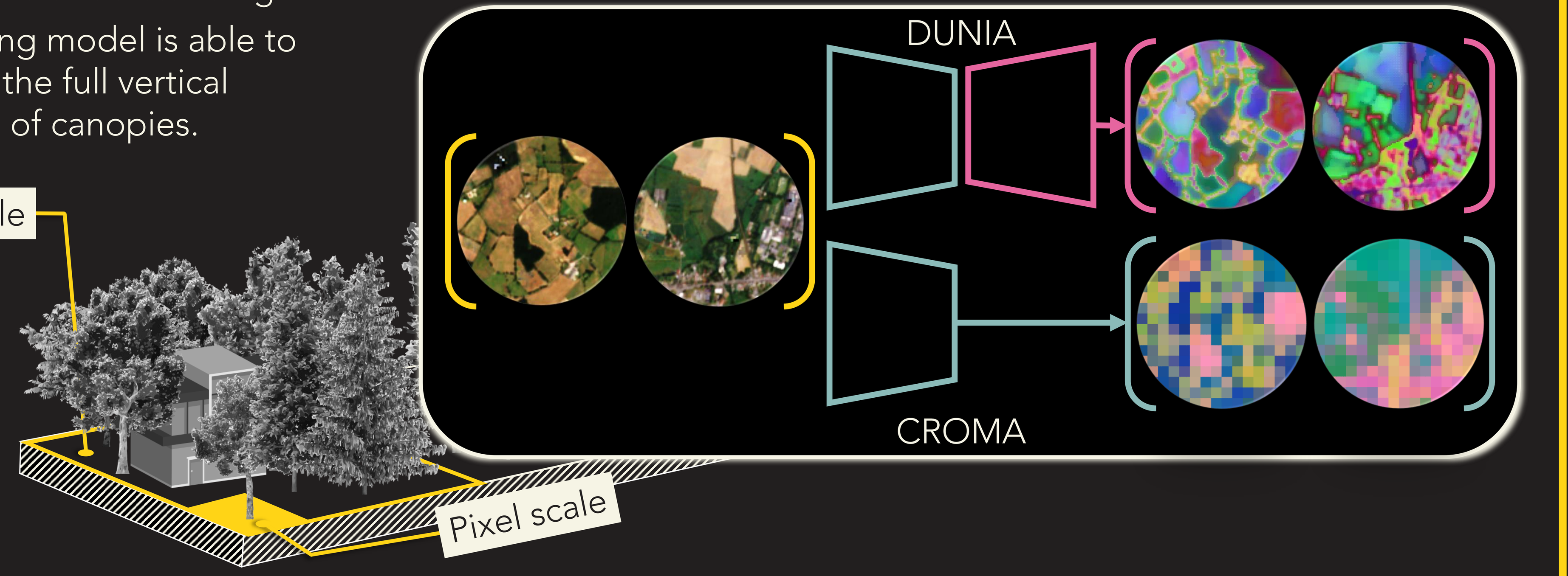
DUNIA: Pixel-Sized Embeddings via Cross-Modal Alignment for Earth Observation Applications

Ibrahim Fayad[1,2], Max Zimmer[3], Martin Schwartz[1], Fabian Gieseke[4], Philippe Ciais[1], Gabriel Belouze[1], Sarah Brood[5], Aurélien De Truchis[2], Alexandre d'Aspremont[5,2]

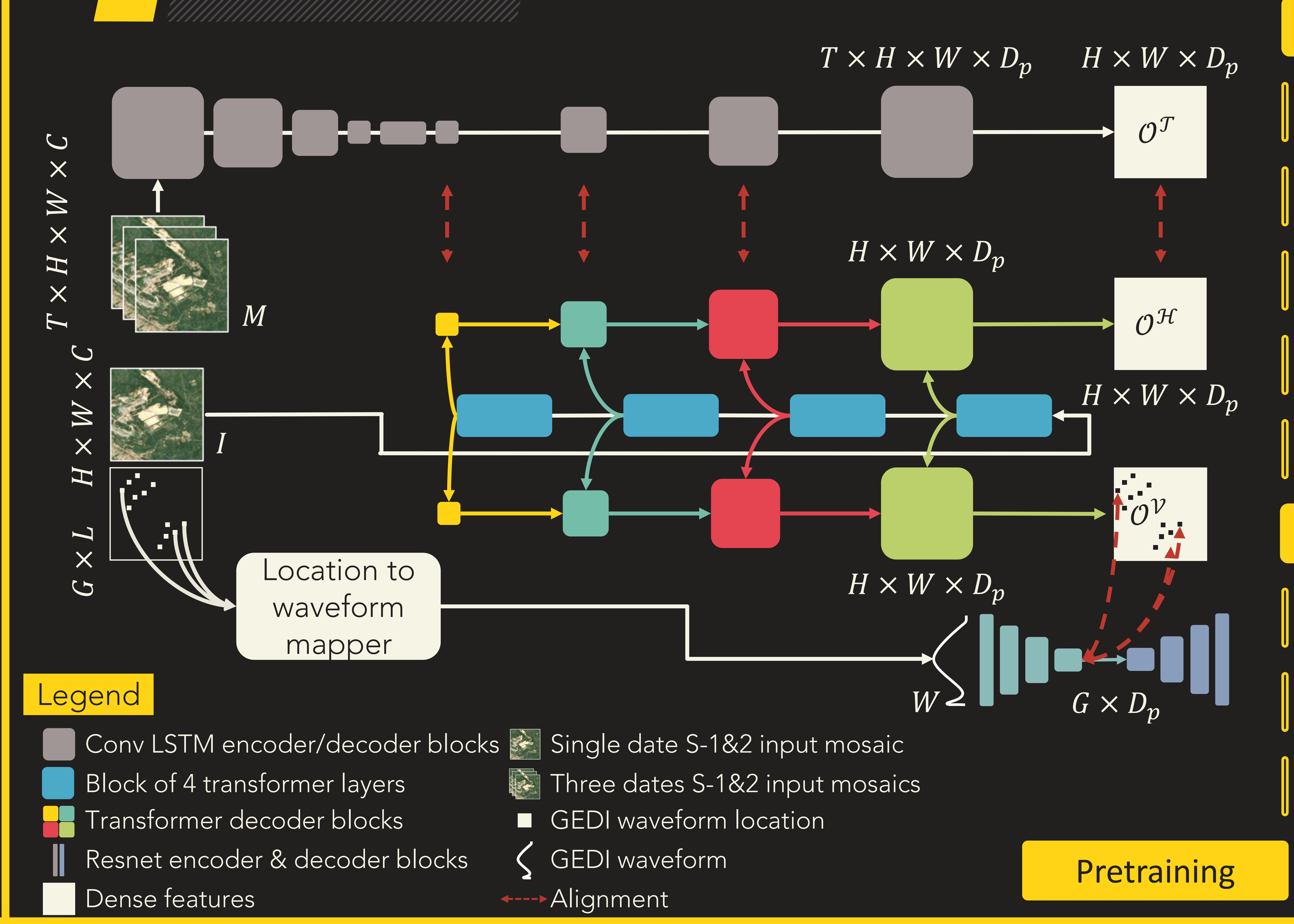
1 LSCE 2 KAYRROS 3 ZIB 4 WWU Münster 5 Inria

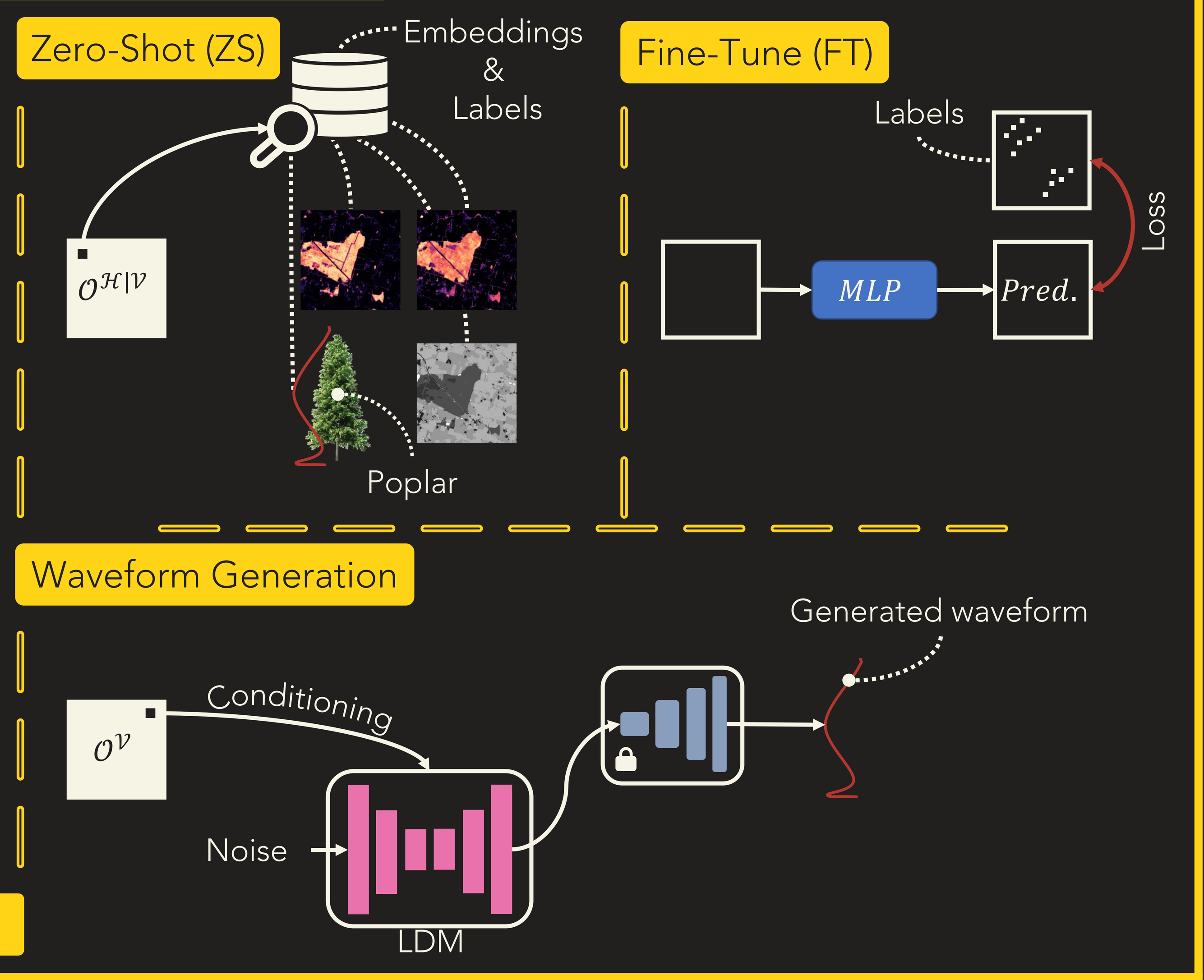
1. Motivation

1. Labels which are a fundamental part for training any ML model, are few and far between for EO applications.
2. Several self-supervised earth observation models currently exist, however:
   - 2.1 The majority operates at the patch scale, which results in embeddings that are not readily useable.
   - 2.2 They do not/cannot leverage LiDAR data (e.g., GEDI), essential for applications requiring vertical structure understanding.
3. No existing model is able to estimate the full vertical Structure of canopies.

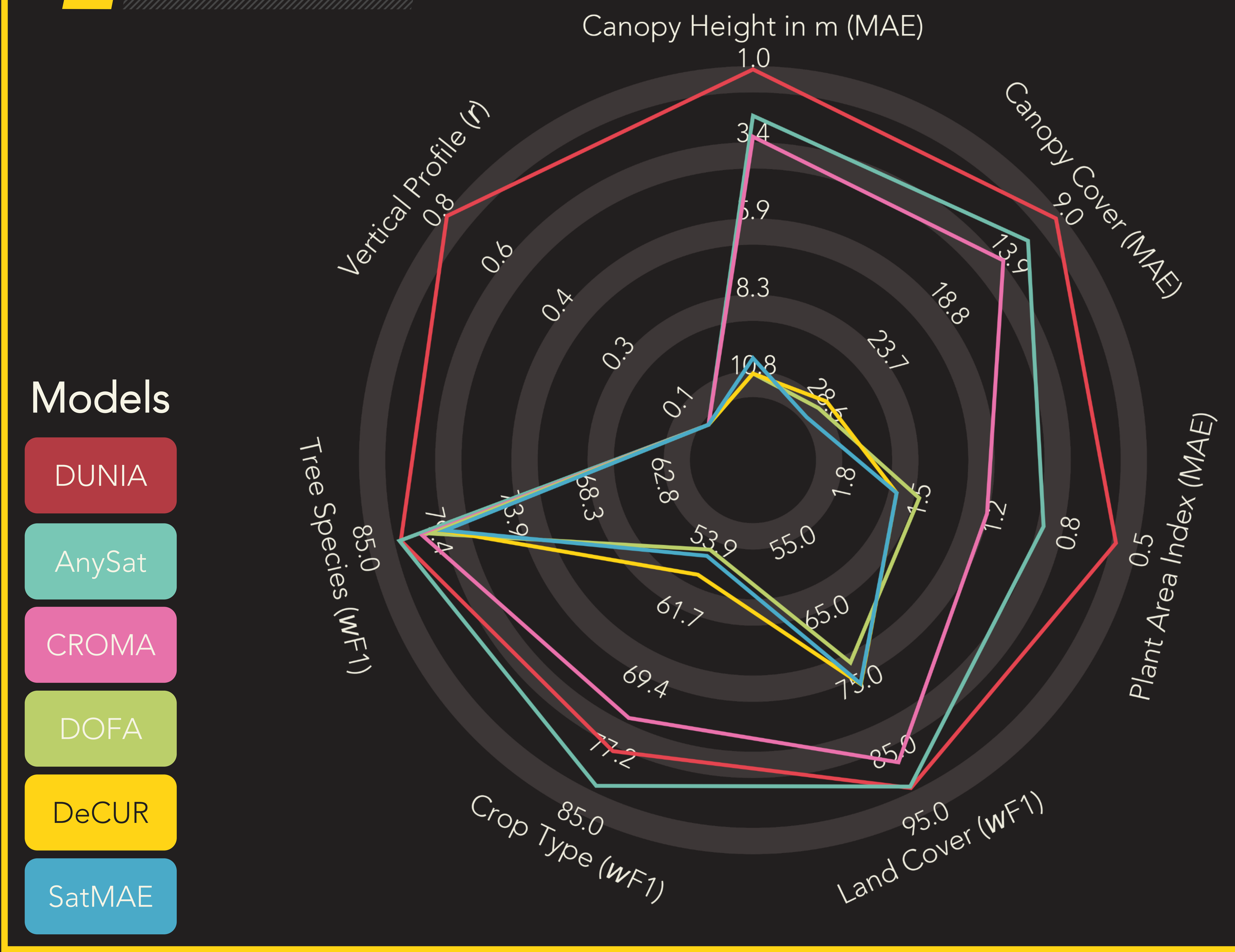


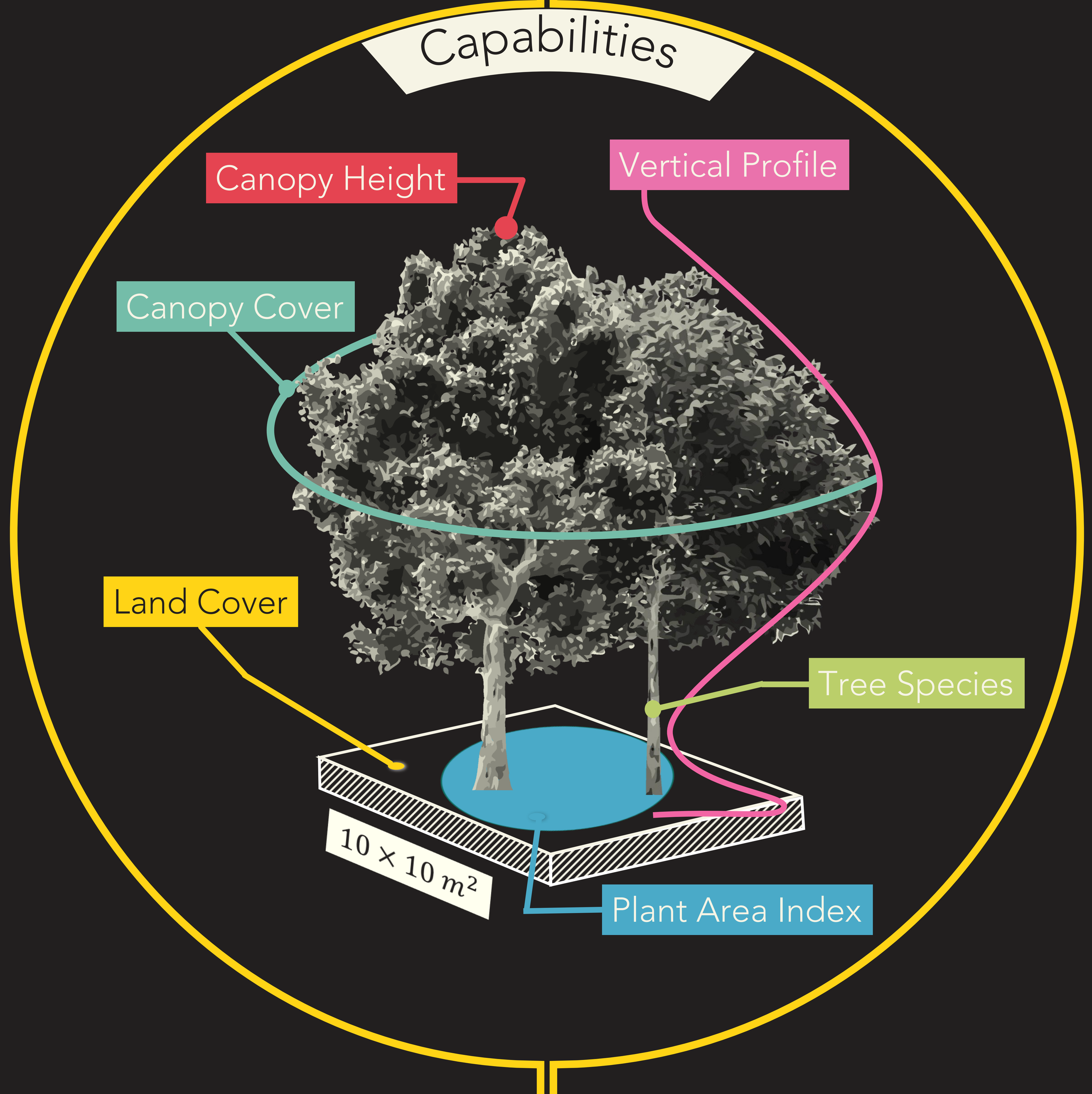
2. Method

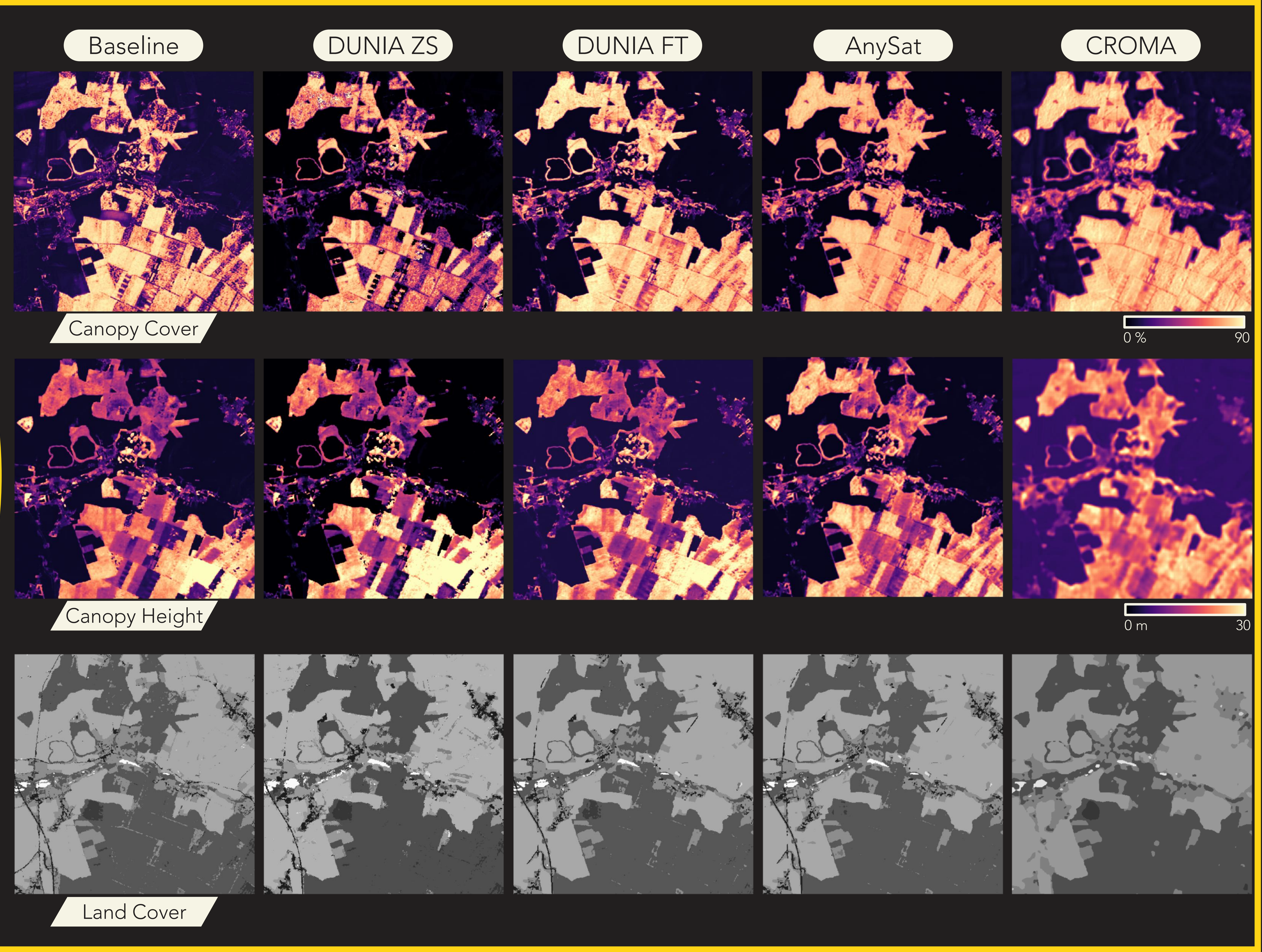




3. Results







4. Conclusions

1. Contrasting at the pixel level enables zero-shot capabilities that in many cases outperform specialist models.
2. By integrating LiDAR data, we recover the full vertical vegetation profile; our approach is the first to achieve complete canopy vertical profiling in a self-supervised framework.
3. Dual vertical and horizontal alignment equips the model for tasks demanding both spectral and structural understanding (e.g., biomass estimation, species delineation).
4. Compositing cuts out the need for time-series storage, lightening data burdens, though certain applications (e.g., crop type mapping) would greatly benefit from temporal dynamics.
5. The architecture is computationally efficient and can be trained with limited compute.

5. References

- Astruc et al. 2025. AnySat: One Earth Observation Model for Many Resolutions, Scales, and Modalities. (CVPR 2025)
- Fuller et al. 2023. Croma: Remote sensing representations with contrastive radar-optical masked autoencoders. (NeurIPS 2023)
- Xiong et al. 2024. Neural Plasticity-Inspired Multimodal Foundation Model for Earth Observation. (arXiv:2403.15356)
- Wang et al. 2024. Decoupling Common and Unique Representations for Multimodal Self-supervised Learning. (ECCV 2024)
- Cong et al. 2022. SatMAE: Pre-training Transformers for Temporal and Multi-Spectral Satellite Imagery. (NeurIPS 2022)

6. Acknowledgements

This work is supported via the AI4Forest project, which is funded by the French National Research Agency (ANR; grant number ANR-22-FAI1-0002) and the German Federal Ministry of Education and Research (BMBF; grant number 01IS23025A).