robust panoptic lifting with 3d gaussian splatting. *IEEE Transactions on Image Processing*.

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.

Zhi, S.; Laidlow, T.; Leutenegger, S.; and Davison, A. J. 2021. In-place scene labelling and understanding with implicit scene representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15838–15847.

Zhou, C.; Loy, C. C.; and Dai, B. 2022. Extract Free Dense Labels from CLIP. In *European Conference on Computer Vision (ECCV)*.

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.