MAESTRO:
Masked AutoEncoders for Multimodal, Multitemporal, and Multispectral Earth Observation Data

Antoine Labatie, Michael Vaccaro, Nina Lardière, Anatol Garioud, Nicolas Gonthier

Code: https://github.com/ignf/maestro

WACV 2026
Tucson, Arizona • Mar 6 - 10, 2026

SSL in Earth Observation

SSL is promising for Earth observation (EO):

- Lot of unlabeled data available
- High labeling cost
- SSL pre-training improves label efficiency on fine-tuning tasks

However, off-the-shelf SSL approaches must be adapted to EO data:

- Point of view
- Scale of objects
- Data heterogeneity: multi-modality, multi-temporality, multi-spectrality → **focus of this work**

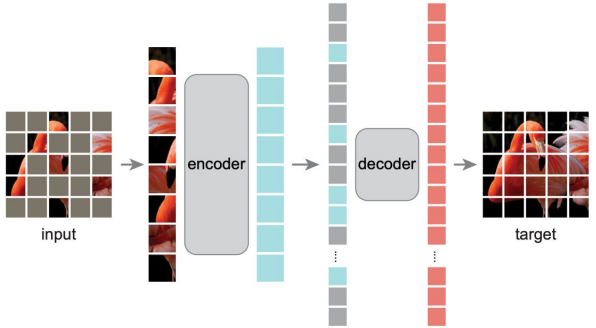
Existing work:

- Multi-modality and multi-temporality:
  - Recent works have begun to address multi-modal and multi-temporal data, but rarely both simultaneously
  - Multi-modality/temporality often handled through parameter sharing → imposes late fusion at fine-tuning
  - Notable exceptions: Presto, Galileo, OmniSat, AnySat, SeaMo, EarthMAE, SkySense
- Multi-spectrality:
  - Rarely accounted for specifically, apart from a few works implementing token-based multi-spectral fusion (band groups)

Chosen Approach

We adapt the Masked AutoEncoder (MAE) = generative SSL in pixel space

- High computational efficiency
- Close to SOTA on natural images (see Pixio vs DINO-v3)
- But designed for mono-modal and mono-temporal RGB data
  → **how to adapt it multi-modal/temporal/spectral data?**

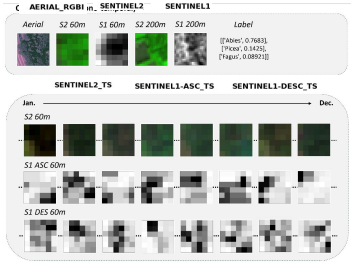


Source: Masked Autoencoders Are Scalable Vision Learners

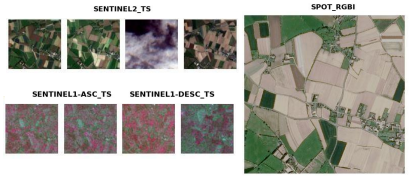
Evaluation datasets:

- We consider 4 evaluation datasets with multi-modal, multi-temporal and multi-spectral inputs

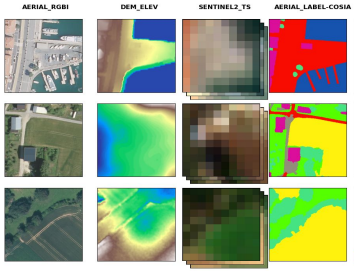
TreeSatAI-TS
Tree species classification



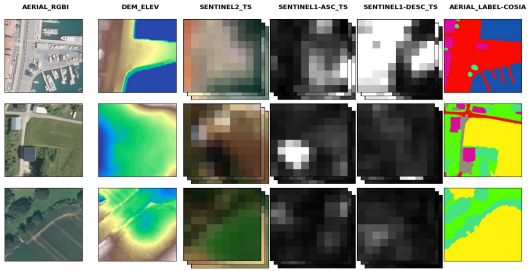
PASTIS-HD
Agriculture crop semantic segmentation



FLAIR#2
Land cover semantic segmentation



FLAIR-HUB
Land cover semantic segmentation



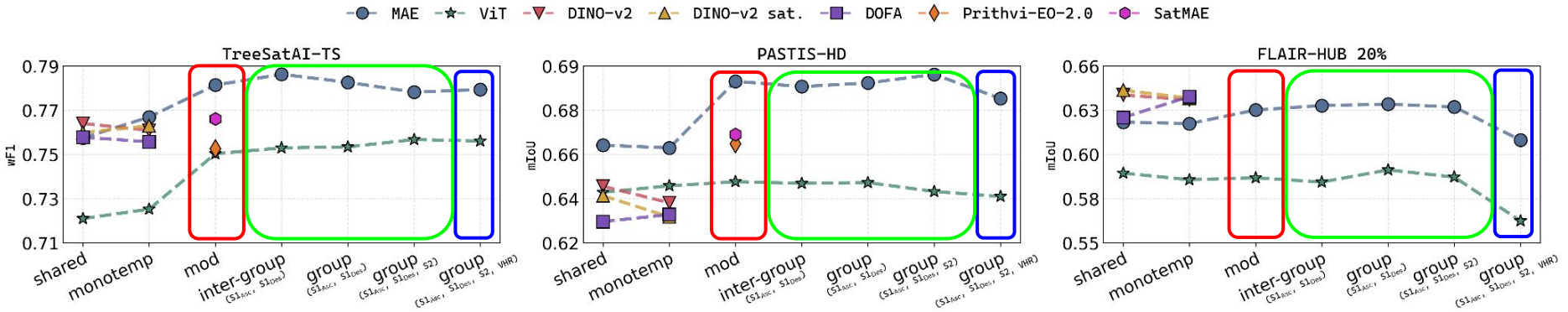
Best choice of multi-modal fusion?

Benchmark:

- Late fusion
- Early fusion across similar modalities, intermediate or late fusion across dissimilar modalities
- Early fusion across all modalities

Conclusions:

- Avoid **early fusion among dissimilar modalities**
- Other than that, choices of multi-modal fusion do not have huge impact



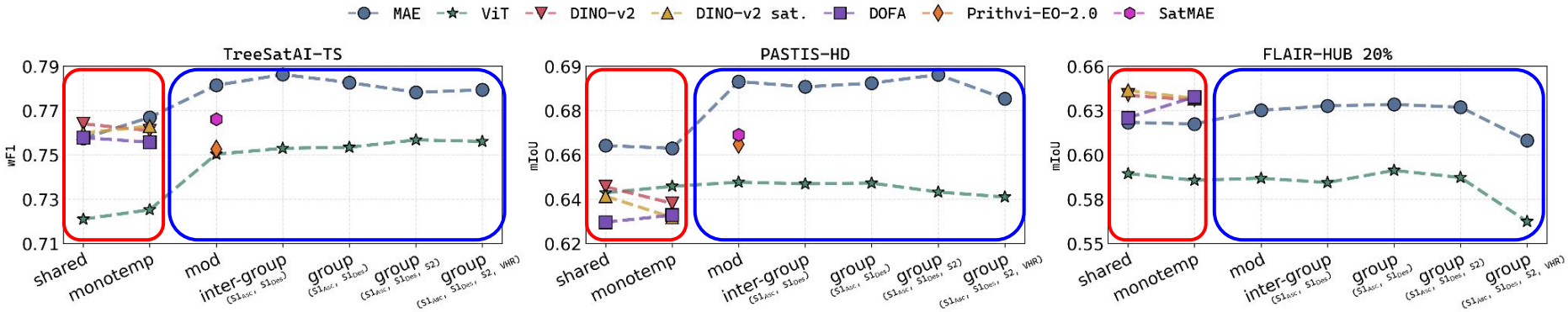
Best choice of multi-temporal fusion?

Benchmark:

- **Late fusion**
- **Early fusion**

Conclusions:

- **Early fusion** ≫ **Late fusion**
- Significance: most off-the-shelf foundation models are pre-trained on single images → they impose late fusion at fine-tuning

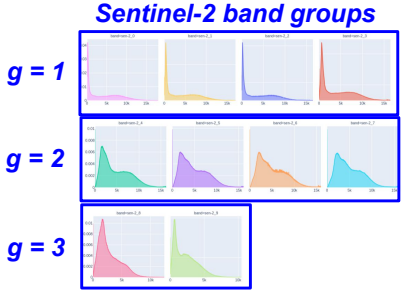


Best choice of target normalization

Benchmark:

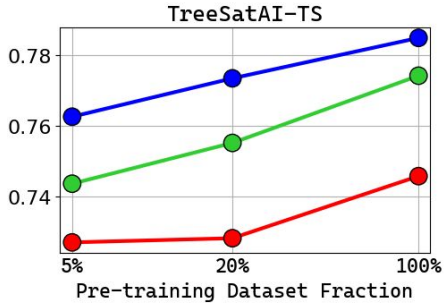
- **no normalization** → $\mathcal{L} = \frac{1}{|\mathcal{P}|} \sum_p \left\| \mathbf{x}_p - \mathbf{x}_p^{\text{rec}} \right\|_1$
- **patch-wise normalization** (MAE, Spectral GPT, EarthView) → $\hat{\mathcal{L}} = \frac{1}{|\mathcal{P}|} \sum_p \left\| \underbrace{\hat{\mathbf{x}}_p}_{\frac{\mathbf{x}_p - \mu(\mathbf{x}_p)}{\sigma(\mathbf{x}_p)}} - \mathbf{x}_p^{\text{rec}} \right\|_1$
- **patch-group-wise normalization (ours)** → $\hat{\mathcal{L}}^{\text{grp}} = \frac{1}{|\mathcal{P}||\mathcal{G}|} \sum_{p,g} \left\| \underbrace{\hat{\mathbf{x}}_{p,g}^{\text{grp}}}_{\frac{\mathbf{x}_{p,g} - \mu(\mathbf{x}_{p,g})}{\sigma(\mathbf{x}_{p,g})}} - \mathbf{x}_{p,g}^{\text{rec}} \right\|_1$

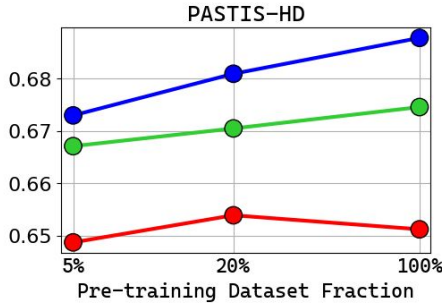
  We partition the set of bands into groups of strongly correlated bands



Conclusions:

- **patch-group-wise** > **patch-wise** > **no normalization**





Best choice of multi-spectral fusion?

Benchmark w/ patch-group-wise normalization:

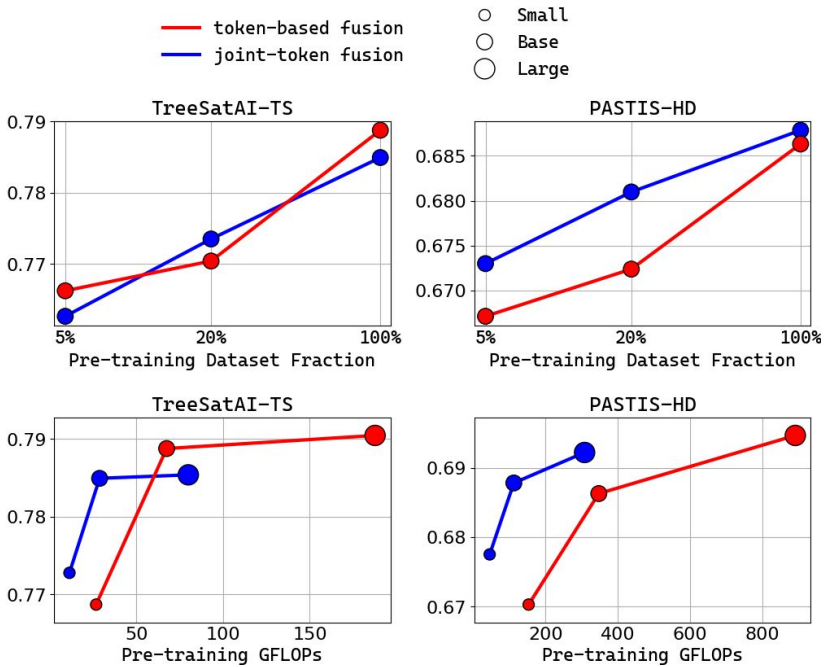
- **joint-token fusion** (all bands projected to the same token)
- **token-based fusion** (band groups projected to different tokens)

Conclusions:

- For a fixed target normalization, **joint-token fusion** matches **token-based fusion**
- Target normalization is the most important factor
  → no need to increase compute budget with **token-based fusion**

Possible interpretation:

- Target normalization improves per-patch balancing of the reconstruction task



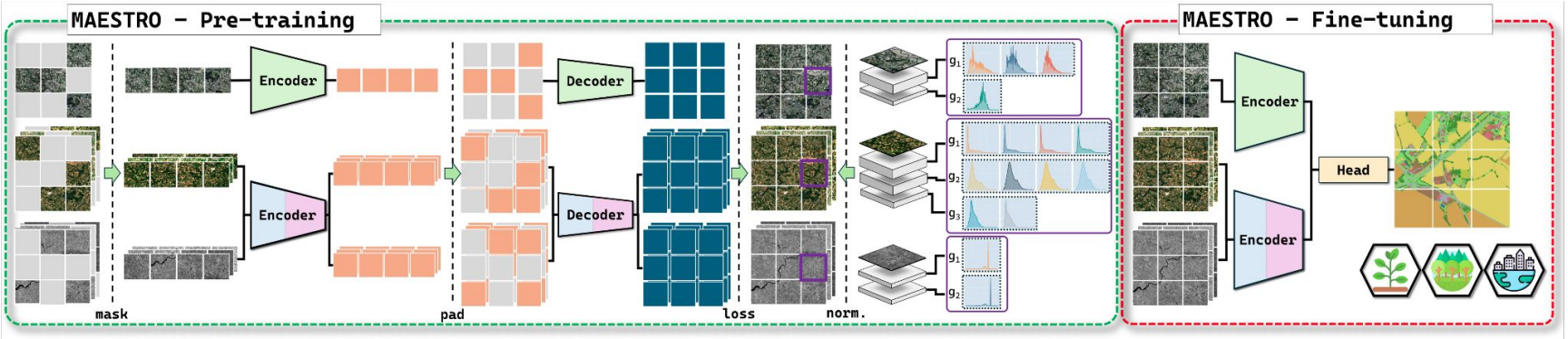
Our approach: MAESTRO

MAESTRO = “orchestration” of the MAE for multi-modal/temporal/spectral data

Adaptation to multi-modal/temporal data:

- Early multi-modal fusion among similar modalities ($S1_{ASC}$/$S1_{DES}$)
- Intermediate fusion across dissimilar modalities
- Early multi-temporal fusion

Adaptation to multi-spectral data = joint-token fusion with patch-group-wise target normalization



Evaluation of MAESTRO

	Tasks with strong multi-temporal components	Tasks dominated by a single mono-temporal modality
Intra-dataset setting	Strongly beats previous SOTA: **+3.8%** wF1 on TreeSatAI-TS **+2.5%** mIoU on PASTIS-HD	Roughly matches previous SOTA **-0.8%** mIoU on FLAIR#2 **+1.6%** mIoU on FLAIR-HUB
Cross-dataset setting	Beats best adapted baseline FMs **+2.7%** wF1 on TreeSatAI-TS **+1.4%** mIoU on PASTIS-HD	Mildly underperforms best adapted baseline FMs **-1.6%** mIoU on FLAIR#2 **-1.4%** mIoU on FLAIR-HUB

Conclusions

Strong benefit at adapting off-the-shelf SSL approaches to multi-modal/temporal/spectral EO data:

- Multi-modality: avoid early fusion among dissimilar modalities
- Multi-temporality: early fusion is highly preferable → **multi-temporal SSL is an underexplored opportunity**
- Multi-spectrality: joint-token fusion with a proper target normalization matches token-based fusion

MAESTRO = novel adaptation of the MAE to multi-modal/temporal/spectral data:

- Significant improvement over SOTA in tasks with strong multi-temporal components
- Competitive on tasks dominated by a single mono-temporal modality