# Multi-modal/temporal/spectral Masked AutoEncoders for Earth Observation

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

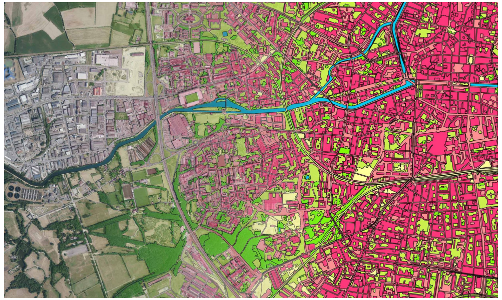
# Context

**Context of this work**

- IGN (French mapping agency) develops AI models for diverse applications (land cover/land use, agriculture, forestry, etc)

- IGN has published multi-modal/temporal/spectral datasets:
  - FLAIR
  - PASTIS/PASTIS-HD
  - PureForest

INSTITUT NATIONAL DE L'INFORMATION GÉOGRAPHIQUE ET FORESTIÈRE

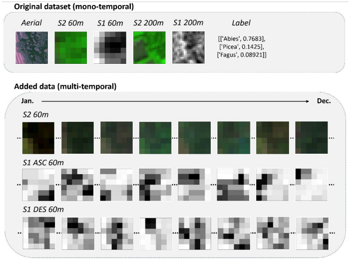


**Questions:**

- **Q1:** Can SSL improve performance/data efficiency on multi-modal/temporal/spectral datasets?
- **Q2:** Can we rely on off-the-shelf foundation models for multi-modal/temporal/spectral datasets?

# Approach : Intra-dataset MAE

**TreeSatAI-TS**
Tree species classification




MAE pretraining
Probing/ Fine-tuning
SSL workflow

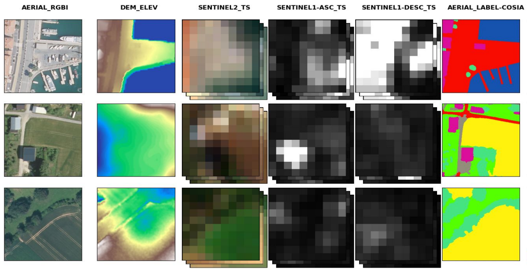
Fine-tuning
no SSL workflow

**PASTIS-HD**
Agriculture crop semantic segmentation


MAE pretraining
Probing/ Fine-tuning
SSL workflow

Fine-tuning
no SSL workflow

**FLAIR-HUB**
Land cover semantic segmentation




MAE pretraining
Probing/ Fine-tuning
SSL workflow

Fine-tuning
no SSL workflow

# Approach : Adapting to multi-modality/temporality

**Encoders "shared"**

- Encoders shared across modalities
- Late multi-modal fusion
- Late multi-temporal fusion

**Encoders "monotemp"**

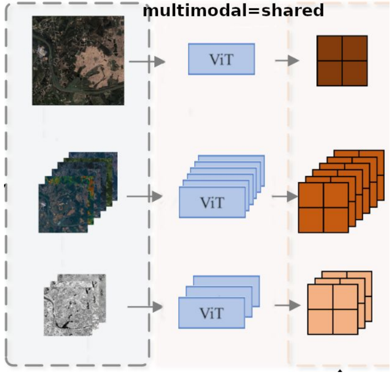
- Encoders specific to each modality
- Late multi-modal fusion
- Late multi-temporal fusion

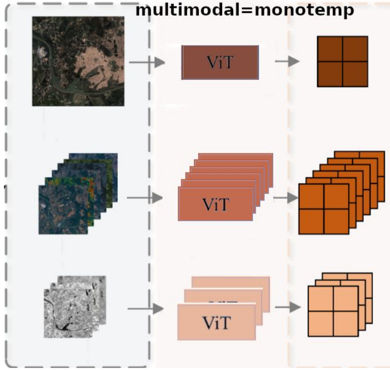
**Encoders "mod"**

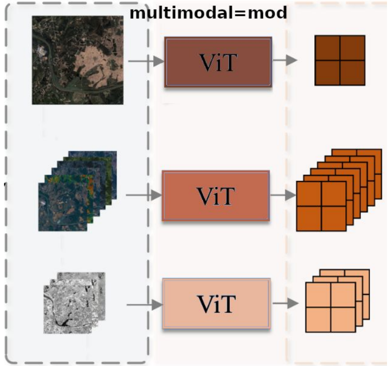
- Encoders specific to each modality
- Late multi-modal fusion
- Early multi-temporal fusion

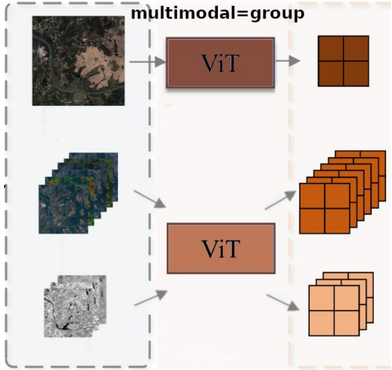
**Encoders "group"**

- Encoders specific to each group of modalities
- Early multi-modal fusion for grouped modalities
- Early multi-temporal fusion



# Approach : Adapting to multi-spectrality

**Targets unnormalized** $\longrightarrow$

$$\mathcal{L} = \frac{1}{|\mathcal{M}||\mathcal{P}|} \sum_{m,p} \left\| \mathbf{x}_{m,p} - \mathbf{x}^{rec}_{m,p} \right\|_1$$

**Targets normalized per patch** $\longrightarrow$

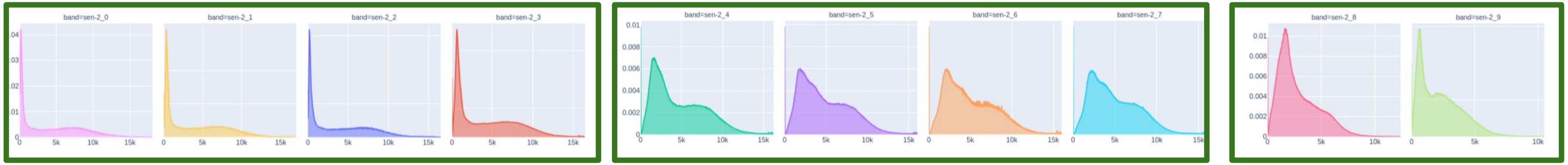
([MAE](), [SpectralGPT](), [EarthView]())

$$\mathcal{L}_{norm} = \frac{1}{|\mathcal{M}||\mathcal{P}|} \sum_{m,p} \left\| \underbrace{\mathbf{x}^{norm}_{m,p}}_{\frac{\mathbf{x}_{m,p} - \mu(\mathbf{x}_{m,p})}{\sigma(\mathbf{x}_{m,p})}} - \mathbf{x}^{rec}_{m,p} \right\|_1$$

**Targets normalized per patch and channel group** $\longrightarrow$

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

$$\mathcal{L}_{norm,\mathcal{G}} = \frac{1}{|\mathcal{M}||\mathcal{P}||\mathcal{G}|} \sum_{m,p,g} \left\| \underbrace{\mathbf{x}^{norm,\mathcal{G}}_{m,p,g}}_{\frac{\mathbf{x}_{m,p,g} - \mu(\mathbf{x}_{m,p,g})}{\sigma(\mathbf{x}_{m,p,g})}} - \mathbf{x}^{rec}_{m,p,g} \right\|_1$$




*Sentinel-2 band histograms on TreeSatAI-TS*

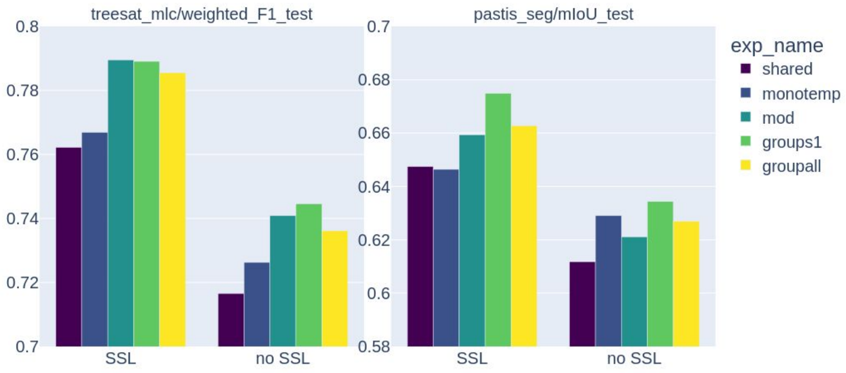
# Results on multi-modality/temporality

**Multi-modality**

- *Early fusion* > *Late fusion* for similar modalities *(e.g. S1A/S1D)*
- *Late fusion* > *Early fusion* for dissimilar modalities (e.g. *VHR* vs *S1A/S1D/S2*)

**Multi-temporality**

- *Early fusion* ≫ *Late fusion*



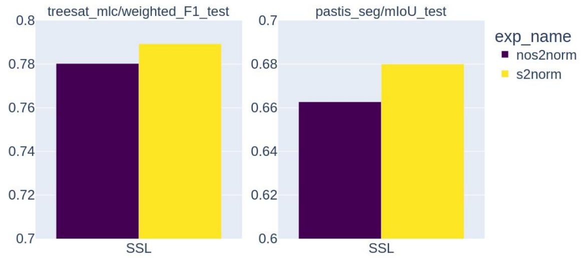

**Significance** : most off-the-shelf foundation models are pre-trained on single images

- They are compatible only with *late fusion*
- This results in a significant performance deficit:
  - -**3%** weighted F1 on TreeSatAI-TS
  - **-3.6%** mIoU on PASTIS-HD

# Results on multi-spectrality

**Normalizing targets with multiple groups**

- Strong benefit at normalizing S2 targets with 3 distinct groups
  - **+0.9%** weighted F1 on TreeSatAI-TS
  - **+1.7%** mIoU on PASTIS-HD
- Not much benefit at normalizing aerial targets with 2 distinct groups (*NIR* vs *RGB*)




**Grouping in separate tokens vs grouping only in target normalization**

- Similar performance with grouping in separate tokens vs only in target normalization
- Target normalization is the most important factor at play → no need to increase compute budget with separate tokens

**Interpretation**

- Per-patch normalization’s benefit might stem from per-patch balancing of reconstruction task
- If band groups have little histogram overlap, per-patch normalization approaches a constant normalization (no patch dependence)

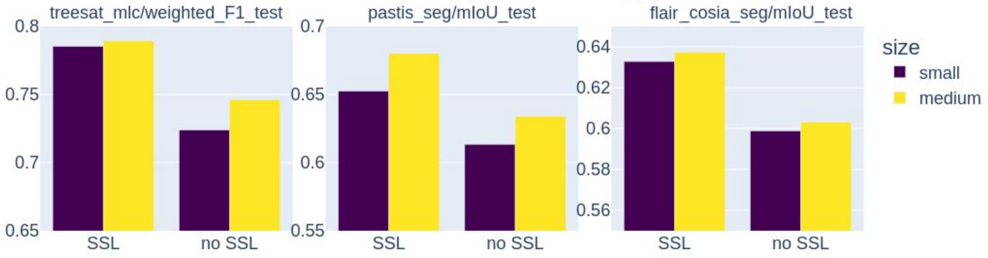
# Final results — Gains with SSL vs no SSL

**Unified approach selected for all datasets : multi-modal/temporal/spectral MAE**

- Multi-temporality: *early fusion*
- Multi-modality: *early fusion* for similar modalities, *late fusion* for dissimilar modalities
- Multi-spectrality: grouping only in target normalization

**SSL gains with this unified approach:**

- **+5-6%** weighted F1 on TreeSatAI-TS
- **+4.5%** mIoU on PASTIS-HD
- **+3.5%** mIoU on FLAIR-INC



# Conclusions

**Q1:** Can SSL improve performance/data efficiency on multi-modal/temporal/spectral datasets?

- SSL with our unified approaches leads to +**3.5-6%** performance gains (F1/mIoU) in fine-tuning
- Even on large datasets (e.g. FLAIR-HUB), gains are significant!

**Q2:** Can we rely on off-the-shelf foundation models on multi-modal/temporal/spectral datasets?

- Most foundation models are only compatible with *late fusion* for multi-modality/temporality
- This results in a significant performance deficit (**>-3%** performance deficit)

**Novelty of these results:**

- Benchmarked different choices of multi-modal/temporal/spectral fusion
- Adapted MAE to multi-modal/temporal/spectral datasets
- Conference paper on its way

# Thank you for your attention!