

Few-shot semantic segmentation of UAV imagery for plant conservation: A systematic comparison of foundation models for *Euphorbia resinifera* mapping

Ilias Jennaoui ^{1*}; El Mostafa Bachaoui ¹; Abderrahmane El Ghmari ¹; Mohamed Biniz ²; Abderrazak El Harti ¹; Mostafa Bousbaa ³; Imad El-Jamaoui ⁴; Ayoub Yazini ⁵; Atika Mouaddine ¹

1, Geomatics, Georesources and Environment Laboratory, Faculty of Sciences and Techniques, Sultan Moulay Slimane University, Beni Mellal, Morocco

2, Information Processing and Decision Support Laboratory (TIAD), Multidisciplinary Faculty of Beni Mellal, Sultan Moulay Slimane University, Beni Mellal, Morocco

3, Center for Remote Sensing Applications (CRSA), Mohammed VI Polytechnic University (UM6P), Benguerir 43150, Morocco

4, Department of Agricultural Chemistry, Geology and Pedology, Faculty of Chemistry, University of Murcia, 30100 Murcia, Spain

5, Laboratory of Landscape Dynamics, Risk, and Heritage, Sultan Moulay Slimane University, Beni Mellal, Morocco

Abstract

E-mail:

ilias.jennaoui@usms.ma

Received: 15/06/2026

Revised: 16/10/2026

Acceptance: 19/10/2026

Available Online: 21/10/2026

Published: 01/01/2027

Keywords: Few-shot, Semantic segmentation, UAV, *Euphorbia resinifera*, Conservation

Deep learning has transformed vegetation mapping from UAV imagery, yet state-of-the-art segmentation models require hundreds to thousands of expertly annotated images, creating prohibitive labor costs for conservation applications targeting endangered flora. This study addresses this bottleneck by systematically comparing three foundation models namely: DINOv2 (self-supervised), SegGPT (supervised), and SAM (promptable), for binary segmentation of *Euphorbia resinifera*, an endemic Moroccan succulent of conservation concern, using patches extracted from a single annotated UAV orthomosaic. A single-image training framework is presented where spatial diversity within one high-resolution orthomosaic substitutes for diversity across multiple images, combined with sliding window inference with vote-based fusion for full-resolution deployment. SegGPT achieved $83.74 \pm 2.55\%$ mean IoU across 20 spatially independent test images spanning a 10.5-fold vegetation density gradient (2.01 to 21.11% coverage), outperforming DINOv2 ($78.09 \pm 6.63\%$) and SAM ($47.03 \pm 20.64\%$). The observed collapse in SAM's performance, despite billion-mask pretraining, can be attributed to its architectural constraints. SAM's fixed 1024×1024 input reduced training patches by 76.2%, and its frozen encoder limited trainable parameters to 0.1M versus SegGPT's 316M. These findings establish that fine-tuned encoders substantially outperform frozen-encoder architectures for specialized vegetation (*Euphorbia resinifera*) segmentation, and that single-image foundation model training offers a viable paradigm for conservation deployment, reducing annotation costs from tens of thousands to hundreds of dollars. While demonstrated for a morphologically distinctive species, these principles may inform broader conservation remote sensing applications pending validation across diverse plant forms.



1. Introduction

Deep learning has revolutionized vegetation mapping from unmanned aerial vehicle (UAV) imagery, with state-of-the-art architectures such as U-Net, DeepLabV3+, and Mask R-CNN achieving 85–95% segmentation accuracy [1][2]. However, this performance critically depends on extensive supervised training requiring 100–1000+ expertly annotated images [3][4]. For high-resolution UAV orthomosaics, manual annotation by domain experts requires substantial time investment per image, with operational deployment requiring hundreds to thousands of annotated images [3][4], creating prohibitive labor costs that severely constrain conservation applications targeting endangered or endemic flora that requires urgent monitoring [5].

Mediterranean ecosystems harbor numerous threatened plant species facing climate change, habitat fragmentation, and overexploitation pressures [6]. *Euphorbia resinifera* Berger, an endemic Moroccan cushion-forming succulent, exemplifies this conservation challenge. The species is vulnerable to overexploitation for resin extraction [7], and its scattered distribution across mountainous terrain (1200–1800 m elevation) demands cost-effective mapping solutions. Current supervised deep learning approaches cannot economically scale across *Euphorbia resinifera*'s geographic range given prohibitive annotation requirements.

Foundation models pretrained on large-scale datasets through self-supervised or supervised learning offer promising transfer learning capabilities for low-data regimes [8]. Prominent examples include DINOv2 [9], SegGPT [10], and the Segment Anything Model (SAM) [11], which have demonstrated strong few-shot and zero-shot performance on generic computer vision benchmarks. However, the extent to which these general-purpose models can be effectively transferred to fine-grained vegetation segmentation from UAV imagery remains insufficiently explored.

Recent work has begun exploring foundation models for remote sensing applications with encouraging results [12][13]. However, critical gaps persist. First, most studies employ traditional few-shot learning where training samples come from multiple independent images [14][15], whereas operational conservation scenarios often provide only one or a few expertly annotated exemplar images from initial field campaigns. Second, vegetation segmentation presents unique challenges such as extreme class imbalance, spectral similarity to background, variable morphology, and irregular spatial distribution that differ fundamentally from generic object detection benchmarks [16][17]. Third, no systematic comparison exists between self-supervised, supervised, and promptable foundation model paradigms under controlled conditions for ecological applications. Fourth, the relationship between training sample quantity (controlled by patch extraction strategy) and model performance remains poorly understood, particularly for architecturally constrained models like SAM.

Recent applications demonstrate diverse adaptation strategies across ecological monitoring domains. A previous study achieved zero-shot motion segmentation by combining DINOv2 features with geometric model fusion, eliminating training requirements but necessitating complex algorithmic design [18]. Another research demonstrated that DINOv2 with K-Nearest Neighbors enables multi-species coral segmentation from sparse point labels (5–300 per image) without fine-tuning, achieving 52.60% mIoU with human-in-the-loop selection for 5 points [19]. Furthermore, researchers introduced learnable prompts to adapt SegGPT for few-shot land cover classification, preventing catastrophic forgetting while achieving 35.08% mIoU on combined base and novel classes [14].

Despite the aforementioned research, critical gaps persist. Most studies employ traditional few-shot learning where training samples come from multiple independent images, whereas operational conservation scenarios often provide only one annotated exemplar image [20]. Vegetation segmentation presents unique challenges extreme class imbalance, spectral similarity, variable morphology that differ from generic benchmarks. No systematic comparison exists between self-supervised, supervised, and promptable foundation models under controlled conditions for binary vegetation segmentation.

This study addresses these gaps by systematically comparing DINOv2, SegGPT, and SAM for binary *Euphorbia resinifera* segmentation using patches extracted from a single training image. Methodologically, a single-image training framework is presented where spatial diversity within one high-resolution UAV orthomosaic (extracted as overlapping patches) substitutes for diversity across multiple images, combined with sliding window inference for operational full-image deployment. Operationally, sliding window inference with vote-based fusion will be evaluated for deploying patch-trained models on full-resolution imagery [21], substantially reducing variance compared to

patch-based evaluation. Consequently, the current work aims to establish a viable single-image foundation model training for *Euphorbia resinifera* mapping while illuminating fundamental trade-offs between model architecture, pretraining scale, and fine-tuning data availability. While demonstrated here for a morphologically distinctive species, these principles may inform broader conservation remote sensing applications pending validation across diverse plant forms.

2. Materials and Methods

2.1. Study area and target species

The study was conducted in the Dir El Ksiba and Timoulit regions of the Middle Atlas Mountains, Morocco (Fig. 1), spanning elevations from 1200 to 1800 m above sea level. The region experiences Mediterranean semi-arid climate with precipitation concentrated in winter (November–March) and pronounced summer drought (June–September). Terrain comprises mountainous rocky slopes with heterogeneous vegetation. *Euphorbia resinifera*, commonly known as the resin spurge, is an endemic Moroccan cushion-forming succulent shrub of conservation and ethnobotanical significance. The species forms dense hemispherical mounds 30–60 cm in height and 0.5–2 m in diameter, with characteristic green-gray photosynthetic stems arranged in tetrahedral patterns. Plants contain a toxic latex rich in resiniferatoxin, historically harvested for pharmaceutical applications, leading to overexploitation and vulnerable conservation status [22]. In the selected study area, *Euphorbia resinifera* exhibits irregular spatial distribution with coverage ranging from 2.01% to 21.11% of total image area across sampled locations, creating ideal conditions for testing model generalization across variable target prevalence.

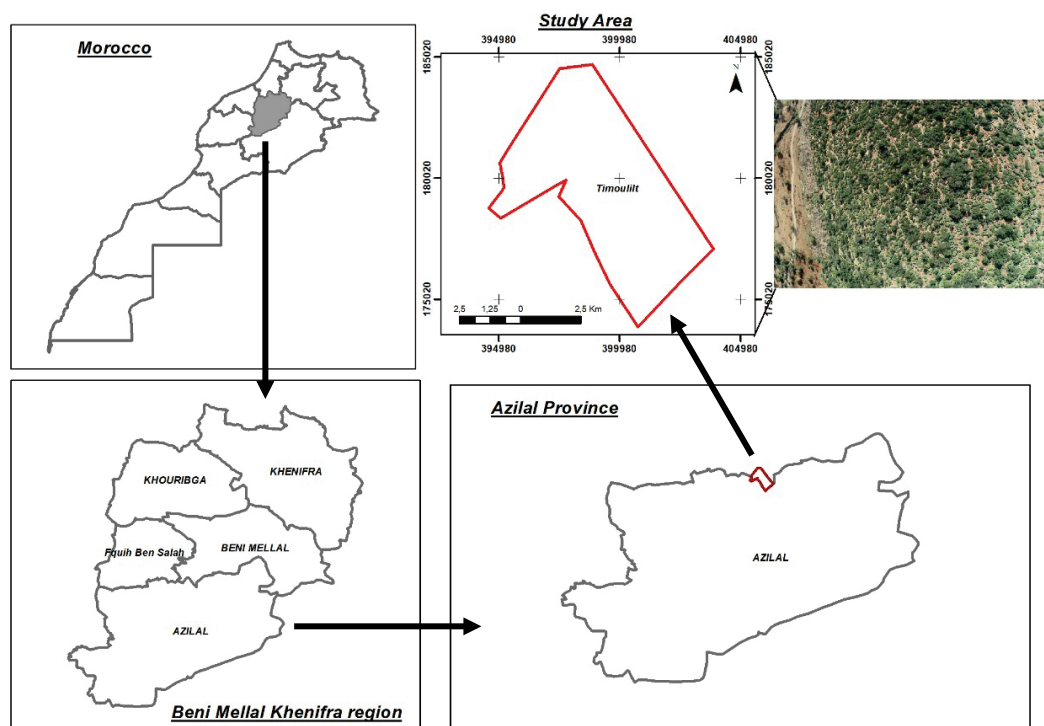


Figure 1. Study area map location within Morrocco with a representative UAV orthomosaic sample of study area

2.2. UAV data acquisition and orthomosaic generation

UAV imagery was acquired using a DJI Phantom 4 RTK quadcopter equipped with a 20-megapixel RGB camera featuring a 1-inch CMOS sensor capturing 3-band visible spectrum imagery [2][21]. Flight missions were conducted during cloud-free conditions in late spring (May–June) between 10:00–14:00 local time to minimize shadow artifacts and vegetation phenology variation. Flight altitude ranged from 50 to 80 m above ground level, yielding ground sample distances (GSD) of 1.3–2.0 cm/pixel. Flight plans followed systematic grid patterns with 75% forward overlap

and 70% side overlap ensuring complete coverage for structure-from-motion reconstruction. The DJI Ground Station Pro application was used to automatically generate flight plans for image acquisition with predefined parameters [2].

Orthomosaic generation employed Agisoft Metashape Professional through the following pipeline: (1) image alignment via feature matching and bundle adjustment, (2) dense point cloud generation, (3) digital elevation model (DEM) construction, (4) orthomosaic projection onto DEM with color blending. Geometric accuracy was verified through differential GPS ground control points (horizontal accuracy ± 5 cm), yielding orthomosaic RMSE < 10 cm. Radiometric processing included vignetting correction and seamline color balancing. Final orthomosaics were exported as 8-bit RGB JPEG at native resolution (3648×5472 pixels, 19.96 megapixels per tile), coordinate system WGS 84 / UTM Zone 29N.

2.3. Dataset composition and ground truth annotation

From orthomosaic tiles covering the study area, 23 images were selected based on: (1) complete *Euphorbia resinifera* annotation coverage verified by expert botanists, and (2) vegetation density diversity for testing generalization. Dataset partition employed a 1-1-1-20 split: one training image, one validation image, one patch-based test image, and twenty full-resolution test images. Table 1 shows ground truth statistics for the three images used in patch-based training, validation, and testing. Each image represents a distinct geographic location with zero spatial overlap, eliminating autocorrelation that could inflate performance estimates.

Table 1. Dataset composition and vegetation coverage statistics

Image ID	Split	Resolution	Euphorbia Pixels	Total Pixels	Coverage (%)
training_image.jpg	Train	3648×5472	4,055,808	19,961,856	20.34
validation_image.jpg	Validation	3648×5472	2,189,675	19,961,856	10.97
test_image_01.jpg	Test	3648×5472	622,890	19,961,856	3.12

Ground truth annotation was performed using the Supervisely platform, an advanced computer vision annotation tool designed for semantic segmentation tasks [2][21]. The annotation protocol involved: (1) automated NDVI-based pre-segmentation to identify vegetation pixels, (2) manual polygon delineation of *Euphorbia resinifera* individuals through expert visual interpretation using Supervisely's polygon annotation interface, (3) field validation via GPS waypoint navigation for random sample verification. The Supervisely platform enabled efficient creation of semantic segmentation masks with its intuitive interface supporting polygon-based annotations, color-coded class visualization, and direct PNG mask export functionality. Each image required 20–30 minutes expert time. Binary masks were generated by exporting polygons as rasterized PNG files at 3648×5472 resolution with pixel values 0 (background: soil, rocks, other vegetation, shadows) or 1 (*Euphorbia resinifera*).

The segmentation task was formulated as binary classification distinguishing *Euphorbia resinifera* (positive class) from all other land cover types including other vegetation species such as bare soil, rocks, and shadows (negative class) (Fig. 2). This binary formulation was a deliberate design choice aligned with conservation priorities, not a technical limitation. It reflects the operational conservation monitoring objective where presence-absence mapping of the target endangered species is the primary management concern, rather than comprehensive vegetation inventory. Conservation managers fundamentally need to answer: where does *Euphorbia resinifera* grow? and how do populations change over time? Multi-class segmentation distinguishing all vegetation types would require substantially more training data and annotation effort, contradicting the single-image few-shot learning paradigm that enables rapid deployment for conservation applications. The binary approach directly addresses the critical operational question while maintaining the annotation efficiency that makes foundation model deployment practical for endangered species monitoring.

2.4. Patch extraction strategy

Patch-based training addresses two constraints. First, full 19.96-megapixel images exceed GPU memory capacity. Second, spatial diversity within a single high-resolution orthomosaic can substitute for diversity across multiple images through systematic patch extraction.

For DINOv2 and SegGPT extraction, patch size was 518×518 pixels, which at 1.5 cm GSD covers approximately 7.8 meters by 7.8 meters, capturing typical *Euphorbia* individuals plus surrounding context. Patches were extracted using a sliding window approach with 50% overlap, corresponding to a stride of $s=518 \times (1-0.5)=259$ pixels. This overlap strategy maximizes spatial coverage and sample diversity from the single training image while ensuring adequate representation of *Euphorbia* morphology and context. Edge handling ensured complete coverage: when image dimensions were not evenly divisible by the stride, additional patches were extracted from right and bottom edges flush with boundaries, and corner patch were added if both dimensions required completion. This resulted in 294 patches per image (approximately 14 columns × 21 rows from the 3648×5472 training image).

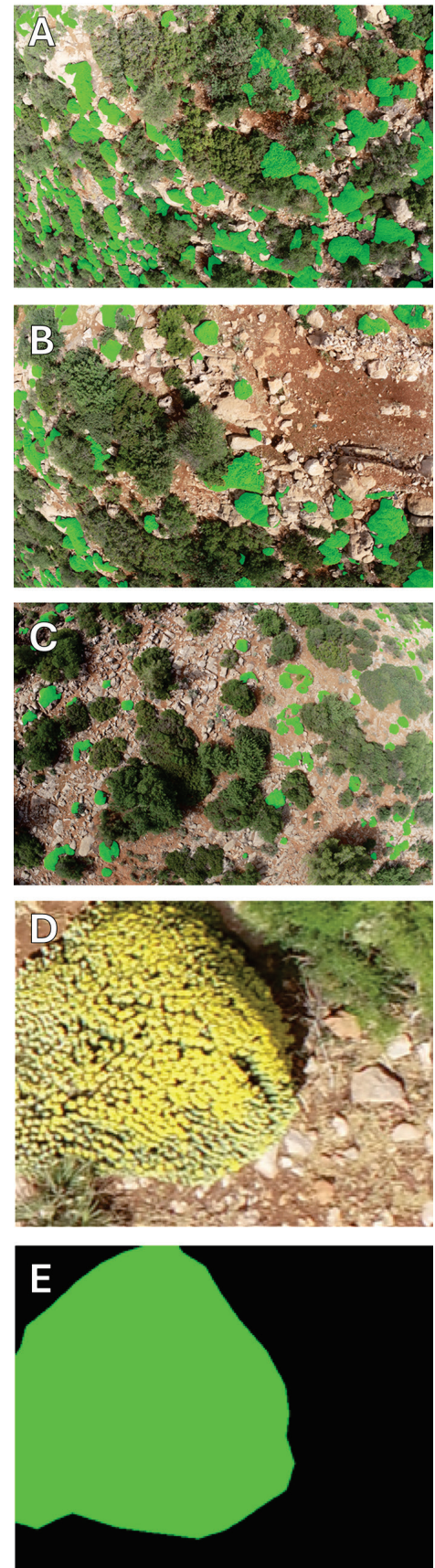
For Segment Anything Model (SAM) extraction, architectural constraints impose fixed 1024×1024 pixel input tied to learned positional embeddings, precluding arbitrary resize without substantial model modification. The same sliding window approach was applied with 50% overlap, corresponding to a stride of $s=1024 \times (1-0.5)=512$ pixels. This resulted in 70 patches per image (approximately 7 columns × 10 rows from the 3648×5472 training image), representing a 76.2% reduction compared to DINOv2 and SegGPT.

The training paradigm involved training models on 294 patches for DINOv2 and SegGPT or 70 patches for SAM, all extracted from single image training_image.jpg. This represents few-shot learning where sample diversity comes from spatial variation within one orthomosaic rather than multiple independent images. Models were validated on 294 or 70 patches (Fig. 2 D and E), respectively from validation_image.jpg, then tested on full-resolution images using sliding window inference rather than patches.

2.5. Data augmentation and model architectures

Heavy augmentation was applied exclusively to training patches to increase effective diversity and prevent overfitting to single-image characteristics (Table 2). The Albumentations library version 1.3 [23] provided the augmentation pipeline. Geometric transformations included horizontal flip with probability 0.5, vertical flip with probability 0.5, random 90-degree rotation with probability 0.5, and shift-scale-rotate with shift limits of ±10%, scale limits of ±20%, and rotation limits of ±45 degrees at probability 0.5.

Figure 2. Training (A), validation (B), and test (C) images with ground truth annotations. An example of patch extraction grid visualization: extracted patch (D) and ground truth annotation (E)



Photometric adjustments included random brightness-contrast with $\pm 30\%$ variation at probability 0.5, hue-saturation-value transformation with hue shift of ± 20 degrees, saturation variation of $\pm 30\%$, and value variation of $\pm 20\%$ at probability 0.5, along with RGB channel shift of ± 25 per channel at probability 0.5.

Degradation effects simulated sensor and environmental variability through Gaussian noise with variance from 10 to 50 at probability 0.3, Gaussian blur with kernel sizes from 3 to 7 at probability 0.3, median blur with kernel sizes from 3 to 7 at probability 0.3, random shadow at probability 0.2, and JPEG compression with quality from 75% to 100% at probability 0.3. This comprehensive augmentation strategy simulates natural UAV acquisition variability including solar illumination angles, atmospheric turbidity, sensor noise, and operational compression artifacts.

Table 2 provides a comprehensive comparison of the architectural specifications for all three foundation models. The DINOv2 architecture employed a `vit_base_patch14_dinov2` encoder from the `timm` library [24], with 86 million parameters pretrained via self-supervised learning using student-teacher distillation on 142 million curated internet images [9]. For 518×518 pixel inputs, the vision transformer divides images into 14×14 pixel patches yielding 37×37 spatial tokens. Each token is projected to 768-dimensional embeddings through 12 transformer layers with 12 attention heads. The encoder output is reshaped from dimensions batch by 1369 by 768 to batch by 768 by 37 by 37 after removing the CLS token. The custom decoder consists of Conv2d layer transforming 768 to 512 channels with 3×3 kernel and padding 1, followed by BatchNorm2d, ReLU, and Dropout with probability 0.3, then Conv2d layer transforming 512 to 256 channels with the same configuration, followed by Conv2d layer reducing to 2 output channels with 1×1 kernel, and finally bilinear up-sampling from 37×37 to 518×518 pixels. The encoder was fine-tuned rather than frozen in final experiments. The total architecture contains 96 million parameters with all parameters trainable, comprising 86 million in the encoder and 10 million in the decoder.

The SegGPT architecture utilized a `vit_large_patch16_224` encoder with 304 million parameters pretrained on ImageNet-21k via supervised classification across 21,841 categories [10]. For 518×518 pixel inputs, the transformer divides images into 16×16 pixel patches producing 32×32 spatial tokens. The architecture employs 24 transformer layers, 16 attention heads, and 1024-dimensional hidden states, producing output with dimensions batch by 1024 by 32 by 32. The decoder follows similar design with Conv2d transforming 1024 to 512 channels with 3×3 kernel, BatchNorm, ReLU, and Dropout with probability 0.3, followed by Conv2d transforming 512 to 256 channels with the same configuration, Conv2d reducing to 2 output channels with 1×1 kernel, and bilinear up-sampling from 32×32 to 518×518 pixels. The total architecture contains 316 million parameters all trainable via end-to-end fine-tuning, comprising 304 million in the encoder and 12 million in the decoder.

The SAM architecture comprises three components [11]. The image encoder uses a ViT-B backbone with 89 million parameters processing 1024×1024 pixel inputs to 256-channel embeddings at 64×64 resolution. The prompt encoder accepts sparse prompts such as points or boxes and dense prompts such as masks. The mask decoder is a transformer-based module generating 256×256 mask logits via cross-attention between image and prompt embeddings. For semantic segmentation without interactive prompts, empty sparse prompts were provided as zero-length tensors and zero-initialized dense prompts to enable training. This zero-prompt configuration was chosen to evaluate SAM's capacity for fully automated semantic segmentation suitable for operational deployment at scale, where human-in-the-loop prompting would be impractical for monitoring hundreds of images across large geographic areas. The image encoder and mask decoder were frozen to preserve SA-1B pretraining. A lightweight trainable head was added consisting of Conv2d transforming 1 to 64 channels with 3×3 kernel, BatchNorm, and ReLU, followed by Conv2d transforming 64 to 32 channels with the same configuration, Conv2d reducing to 2 output channels with 1×1 kernel, and bilinear up-sampling from 256×256 to 1024×1024 pixels. The total architecture contains 93 million parameters with only 0.1 million trainable, comprising 89 million in the frozen image encoder, 4 million in the frozen mask decoder, and 0.1 million in the trainable segmentation head. This minimal-adaptation design reflects SAM's optimization for zero-shot generalization rather than task-specific fine-tuning.

For preprocessing, DINOv2 and SegGPT used ImageNet normalization where pixels were first scaled to the range 0 to 1 then normalized per-channel using the formula x minus mean divided by standard deviation, with means of 0.485, 0.456, and 0.406 and standard deviations of 0.229, 0.224, and 0.225 for RGB channels respectively. SAM used simple scaling to the range 0 to 1 via division by 255.

Table 2. Architecture comparison of foundation models

Component	DINOv2	SegGPT	SAM
Encoder Architecture	ViT-Base (vit_base_patch14_dinov2)	ViT-Large (vit_large_patch16_224)	ViT-B (SAM image encoder)
Encoder Parameters	86M	304M	89M
Encoder Pretraining	Self-supervised (142M images)	Supervised ImageNet-21k (21,841 classes)	Supervised SA-1B (1.1B masks)
Encoder Status	Fine-tuned (trainable)	Fine-tuned (trainable)	Frozen
Decoder Architecture	Custom CNN (3 conv layers)	Custom CNN (3 conv layers)	Transformer mask decoder + lightweight head
Decoder Parameters	10M	12M	4M (frozen) + 0.1M (trainable head)
Decoder Status	Trainable	Trainable	Frozen (mask decoder) + Trainable (head only)
Input Resolution	518×518 pixels	518×518 pixels	1024×1024 pixels (fixed)
Patch Size	14×14 pixels	16×16 pixels	16×16 pixels
Spatial Tokens	37×37	32×32	64×64
Embedding Dimension	768	1024	256
Transformer Layers	12	24	12
Attention Heads	12	16	12
Total Parameters	96M	316M	93M
Trainable Parameters	96M (100%)	316M (100%)	0.1M (0.1%)
Training Patches per Image	294	294	70
Preprocessing	ImageNet normalization	ImageNet normalization	[0,1] scaling

2.6. Training configuration

The optimizer was AdamW [25] with beta-1 equal to 0.9, beta-2 equal to 0.999, and epsilon equal to 10^{-8} . Learning rates were set to 5×10^{-5} for DINOv2 and SegGPT, chosen conservatively for end-to-end fine-tuning, while SAM used 10^{-4} , a higher rate acceptable given the frozen encoder. Weight decay was 10^{-3} for DINOv2 and SegGPT, and 10^{-4} for SAM.

The learning rate schedule employed cosine annealing from the initial rate to zero over 100 epochs, following the formula:

$$\eta = \eta_{min} + 0.5 \times (\eta_{max} - \eta_{min}) \times \left(1 + \cos\left(\frac{\pi \times t}{T}\right)\right)$$

Where η is the current learning rate at the current epoch t , and T equals 100. Batch size was 4 for DINOv2 and SegGPT due to memory limitations with 518×518 pixel patches, and 2 for SAM with its larger 1024×1024 pixel patches. Cross-entropy loss provided pixel-wise supervision as the loss function.

Early stopping monitored validation IoU with patience of 15 epochs, terminating training if no improvement occurred for 15 consecutive epochs. Reproducibility was ensured through random seed 42 for NumPy, PyTorch, and CUDA. Hardware consisted of NVIDIA RTX 4090 GPU with 24GB VRAM, AMD Ryzen CPU, 32GB RAM, and SSD storage. DINOv2 achieved best validation IoU at epoch 25, SegGPT at epoch 45, and SAM at epoch 30. Training continued until no improvement occurred for 15 consecutive epochs (early stopping patience), after which training terminated.

2.7. Sliding window inference for full-image testing

Operational deployment requires segmenting full-resolution test images (3648×5472 pixels); however, the models were trained using patches of 518×518 or 1024×1024 pixels. The sliding window inference algorithm with vote-based fusion provides the solution for this issue [21]. The algorithm initializes zero-filled prediction accumulator and count map with the same dimensions as the test image. The window slides across the image with 30% overlap, using stride of 362 pixels for DINOv2 and SegGPT, and stride of 717 pixels for SAM. At each position, the patch is extracted and model inference generates a probability map with dimensions batch by 2 by height by width. The *Euphorbia* class probabilities are extracted from this map and added to the prediction accumulator at the patch location, while the count map is incremented at the same location. After complete traversal, which generates approximately 150 patches for DINOv2 and SegGPT or 40 patches for SAM, the final prediction is computed by dividing the prediction map by the count map to obtain averaged probabilities. A threshold of 0.5 is then applied to generate the binary mask, which is saved as the final result.

Vote-based fusion provides substantial benefits for prediction quality [21]. Pixels near patch centers receive predictions from multiple overlapping windows, and averaging these predictions reduce variance and smooth boundary artifacts compared to non-overlapping tiling. The reduced overlap strategy balances competing objectives. Training used 50% overlapping patches (stride=259 pixels for DINOv2/SegGPT, stride=512 pixels for SAM) to maximize spatial coverage and sample diversity from the single training image, while inference uses 30% overlap (stride=362 pixels for DINOv2/SegGPT, stride=717 pixels for SAM) to balance prediction smoothness against computational efficiency.

2.8. Evaluation metrics and statistical analysis

Performance was evaluated using standard semantic segmentation metrics computed per-pixel across entire test images. The primary metric was Intersection over Union (IoU), defined as:

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}$$

where TP represents true positives, FP represents false positives, and FN represents false negatives for the *Euphorbia* class. Additional metrics included the F1-score:

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

where Precision and Recall are defined below. Precision was calculated as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

where TP represents true positives and FP represents false positives for the *Euphorbia* class. Recall was calculated as:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

where TP represents true positives and FN represents false negatives for the *Euphorbia* class. IoU was selected as the primary metric due to its standard adoption in the segmentation literature and its balanced sensitivity to both false positives and false negatives in imbalanced scenarios.

to compare model performance. The coefficient of variation (CV) was calculated as:

$$\text{CV} = \frac{\sigma}{\mu} \times 100$$

where σ represents the standard deviation and μ represents the mean of the performance metric across test images, to quantify performance stability across test images.

3. Results

3.1. Validation performance (patch-based evaluation)

Validation performance on patch-based evaluation (Table 3) demonstrated SegGPT's superiority with an IoU of 88.37%, surpassing DINOv2 (86.80%) and SAM (70.82%) by substantial margins. SegGPT achieved exceptional recall (96.03%), indicating minimal false negative rate for *Euphorbia* detection, balanced with high precision (91.72%) controlling false alarms. DINOv2 exhibited balanced precision-recall trade-off (92.62%/93.26%) with slightly lower overall accuracy. SAM's validation performance revealed significant limitations even on patch-level evaluation with an IoU of 70.82%, reflecting both lower precision (78.18%) and recall (88.27%) compared to vision transformer alternatives.

Table 3. Validation results on patch-based evaluation (294 patches from validation_image.jpg for DINOv2/SegGPT, 70 patches for SAM)

Model	Best Epoch	Val Loss	Val IoU (%)	Val F1 (%)	Precision (%)	Recall (%)
SegGPT	45	0.0354	88.37	93.82	91.72	96.03
DINOv2	25	0.2150	86.80	92.94	92.62	93.26
SAM	30	0.1309	70.82	82.92	78.18	88.27

Training dynamics (Fig. 3) illustrated convergence patterns typical of extreme few-shot learning scenarios. All three models exhibited training instability to varying degrees, which is expected and acceptable when training on limited patches from a single image, as the empirical risk minimizer becomes unreliable with insufficient supervised information in such scenarios [26]. SegGPT showed oscillating validation performance across epochs with some fluctuations, though ultimately achieving stable convergence at 88.37% validation IoU. DINOv2 exhibited pronounced instability with validation IoU variance reaching $\pm 15\%$ between epochs, and counterintuitive behavior where validation occasionally exceeded training performance, attributable to heavy augmentation making training more challenging than clean validation. This phenomenon is recognized in few-shot learning where aggressive regularization induces temporary training difficulty while improving generalization [27]. SAM demonstrated moderate training fluctuations with final convergence showing slight overfitting gap between training and validation performance. Critically, this training instability does not indicate learning failure or predict poor generalization. Test performance validates successful learning despite optimization challenges: SegGPT achieved 83.74% mean test IoU across twenty independent test images, DINOv2 attained 78.09%, demonstrating that models learned generalizable features despite unstable training curves. The observed instability reflects the fundamental challenge of learning from 294 patches extracted from a single image, where limited training diversity necessitates careful regularization and early stopping. For operational deployment, the final model quality on held-out data is more important compared to training curve smoothness. Foundation models' rich pretrained representations provide inductive biases that regularize learning and enable strong generalization even when optimization landscapes are difficult to navigate with limited data. Future work could explore learning rate tuning, alternative optimizers, or gradient clipping to reduce training instability.

3.2. Test performance on full-resolution images (sliding window inference)

Test performance via sliding window inference on 20 spatially independent test images (Table 4) revealed substantial model differentiation under operational deployment conditions. SegGPT achieved mean IoU of $83.74 \pm 2.55\%$ with coefficient of variation of 3.05%, demonstrating remarkable consistency across vegetation coverage ranging from 2.01% to 21.11% (mean $10.22 \pm 6.45\%$). Performance ranged from 79.84% on sparsest vegetation to 87.80% on low-density imagery. DINOv2 attained $78.09 \pm 6.63\%$ mean IoU with CV of 8.49%, performing strongly on some test images but exhibiting moderate variability across the density spectrum. SAM exhibited severe performance collapse with $47.03 \pm 20.64\%$ mean IoU and CV of 43.89%, indicating highly inconsistent performance ranging from acceptable on dense vegetation (75.23% on one test image at 16.51% coverage, 74.41% on another at 21.11% coverage) to

operational failure on sparse scenarios (22.02% on the sparsest test image at 2.01% coverage, 43.46% at 3.12% coverage).

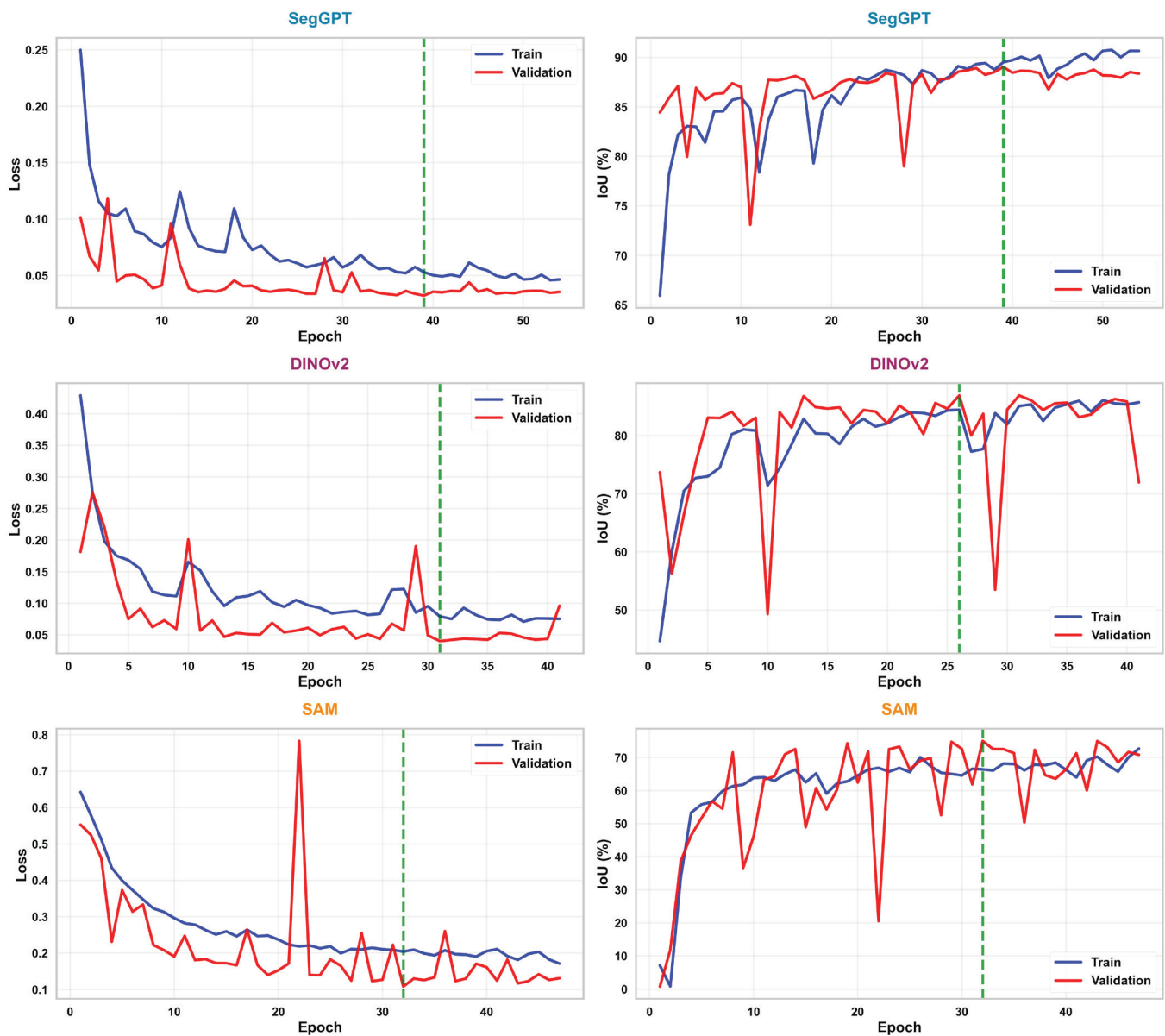


Figure 3. Training and validation curves for loss and IoU

Representative examples across the density spectrum (Fig. 4) illustrate performance trends. SegGPT achieved consistent performance across density gradients, with IoU values of 87.80% on low-density imagery (4.58% coverage) (Fig. 4 A), 85.22% on high-density conditions (16.51% coverage) (Fig. 4 B), and 79.84% on very sparse vegetation (2.01% coverage) (Fig. 4 C). DINOv2 exhibited moderate density dependence, ranging from 85.49% (Fig. 4 A) to 69.55% (Fig. 4 C). SAM demonstrated severe performance degradation on sparse vegetation, collapsing from 75.23% (Fig. 4 B) to 22.02% (Fig. 4 C), with extensive false negatives visible in Panel C where entire *Euphorbia* patches were completely missed.

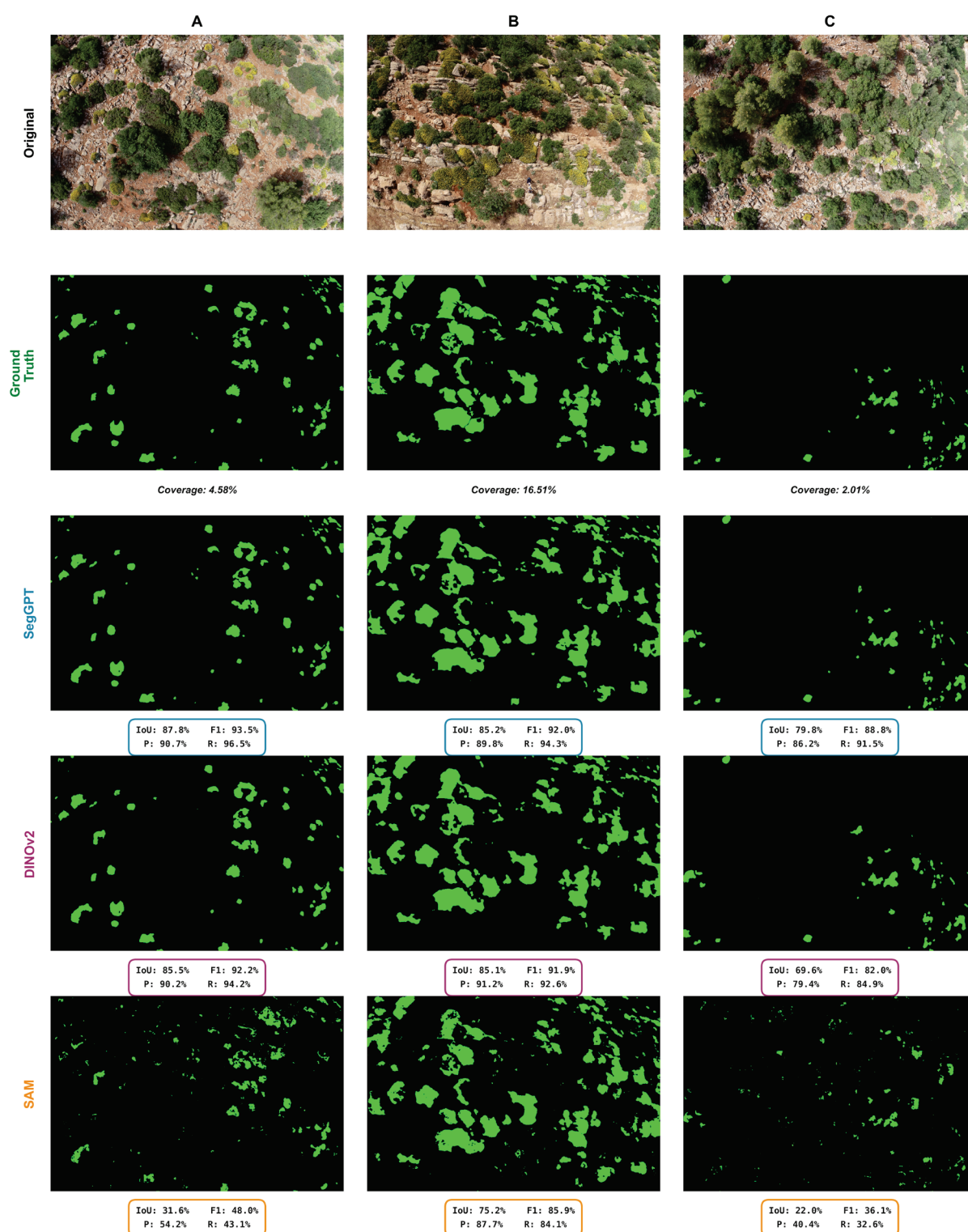
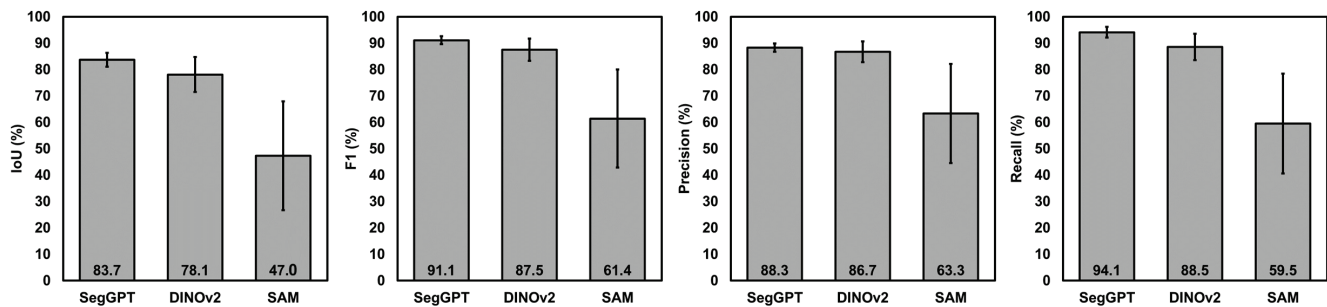


Figure 4. Visual comparison on example panels showing model predictions for low (A), high (B), and sparse (C) vegetation density

Table 4. Representative full-resolution test results via sliding window inference (6 of 20 test images shown; summary statistics computed from all 20 images, 3648×5472 pixels each)

Image ID	Coverage (%)	SegGPT IoU (%)	DINOv2 IoU (%)	SAM IoU (%)	Best Model
Test-01	2.01	79.84	69.55	22.02	SegGPT
Test-02	3.12	81.51	72.16	43.46	SegGPT
Test-03	4.58	87.80	85.49	31.58	SegGPT
Test-04	5.05	83.68	73.03	35.44	SegGPT
Test-05	16.51	85.22	85.07	75.23	SegGPT
Test-06	21.11	84.35	83.22	74.41	SegGPT
Mean ± SD (n=20)	10.22 ± 6.45	83.74 ± 2.55	78.09 ± 6.63	47.03 ± 20.64	
CV (%)		3.05	8.49	43.89	
Range	2.01–21.11	79.84–87.80	69.55–85.49	22.02–78.01	

Comprehensive metric evaluation (Fig. 5) confirmed SegGPT's superior performance across all measures. SegGPT achieved 91.13% F1-score, 88.33% precision, and 94.13% recall, demonstrating balanced and robust segmentation. DINOv2 attained 87.54% F1-score with 86.68% precision and 88.49% recall, showing good but more variable performance. SAM obtained only 61.37% F1-score with 63.83% precision and 59.50% recall, reflecting fundamental limitations in few-shot adaptation with frozen encoders.

**Figure 5. Performance comparison bar charts for IoU, F1-Score, Precision, and Recall with standard error bars**

4. Discussion

Foundation models trained on patches extracted from a single UAV orthomosaic achieve operationally viable *Euphorbia resinifera* segmentation, validating the core premise of this single-image training framework. SegGPT attained 83.74% mean IoU from 294 patches drawn from one annotated image, approaching the lower bound of the 85–95% accuracy range reported for fully-supervised deep learning methods trained on hundreds of annotated images [1][2][28]. This result is particularly significant given that conventional supervised segmentation pipelines typically require 100–1000+ expertly annotated images to achieve competitive accuracy [3][4], creating prohibitive annotation costs that severely constrain conservation applications targeting endangered or endemic vegetation species [5]. The findings demonstrate that, for *Euphorbia resinifera*, spatial diversity within one high-resolution orthomosaic (19.96 megapixels, approximately 130×195 meters at 1.5 cm GSD) can substitute for diversity across multiple independent images. The 294 overlapping patches capture substantial variability in *Euphorbia* morphology, illumination, surrounding context, and spatial arrangements, providing sufficient training signal for foundation models to generalize across a 10-fold vegetation density gradient (2.01–21.11%, mean 10.22% ± 6.45%). The strict 1-1-1-20 experimental split, with all 20 test images representing completely independent geographic locations with zero spatial overlap to training data, ensures that performance estimates are conservative and ecologically valid, directly relevant to operational deployment scenarios where models must generalize to previously unvisited sites [20].

SegGPT's superiority over DINOv2 (83.74% versus 78.09% mean IoU) derives from a synergistic combination of architectural capacity, pretraining paradigm, and fine-tuning flexibility. SegGPT's ViT-Large encoder (304 million parameters, 24 transformer layers) provides substantially greater representational capacity than DINOv2's ViT-Base encoder (86 million parameters, 12 transformer layers), consistent with established findings that larger vision transformers exhibit stronger transfer performance on specialized downstream tasks, particularly when fine-tuned end-to-end [8]. SegGPT's supervised pretraining on ImageNet-21k across 21,841 semantic categories encodes rich hierarchical features textures, shapes, object boundaries that facilitate rapid adaptation to fine-grained vegetation classes exhibiting irregular morphology and variable spectral signatures [10]. Supervised pretraining has been consistently shown to confer faster convergence and higher task-specific accuracy compared to random initialization in remote sensing segmentation applications [1][2]. End-to-end fine-tuning of all 316 million trainable parameters enables vegetation-specific specialization across both encoder and decoder, which is critical for adapting generic visual representations to domain-specific remote sensing characteristics such as top-down viewing angle, high spatial resolution, and species-specific morphological signatures [12][13]. DINOv2's competitive performance despite its smaller capacity (78.09% mean IoU, peaking at 85.49% on high-density images) demonstrates the generalization advantages of self-supervised pretraining on 142 million curated internet images through student-teacher distillation [9]. Self-supervised objectives encourage learning of transferable structural and textural representations without relying on labeled supervision, which can be particularly advantageous in low-data regimes where labeled diversity is limited [18][19]. The statistical significance of SegGPT's advantage confirms genuine superiority rather than sampling artifact, yet DINOv2's strong performance establishes self-supervised models as viable and resource-efficient alternatives when computational or memory constraints preclude deployment of larger architectures.

SAM's severe performance degradation (47.03% mean IoU, 36.71 percentage points below SegGPT) despite pretraining on 1.1 billion masks [11] constitutes the most analytically significant finding of this study, revealing that pretraining scale alone does not guarantee transfer success in specialized few-shot scenarios. Two interacting architectural factors explain this failure. First, SAM's fixed 1024×1024 pixel input resolution, tied to learned positional embeddings, is a structural constraint that cannot be modified without retraining the model from scratch [11]. Applied to our 3648×5472 pixel training image with 50% sliding window overlap, this constraint generated only 70 training patches versus 294 for DINOv2 and SegGPT a 76.2% reduction. This severe reduction in within-image sample diversity prevented adequate learning of *Euphorbia*'s morphological and contextual variation across different spatial locations, consistent with findings in few-shot learning literature that within-class sample diversity critically determines the generalization ability of learned classifiers [26]. The problem of insufficient training diversity is further compounded by class imbalance, as sparse vegetation images (below 3% coverage) provide very few positive pixels per patch, a known challenge for deep learning-based vegetation segmentation from remote sensing imagery [16][17]. Second, SAM's design philosophy for interactive, prompt-guided segmentation resulted in near-total architectural freezing: only 0.1 million of 93 million total parameters were trainable, compared to 316 million in SegGPT a 3,160-fold difference in fine-tuning capacity. This minimal-adaptation design reflects SAM's optimization for zero-shot generalization through geometric and semantic prompt engineering [11], but proves fundamentally unsuitable for fully automated semantic segmentation required in operational UAV monitoring, where human-in-the-loop prompting is impractical for processing hundreds of images across large geographic areas. These two factors 76% fewer training patches and 3,160 times fewer trainable parameters create a multiplicative disadvantage that SAM's billion-mask pretraining cannot overcome. This finding extends and empirically confirms broader observations in domain adaptation and remote sensing literature that frozen encoders consistently underperform fine-tuned counterparts on specialized segmentation tasks, even when pretrained on substantially larger datasets [12][13][14]. For conservation remote sensing practitioners, this establishes a critical model selection principle: architectural flexibility enabling end-to-end fine-tuning should be prioritized over pretraining dataset size when deploying foundation models for specialized fine-grained vegetation segmentation tasks.

Training dynamics exhibited instability across all three models, manifesting as oscillating validation loss curves and high epoch-to-epoch IoU variance in the extreme few-shot regime. This behavior is both expected and acceptable when training on patches extracted from a single image, as the empirical risk minimizer becomes unreliable with insufficient supervised diversity [26]. DINOv2 exhibited the most pronounced instability, with validation IoU variance reaching ±15% between consecutive epochs and counterintuitive episodes where validation performance exceeded training performance a well-documented phenomenon in few-shot and heavily regularized training regimes where

augmentation renders training examples substantially harder than unaugmented validation examples [27]. This pattern reflects a productive form of training difficulty: aggressive augmentation temporarily inflates the apparent training loss while simultaneously improving generalization by preventing overfitting to the limited morphological variation captured in 294 patches from a single image [26][27]. SegGPT exhibited moderate oscillation with ultimately stable convergence at epoch 45, benefiting from its larger capacity to absorb augmentation variance. SAM demonstrated a narrower training-validation gap consistent with its frozen encoder providing strong but inflexible inductive biases that limit both overfitting and adaptation. Critically, optimization instability during training does not imply poor generalization on held-out data. Operational utility is determined by final model quality on spatially independent test images, not training curve smoothness. Foundation models' rich pretrained representations provide strong inductive biases that regularize learning and enable generalization even when the optimization landscape is noisy and irregular under limited supervised signal [8][9]. Future work could explore gradient clipping, warmup learning rate schedules, or stochastic weight averaging to reduce training instability without compromising final performance.

Sliding window inference with vote-based fusion substantially reduced prediction variance compared to direct patch-based evaluation, and proved essential for operational full-resolution deployment [21]. The 30% overlap strategy at inference (stride=362 pixels for DINOv2/SegGPT, stride=717 for SAM) ensures that most image pixels receive predictions from multiple overlapping windows, with averaged probability maps reducing boundary artifacts and suppressing spatially localized errors that would otherwise propagate into the final binary mask [21]. This ensemble-averaging mechanism is conceptually analogous to test-time augmentation techniques widely used in medical image segmentation and remote sensing to improve prediction robustness without additional training [1]. The deliberate asymmetry between training overlap (50%) and inference overlap (30%) reflects a principled trade-off: during training, maximum patch overlap maximizes the spatial diversity extracted from a single image to counteract data scarcity; during inference, reduced overlap balances prediction smoothness against computational efficiency, with full-resolution image processing requiring approximately 5–6 seconds per image on an RTX 4090 GPU. The vote-based fusion approach proved particularly beneficial for SegGPT, contributing to its narrow IoU interquartile range of 2.95 percentage points across 20 test images, demonstrating that combining multiple overlapping predictions produces stable, consistent outputs suitable for conservation monitoring applications requiring repeatable measurements [5].

All three models exhibit performance degradation on very sparse vegetation (below 3% coverage), with SAM most severely affected (22.02% IoU at 2.01% coverage, 43.46% at 3.12% coverage), DINOv2 moderately impacted (69.55% at 2.01% coverage), and SegGPT relatively robust (79.84% minimum, only 7.96 percentage point range across the full 2.01–21.11% density spectrum). This density-dependent performance gradient is consistent with known challenges of extreme foreground class imbalance in semantic segmentation, where sparse target classes provide insufficient gradient signal during training and produce high false negative rates during inference [16][29]. For SAM, sparse imagery is doubly problematic: low-coverage patches from the already-reduced 70-patch training set contain very few positive pixels, creating severe class imbalance at the individual patch level [17]. SegGPT's relative robustness to density variation is attributable to its larger encoder capacity (304 million parameters) enabling richer feature learning from the available positive pixels, combined with high recall (94.13% mean across test images) that minimizes false negatives even at low vegetation prevalence. Future work should explore class-balanced loss functions such as focal loss [29], weighted cross-entropy with dynamic positive-class weighting, or oversampling strategies that preferentially select high-coverage patches during training to mitigate sparse-vegetation degradation. Multi-scale training approaches, where patches at different resolutions capture both individual plant morphology and broader spatial context, could additionally improve performance on sparse scattered distributions characteristic of *Euphorbia resinifera* across its natural habitat [6][22].

Our experimental design prioritized internal validity through controlled comparison under identical conditions, creating acknowledged limitations that bound the generalizability of these conclusions. Results pertain specifically to binary segmentation of *Euphorbia resinifera*, a morphologically distinctive cushion-forming succulent with characteristic hemispherical mounds (30–60 cm height, 0.5–2 m diameter) and green-gray photosynthetic stems that provide relatively high visual contrast against rocky, semi-arid backgrounds [7][22]. This morphological distinctiveness may render the species more discriminable than vegetation with variable or irregular growth forms, such as tree canopies with complex branching, fine-textured grasses, or forbs with seasonal phenological variation [3][4]. Future research must systematically validate this single-image training methodology across species with

fundamentally different visual characteristics to establish general applicability beyond morphologically distinctive target species. The binary formulation *Euphorbia* versus all other land cover was a deliberate design choice aligned with conservation monitoring priorities (presence-absence mapping of the target endangered species) rather than a technical limitation [5]. Multi-class segmentation distinguishing multiple vegetation species simultaneously would require substantially more training data and annotation effort, contradicting the extreme few-shot premise of this framework. Twenty spatially independent test images spanning a 10.5-fold density range (2.01–21.11%) provide meaningful evaluation diversity, though this sample size remains modest compared to large-scale computer vision benchmarks and may not capture the full environmental variability across *Euphorbia resinifera*'s geographic range in the Middle Atlas Mountains [6]. All imagery was acquired under specific seasonal conditions (May–June), flight parameters (50 m altitude, 1.3 cm GSD), and sensor characteristics (DJI Phantom 4 RTK RGB), and model performance under different phenological stages, sensor types, or acquisition conditions remains untested [2].

Practical implications for *Euphorbia resinifera* conservation monitoring are substantial and directly actionable. Training on a single annotated orthomosaic dramatically reduces both annotation time (20–30 minutes per image versus hundreds of hours for full supervised datasets) and financial cost (from tens of thousands to hundreds of dollars), enabling rapid deployment by conservation practitioners with limited resources [3][5]. The operational workflow requires no specialized machine learning expertise beyond standard deep learning frameworks: conduct UAV survey, annotate one representative orthomosaic using standard polygon annotation tools such as Supervisely, extract overlapping patches, fine-tune the foundation model with the provided training configuration, and deploy via sliding window inference on new images to generate quantitative vegetation maps for temporal change monitoring. This accessibility is critical for Mediterranean conservation contexts where endangered endemic species such as *Euphorbia resinifera* require urgent monitoring but institutional resources are severely limited [6][7][22]. The framework's demonstrated generalization across 20 geographically independent sites spanning 10-fold density variation provides initial evidence that models trained on one representative site may transfer to broader geographic ranges, though systematic cross-regional validation remains an essential future research priority before operational large-scale deployment.

Despite the promising results presented in this study, several limitations should be acknowledged. First, the study employed a single-image training paradigm, which, while reducing annotation effort, inherently constrains the diversity of training samples. Although patch extraction and heavy augmentation partially compensate for spatial variability, the developed model under these conditions may not fully capture rare morphological variants, phenological stages, or extreme lighting conditions present across broader geographic ranges or different seasons. Second, SAM's architectural constraints (1024×1024 fixed input and frozen image encoder) limited the number of training patches to 70, resulting in substantial performance degradation on sparse vegetation. This highlights that architectural rigidity can critically restrict few-shot adaptation, suggesting that results with SAM may not generalize to other foundation models or larger-scale deployments without significant modification. Third, the study focuses exclusively on binary segmentation of *Euphorbia resinifera*, which simplifies operational conservation monitoring but does not account for multi-species interactions, understory vegetation, or soil heterogeneity that may confound pixel-level classification in more complex ecological settings. Fourth, all imagery was acquired under specific seasonal, lighting, and weather conditions in the Middle Atlas Mountains. Consequently, model performance under different environmental conditions, sensor types, or flight parameters remains untested. Fifth, while sliding window inference with vote-based fusion reduces variance, it introduces computational overhead, and model inference time on full-resolution orthomosaics may limit real-time or large-scale operational applicability. Sixth, while the used 1-1-1-20 split (one training image, one validation image, one patch-level test image, twenty full-resolution test images) provides meaningful evaluation across spatially independent locations with 10.5-fold density variation, this sample size is modest compared to large-scale computer vision benchmarks, though adequate for establishing proof-of-concept for conservation applications. Finally, the study evaluates only three foundation models, leaving open the question of whether alternative vision transformers, CNN-based backbones, or emerging multimodal models could further improve few-shot vegetation segmentation. Collectively, these limitations indicate that while single-image foundation model training is promising for morphologically distinctive species, careful consideration of training diversity, architectural constraints, environmental variability, and computational requirements is necessary before broader operational deployment.

5. Conclusions

Foundation models trained on patches from a single UAV orthomosaic achieve operationally viable *Euphorbia resinifera* segmentation. SegGPT attained $83.74\% \pm 2.55\%$ mean IoU from 294 patches, outperforming DINOv2 ($78.09 \pm 6.63\%$) and SAM ($47.03 \pm 20.64\%$) with superior stability (coefficient of variation = 3.05%). Evaluation across 20 spatially independent test images spanning 2.01–21.11% vegetation coverage ($10.22 \pm 6.45\%$) demonstrates robust generalization under operational conditions. Training sample quantity critically impacts performance where SAM's 76% patch reduction corresponded to 36.71 percentage point IoU decrease despite billion-mask pretraining. Fine-tuned encoders substantially outperform frozen-encoder architectures for specialized vegetation segmentation. Sliding window inference substantially reduces variance. Single-image training with 50% overlapping patch extraction offers a promising paradigm for conservation deployment of morphologically distinctive species. The framework consists of acquiring one annotated exemplar, extracting 518×518-pixel patches via regular grid, applying heavy augmentation, fine-tuning encoder and decoder jointly, and deploying via sliding window inference with 30% overlap. For *Euphorbia resinifera*, this reduces costs from tens of thousands to hundreds of dollars. Future priorities include validation on species with different morphologies (tree canopies, grasses, forbs) essential to establish general applicability, evaluation across three to five training images, cross-regional testing, multi-class extension, multispectral integration, and active learning. While this work establishes proof-of-concept for *Euphorbia*'s distinctive morphology, empirical validation across diverse plant species is absolutely essential. Until such validation, practitioners should exercise caution applying this to species with less distinctive visual characteristics.

6. Acknowledgements

The authors would like to thank the Geomatics, Georesources and Environment Laboratory at Sultan Moulay Slimane University for providing the research infrastructure and computational resources necessary for this study. The authors also thank the expert botanists who assisted with ground truth annotation and field validation of *Euphorbia resinifera* populations in the Middle Atlas Mountains study area.

Conflict of interest statement

The authors declared no conflict of interest.

Funding statement

The authors declared that no funding was received in relation to this manuscript.

Data availability statement

The authors declared that the UAV imagery dataset and annotations as well as the source code for models training are available from the corresponding author upon reasonable request. All experiments were implemented using PyTorch 2.0 [30], with data augmentation via Albumentations 1.3 [23] and pretrained encoders from the timm library [24]. Foundation model weights are available from original sources: DINOv2 (Meta AI Research, <https://github.com/facebookresearch/dinov2>), SegGPT (<https://github.com/baaivision/Painter>), and SAM (Meta AI, <https://github.com/facebookresearch/segment-anything>). All methodological details necessary for reproduction are provided in the Materials and Methods section.

References

1. Chen L, Zhu Y, Papandreou G, Schroff F, Adam H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In: Lecture Notes in Computer Science. Springer International Publishing. 2018. [DOI](#)
2. Badrouss S, Daiaeddine MJ, Bachaoui EM, Biniz M, Mouncif H, Ghmari AE, Harti AE, Boulli A. Modeling plant species segmentation using an advanced U-Net and UAV remote sensing: a case study in the High Atlas Mountains of Morocco. *Model. Earth Syst. Environ.* 2024;11(1):23. [DOI](#)
3. Diez Y, Kentsch S, Fukuda M, Caceres MLL, Moritake K, Cabezas M. Deep Learning in Forestry Using UAV-Acquired RGB Data: A Practical Review. *Remote Sens.* 2021;13(14):2837. [DOI](#)
4. Egli S, Höpke M. CNN-Based Tree Species Classification Using High Resolution RGB Image Data from Automated UAV Observations. *Remote Sens.* 2020;12(23):3892. [DOI](#)
5. Schöttker O, Hütt C, Jauker F, Witt J, Bareth G, Wätzold F. Monitoring costs of result-based payments for biodiversity conservation: Will UAV-assisted remote sensing be the game-changer?. *J. Nat. Conserv.* 2023;76:126494. [DOI](#)
6. Medail F, Quezel P. Hot-Spots Analysis for Conservation of Plant Biodiversity in the Mediterranean Basin. *Ann. Mo. Bot. Gard.* 1997;84(1):112. [DOI](#)
7. Hammad D, Ourhzif E, Hamri S, Abderrafia H, Khouili M, Pujol MD. Chemical Components of the Dried Latex of *Euphorbia resinifera* Berg and Its Medicinal Features. *Chem. Biodivers.* 2025;22(7):e202403067. [DOI](#)
8. Wang X, Chen G, Qian G, Gao P, Wei X, Wang Y, Tian Y, Gao W. Large-scale Multi-modal Pre-trained Models: A Comprehensive Survey. *Mach. Intell. Res.* 2023;20(4):447-82. [DOI](#)
9. Oquab M, Darcet T, Moutakanni T, Vo H, Szafraniec M, Khalidov V, Fernandez P, Haziza D, Massa F, El-Nouby A, Assran M. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193.* 2023. [DOI](#)
10. Wang X, Zhang X, Cao Y, Wang W, Shen C, Huang T. SegGPT: Segmenting everything in context. *arXiv preprint arXiv:2304.03284.* 2023. [DOI](#)
11. Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, Xiao T, Whitehead S, Berg AC, Lo WY, Dollár P. Segment anything. In: 2023 IEEE/CVF international conference on computer vision (ICCV). IEEE. 2023:3992-4003. [DOI](#)
12. Klemmer K, Rolf E, Robinson C, Mackey L, Rußwurm M. Satclip: Global, general-purpose location embeddings with satellite imagery. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 2025;39(4):4347-55. [DOI](#)
13. Wang D, Zhang J, Du B, Xu M, Liu L, Tao D, Zhang L. SAMRS: Scaling-up remote sensing segmentation dataset with segment anything model. *Adv. Neural Inf. Process. Syst.* 2023;36:8815-27. [DOI](#)
14. Immanuel SA, Sinulingga HR. Learnable prompt for few-shot semantic segmentation in remote sensing domain. In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE. 2024:2755-61. [DOI](#)
15. Li Z, Lu F, Zou J, Hu L, Zhang H. Generalized few-shot meets remote sensing: Discovering novel classes in land cover mapping via hybrid semantic segmentation framework. In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE. 2024:2744-54. [DOI](#)
16. Bazi Y, Melgani F. Toward an Optimal SVM Classification System for Hyperspectral Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* 2006;44(11):3374-85. [DOI](#)
17. Huang H, Lan Y, Yang A, Zhang Y, Wen S, Deng J. Deep learning versus Object-based Image Analysis (OBIA) in weed mapping of UAV imagery. *Int. J. Remote Sens.* 2020;41(9):3446-79. [DOI](#)
18. Huang Y, Chen Y, Zelek J. Zero-shot monocular motion segmentation in the wild by combining deep learning with geometric motion model fusion. In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE. 2024:2733-43. [DOI](#)

19. Raine S, Marchant R, Kusy B, Maire F, Sunderhauf N, Fischer T. Human-in-the-loop segmentation of multi-species coral imagery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE. 2024;2723-32.
20. Lee GY, Dam T, Ferdaus MM, Poenar DP, Duong VN. Unlocking the capabilities of explainable few-shot learning in remote sensing. *Artif. Intell. Rev.* 2024;57(7):169. [DOI](#)
21. Yu C, Liu Y, Hu Z, Xia X. Accurate segmentation of remote sensing cages based on U-Net and voting mechanism. In: Seventh Asia Pacific Conference on Optics Manufacture and 2021 International Forum of Young Scientists on Advanced Optical Manufacturing (APCOM and YSAOM 2021). SPIE. 2022;12166:727-32. [DOI](#)
22. Hmidouche O, Bouftini K, Chafik A, Khouri S, Rchid H, Rahimi A, Mimouni M, Maarouf E, Zaakour F, Nmila R. Ethnomedicinal Use, Phytochemistry, Pharmacology, and Toxicology of *Euphorbia resinifera* O. Berg. (B): A Review. *J. Zool. Bot. Gard.* 2023;4(2):364-95. [DOI](#)
23. Buslaev A, Iglovikov VI, Khvedchenya E, Parinov A, Druzhinin M, Kalinin AA. Albumentations: Fast and Flexible Image Augmentations. *Information.* 2020;11(2):125. [DOI](#)
24. Wightman R. PyTorch image models. GitHub repository; 2019. Available from: [Link](#)
25. Loshchilov I, Hutter F. Decoupled weight decay regularization. In: International Conference on Learning Representations (ICLR). 2019.
26. Wang Y, Yao Q, Kwok JT, Ni LM. Generalizing from a Few Examples. *ACM Comput. Surv.* 2020;53(3):1-34. [DOI](#)
27. Zhang C, Bengio S, Hardt M, Recht B, Vinyals O. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM* 2021;64(3):107-15. [DOI](#)
28. Cano-Solis M, Ballesteros JR, Sanchez-Torres G. VEPL-Net: A Deep Learning Ensemble for Automatic Segmentation of Vegetation Encroachment in Power Line Corridors Using UAV Imagery. *ISPRS Int. J. Geo-Inf.* 2023;12(11):454. [DOI](#)
29. Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. IEEE. 2017:2980-8. [DOI](#)
30. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A. Pytorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems 32 (NeurIPS 2019); 2019:8024–35.