Time2Agri: Temporal Pretext Tasks for Agricultural Monitoring

Moti Rattan Gupta [1], **Anupam Sobti** [1]

[1]Plaksha University, Mohali, Punjab, India
moti.gupta@plaksha.edu.in, anupam.sobti@plaksha.edu.in

Abstract

Self Supervised Learning (SSL) has emerged as a prominent paradigm for label-efficient learning, and has been widely utilized by remote sensing foundation models (RSFMs). Recent RSFMs including SatMAE and DoFA primarily rely on masked autoencoding (MAE), contrastive learning or some combination of them. However, these pretext tasks often overlook the unique temporal characteristics of agricultural landscape, namely nature's cycle of sowing, growth, and harvest. Motivated by this gap, we propose three novel agriculture-specific pretext tasks, namely Time-Difference Prediction (TD), Temporal Frequency Prediction (FP), and Future-Frame Prediction (FF). Comprehensive evaluation on SICKLE dataset shows FF achieves 69.6% IoU on crop mapping and FP reduces yield prediction error to 30.7% MAPE, outperforming all baselines, and TD remains competitive on most tasks. Further, we also scale FF to the national scale of India, achieving 54.2% IoU outperforming all baselines on field boundary delineation on FTW India dataset.

Code — https://github.com/Geospatial-Computer-Vision-Group/agri-pretext

1 Introduction

With over 582 million people projected to be chronically undernourished by 2030, of which 53% would be concentrated in Africa, the world is significantly off-track in achieving SDG 2 - Zero Hunger (Unicef et al. 2024). These are further exacerbated by climate extremes like droughts and heatwaves affecting crop yields, and crop failures for farmers (TORETI et al. 2024). Remote sensing plays a critical role in addressing the growing challenges of food security and agricultural sustainability by enabling improved crop monitoring, yield forecasting, and resource optimization (Mehedi et al. 2024). However, the limited availability of labeled data, especially in developing regions, hinders the full potential of remote sensing.

Remote Sensing Foundation Models (RSFMs), built upon the principle of self-supervised learning, have shown remarkable promise in learning from the abundant unlabelled satellite imagery data, demonstrating superior performance on downstream tasks including landcover mapping, crop type mapping, and field boundary delineation (Lu et al. 2025). However, we argue most of them are suboptimal for agricultural monitoring due to the following reasons:

1. *Lack of domain-specific knowledge*: RSFMs trained primarily using masked autoencoding and/or contrastive learning do not incorporate domain-specific knowledge for agricultural monitoring;
2. *Lack of regional focus*: RSFMs trained on global datasets may not be optimal for regional agricultural monitoring due to differences in crop types, farming practices, and climate.

Inspired by work of Rolf et al. (2024), who argue that satellite imagery constitutes a distinct modality within computer vision due to its unique spatial, spectral, and temporal

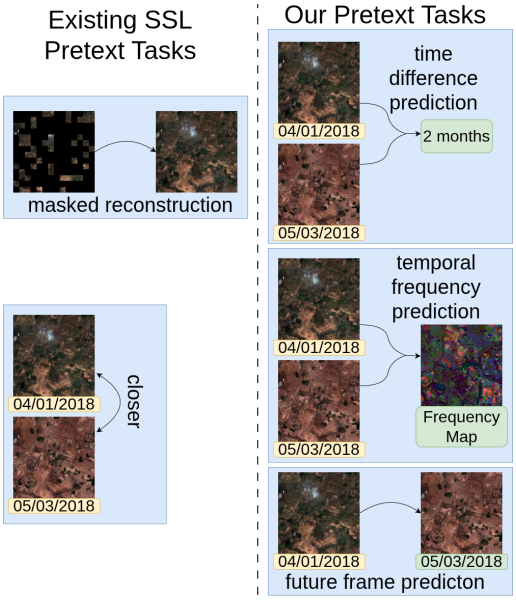


Figure 1: Overview of the proposed pretext tasks for self-supervised learning in agricultural remote sensing. On the left side, we have traditional pretext tasks focusing on masked reconstruction and invariance to augmentation (in this case, temporal), while on the right we have our temporally inspired pretext tasks.

Figure 2: Comparison of a region across two different acquisitions. Agricultural areas exhibit large, coherent frequency clusters—represented by different colors—that correspond to varying crop growth rates and naturally delineate farm parcels. In contrast, urban regions (highlighted in red) do not display such structured clustering, indicating distinct temporal dynamics.

properties, we extend this perspective to agricultural monitoring. Specifically, we argue that agricultural landscapes exhibit characteristic temporal dynamics driven by the natural cycle of sowing, growth, and harvest, leading to predictable spatial, spectral and temporal patterns. For instance, agricultural landscapes exhibit strong spatial clustering, in contrast to urban landscapes (Fig 2). This inherent regularity serves as a strong inductive bias to inform the design of more effective pretext tasks enabling the development of rich task-relevant representations for agricultural monitoring.

To this end, we propose three novel pretext tasks that leverage the seasonal nature of agricultural land and explain our intuitions behind them,

1. *Temporal Difference Prediction*: This task predicts the time difference between two images from a bitemporal pair of satellite images. By learning to estimate temporal gaps, the network implicitly learns to recognize phenological progression, from bare soil to seedlings to mature crops to harvest.
2. *Temporal Frequency Prediction*: This task predicts the per-region dominant temporal frequencies from a bitemporal satellite image pair. These temporal frequencies capture diverse cropping patterns, such as annual cycles for most crops, biannual cycles for double-cropping systems, or multi-year rotations. By learning to infer these temporal dynamics, the model captures information about crop types and farming practices that extends well beyond the observed dates.
3. *Future Frame Prediction*: Given a bitemporal pair of satellite images, this task predicts the future image from the past image. By exploiting the predictable nature of agricultural cycles, this task forces the network to understand the causal relationships between current conditions and future states, learning rich representations of crop phenology and management practices.

We evaluate our proposed pretext tasks across a range of key agricultural monitoring tasks at regional scale of Tamil Nadu, on the SICKLE benchmark (Sani et al. 2024) and assess the performance of one of our best performing pretext task on the field boundary delineation at the national scale of India, on FTW India (Kerner et al. 2024). We have observed consistent improvements over both state-of-the-art remote sensing foundation models DoFA (Xiong et al. 2024) and CopernicusFM (Wang et al. 2025), and a self-supervised MAE baseline. Our results suggest that incorporating domain-specific cues in the pretext task design can result in rich representations for agricultural monitoring tasks.

Beyond pretext task design, we have also investigated the impact of geographical scale for training agricultural models. Specifically, we compare models trained at a national scale with those trained at a regional scale. Our findings reveal that regional scale pretraining outperform national scale pretraining on most regional downstream tasks, emphasizing the importance of regional agricultural representations.

To summarize, our contributions are fourfold:

1. To the best of our knowledge, we are first to propose novel domain-inspired pretext tasks that exploit the seasonality of agriculture.
2. We demonstrate their effectiveness across multiple agricultural monitoring tasks at regional (Tamil Nadu) and national (India) scales.
3. We show that region-specific pretraining outperforms national-scale pretraining for regional agricultural challenges, highlighting the value of geographic focus.
4. We create two novel dense satellite-image time series (SITS) pretraining datasets for Tamil Nadu and India.

2 Related Work

Pretext Tasks in Satellite Imagery Early remote sensing SSL pretext tasks adapted their natural image counterparts, focusing on invariance of spatially closer tiles (Jean et al. 2018) or patches (Kang et al. 2021), while others extended colorization (Vincenzi et al. 2020). Following work, GASSL (Ayush et al. 2021) and SeCo (Mañas et al. 2021) emphasized invariance to temporal evolution, while CaCo (Mall, Hariharan, and Bala 2023) restricted invariance to

seasonal changes. However, seasonal invariance is suboptimal for agriculture, seasonality being an important marker of agricultural land, and can lead to loss of rich phenological features.

Masked Autoencoders in Remote Sensing Recent RSFMs have built upon the MAE (He et al. 2021) framework, including SatMAE (Cong et al. 2022), ScaleMAE (Reed et al. 2023), SelectiveMAE (Wang et al. 2024) and M3AE (Li et al. 2024), while some combine masked image modelling (MIM) with contrastive objectives (Zhang, Liu, and Wang 2024). Although they have promising performance on several remote sensing tasks including change detection, landcover classification, and object detection, they neglect the temporal dimension of remote sensing data.

Domain Specific Features While generic SSL methods often focus on raw pixel reconstruction or learning invariance to different augmentations, agricultural applications benefit from incorporating domain knowledge. For example, FGMAE (Wang et al. 2023b) has shown that reconstructing agricultural indices (e.g., NDVI) can improve downstream performance in MAE, compared to pixel-based reconstruction. Instead of focusing on reconstructing spectral indices, which primarily focus on spatial and spectral features, we focus on integrating both change related temporal features along with rich semantic spatial and spectral features, which provide rich spatial, temporal and spectral signal for agricultural monitoring.

Temporal Pretext Tasks in Other Modalities Future frame prediction extends sequence modeling in time series data, aiming to generate future observations based on historical inputs. This task has been extensively studied across domains such as video understanding (Jang et al. 2024) and weather forecasting (Bodnar et al. 2024; Nguyen et al. 2023). Similarly, temporal difference prediction has been studied as playback rate prediction in videos (Schiappa, Rawat, and Shah 2023) but has not been explored in the context of remote sensing imagery. Recently, Ravirathinam et al. (2024) explored predicting future frames from past visual sequences conditioned on weather data. In contrast, our work proposes a novel approach to future frame prediction in agricultural landscapes by leveraging only a single past frame, without relying on any external data. We hypothesize that this is feasible due to the inherently cyclical nature of agricultural patterns.

To the best of our knowledge, this is the first work to introduce temporal difference prediction and temporal frequency prediction as pretext tasks for remote sensing imagery. Furthermore, we are the first to explore future frame prediction from a single past frame without incorporating any external information—an approach made viable by the cyclical patterns inherent in agricultural landscapes.

3 Methodology

3.1 Pretraining Data

Our pretext tasks, as will be introduced later, require a dense SITS over the target agricultural region. Existing large SITS pretraining datasets including SeCo (Mañas et al. 2021) and SSL4EO (Wang et al. 2023a) lack a definitive agricultural focus, and have sparse temporal sampling. To capture rich agricultural representations at both regional and national scales, we developed two novel dense SITS pretraining datasets for Tamil Nadu and India, ensuring extensive temporal coverage.

Regional Pretraining

SICKLE SICKLE dataset provides multi-sensor imagery (Sentinel-2, Sentinel-1, and Landsat-8) from January 2018 till March 2021 over Cauvery Delta region in Tamil Nadu. They provide rich annotations for key agricultural tasks including crop type mapping, crop yield estimation, sowing date, transplanting date and harvesting date prediction. This dataset's agricultural focus and geographical specificity not only establish it as our primary downstream evaluation benchmark but also provide the empirical justification for concentrating our regional pretraining efforts on Tamil Nadu

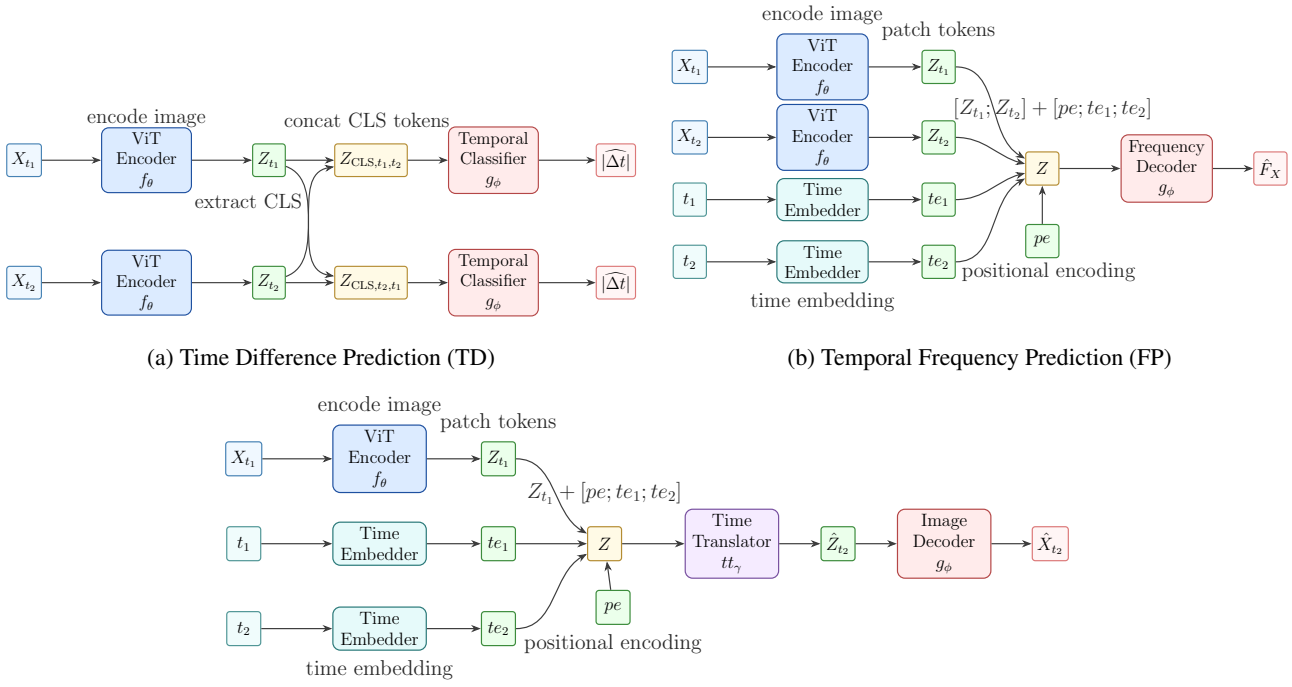
Data Construction To construct our regional scale pretraining dataset, we systematically identified all Sentinel-2 tiles that intersect with SICKLE's spatial extent, specifically tiles: 44PKS, 44PKT, 44PLS, and 44PLT, with 44PLS and 44PLT having significant ocean coverage. After application of a water body mask, all 224x224 spatial chips were sampled and a dense SITS for each chip was constructed with a monthly resolution from the dataset's original acquisition period.

Dominant Frequency Map We additionally augment our regional dataset with a per-pixel dominant frequency map derived from the per-pixel NDVI time series in our data. Each per-pixel NDVI time series is interpolated to a regular monthly grid, and is smoothed using Savitzky-Golay filter (Savitzky and Golay 1964). Dominant per-pixel Fourier frequencies are calculated and stored as pixel values for each pixel sorted by their power. As discussed in Section 1, these frequencies capture different crop growth rates for crops grown in an agricultural parcel, and can naturally identify different parcels. We refer the reader to Appendix for more details.

National Pretraining

Fields of the World (FTW) FTW is a field instance segmentation benchmark across multiple countries including India, and has over 70,000 chips globally and 1,960 chips specifically from India. Each chip consists of a bitemporal harmonized Sentinel 2 RGB+NIR imagery paired with instance and semantic segmentation masks, making it ideal to evaluate effectiveness of our methods at a national scale.

National Pretraining Data To construct our national scale pretraining datasets, we referred to India chips from FTW (FTW India) and for each chip, we construct a 3x3 grid with chip at center and additionally sample its eight spatial neighbors, since an agricultural area is likely to be surrounded by other agricultural areas. Similar to our regional pipeline, we keep a monthly resolution but chose acquisition period from January 2016 to January 2019.



(a) Time Difference Prediction (TD)

(b) Temporal Frequency Prediction (FP)

(c) Future Frame Prediction (FF)

Figure 3: Overview of proposed pretext tasks. (a) **Time Difference Prediction (TD)**: Predicts absolute time gap $|\Delta t|$ from stacked CLS tokens $[(Z_{t_1})_{\text{CLS}}; (Z_{t_2})_{\text{CLS}}]$ via temporal classifier g_ϕ. (b) **Temporal Frequency Prediction (FP)**: Predicts per-pixel dominant frequency map $\hat{F}_X$ from concatenated latents $[Z_{t_1}; Z_{t_2}] + [pe; te_1; te_2]$ using decoder g_ϕ. (c) **Future Frame Prediction (FF)**: Predicts future frame $\hat{X}_{t_2}$ by generating latent $\hat{Z}_{t_2}$ from Z_{t_1} and $[pe; te_1; te_2]$ using time translator tt_γ, followed by decoder g_ϕ. Here, f_θ is a ViT encoder, g_ϕ is a transformer decoder, and tt_γ is a time translator module. pe and te_i denote positional and temporal encodings respectively.

3.2 Pretext Tasks

Setup We define our pretraining dataset as $S = (X_{t_i}, X_{t_j}, F_X)$ consisting of bitemporal image pairs , where X_{t_i} and X_{t_j} are satellite images at t_i and t_j with $t_i <= t_j$, and F_X is per-pixel dominant frequency map constructed from entire time series. We define f_θ as a ViT encoder which extracts $Z_{t_i} \in \mathbb{R}^{(N+1)\times D^{in}}$ from $X_{t_i} \in \mathbb{R}^{H\times W\times C}$,where N is the number of patches. From Vaswani et al. (2017), we define positional encoding $PE(i, .)$

$$PE(i, 2j) = \sin(\frac{i}{B^{\frac{2j}{D}}}) \tag{1}$$

$$PE(i, 2j+1) = \cos(\frac{i}{B^{\frac{2j}{D}}}) \tag{2}$$

where i, j denote position and feature index, D denotes length of the encoding and B is a constant. Similarly, we define time encoding of timestamp t_i as te_i as

$$te_i = [PE(m_i); PE(y_i)] \tag{3}$$

where m_i, y_i denote month, and year of t_i, $[;]$ denotes concatenation. We avoid using day-of-year (Cong et al. 2022) or month-of-year (Tseng et al. 2023) encodings, as we believe they undermine inter-year differences.

Time Difference Prediction - TD In this task (Fig 3a), we predict the absolute monthly time difference $|\Delta t| = |t_2 - t_1|$ from latent representations Z_{t_1} and Z_{t_2} by applying a lightweight temporal classifier (e.g., a three-layer MLP) on top of stacked CLS tokens from both representations

$$Z_{CLS,t_1,t_2} = [(Z_{t_1})_{CLS}; (Z_{t_2})_{CLS}] \tag{4}$$

where $(Z_{t_i})_{CLS}$ denote the CLS token of Z_{t_i}. The classifier is trained to predict $|\Delta t|$ discretized into C classes, and trained using a cross-entropy loss.

To ensure robustness to temporal order, we use a symmetric loss function considering stacking Z_{CLS,t_1,t_2} and Z_{CLS,t_2,t_1} as input to the classifier.

TD exploits temporal dimension of agricultural landscape by encoding temporal difference between sowing and harvest, or between two harvests in its representation. This is useful for learning crop phenological stages and growth patterns, which is important for agricultural monitoring.

Temporal Frequency Prediction - FP FP (Fig 3b) effectively addresses the limitations of TD, namely lack of dense supervision and more rich temporal features. By predicting a per-pixel temporal frequency map F_X from the latent representations Z_{t_1} and Z_{t_2}, it allows encoder f_θ to learn dominant temporal features over each pixel from the entire time

	Model	Pretrain Region	Crop Type (IoU ↑)	Crop Yield (MAPE ↓)	Crop Yield In-Season (MAPE ↓)	Sowing Date (MAPE ↓)	Transpl. Date (MAPE ↓)	Harvest Date (MAPE ↓)
Sup.	ViT-S	N/A	0.65523	0.36357	0.33687	0.016278	0.025533	0.069594
	UNet3D	N/A	0.65442	0.38805	0.37964	**0.012484**	**0.019249**	**0.047596**
Existing	MAE	TN	0.67138	0.37833	0.32701	0.017487	0.023406	0.065195
	DoFA	GL	0.58363	0.36144	0.34571	0.016849	0.022475	0.071494
	CopernicusFM	GL	0.57229	**0.32334**	0.33706	0.021345	0.028231	0.068336
Proposed	FF	TN	**0.69595**	0.36765	0.32516	0.015821	0.028394	0.068059
	FP	TN	0.61976	0.34338	**0.30727**	0.016081	0.022514	0.066166
	TD	TN	0.66628	0.36875	0.32877	0.017842	0.024985	0.065945
	India FF[1]	IN	0.60483	0.39750	0.37407	0.016549	0.022403	0.064496

Table 1: Performance comparison of our pretext tasks with MAE, DoFA, CopernicusFM, supervised ViT-S and UNet3D on the SICKLE dataset. ↓ indicates lower is better, ↑ indicates higher is better. Bold values indicate the best performance for each task. Note: Pretext tasks FF, FP, TD and MAE are executed on our Tamil Nadu pretraining data (TN), while DoFA and CopernicusFM use global pretrained checkpoints (GL), from Xiong et al. (2024) and Wang et al. (2025) respectively.

Model	Crop Type (IoU ↑)	Crop Yield (MAPE ↓)	Crop Yield In-Season (MAPE ↓)	Sowing Date (MAPE ↓)	Transpl. Date (MAPE ↓)	Harvest Date (MAPE ↓)
FF	**0.69595**	0.36765	**0.32516**	0.015821	0.028394	0.068059
FF-(3)	0.58737	0.34233	0.34674	0.01515	0.027323	0.066392
FF-(3,6)	0.64414	**0.33411**	0.36291	**0.015033**	0.025032	0.057168
FF-(3,6,9)	0.59340	0.34979	0.34940	0.01544	**0.021851**	**0.05391**

Table 2: Performance comparison of different temporal gap between bitemporal pairs on FF (executed on regional data). ↓ indicates lower is better, ↑ indicates higher is better. Bold values indicate the best performance for each task. FF denotes 0–3 months apart; FF-(3) denotes 3 months apart; FF-(3,6) denotes 3 and 6 months apart; FF-(3,6,9) denotes 3, 6, and 9 months apart.

series. We utilize MAE inspired transformer decoder g_ϕ to predict frequency map $\hat{F}_X$, by using concatenated latents (without CLS tokens) with summation of concatenated temporal encodings te_1 and te_2, and positional encoding pe.

$$Z = [Z_{t_1}; Z_{t_2}] + [pe; te_1; te_2] \tag{5}$$

$$\hat{F}_X = g_\phi(Z) \tag{6}$$

This results in a rich temporal representation capturing temporal patterns (e.g., crop rotation) beyond the immediate observable window. This is a regression task, and is optimized using normalized pixel-wise MSE (He et al. 2021).

Future Frame Prediction - FF Unlike FP, FF (Fig 3c) is a generative task, where we predict the future frame X_{t_2} from the past frame X_{t_1}. This addresses the limitation of dense temporal supervision, without the need of pre-computed frequency maps over the entire time series. We retain the same transformer decoder g_ϕ as in FP, but additionally use another transformer decoder tt_γ, called time translator, to predict future latent representation $\hat{Z}_{t_2}$ from summation of Z_{t_1} and concatenation of te_1, te_2 and pe.

$$\hat{Z}_{t_2} = tt_\gamma([Z_{t_1}] + [pe; te_1; te_2]) \tag{7}$$

$$\hat{X}_{t_2} = g_\phi(\hat{Z}_{t_2}) \tag{8}$$

Without the computational overhead of computing frequency maps, this results in rich temporal representation encoding causal relationship between current state and future state of an agricultural landscape. The network is similarly optimized using normalized pixel-wise MSE (He et al. 2021) loss between predicted future frame $\hat{X}_{t_2}$ and ground truth X_{t_2}.

4 Experiments & Results

We aim to answer two fundamental questions that emerge from our core hypothesis that agricultural landscape's cyclical nature can inform better self-supervised pretext tasks. Specifically, we aim to answer: (i) do temporal pretext tasks learn more agriculturally relevant representations, and (ii) which temporal characteristics are most valuable for different agricultural monitoring tasks. Additionally, we also seek to understand the role of national scale pretraining on tackling regional challenges.

[1]"India FF" refers to FF trained over national-scale India dataset (IN). For further details, see Section 4.5.

Sup. Pretrain Region	Model	IoU (↑)	Pixel Recall (↑)	Object Recall (↑)
FTW - India	FF	**0.541998**	**0.637823**	**0.029626**
	MAE	0.527783	0.628631	0.022341
	ViT-S	0.493208	0.577795	0.022827
	DoFA	0.38059	0.442879	0.002428
	CopernicusFM	0.414409	0.485358	0.007771
Ai4boundaries	FF	**0.530193**	**0.626151**	**0.023798**
	MAE	0.517006	0.615881	0.019427
	ViT-S	0.415647	0.488697	0.0034
	DoFA	0.368869	0.427868	0.001943
	CopernicusFM	0.409614	0.480632	0.002914
France	FF	**0.490628**	**0.577658**	**0.015542**
	MAE	0.465626	0.55264	0.014085
	ViT-S	0.476051	0.573084	0.003885
	DoFA	0.37447	0.439183	0.003885
	CopernicusFM	0.388325	0.452524	0.005342
India	FF	0.492768	0.591907	**0.024284**
	MAE	**0.497104**	**0.60582**	0.009228
	ViT-S	0	0	0
	DoFA	0.430635	0.513778	0.004371
	CopernicusFM	0.398615	0.468785	0.009228

Table 3: Performance comparison of FF, MAE, DoFA, CopernicusFM and supervised ViT-S on the India subset of the FTW dataset. FTW-India (all FTW countries except India), Ai4boundaries, France, and India refer to supervised pretraining regions (Kerner et al. 2024), while metrics are reported on the India test set from FTW. Bold indicates the best performance per each supervised pretraining region. Note: Pretext tasks FF and MAE are executed on our national scale pretraining data, while DoFA and CopernicusFM use checkpoints from Xiong et al. (2024) and Wang et al. (2025), respectively.

4.1 Downstream Evaluations

SICKLE SICKLE benchmark, as discussed earlier in Section 3.1, is a comprehensive agricultural monitoring benchmark over Cauvery Delta region of Tamil Nadu, consisting of SITS with annotations for crop type mapping, crop yield estimation, and sowing, transplanting and harvesting date prediction. Crop type mapping is a binary segmentation task for segmenting paddy pixels, measured by IoU, while crop yield estimation estimates the yield of a paddy parcel, measured by mean-absolute-percentage-error (MAPE). Similar to Sani et al. (2024), we consider using both satellite images during actual growing season, referred to as in-season, as well as using the whole time series available for estimating crop-yield. Date prediction tasks are carried out using in-season time series, where the task is to predict how many days apart from reference does sowing, transplanting and harvesting occur. Date prediction, similar to yield estimation, is a regression tasks and evaluated using MAPE.

Fields of the World FTW, as discussed earlier in Section 3.1, consists of over 70,000 chips with annotations for instance and semantic segmentation mask. Similar to Kerner et al. (2024), we consider field instance segmentation performance on India for 3-class setting: field interior, field exterior, and field boundary. Our downstream tasks use IoU, Pixel Recall and Object Recall as our metrics due to availability of presence-only labels in FTW India. We consider evaluation strategy similar to Kerner et al. (2024), which involved three settings: training on all FTW countries following by finetuning only on India, training on Ai4boundaries countries followed by finetuning on India, and training on France followed by finetuning on India. In addition, we also consider training on India only and evaluation on India test set of FTW.

4.2 Implementation Details

We use a ViT-S encoder for pretraining, with a 3-layer MLP for TD, 6-layer Transformer decoder for FP, and a 2-layer Time Translator + 4-layer Transformer decoder for FF. Bitemporal pairs are sampled with a max gap of 3 months. As a baseline, we train an MAE on single-frame reconstruction. All experiments use 4-band (RGB+NIR) imagery, consistent with the FTW dataset. We adopt DoFA and CopernicusFM as our RSFM baselines due to their native 4-band support and strong downstream performance. For the SICKLE task, we use a UPerNet segmentation head and compare with supervised ViT-S and UNet3D. For FTW India, we use a lightweight segmentation head (three ConvTranspose–Conv–BN–ReLU–Dropout blocks, followed by ConvTranspose–Conv), and compare with supervised ViT-S. SSL experiments use AdamW with cosine annealing, trained for 100 epochs. Regional-scale pretraining uses a batch size of 256 on a single RTX A6000 (48GB); national-scale uses 2×A100 (40GB) GPUs with effective batch size 640 and 10-epoch warmup.

4.3 Regional Scale Evaluation on SICKLE

Do Temporal Pretext Tasks Perform Better? Our Temporal pretext tasks demonstrate clear advantages over existing methods (Table 1). FF achieves the highest IoU on crop type (69.6%) mapping, substantially outperforming MAE (67.1%), DoFA (58.4%), CopernicusFM (57.2%) and supervised baselines (65.5%). FP achieves lowest in-season crop yield MAPE of 31%, outperforming all baselines, and remains competitive for full time series yield estimation, achieving 34% MAPE, closely following CopernicusFM (32%). While temporal tasks remain competitive on date prediction tasks, they show consistent advantages across the broader spectrum of agricultural monitoring applications. This suggests that learning temporal characteristics of agricultural landscapes leads to more informative and transferable agricultural representations than spatial-only or globally-pretrained alternatives.

Which Temporal Pretext Task for Agriculture? Among our temporal pretext tasks, FF and FP show complementary strengths aligned with different agricultural monitoring objectives. FF outperforms FP and TD by 7.6% and 3.0% respectively on crop type mapping, while FP achieves superior performance for crop yield prediction, outperforming both FF and TD by 2.4% and 2.5% MAPE points respectively on both in-season and complete season prediction. This performance differentiation reveals an important principle: FF learns comprehensive spatial, temporal and spectral crop characteristics suitable for identification tasks, while FP leverages its global temporal horizon to capture growth dynamics critical for yield estimation. TD shows intermediate performance across tasks, suggesting that simple temporal differences provide valuable but less specialized representations compared to generative (FF) or frequency-based (FP) approaches.

Why Different Downstream Favor Different Temporal Pretexts? The task-specific performance patterns reflect fundamental differences in different agricultural monitoring application. While FF's comprehensive understanding of temporal evolution over shorter horizon allows it to capture and anticipate rich spatial, phenological, and spectral patterns suitable for crop mapping, it remains suboptimal for anticipating crop stresses. On the other hand FP's frequency analysis approach captures both seasonal growth patterns and stress events over a much longer horizon, allowing it to learn robust representations for detecting crop stresses and yield estimation. Although neither beats supervised baselines on date prediction tasks, suggesting supervised approaches still remain superior for precise event detection.

Does Time Scale Matter in Agriculture? Table 2 reveals a clear principle: for FF, different agricultural monitoring tasks align with specific temporal scales in agricultural systems. While crop type mapping achieves best performance (69.6% IoU) with FF-(0,1,2,3), suggesting that crop identification benefits from observing phenological changes within the growing season, crop yield prediction performs optimally (33.4% MAPE) with FF-(3,6) configuration, indicating that yield estimation requires understanding seasonal to inter-seasonal patterns. Date prediction tasks show progressive improvement with longer temporal contexts, achieving best performance (3.01% MAPE) with FF-(3,6,9), suggesting that accurate agricultural timing prediction benefits from understanding multi-season planting patterns.

4.4 National Scale Evaluation on FTW India

Which Pretext Task to Scale? Our regional scale experiments (Table 1) suggests FF leads to better crop type representations, while FP performs the best on crop yield estimation. Our intuition that field boundary delineation is more closer to crop type mapping instead of yield estimation led us to go forward with FF for our national scale training.

Do Temporal Pretext Tasks Improve Cross-Region Transfer? FF consistently outperforms MAE across different geographic supervised pretraining sources (Table 3), with advantages ranging from 1.3 %(Ai4boundaries) to 2.5 %(France), suggesting that FF learn more fundamental agricultural principles that transcend specific geographic contexts.

Does SSL Eliminate the Need for Cross-Region Data? FF and MAE on India supervised pretraining only, recovers 99.7% and 100.8% of the supervised performance of FTW-India suggesting SSL pretraining alleviate the need of cross-region labelled data in label sparse setting.

4.5 Role of Geographical Scale

While *bigger is better* paradigm suggests larger, more diverse datasets produce superior representations, agricultural systems are inherently location-specific due to climate, soil, crop varieties, and management practices. We evaluated our national-scale FF representations (India FF) on the SICKLE benchmark to test this assumption.

Regional Representations Substantially Outperform National Representations Table 1 reveals a striking result that challenges foundation model conventions. Regional pretraining achieves 9.1% advantage for crop type mapping, and 3.0% advantage for crop yield prediction. This finding suggests that agricultural foundation models should prioritize geographic specialization over scale, highlighting *agriculture's location-dependent nature* that deviates from *bigger is better* paradigm.

5 Conclusion

This work demonstrates that agricultural landscapes' inherent temporal structure provides powerful supervisory signals for self-supervised learning. By designing three novel pretext tasks that leverages this structure, we achieve substantial improvements over diverse agricultural monitoring tasks. Further, this work shows that regional pretraining substantially outperforms national scale for regional downstream tasks, emphasizing on development of regional agricultural representations instead of global representations for tackling regional agricultural challenges. In future, we want to explore agro-climatic learning to enable generalized representation learning at global scale.

Acknowledgments

We are grateful to Dr. Shashank Tamaskar for his valuable feedback throughout various stages of this work. This research was supported by the Plaksha University Startup Research Grant and partially supported under Project Pahal, a CSR initiative by CNH Industrial.

References

Ayush, K.; Uzkent, B.; Meng, C.; Tanmay, K.; Burke, M.; Lobell, D.; and Ermon, S. 2021. Geography-aware self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10181–10190.

Bodnar, C.; Bruinsma, W. P.; Lucic, A.; Stanley, M.; Vaughan, A.; Brandstetter, J.; Garvan, P.; Riechert, M.; Weyn, J. A.; Dong, H.; Gupta, J. K.; Thambiratnam, K.; Archibald, A. T.; Wu, C.-C.; Heider, E.; Welling, M.; Turner, R. E.; and Perdikaris, P. 2024. A Foundation Model for the Earth System. arXiv:2405.13063.

Cong, Y.; Khanna, S.; Meng, C.; Liu, P.; Rozi, E.; He, Y.; Burke, M.; Lobell, D.; and Ermon, S. 2022. SatMAE: Pre-training Transformers for Temporal and Multi-Spectral Satellite Imagery. *ArXiv*, abs/2207.08051.

He, K.; Chen, X.; Xie, S.; Li, Y.; Doll'ar, P.; and Girshick, R. B. 2021. Masked Autoencoders Are Scalable Vision Learners. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15979–15988.

Jang, H.; Kim, D.; Kim, J.; Shin, J.; Abbeel, P.; and Seo, Y. 2024. Visual Representation Learning with Stochastic Frame Prediction. arXiv:2406.07398.

Jean, N.; Wang, S.; Samar, A.; Azzari, G.; Lobell, D.; and Ermon, S. 2018. Tile2Vec: Unsupervised representation learning for spatially distributed data. *ArXiv*, abs/1805.02855.

Kang, J.; Fernandez-Beltran, R.; Duan, P.; Liu, S.; and Plaza, A. J. 2021. Deep Unsupervised Embedding for Remotely Sensed Images Based on Spatially Augmented Momentum Contrast. *IEEE Transactions on Geoscience and Remote Sensing*, 59(3): 2598–2610.

Kerner, H.; Chaudhari, S.; Ghosh, A.; Robinson, C.; Ahmad, A.; Choi, E.; Jacobs, N.; Holmes, C.; Mohr, M.; Dodhia, R.; Ferres, J. M. L.; and Marcus, J. 2024. Fields of The World: A Machine Learning Benchmark Dataset For Global Agricultural Field Boundary Segmentation. arXiv:2409.16252.

Li, Z.; Hou, B.; Ma, S.; Wu, Z.; Guo, X.; Ren, B.; and Jiao, L. 2024. Masked Angle-Aware Autoencoder for Remote Sensing Images. arXiv:2408.01946.

Lu, S.; Guo, J.; Zimmer-Dauphinee, J. R.; Nieusma, J. M.; Wang, X.; VanValkenburgh, P.; Wernke, S. A.; and Huo, Y. 2025. Vision Foundation Models in Remote Sensing: A Survey. arXiv:2408.03464.

Mañas, O.; Lacoste, A.; Giró-i Nieto, X.; Vazquez, D.; and Rodríguez, P. 2021. Seasonal Contrast: Unsupervised Pre-Training From Uncurated Remote Sensing Data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 9414–9423.

Mall, U.; Hariharan, B.; and Bala, K. 2023. Change-Aware Sampling and Contrastive Learning for Satellite Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5261–5270.

Mehedi, I. M.; Hanif, M. S.; Bilal, M.; Vellingiri, M. T.; and Palaniswamy, T. 2024. Remote sensing and decision support system applications in precision agriculture: Challenges and possibilities. *Ieee Access*, 12: 44786–44798.

Nguyen, T.; Brandstetter, J.; Kapoor, A.; Gupta, J. K.; and Grover, A. 2023. ClimaX: A foundation model for weather and climate. arXiv:2301.10343.

Ravirathinam, P.; Khandelwal, A.; Ghosh, R.; and Kumar, V. 2024. A Causally Informed Pretraining Approach for Multimodal Foundation Models: Applications in Remote Sensing.

Reed, C. J.; Gupta, R.; Li, S.; Brockman, S.; Funk, C.; Clipp, B.; Keutzer, K.; Candido, S.; Uyttendaele, M.; and Darrell, T. 2023. Scale-MAE: A Scale-Aware Masked Autoencoder for Multiscale Geospatial Representation Learning. arXiv:2212.14532.

Rolf, E.; Klemmer, K.; Robinson, C.; and Kerner, H. 2024. Position: Mission Critical – Satellite Data is a Distinct Modality in Machine Learning. In Salakhutdinov, R.; Kolter, Z.; Heller, K.; Weller, A.; Oliver, N.; Scarlett, J.; and Berkenkamp, F., eds., *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 42691–42706. PMLR.

Sani, D.; Mahato, S.; Saini, S.; Agarwal, H. K.; Devshali, C. C.; Anand, S.; Arora, G.; and Jayaraman, T. 2024. SICKLE: A Multi-Sensor Satellite Imagery Dataset Annotated With Multiple Key Cropping Parameters. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 5995–6004.

Savitzky, A.; and Golay, M. J. E. 1964. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry*, 36(8): 1627–1639.

Schiappa, M. C.; Rawat, Y. S.; and Shah, M. 2023. Self-Supervised Learning for Videos: A Survey. *ACM Computing Surveys*, 55(13s): 1–37.

TORETI, A.; BAVERA, D.; ACOSTA, N. J.; ACQUAFRESCA, L.; AZAS, K.; BARBOSA, P.; DE, J. A.; FICCHI, A.; FIORAVANTI, G.; GRIMALDI, S.; et al. 2024. Global Drought Overview September 2024.

Tseng, G.; Zvonkov, I.; Purohit, M.; Rolnick, D.; and Kerner, H. R. 2023. Lightweight, Pre-trained Transformers for Remote Sensing Timeseries. *ArXiv*, abs/2304.14065.

Unicef; et al. 2024. The state of food security and nutrition in the world 2024.

Vaswani, A.; Shazeer, N. M.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *Neural Information Processing Systems*.

Vincenzi, S.; Porrello, A.; Buzzega, P.; Cipriano, M.; Fronte, P.; Cuccu, R.; Ippoliti, C.; Conte, A.; and Calderara, S. 2020. The color out of space: learning self-supervised representations for Earth Observation imagery. arXiv:2006.12119.

Wang, F.; Wang, H.; Wang, D.; Guo, Z.; Zhong, Z.; Lan, L.; Zhang, J.; Liu, Z.; and Sun, M. 2024. Scaling Efficient Masked Image Modeling on Large Remote Sensing Dataset. arXiv:2406.11933.

Wang, Y.; Braham, N. A. A.; Xiong, Z.; Liu, C.; Albrecht, C. M.; and Zhu, X. X. 2023a. SSL4EO-S12: A Large-Scale Multi-Modal, Multi-Temporal Dataset for Self-Supervised Learning in Earth Observation. arXiv:2211.07044.

Wang, Y.; Hernández, H. H.; Albrecht, C. M.; and Zhu, X. X. 2023b. Feature Guided Masked Autoencoder for Self-supervised Learning in Remote Sensing. arXiv:2310.18653.

Wang, Y.; Xiong, Z.; Liu, C.; Stewart, A. J.; Dujardin, T.; Bountos, N. I.; Zavras, A.; Gerken, F.; Papoutsis, I.; Leal-Taixé, L.; and Zhu, X. X. 2025. Towards a Unified Copernicus Foundation Model for Earth Vision. arXiv:2503.11849.

Xiong, Z.; Wang, Y.; Zhang, F.; Stewart, A. J.; Hanna, J.; Borth, D.; Papoutsis, I.; Saux, B. L.; Camps-Valls, G.; and Zhu, X. X. 2024. Neural Plasticity-Inspired Multimodal Foundation Model for Earth Observation. arXiv:2403.15356.

Zhang, M.; Liu, Q.; and Wang, Y. 2024. CtxMIM: Context-Enhanced Masked Image Modeling for Remote Sensing Image Understanding. arXiv:2310.00022.