DINOv3-Powered Multi-Task Foundation Model for Quantitative Remote Sensing Estimation (Student Abstract)

Zhenyu Yu[1], **Mohd Yamani Idna Idris**[1], **Pei Wang**[2,*], **Rizwan Qureshi**[3]
[1]Universiti Malaya
[2]Kunming University of Science and Technology
[3]University of Central Florida
yuzhenyuyxl@foxmail.com, yamani@um.edu.my, eiwang@kust.edu.cn, rrizwan2-c@my.cityu.edu.hk

Abstract

Quantitative remote sensing estimation is critical for environmental monitoring, providing continuous measures of vegetation indices, canopy height, and carbon stock. Traditional radiative-transfer models and empirical regressions require expert knowledge and generalize poorly, while deep learning methods remain task-specific. We propose *SatelliteCalculator+*, a DINOv3-powered multi-task foundation model for continuous regression of spectral and structural variables. The framework combines prompt-driven cross-attentive adapters with lightweight MLP decoders, enabling efficient dense prediction from frozen features. To overcome limited supervision, we synthesize over one million paired samples from SPOT 6/7 imagery using physically defined formulas. On the Open-Canopy dataset, *SatelliteCalculator+* achieves competitive accuracy across eight ecological variables while reducing inference cost, demonstrating the promise of self-supervised transformers and scalable multi-task learning for large-scale Earth observation.

Code — https://github.com/YuZhenyuLindy/SC2
Extended version — https://arxiv.org/pdf/2504.13442

Introduction

Quantitative remote sensing estimation is vital for ecosystem monitoring, providing continuous measures of vegetation indices, canopy height, and carbon stock. Traditional radiative-transfer models and empirical regressions require expert knowledge and generalize poorly across regions . Although deep learning has advanced classification and segmentation in remote sensing, no foundation model exists for *continuous multi-task regression* of ecological variables (Ren et al. 2025), which is essential for robust, transferable monitoring.

Unlike natural images, multi-spectral data include additional bands beyond RGB, requiring models to handle multi-band inputs, transfer pretrained weights, and remain consistent across diverse regression targets (Sarkar, Idris, and Yu 2025). We address this with *SatelliteCalculator+*, a DINOv3-powered multi-task foundation model for quantitative estimation. From four SPOT 6/7 bands, it jointly predicts eight variables—NDVI, GNDVI, SAVI, EVI, NDWI, canopy height (H), aboveground biomass (AGB), and carbon stock (CS). By synthesizing over one million paired samples with physically defined formulas, we alleviate annotation bottlenecks and enable scalable supervised training.

Our **contributions** are: (1) We present *SatelliteCalculator+*, the first DINOv3-powered multi-task foundation model for continuous regression of eight ecological variables from imagery. (2) We design prompt-driven cross-attentive adapters that inject task semantics into frozen features, enabling flexible and accurate multi-task regression. (3) We adopt lightweight task-specific MLP decoders, achieving an effective balance of accuracy.

Related Work

Traditional estimation approaches rely on physical RTMs (e.g., PROSAIL) or index-based regressions (NDVI, EVI, SAVI, NDWI) coupled with statistical models like PCR and PLSR (Yu, IDRIS, and Wang 2025). While interpretable, these methods require expert knowledge, generalize poorly across regions, and cannot jointly estimate multiple variables. Deep learning models, from CNNs and UNets to RNNs and Transformers (Ren et al. 2025), enable end-to-end prediction but remain task-specific, data-intensive, and lack scalable multi-indicator support. Foundation models such as MAE (Yu et al. 2025b), DINO , and SAM (Yu et al. 2023) exhibit strong transferability in vision, and remote-sensing adaptations (SatMAE, RS-FM, RemoteCLIP, GeoGPT (Yu et al. 2025a)) extend these capabilities. However, they focus on classification or retrieval rather than continuous regression. No existing foundation model addresses multi-band adaptation, regression-specific outputs, and multi-task scalability simultaneously.

Methodology

We formulate quantitative remote sensing estimation as a prompt-guided, multi-task regression problem (Fig.1). Given a four-band image and a task prompt, *SatelliteCalculator+* produces dense response maps.

Prompt embedding. Each task is represented by a learnable prompt token, which is mapped into a query vector conditioning downstream modules. This mechanism enables the model to dynamically adapt shared features to diverse regression targets, such as vegetation indices or biomass-related variables, without retraining the backbone.

*Corresponding author.

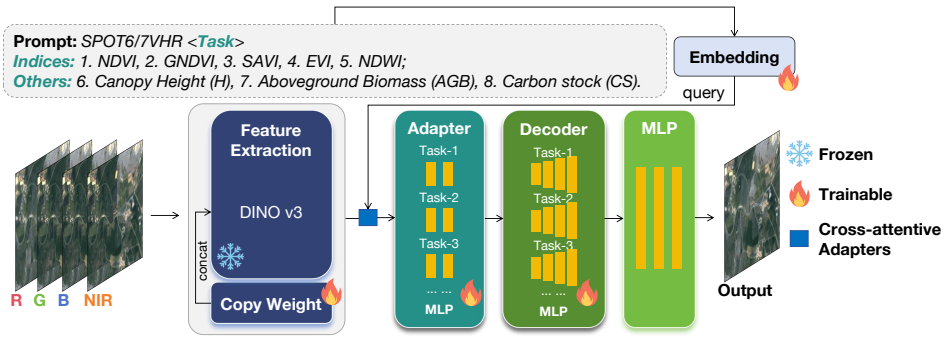

Figure 1: Overview of the *SatelliteCalculator+* framework.

Feature extraction. We adopt DINOv3 as the backbone and adapt it to four-band inputs using a copy-weight strategy for the near-infrared channel. The backbone is kept frozen, leveraging strong self-supervised representations while avoiding overfitting.

Cross-attentive adapters and decoders. Task-specific adapters fuse prompt queries with visual tokens through cross-attention, followed by lightweight MLP decoders that generate continuous prediction maps. This modular design allows adding new variables by introducing new adapters/decoders without modifying the backbone.

Overall, *SatelliteCalculator+* combines the generalization ability of DINOv3 with parameter-efficient task modules, achieving scalable multi-task inversion while keeping computation and training cost low.

Experiments

Data. We evaluate *SatelliteCalculator+* on the Open-Canopy dataset (Fogel et al. 2025), covering $>$87,000,km^2 of French forests with VHR imagery and LiDAR canopy height. From four SPOT 6/7 bands (B, G, R, NIR), we derive five indices (NDVI, GNDVI, SAVI, EVI, NDWI) and include three structural targets: canopy height (H), aboveground biomass (AGB), and carbon stock (CS). This yields $\sim$1M paired samples with consistent multi-task labels.

Settings. We train on 224×224 crops with scale/rotation augmentation using Adam (10^{-4}) and early stopping. A weighted ℓ_1 loss balances structural (H, AGB, CS) and spectral index tasks. Only prompts, adapters, and decoders are updated, while the DINOv3 backbone remains frozen.

Results. Spectral indices yield low errors and strong correlations (e.g., NDVI RMSE $\approx$0.22, $R^2 > 0.85$)(see Table 1). Structural tasks remain more challenging but competitive (e.g., H MAE 2.55,m, RMSE 4.02,m, $R^2 = 0.61$). Compared with single-task baselines, *SatelliteCalculator+* preserves per-task accuracy while enabling efficient multi-task regression. The 4-layer MLP decoder achieves the best speed–accuracy balance, confirming the framework's scalability and efficiency.

Task	MAE	RMSE	R^2	PSNR
NDVI	0.05	0.22	0.85	29.3
GNDVI	0.06	0.22	0.86	29.0
EVI	0.23	0.51	0.66	22.0
H (m)	2.55	4.02	0.61	21.1
AGB (t/ha)	21.34	26.04	0.55	21.4
CS (Mg/ha)	16.80	21.84	0.52	21.1

Table 1: Representative results on Open-Canopy.

Conclusion

We introduced *SatelliteCalculator+*, the first DINOv3-powered multi-task foundation model for quantitative remote sensing estimation. By generating over one million paired samples from physically defined formulas, the framework supports scalable training across eight ecological variables. Prompt-guided adapters and lightweight MLP decoders deliver competitive accuracy, efficient inference, and extensibility. Future work will extend to additional sensors and integrate physics-informed constraints.

References

Fogel, F.; et al. 2025. Open-Canopy: A country-scale benchmark for canopy height estimation at very high resolution. In *CVPR*, 1–10.

Ren, J.; et al. 2025. Estimating forest carbon stock using enhanced resnet and sentinel-2 imagery. *Forests*, 16(7): 1198.

Sarkar, A.; Idris, M. Y. I.; and Yu, Z. 2025. Reasoning in computer vision: Taxonomy, models, tasks, and methodologies. *arXiv preprint arXiv:2508.10523*.

Yu, Z.; Idris, M. Y. I.; Wang, H.; Wang, P.; Chen, J.; and Wang, K. 2025a. From physics to foundation models: A review of ai-driven quantitative remote sensing inversion. *arXiv preprint arXiv:2507.09081*.

Yu, Z.; IDRIS, M. Y. I.; and Wang, P. 2025. Physics-Constrained Symbolic Regression from Imagery. In *2nd AI for Math Workshop@ ICML 2025*.

Yu, Z.; Wang, J.; Chen, H.; and Idris, M. Y. I. 2025b. Qrs-trs: Style transfer-based image-to-image translation for carbon stock estimation in quantitative remote sensing. *IEEE Access*.

Yu, Z.; Wang, J.; Yang, X.; and Ma, J. 2023. Superpixel-based style transfer method for single-temporal remote sensing image identification in forest type groups. *Remote Sensing*, 15(15): 3875.