

TiMo: Spatiotemporal Foundation Model for Satellite Image Time Series

Xiaolei Qin^{1*}, Di Wang^{1,2*}, Jing Zhang^{1†}, Fengxiang Wang³, Xin Su¹, Bo Du^{1,2}, Liangpei Zhang¹

¹Wuhan University ²Zhongguancun Academy ³National University of Defense Technology
 {qinxlei, dwang, xinsu.rs, dubo, zlp62}@whu.edu.cn
 jingzhang.cv@gmail.com, wfx23@nudt.edu.cn

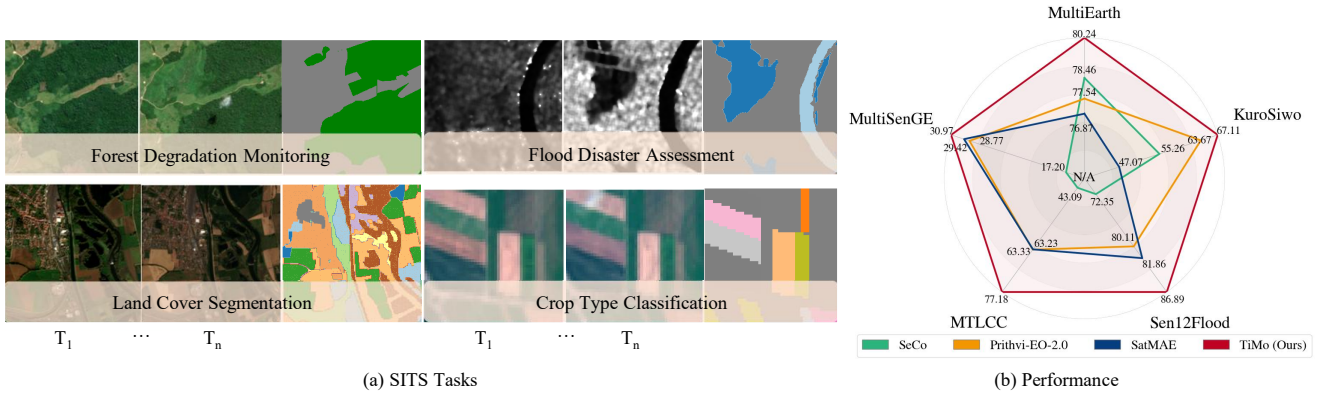


Figure 1. TiMo surpasses existing spatiotemporal RSFMs, delivering superior performance across diverse SITS tasks, including forest monitoring, disaster assessment, ground-object recognition, and agricultural identification.

Abstract

Satellite image time series (SITS) provide continuous observations of the Earth’s surface, making them essential for applications such as environmental management and disaster assessment. However, existing spatiotemporal foundation models rely on plain vision transformers, which encode entire temporal sequences without explicitly capturing multiscale spatiotemporal relationships between land objects. This limitation hinders their effectiveness in downstream tasks. To overcome this challenge, we propose TiMo, a novel hierarchical vision transformer foundation model tailored for SITS analysis. At its core, we introduce a spatiotemporal gyroscope attention mechanism that dynamically captures evolving multiscale patterns across both time and space. For pre-training, we curate MillionST, a large-scale dataset of one million images from 100,000 geographic locations, each captured across 10 temporal phases over five years, encompassing diverse geospatial changes and seasonal variations. Leveraging this dataset, we adapt

masked image modeling to pre-train TiMo, enabling it to effectively learn and encode generalizable spatiotemporal representations. Extensive experiments across multiple spatiotemporal tasks—including deforestation monitoring, land cover segmentation, crop type classification, and flood detection—demonstrate TiMo’s superiority over state-of-the-art methods. Code, model, and dataset will be released at [TiMo](#).

1. Introduction

Recent advances in remote sensing and satellite missions have enabled higher-resolution satellite imagery with more frequent coverage, offering unprecedented opportunities to monitor Earth’s surface [1]. These repeated observations yield multi-temporal views, establishing satellite image time series (SITS) as vital tools for applications such as land cover monitoring, vegetation identification, and change detection [2–5]. Automated extraction of spatiotemporal patterns from SITS is essential to derive actionable insights and advance these critical domains.

*: Equal contribution; †: Corresponding author.

Deep learning (DL) has become a cornerstone of satellite image time series (SITS) modeling due to its automated feature learning capabilities. However, conventional DL methods depend on large labeled datasets, which are frequently scarce [6] due to annotation costs. Self-supervised learning (SSL) has gained traction as an efficient alternative, constructing supervisory signals from unlabeled data via pretext tasks that exploit inherent spatiotemporal structures [7–9]. This paradigm minimizes dependence on manual annotations while maintaining strong downstream task performance [6].

Recently, SSL has driven the development of foundation models (FMs) in remote sensing, which learn universal representations from large-scale unlabeled data. FMs generalize across diverse downstream applications—even with limited labels—by pretraining on expansive datasets [10]. For SITS, masked image modeling frameworks such as Masked Autoencoders (MAE) [11, 12] and VideoMAE [13] have been adapted [6, 14], leveraging vision transformers (ViTs) [15] to compute spatiotemporal self-attention across flattened token sequences. However, these models often overlook the inherent spatial alignment of SITS imagery and struggle to capture multiscale spatiotemporal patterns essential for land object analysis. On the other side, pre-training data also shapes FM performance. While datasets such as FMoW [16] and Sentinel-2 temporal composites [8] have been used for SITS pretraining [6, 8, 9], their limited temporal cardinality (typically ≤ 5 timestamps) constrains the diversity of spatiotemporal change representation, reducing their effectiveness on downstream tasks.

To address these limitations, we propose TiMo, a novel hierarchical vision transformer foundation model for SITS analysis. TiMo introduces a spatiotemporal gyroscope attention (STGA) mechanism, which leverages the spatial alignment of SITS to capture correlations between tokens with identical temporal or spatial positions, where the term “gyroscope” refers to the geometric structure of attention regions. For pretraining, we curate MillionST, a million-scale dataset comprising 1 million Sentinel-2 images sampled from 100,000 geographic locations, each captured across 10 temporal phases over five years. This dataset encompasses diverse geospatial changes and seasonal variations. Inspired by [17], we pretrain TiMo via spatiotemporal masked image modeling. To maximize the potential of FMs in capturing complex patterns within SITS [18–20], we scale TiMo to 300M parameters, enhancing its capacity to learn generalizable spatiotemporal representations. Extensive experiments across multiple spatiotemporal tasks, including deforestation monitoring, land cover segmentation, and flood detection, show that TiMo outperforms existing state-of-the-art methods, as shown in Figure 1.

Our contributions are summarized as follows:

- (1) We introduce TiMo, a novel spatiotemporal founda-

tion model built on a hierarchical vision transformer tailored for unified SITS analysis. TiMo is pretrained via masked image modeling to learn multiscale spatiotemporal representations for diverse downstream tasks.

- (2) We design a spatiotemporal gyroscope attention mechanism in TiMo, which leverages SITS’ inherent spatial alignment to capture spatiotemporal relationships across satellite imagery.

- (3) We curate MillionST, a large-scale pretraining dataset with 1 million samples from 100,000 geographic regions, each captured across 10 temporal phases over 5 years. This diversity enhances TiMo’s ability to learn generalizable spatiotemporal features.

- (4) Experiments demonstrate TiMo’s superiority over existing SITS foundation models across tasks like land-cover classification, disaster assessment, and agricultural mapping. Additionally, TiMo exhibits scalability, sample efficiency, and interpretability, further validating its utility.

2. Related work

2.1. Remote Sensing Foundation Models

Inspired by the success of FMs in the natural image domain [21–25], researchers have increasingly explored the development of remote sensing FMs (RSFMs), with vision transformers emerging as the predominant model architecture. For instance, some studies have leveraged large-scale labeled dataset to pre-train RSFMs [26, 27] using advanced vision transformer architectures [28, 29]. In addition, domain-specific features, such as spectral information [19, 30, 31], geographic locations [32–34], geometric and physical attributes of land objects [35–39], and modal properties [40–44], have been extensively explored to refine pre-training strategies and model design [45–50]. However, these pioneering RSFMs mainly focused on learning generalizable feature representations by pre-training on single-temporal RS images. As a result, they are better suited for time-independent downstream tasks such as scene classification, semantic segmentation, and object detection.

2.2. SITS Foundation Models

Unlike those FMs that focus solely on single-temporal imagery, recent studies have begun incorporating temporal information from revisited satellite images [6, 8–10, 14, 51–54]. For example, some approaches leveraged contrastive learning (CL) by constructing positive and negative pairs from temporal remote sensing images [8, 9, 51]. To simplify the complex process of generating negative pairs, SkySense [10] employs a teacher-student framework to apply CL to augmented temporal inputs. While these CL-based FMs rely heavily on data augmentation [11], alternative approaches [6, 14] adopt masked image modeling [12] for pre-training. These SITS FMs typically extend the MAE [11]

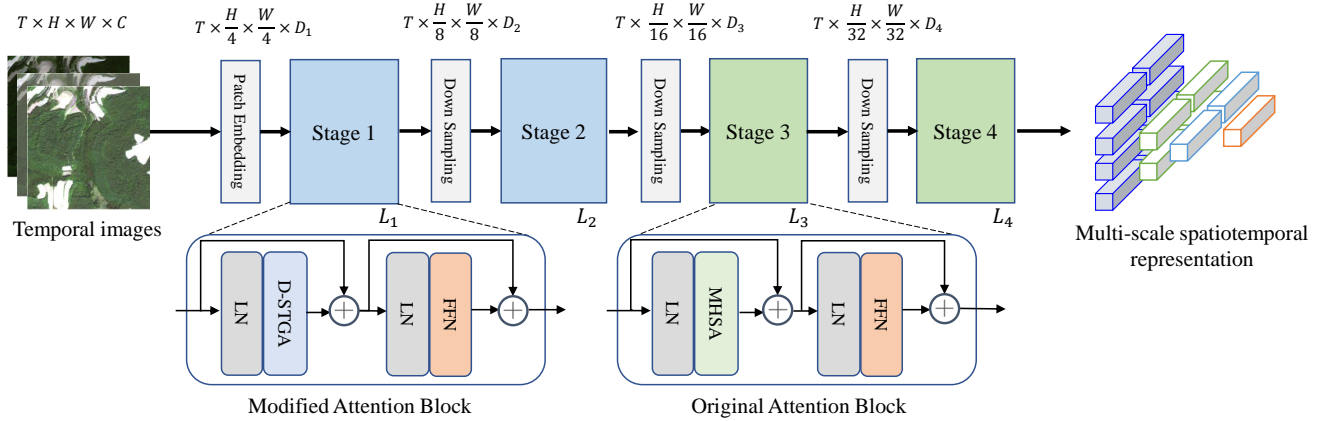


Figure 2. **Overview of TiMo’s Architecture:** TiMo follows a four-stage hierarchical design, where each stage comprises attention layers and feed-forward networks following a downsampling layer. To enhance spatio-temporal representation learning, TiMo replaces Multi-head Self-Attention (MHSA) with a novel and efficient Differential Spatiotemporal Gyroscope Attention (D-STGA) in the first two stages.

and plain ViT [15] architectures to encode patches across both spatial and temporal dimensions. However, they often overlook the underlying spatiotemporal relationships within SITS data. In contrast, we introduce TiMo, a novel hierarchical vision transformer designed for multiscale spatiotemporal representation learning. TiMo builds upon the STGA module, which simultaneously captures contextual relationships within the same spatial or temporal dimensions across all timestamps. Given the progressive downsampling nature of its hierarchical structure, we adopt the self-supervised learning paradigm of Hiera [17] and apply masked image modeling to multi-temporal images for pre-training.

2.3. Pre-training Data for SITS Foundation Models

Satellite revisit cycles generate vast amounts of global-scale SITS, significantly enhancing the SSL pre-training of spatiotemporal FMs. The Sentinel-2 mission¹, with its twin satellites revisiting the same location every five days, provides high-temporal-resolution multispectral imagery. SeCo [8] curated Sentinel-2 SITS from circular regions with a 100 km radius across 10,000 cities worldwide, with each location contributing five images representing different seasons over a year. CACo [51] further refined this sampling approach by reducing spatial distances relative to city centers. The widely recognized FMoW dataset [16] has played a crucial role in advancing RSFMs for SITS [6, 9]. It consists of high-resolution satellite image time series designed to classify images into 62 distinct categories. However, many existing pre-training datasets suffer from limited temporal coverage. For instance, approximately 70% of FMoW samples contain time series with five or fewer timestamps [9], which may constrain the effectiveness of pre-trained RSFMs in downstream tasks. To address

these limitations, this paper introduces MillionST, a large-scale SITS dataset specifically designed for spatiotemporal pre-training. Collected from 100,000 geographic locations, MillionST comprises approximately one million samples, with each location contributing images from 10 timestamps spanning five years. This extensive spatiotemporal diversity enhances RSFMs pre-training, *e.g.*, our TiMo series.

3. Method

3.1. Overall Architecture of TiMo

To more effectively capture the spatiotemporal variations—such as differing scales and shapes—inherent in SITS data, we employ a hierarchical vision transformer design in our TiMo model, which is capable of generating multiscale pyramid features. As shown in Figure 2, TiMo consists of four stages separated by downsampling layers. In the early stages, we replace the standard multi-head self-attention (MHSA) with a novel Spatiotemporal Group Attention to enhance the extraction of spatiotemporal features. However, in the deeper stages, where progressive downsampling results in more abstract, low-resolution features, the STGA module may be less effective, so we retain the original attention mechanism. The optimal configuration for each stage will be validated via empirical study.

Patch Embedding and Downsampling Layers Given a group of input SITS with dimensions $T \times H \times W \times C$, where T represents the temporal length and H, W , and C denote the height, width, and number of channels, respectively, we begin by applying patch embedding to each image across different timestamps. The patch embedding layer consists of two convolutional operations: a 7×7 convolution with a stride of 2 and padding of 3, followed by a 2×2 convolution with a stride of 2. As a result, each image is

¹<https://sentiwiki.copernicus.eu/web/s2-mission>

Table 1. Configuration for TiMo variants. S_i denotes the i th stage.

Version	Dims (D)				Heads (h)				Depths (L)			
	S_1	S_2	S_3	S_4	S_1	S_2	S_3	S_4	S_1	S_2	S_3	S_4
TiMo-Base	128	256	512	1024	4	8	16	32	2	2	18	2
TiMo-Large	384	768	960	1536	6	12	24	48	2	2	18	2
TiMo-Huge	512	1024	1280	2048	8	16	32	64	3	3	22	3

transformed into $\frac{H}{4} \times \frac{W}{4}$ tokens with a feature dimension of D_1 . For simplicity, we obtain the spatiotemporal positional encoding of each token by concatenating its temporal and spatial positional encodings—both represented as 1D vectors—along the channel dimension. The temporal positional embedding follows the approach introduced by [55], while the spatial positional encoding is adopted from [11]. To facilitate hierarchical processing, the model consists of four stages connected by downsampling layers. Each downsampling layer applies a 3×3 convolution with a stride of 2 and padding of 1, ensuring a downsampling ratio of 2 between consecutive stages. Additionally, the output feature maps of each downsampling layer have an expanded channel dimension, doubling in size at each stage.

Transformer Blocks TiMo employs two types of attention blocks: one with the proposed attention module and the other using the original MHSA. Aside from the attention mechanism, both blocks share the same structure, consisting of a layer normalization (LN) layer, a feed-forward network (FFN), and residual connections. The details of these components follow that of the Swin Transformer [28].

Multi-head Self-Attention MHSA is a fundamental component of vision transformers, comprising several parallel self-attention mechanisms—each commonly referred to as a head. It is defined by:

$$SA(\mathbf{F}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{D'}}\right)\mathbf{V}, \quad (1)$$

where $\mathbf{Q} \in \mathbb{R}^{N \times D'}$, $\mathbf{K} \in \mathbb{R}^{N \times D'}$, $\mathbf{V} \in \mathbb{R}^{N \times D'}$ represent the query, key, and value matrices, respectively, which are derived from $\mathbf{F} \in \mathbb{R}^{N \times D}$ by three linear layers. Here, N denotes the total number of tokens across all temporal dimensions, and D is the channel dimension of the tokens. The head dimension D' is defined such that $D = hD'$, with h being the number of heads. Specifically, for the i th stage, N is calculated as: $N = \frac{H}{2^{i+1}} \cdot \frac{W}{2^{i+1}} \cdot T$, $i \in \{1, 2, 3, 4\}$. In the TiMo architecture, \mathbf{F} is the feature output from the first LN layer in MHSA blocks. In subsequent sections, we omit different heads for convenience.

Parameter Scaling Following [28], we design a series of TiMo with various sizes, *i.e.*, TiMo-Base, TiMo-Large and TiMo-Huge. These versions differ in the dimensions of features, number of heads, and the depths in each stage, as shown in Table 1. Other configurations between these ver-

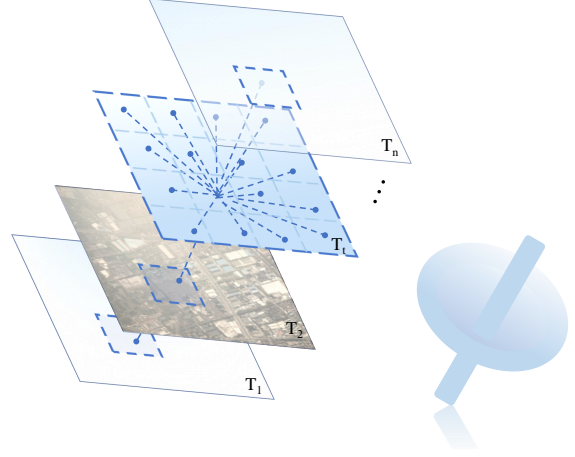


Figure 3. Illustration of Spatiotemporal Gyroscopic Attention.

sions are always the same. For example, the MLP expansion ratio is set to 4.

3.2. STGA: Spatiotemporal Gyroscopic Attention

In this section, we introduce the STGA mechanism. Rather than calculating similarities between all tokens as in traditional MHSA used in standard vision transformers, STGA computes correlations only among tokens that share the same spatial or temporal positions, thereby enhancing the model’s ability to capture spatiotemporal dynamics.

Let $\mathbf{Q} \in \mathbb{R}^{T \times N_p \times N_p \times D'}$, $\mathbf{K} \in \mathbb{R}^{T \times N_p \times N_p \times D'}$, and $\mathbf{V} \in \mathbb{R}^{T \times N_p \times N_p \times D'}$, where $N_p = \sqrt{N/T}$ denotes the number of tokens along the height or width of an image (assuming $H = W$). Consider a query vector $q_p \in \mathbb{R}^{1 \times D'}$ from \mathbf{Q} at position $p(x_0, y_0, z_0)$, where x and y lie in the plane \mathcal{P} and z represents the time axis \mathcal{T} . Define the set of positions $\mathbf{u} = \{(x, y, z) \mid z = z_0 \vee (x = x_0 \wedge y = y_0)\}$ within the three-dimensional spatiotemporal space $\mathcal{S} = \mathcal{P} \times \mathcal{T}$ (with \times denoting the Cartesian product), ensuring that each element in \mathbf{u} shares either the same temporal or spatial coordinate as p . From \mathbf{K} , we then extract a vector set $\mathbf{K}_u \in \mathbb{R}^{(T+N_p \times N_p - 1) \times D'}$ corresponding to these positions. Finally, the attention scores for each spatiotemporal location are computed as follows:

$$\mathbf{A}_p = q_p \mathbf{K}_u^\top, \quad (2)$$

where $\mathbf{A}_p \in \mathbf{A}$ denotes the affinity between q_p and \mathbf{K}_u . Similarly, the overall attention matrix is given by $\mathbf{A} \in \mathbb{R}^{(T \times N_p \times N_p) \times (T + N_p \times N_p - 1)}$. We then apply a softmax operation along the last dimension of \mathbf{A} to compute the attention map \mathbf{A}' , where each $\mathbf{A}'_p \in \mathbb{R}^{1 \times (T + N_p \times N_p - 1)}$ determines the contribution of each location to the representation at position p .

Following a similar procedure, we extract a set of vectors $\mathbf{V}_u \in \mathbb{R}^{(T + N_p \times N_p - 1) \times D'}$ from \mathbf{V} . The output for position

p , denoted as $\mathbf{O}_p \in \mathbb{R}^{1 \times D'}$, is then computed according to:

$$\mathbf{O}_p = \mathbf{A}'_p \mathbf{V}_u. \quad (3)$$

The final feature $\mathbf{F}_p \in \mathbb{R}^{1 \times D}$ is obtained by concatenating the outputs \mathbf{O}_p from different heads and passing the result through a linear layer. Consequently, for all spatiotemporal positions, we obtain $\mathbf{F} \in \mathbb{R}^{T \times N_p \times N_p \times D}$.

This process reveals that the shape of the core variable \mathbf{u} resembles an extreme gyroscopic form in the spatiotemporal domain, which is the inspiration behind the name ‘‘STGA’’, as illustrated in Fig. 3.

3.3. D-STGA: Differential Spatiotemporal Gyro-scope Attention

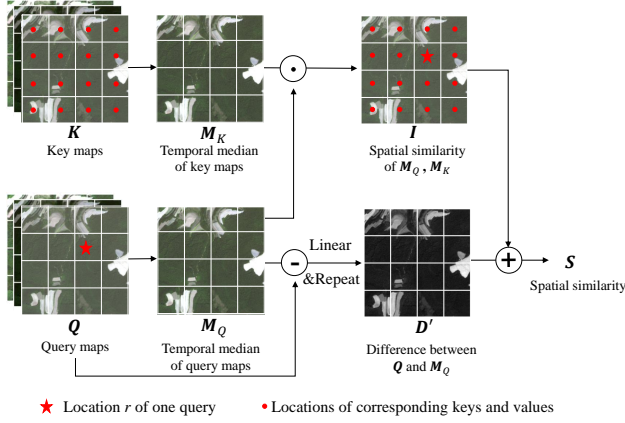


Figure 4. Diagram of calculating spatial similarity in D-STGA.

Although STGA can model spatiotemporal variations, its computational cost increases rapidly with the temporal length T , particularly when computing the attention matrix. With a fixed input size, the theoretical computational complexity of obtaining \mathbf{A} in STGA is $O((T \times N_p \times N_p) \times (T + N_p \times N_p - 1)) \approx O(T^2)$, which motivates us to develop a more lightweight implementation.

We observe that multi-temporal observations contain both variant and invariant regions; that is, the spatial status at any time can be inferred from temporal relationships. This propagation can be approximated by using temporal differences, leading to a novel differential spatiotemporal gyroscope attention (D-STGA).

Given $\mathbf{Q} \in \mathbb{R}^{T \times N_p \times N_p \times D'}$, $\mathbf{K} \in \mathbb{R}^{T \times N_p \times N_p \times D'}$, and $\mathbf{V} \in \mathbb{R}^{T \times N_p \times N_p \times D'}$, we first extract the median temporal features from \mathbf{Q} and \mathbf{K} (using the `pytorch.median` function) to mitigate the effect of abnormal observations (e.g., clouds in SITS). This yields $\mathbf{M}_Q \in \mathbb{R}^{N_p \times N_p \times D'}$ and $\mathbf{M}_K \in \mathbb{R}^{N_p \times N_p \times D'}$. We then compute their similarity:

$$\mathbf{I} = \mathbf{M}_Q \mathbf{M}_K^\top. \quad (4)$$

Here, \mathbf{I} represents the temporal-invariant spatial similarity. To recover spatial information for all timestamps, we compute the difference sequence $\{\mathbf{D}_t \mid t = 1, 2, \dots, T\}$, where $\mathbf{D}_t = \mathbf{Q}_t - \mathbf{M}_Q$, with $\mathbf{Q}_t \in \mathbb{R}^{N_p \times N_p \times D'}$ denoting the query at time t . Next, we apply a linear projection (comprising a Linear layer, BatchNorm, and ReLU) to reduce the channel dimension of \mathbf{D}_t to 1. Repeating the result $N_p \times N_p$ times produces $\mathbf{D}'_t \in \mathbb{R}^{(N_p \times N_p) \times (N_p \times N_p)}$, which captures the spatial discrepancy.

The spatial status for each timestamp is then obtained by adding \mathbf{D}'_t to \mathbf{I} : $\mathbf{S}_{:,t} = \mathbf{D}'_t + \mathbf{I}$, $t = 1, 2, \dots, T$. Thus, the complete spatial status set is $\mathbf{S} = \{\mathbf{S}_{:,1}, \dots, \mathbf{S}_{:,T}\} \in \mathbb{R}^{T \times (N_p \times N_p) \times (N_p \times N_p)}$. In this way, it efficiently calculates the spatial similarity across all timestamps. The above process is illustrated in Figure 4.

In addition to spatial status, temporal relationships must also be estimated. For a spatial location r at time t , we compute the temporal correlation by multiplying its query token $\mathbf{Q}_{r,t} \in \mathbb{R}^{1 \times D'}$ with the key tokens $\mathbf{K}_o \in \mathbb{R}^{(T-1) \times D'}$ from the same spatial position at other timestamps:

$$\mathbf{T}_{r,t} = \mathbf{Q}_{r,t} \mathbf{K}_o^\top, \quad (5)$$

where $\mathbf{T}_{r,t} \in \mathbb{R}^{1 \times (T-1)}$. We then concatenate the spatial affinity map $\mathbf{S}_{r,t} \in \mathbb{R}^{1 \times (N_p \times N_p)}$ with the temporal correlation $\mathbf{T}_{r,t}$ to obtain the spatiotemporal similarity map $\mathbf{A}_{r,t} \in \mathbb{R}^{1 \times (T + N_p \times N_p - 1)}$. The attention score is then derived as described in Sec. 3.2, with subsequent steps—such as applying the softmax, matrix multiplication, and linear projection—following the same procedure as in STGA.

Compared to STGA, D-STGA circumvents the vector inner product, reducing the computational complexity of obtaining \mathbf{A} to $O(T)$.

3.4. Pre-training and Fine-tuning

Pre-training Masked image modeling methods, such as MAE [11], are well-regarded for their performance and efficiency and have been widely adopted in pre-training remote sensing and SITS foundation models [6, 35, 38, 39, 45] that use the vanilla ViT [15]. However, TiMo employs a hierarchical vision transformer network with convolutional downsampling layers, which prevents the direct application of MAE for pre-training. To address this, we follow Hiera [17] to adopt an MAE-like approach specifically designed for hierarchical architectures. Although it was originally used to pre-train MViTv2 [56] on natural images and video, we have successfully applied it to our architecture without modifying the attention mechanism (i.e., we use MHSA in all blocks during pre-training). Notably, while Hiera uses a mask unit of $2 \times 32 \times 32$ pixels for video data to reduce redundancy between frames, SITS data are sparser with longer temporal intervals; hence, we adopt a mask unit of $1 \times 32 \times 32$ pixels. In this way, we successfully pre-trained TiMo, a hierarchical vision transformer foundation model

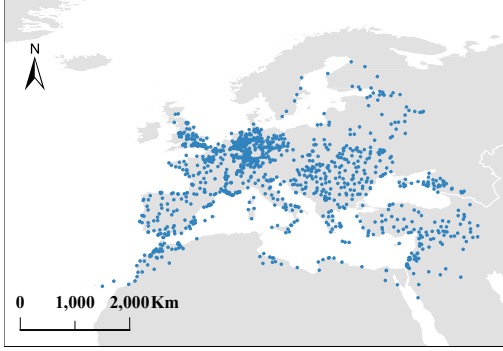


Figure 5. City distribution for MillionST data sampling.

for SITS. Additional details on pre-training are provided in the Supplementary Material.

Fine-tuning During fine-tuning on downstream tasks, we reuse the attention layer weights as in [35, 57] and replace the standard MHSA in the early stages with the proposed STGA (or D-STGA) to better model spatiotemporal contexts. For the parameters introduced by the linear projection used to compute the spatial discrepancy in D-STGA—which lack pre-trained weights—we initialize them randomly. The output features from all four stages of TiMo are used for segmentation tasks, while the feature from the final stage is used for classification.

4. Experiments

4.1. Pre-training Dataset: MillionST

We collected Sentinel-2 multi-spectral SITS data from 100,000 locations spanning 1,317 cities—primarily in Europe, North Africa, and West Asia—all of which rank among the 10,000 most populous cities globally, as shown in Fig. 5. Following the approach of [8], coordinates were sampled from a Gaussian distribution centered on each city with a 50 km standard deviation. For every location, we compiled a temporal sequence of images starting on June 30, 2017, with captures occurring at six-month intervals. Consequently, the dataset comprises SITS data from ten distinct timestamps covering the period from 2017 to 2022. Additional details regarding the sampling cities are provided in the Supplementary Material.

4.2. Deforestation Monitoring

The MultiEarth dataset [58] is a multi-modal RS dataset designed to monitor deforestation in the Amazon rainforest. It utilizes Sentinel-2 RGB temporal images, with the training split comprising 1,565 images that include pixel-level annotations distinguishing forest from deforested areas. The dataset is divided into training, validation, and testing sets in a 1:1:8 ratio. Each image is 256×256 pixels in size, and three timestamps are randomly selected for both train-

Table 2. Comparison of different models on MultiEarth dataset, †: pre-training on MillionST.

Method	Backbone	#Parameter	mIoU
SeCo[8]	ResNet-50	26M	0.7846
CACo[51]	ResNet-50	26M	0.7692
GASSL[9]	ResNet-50	26M	0.7638
Prithvi-EO-2.0[14]	ViT-Large	300M	0.7754
SatMAE[6]	ViT-Large	307M	0.7687
SatMAE[6]†	ViT-Large	307M	0.7828
TiMo-Base	TiMo-Base	91M	0.7860
TiMo-Large	TiMo-Large	298M	0.8024

ing and testing. Model performance is evaluated using the mean intersection over union (mIoU) metric. For further details on the implementation, as well as those in the following sections, please refer to the Supplementary Material.

We compare TiMo against recent SOTA multi-temporal RSFMs, including SeCo [8], CACo [51], GASSL [9], Prithvi-EO-2.0 [14], and SatMAE [6]. To assess the impact of the established pre-training dataset, we also pre-trained SatMAE on our MillionST, which is denoted as SATMAE†. The quantitative results in Table 2 indicate that: (1) MillionST is more effective for pre-training spatiotemporal foundation models—likely due to its richer set of timestamps compared to the FMoW dataset [16] originally used for SatMAE, and (2) TiMo-Base outperforms the competing methods, even with relatively fewer parameters than Prithvi-EO-2.0 and SatMAE. Moreover, TiMo’s performance scales favorably with increased model size.

4.3. Multi-class Land Cover Segmentation

We perform multi-class land cover segmentation using the MultiSenGE dataset [2], which generates a single land cover map from multi-temporal images. The dataset comprises 8,157 time series samples collected from 14 distinct Sentinel-2 tiles, each with 10 bands and a temporal length of 12. Each image has a spatial resolution of 256×256 . For our experiments, we allocate 8 tiles for training, 3 for validation, and the remaining 3 for testing. In all experiments, we randomly select 3 temporal data points from each image time series to train all comparison methods.

The experimental results for multi-class land cover segmentation are presented in Table 3. As shown, SeCo, CACo, and GASSL achieve relatively lower accuracy. Although both Prithvi-EO-2.0 and SATMAE perform well, our proposed TiMo-Base and TiMo-Large models outperform them, underscoring the robust capabilities of our approach. Moreover, as the model’s parameter count increases, its performance continues to improve.

Table 3. Comparison of different models on MultiSenGE dataset.

Method	Backbone	#Parameter	mIoU
SeCo[8]	ResNet-50	26M	0.1720
CACo[51]	ResNet-50	26M	0.1789
GASSL[9]	ResNet-50	26M	0.1712
Prithvi-EO-2.0[14]	ViT-Large	300M	0.2877
SatMAE[6]	ViT-Large	307M	0.2942
TiMo-Base	TiMo-Base	91M	0.2977
TiMo-Large	TiMo-Large	298M	0.3097

Table 4. Comparison of different models on MTLCC dataset.

Method	Backbone	#Parameter	mIoU
TSViT[1]	TSViT	2M	0.7617
SeCo[8]	ResNet-50	26M	0.4309
CACo[51]	ResNet-50	26M	0.3766
GASSL[9]	ResNet-50	26M	0.3798
Prithvi-EO-2.0[14]	ViT-Large	300M	0.6333
SatMAE[6]	ViT-Large	307M	0.6323
TiMo-Base	TiMo-Base	91M	0.7655
TiMo-Large	TiMo-Large	298M	0.7718

4.4. Crop Type Classification

To evaluate the models’ ability to handle longer time series, we conducted an experiment using the MTLCC dataset [59], which is designed for crop type classification. This dataset comprises Sentinel-2 image time series acquired in 2016. Each image has a spatial resolution of 48×48 pixels and contains 13 channels. Every sample consists of a sequence of 30 time steps paired with a single segmentation-labeled image. We adopt the original division of the training, validation, and test sets.

In addition to the previously discussed SITS FMs, we also compare a fully supervised vision transformer network, TSViT [1], specifically developed for crop classification using longer time series. Table 4 presents the experimental results, which show that TSViT achieves higher accuracy than FMs such as SeCo, CACo, and GASSL. This improvement may be attributed to its attention mechanism that effectively encodes temporal information, whereas SeCo, CACo, and GASSL rely solely on a ResNet-50 encoder for extracting spatial context. With pre-training on MillionST, our TiMo significantly outperforms these models by achieving the highest accuracy, highlighting its strong capability to model long time series RS data.

4.5. Flood Disaster Assessment

Finally, we focus on flood detection. We begin by evaluating the methods using the Sen12Flood dataset [60], which was designed for image classification. Sen12Flood includes both synthetic aperture radar (SAR) and multispectral images intended for multi-temporal flood detection, where images are classified as either *flood* or *non-flood*. The dataset comprises 336 sequences from regions in West and South-

Table 5. Comparison of different models on Sen12Flood dataset.

Method	#Parameter	Acc(%)
SeCo[8]	26M	72.35
CACo[51]	26M	75.52
GASSL[9]	26M	74.97
SatMAE[6]	300M	81.86
Prithvi-EO-2.0[14]	307M	80.11
TiMo-Base	91M	86.89
TiMo-Large	298M	86.78

East Africa, the Middle East, and Australia. For our experiments, we use 30 Sentinel-2 optical (RGB) image sequences for training and the remainder for testing. On average, each sequence contains 14 temporal phases; for classification purposes, we randomly select three phases and use the corresponding RGB channels.

Table 5 presents the results on the Sen12Flood dataset. Among all the methods compared, SatMAE demonstrates strong performance by slightly outperforming Prithvi-EO-2.0, while our proposed TiMo achieves the highest accuracy. Nevertheless, we observe that TiMo-Large performs only on par with the base version, possibly due to overfitting caused by the task’s simplicity.

We then refine our analysis to locate the precise pixel-level regions affected by floods by framing the problem as a segmentation task. To further assess cross-domain performance, we utilize the KuroSiWo dataset [61], which consists of SAR images captured by the Sentinel-1 satellite across three temporal phases, with two channels per image. Each sample in KuroSiWo includes two pre-flood images and one post-flood image. The corresponding label is a mono-temporal, pixel-level annotation categorizing each pixel into one of three classes: *No water*, *Permanent waters*, and *Floods*. In our experiments, we use the first fold of the dataset, comprising 7,103 time-series image samples, and split it into training, validation, and testing sets in a 3:1:1 ratio.

The experimental results are presented in Table 6. As shown, TiMo-Base achieves competitive performance compared to other RSFMs. Notably, TiMo-Large significantly improves accuracy, achieving an impressive mIoU of 0.6711, despite the challenging transfer from optical to SAR images. These findings highlight the practicality and effectiveness of the proposed SITS FMs—TiMo.

4.6. Ablation Study

We conduct an ablation study to evaluate the impact of the model’s structure and pre-training strategy. The results, presented in Table 7, examine the effect of different attention mechanisms. We compare six configurations, each defined by a unique combination of attention mechanisms. Our findings show that using MHSA in all blocks leads to high computational complexity and excessive GPU memory us-

Table 6. Comparison of different models on KuroSiwo dataset.

Method	Backbone	#Parameter	mIoU
SeCo[8]	ResNet-50	26M	0.5526
CACo[51]	ResNet-50	26M	0.5223
GASSL[9]	ResNet-50	26M	0.5742
Prithvi-EO-2.0[14]	ViT-Large	300M	0.6367
SatMAE[6]	ViT-Large	307M	0.4707
TiMo-Base	TiMo-Base	91M	0.6680
TiMo-Large	TiMo-Large	298M	0.6711

Table 7. Ablation study on the MultiEarth dataset, where “X” in “X-X-X-X” denotes the attention mechanism used at each stage. S: STGA, D: D-STGA, M: MHSA, and OOM: Out of Memory. Memory was measured on an NVIDIA RTX 4090 GPU.

Method	mIoU	FLOPs (G)	GPU Memory(M)
M-M-M-M	-	642.15	OOM
S-M-M-M	0.7827	436.87	15,284
D-M-M-M	0.7836	402.52	15,332
D-D-M-M	0.7860	372.60	14,222
D-D-D-M	0.7827	339.10	13,129
D-D-D-D	0.7628	333.64	13,122

Table 8. Ablation study on the sampling strategy of pre-training, where the model is fine-tuning on the MultiEarth dataset.

Method	Sampling	mIoU
TiMo-Base	Fixed	0.7839
TiMo-Base	Random	0.7860

age, causing the model to run out of memory. By replacing MHSA with the proposed STGA or D-STGA, we consistently achieve competitive performance compared to existing FMs (as shown in Table 2), validating that the proposed attention mechanisms are more effective at capturing spatiotemporal dynamics. Furthermore, we observe a significant reduction in computational complexity (FLOPs) with our attention mechanisms, particularly compared to the default MHSA used in vanilla vision transformers (*i.e.*, M-M-M-M), demonstrating their higher efficiency. When comparing the second and third configurations, we find that D-M-M-M yields better accuracy, emphasizing the importance of modeling spatiotemporal relationships through the differential design. Additionally, FLOPs are further reduced, proving that this approach saves computational resources by minimizing the number of vector inner products. Finally, as we increase the number of blocks using D-STGA, we identify that the D-D-M-M configuration strikes the best balance between accuracy and efficiency.

Since MillionST consists of 10 time phases, but our limited GPU memory only supports simultaneous processing of images with a temporal length of 3, we also compare different temporal sampling strategies during pre-training. Specifically, we evaluate the effects of using fixed versus

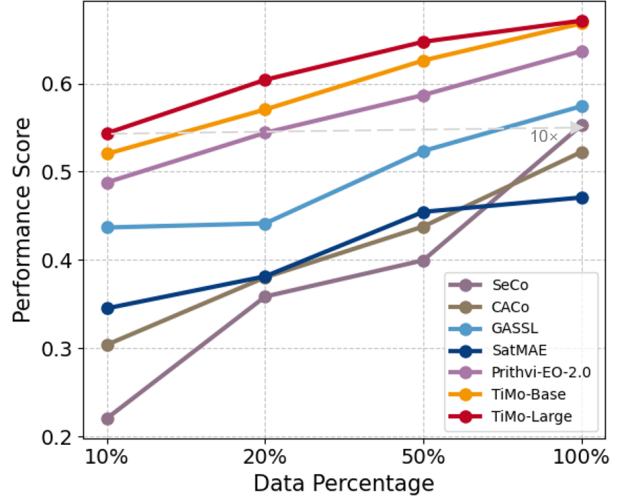


Figure 6. Experimental results of different SITS FMs trained with varying sample sizes on the KuroSiwo dataset.

random temporal sampling. In the random sampling approach, for each sample, 3 temporal images are randomly selected from the time series in every iteration. The results in Table 8 show that random temporal sampling during pre-training leads to better performance compared to fixed sampling. This is likely because random sampling provides more diverse data, helping the model capture dynamic temporal variations in SITS, which in turn improves generalization and performance on downstream tasks.

4.7. Data Efficiency of SITS FMs

Data efficiency is a critical property of FMs [29, 35, 62, 63], especially when fine-tuning on downstream tasks with limited labeled data. To investigate this aspect in the context of TiMo, we conduct fine-tuning experiments on the KuroSiwo dataset [61], utilizing different proportions of the original training data—specifically 10%, 20%, and 50%. We then evaluate the resulting models on the original test set. The experimental results, shown in Figure 6, demonstrate that our proposed methods consistently outperform other models across all training sample settings. Notably, TiMo-Large, trained on 50% of the samples, outperforms Prithvi-EO-2.0, trained on the full dataset. With just 10% of the data, TiMo-Large almost matches GASSL and SeCo and outperforms CaCo and SatMAE (all trained on the full dataset), highlighting its exceptional data efficiency.

4.8. Scalability of TiMo

To evaluate TiMo’s scalability, we conduct fine-tuning experiments using TiMo’s variants of different model sizes (91M, 298M, and 675M for TiMo-Base, TiMo-Large, and TiMo-Huge respectively) on two challenging datasets. As

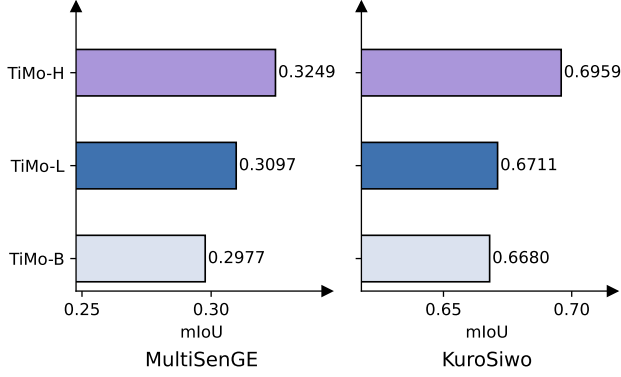


Figure 7. Fine-tuning performance of TiMo’s variants on MultiSenGE and KuroSiwo datasets. TiMo-B, TiMo-L, and TiMo-H are short for TiMo-Base, TiMo-Large, and TiMo-Huge, respectively.

shown in Figure 7, larger TiMo backbones lead to better performance. Specifically, TiMo-Huge demonstrates a 9.14% higher mIoU than TiMo-Base on MultiEarth dataset. Additionally, on the KuroSiwo dataset, TiMo-Huge exhibits a 4.18% mIoU improvement over TiMo-Base. These results underscore TiMo’s scalability across diverse tasks.

5. Conclusion

This study introduces TiMo, a novel hierarchical vision transformer foundation model explicitly designed for satellite image time series analysis. TiMo incorporates a novel spatiotemporal gyroscope attention (STGA) mechanism that effectively captures complex spatiotemporal relationships inherent in aligned SITS data, alongside an efficient D-STGA variant that leverages temporal differences to accelerate computation. To support effective pre-training, we curate MillionST, a large-scale dataset spanning 100,000 geographic locations with 10 temporal phases over five years, capturing diverse geospatial and seasonal variations. By adapting masked image modeling to this spatiotemporal context, TiMo learns generalizable representations for a wide range of downstream tasks. Extensive experiments across deforestation monitoring, land cover segmentation, and flood detection demonstrate TiMo’s consistent superiority over existing models. Ablation studies further confirm the effectiveness of its design choices, highlighting its strengths in data efficiency and model scalability.

References

- [1] Michail Tarasiou, Erik Chavez, and Stefanos Zafeiriou. Vits for sits: Vision transformers for satellite image time series. In *CVPR*, pages 10418–10428, 2023. 1, 7
- [2] Romain Wenger, Anne Puissant, Jonathan Weber, Lhasane Idoumghar, and Germain Forestier. Multisenge: A multimodal and multitemporal benchmark dataset for land use/land cover remote sensing applications. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 3:635–640, 2022. 1, 6, 12
- [3] Jialu Li and Chen Wu. Using difference features effectively: A multi-task network for exploring change areas and change moments in time series remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 218:487–505, 2024.
- [4] Vivien Sainte Fare Garnot and Loic Landrieu. Panoptic segmentation of satellite image time series with convolutional temporal attention networks. In *ICCV*, pages 4872–4881, 2021.
- [5] Xiaolei Qin, Haonan Guo, Xin Su, Zhenghui Zhao, Di Wang, and Liangpei Zhang. Spatiotemporal masked pre-training for advancing crop mapping on satellite image time series with limited labels. *International Journal of Applied Earth Observation and Geoinformation*, 137:104426, 2025. 1
- [6] Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *NeurIPS*, 35:197–211, 2022. 2, 3, 5, 6, 7, 8, 12
- [7] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020. 2
- [8] Oscar Manas, Alexandre Lacoste, Xavier Giró-i Nieto, David Vazquez, and Pau Rodriguez. Seasonal contrast: Unsupervised pre-training from uncured remote sensing data. In *ICCV*, pages 9414–9423, 2021. 2, 3, 6, 7, 8
- [9] Kumar Ayush, Burak Uzkent, Chenlin Meng, Kumar Tanmay, Marshall Burke, David Lobell, and Stefano Ermon. Geography-aware self-supervised learning. In *ICCV*, pages 10181–10190, 2021. 2, 3, 6, 7, 8
- [10] Xin Guo, Jiangwei Lao, Bo Dang, Yingying Zhang, Lei Yu, Lixiang Ru, Liheng Zhong, Ziyuan Huang, Kang Wu, Dingxiang Hu, et al. Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. In *CVPR*, pages 27672–27683, 2024. 2
- [11] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022. 2, 4, 5, 12
- [12] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEit: BERT pre-training of image transformers. In *ICLR*, 2022. 2
- [13] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *NeurIPS*, 35:10078–10093, 2022. 2
- [14] Daniela Szwarzman, Sujit Roy, Paolo Fraccaro, THorsteinn Elí Gíslason, Benedikt Blumenstiel, Rinki Ghosal, Pedro Henrique de Oliveira, Joao Lucas de Sousa Almeida, Rocco Sedona, Yanghui Kang, et al. Prithvi-eo-2.0: A versatile multi-temporal foundation model for earth observation applications. *arXiv preprint arXiv:2412.02732*, 2024. 2, 6, 7, 8
- [15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner,

- Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 2, 3, 5
- [16] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *CVPR*, pages 6172–6180, 2018. 2, 3, 6
- [17] Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, et al. Hiera: A hierarchical vision transformer without the bells-and-whistles. In *ICML*, pages 29441–29454. PMLR, 2023. 2, 3, 5
- [18] Keumgang Cha, Junghoon Seo, and Taekyung Lee. A billion-scale foundation model for remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, pages 1–17, 2024. 2
- [19] Di Wang, Meiqi Hu, Yao Jin, Yuchun Miao, Jiaqi Yang, Yichu Xu, Xiaolei Qin, Jiaqi Ma, Lingyu Sun, Chenxing Li, Chuan Fu, Hongruixuan Chen, Chengxi Han, Naoto Yokoya, Jing Zhang, Minqiang Xu, Lin Liu, Lefei Zhang, Chen Wu, Bo Du, Dacheng Tao, and Liangpei Zhang. Hypersigma: Hyperspectral intelligence comprehension foundation model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–18, 2025. 2
- [20] Philipe Dias, Aristeidis Tsaris, Jordan Bowman, Abhishek Potnis, Jacob Arndt, H. Lexie Yang, and Dalton Lunga. Oreole-fm: Successes and challenges toward billion-parameter foundation models for high-resolution satellite imagery. In *ACM SIGSPATIAL*, 2024. 2
- [21] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution. In *CVPR*, pages 12009–12019, June 2022. 2
- [22] Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. In *NeurIPS*, volume 34, pages 8583–8595, 2021.
- [23] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *CVPR*, pages 12104–12113, June 2022.
- [24] Yufei Xu, Qiming Zhang, Jing Zhang, and Dacheng Tao. ViTAE: Vision transformer advanced by exploring intrinsic inductive bias. In *NeurIPS*, volume 34, pages 28522–28535, 2021.
- [25] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, et al. Internimage: Exploring large-scale vision foundation models with deformable convolutions. In *CVPR*, pages 14408–14419, 2023. 2
- [26] Favyen Bastani, Piper Wolters, Ritwik Gupta, Joe Ferdinando, and Aniruddha Kembhavi. Satlaspretrain: A large-scale dataset for remote sensing image understanding. In *ICCV*, pages 16772–16782, 2023. 2
- [27] Di Wang, Jing Zhang, Bo Du, Gui-Song Xia, and Dacheng Tao. An empirical study of remote sensing pretraining. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–20, 2022. 2
- [28] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, October 2021. 2, 4
- [29] Qiming Zhang, Yufei Xu, Jing Zhang, and Dacheng Tao. Vi-tae2: Vision transformer advanced by exploring inductive bias for image recognition and beyond. *International Journal of Computer Vision*, 131(5):1141–1162, 2023. 2, 8
- [30] Danfeng Hong, Bing Zhang, Xuyang Li, Yuxuan Li, Chenyu Li, Jing Yao, Pedram Ghamisi, Naoto Yokoya, Hao Li, Xiuping Jia, Antonio Plaza, et al. Spectralgpt: Spectral remote sensing foundation model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. DOI:10.1109/TPAMI.2024.3362475. 2
- [31] Jeremy Irvin, Lucas Tao, Joanne Zhou, Yuntao Ma, Langston Nashold, Benjamin Liu, and Andrew Y Ng. Usat: A unified self-supervised encoder for multi-sensor satellite imagery. *arXiv preprint arXiv:2312.02199*, 2023. 2
- [32] Gengchen Mai, Ni Lao, Yutong He, Jiaming Song, and Stefano Ermon. CSP: Self-supervised contrastive spatial pre-training for geospatial-visual representations. In *ICML*, pages 23498–23515. PMLR, 2023. 2
- [33] Vicente Vivanco, Gaurav Kumar Nayak, and Mubarak Shah. GeoCLIP: Clip-inspired alignment between locations and images for effective worldwide geo-localization. In *NeurIPS*, 2023.
- [34] Konstantin Klemmer, Esther Rolf, Caleb Robinson, Lester Mackey, and Marc Rußwurm. SatCLIP: Global, general-purpose location embeddings with satellite imagery. *ArXiv*, abs/2311.17179, 2023. 2
- [35] Di Wang, Qiming Zhang, Yufei Xu, Jing Zhang, Bo Du, Dacheng Tao, and Liangpei Zhang. Advancing plain vision transformer toward remote sensing foundation model. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2022. 2, 5, 6, 8
- [36] Xian Sun, Peijin Wang, Wanxuan Lu, Zicong Zhu, Xiaonan Lu, Qibin He, Junxi Li, Xuee Rong, Zhujun Yang, Hao Chang, et al. RingMo: A remote sensing foundation model with masked image modeling. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–22, 2022.
- [37] Peri Akiva, Matthew Purri, and Matthew Leotta. Self-supervised material and texture representation learning for remote sensing tasks. In *CVPR*, pages 8203–8215, June 2022.
- [38] Maofeng Tang, Andrei Liviu Cozma, Konstantinos Georgiou, and Hairong Qi. Cross-scale mae: A tale of multiscale exploitation in remote sensing. In *NeurIPS*, 2023. 5
- [39] Zhihao Li, Biao Hou, Siteng Ma, Zitong Wu, Xianpeng Guo, Bo Ren, and Licheng Jiao. Masked angle-aware autoencoder for remote sensing images. In *ECCV*, pages 260–278, 2024. 2, 5
- [40] Boran Han, Shuai Zhang, Xingjian Shi, and Markus Reichstein. Bridging remote sensors with multisensor geospatial foundation models. In *CVPR*, pages 27852–27862, June 2024. 2

- [41] Zhitong Xiong, Yi Wang, Fahong Zhang, Adam J Stewart, Joëlle Hanna, Damian Borth, Ioannis Papoutsis, Bertrand Le Saux, Gustau Camps-Valls, and Xiao Xiang Zhu. Neural plasticity-inspired foundation model for observing the earth crossing modalities. *arXiv preprint arXiv:2403.15356*, 2024.
- [42] Jonathan Prexl and Michael Schmitt. SenPa-MAE: Sensor parameter aware masked autoencoder for multi-satellite self-supervised pretraining. *ArXiv*, abs/2408.11000, 2024.
- [43] Guillaume Astruc, Nicolas Gonthier, Clement Mallet, and Loic Landrieu. OmniSat: Self-supervised modality fusion for Earth observation. *ECCV*, 2024.
- [44] Guillaume Astruc, Nicolas Gonthier, Clement Mallet, and Loic Landrieu. AnySat: An Earth observation model for any resolutions, scales, and modalities. *arXiv preprint arXiv:2412.14123*, 2024. 2
- [45] Colorado J Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. In *ICCV*, pages 4088–4099, 2023. 2, 5
- [46] Matías Mendieta, Boran Han, Xingjian Shi, Yi Zhu, and Chen Chen. Towards geospatial foundation models via continual pretraining. In *ICCV*, pages 16806–16816, 2023.
- [47] Di Wang, Jing Zhang, Mingqiang Xu, Lin Liu, Dongsheng Wang, Erzong Gao, Chengxi Han, Haonan Guo, Bo Du, Dacheng Tao, et al. MTP: Advancing remote sensing foundation model via multi-task pretraining. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.
- [48] Tong Zhang, Peng Gao, Hao Dong, Yin Zhuang, Guanqun Wang, Wei Zhang, and He Chen. Consecutive pre-training: A knowledge transfer learning strategy with relevant unlabeled data for remote sensing domain. *Remote Sensing*, 14(22), 2022.
- [49] Chao Tao, Ji Qi, Guo Zhang, Qing Zhu, Weipeng Lu, and Haifeng Li. TOV: The original vision model for optical remote sensing image understanding via self-supervised learning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2023.
- [50] Dilxat Muhtar, Xueliang Zhang, Pengfeng Xiao, Zhenshi Li, and Feng Gu. CMID: A unified self-supervised learning framework for remote sensing image understanding. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–17, 2023. 2
- [51] Utkarsh Mall, Bharath Hariharan, and Kavita Bala. Change-aware sampling and contrastive learning for satellite images. In *CVPR*, pages 5261–5270, 2023. 2, 3, 6, 7, 8
- [52] Lixian Zhang, Yi Zhao, Runmin Dong, Jinxiao Zhang, Shuai Yuan, Shilei Cao, Mengxuan Chen, Juepeng Zheng, Weijia Li, Wei Liu, Wayne Zhang, Litong Feng, and Haohuan Fu. A2-MAE: A spatial-temporal-spectral unified remote sensing pre-training method based on anchor-aware masked autoencoder. *ArXiv*, abs/2406.08079, 2024.
- [53] Yi Wang, Nassim Ait Ali Braham, Zhitong Xiong, Chenying Liu, Conrad M Albrecht, and Xiao Xiang Zhu. SSL4EO-S12: A large-scale multimodal, multitemporal dataset for self-supervised learning in earth observation [software and data sets]. *IEEE Geoscience and Remote Sensing Magazine*, 11(3):98–106, 2023.
- [54] Yohei Nakayama and Jiawei Su. Spatio-temporal swin-mae: A swin transformer based multiscale representation learner for temporal satellite imagery. *arXiv preprint arxiv:2405.02512*, 2024. 2
- [55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017. 4
- [56] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. MVITv2: Improved multiscale vision transformers for classification and detection. In *CVPR*, pages 4804–4814, June 2022. 5
- [57] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *ECCV*, pages 280–296. Springer, 2022. 6
- [58] Miriam Cha, Gregory Angelides, Mark Hamilton, Andy Soszynski, Brandon Swenson, Nathaniel Maidel, Phillip Isola, Taylor Perron, and Bill Freeman. Multiearth 2023—multimodal learning for earth and environment workshop and challenge. *arXiv preprint arXiv:2306.04738*, 2023. 6, 12
- [59] Marc Rußwurm and Marco Körner. Multi-temporal land cover classification with sequential recurrent encoders. *ISPRS International Journal of Geo-Information*, 7(4):129, 2018. 7
- [60] Clément Rambour, Nicolas Audebert, Elise Koeniguer, Bertrand Le Saux, Michel Crucianu, and Mihai Datcu. Sen12-flood: a sar and multispectral dataset for flood detection. *IEEE: Piscataway, NJ, USA*, 2020. 7, 12
- [61] Nikolaos Ioannis Bountos, Maria Sdraka, Angelos Zavras, Andreas Karavias, Ilektra Karasante, Themistocles Herekakis, Angeliki Thanasou, Dimitrios Michail, and Ioannis Papoutsis. Kuro siwo: 33 billion m² under the water. a global multi-temporal satellite dataset for rapid flood mapping. In *NeurIPS*, volume 37, pages 38105–38121, 2024. 7, 8, 12
- [62] Avaniika Narayan, Ines Chami, Laurel Orr, and Christopher Ré. Can foundation models wrangle your data? *Proceedings of the VLDB Endowment*, 16(4):738–746, December 2022. 8
- [63] Shashank Subramanian, Peter Harrington, Kurt Keutzer, Wahid Bhimji, Dmitriy Morozov, Michael W Mahoney, and Amir Gholami. Towards foundation models for scientific machine learning: Characterizing scaling and transfer behavior. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *NeurIPS*, volume 36, pages 71242–71262, 2023. 8

Supplementary Material

A. Pre-training Dataset: MillionST

To build the pre-training dataset, MillionST, we gathered Sentinel-2 multi-spectral Sequential Images Time Series (SITS) data from 100,000 sites across 1,317 cities in Europe, North Africa, and West Asia. These cities are among the world’s 10,000 most populous. Here, we provide a list of 100 cities among them.

[La Louvière, Seraing, Lierre, Binche, Gorna Oryahovitsa, Sandanski, Lida, Lugano, Biel/Bienne, Bellinzona, Köniz, Sion, Most, Frýdek-Místek, Mladá Boleslav, Pířerov, Třinec, Znojmo, Přibram, Trutnov, Norderstedt, Garbsen, Langenhagen, Lingen, Elmshorn, Melle, Oberursel, Rodgau, Neustadt am Rűbenberge, Lehrte, Falkensee, Dreieich, Wunstorf, Laatzen, Bensheim, Weißenfels, Buxtehude, Freital, Völklingen, Maintal, Ilmenau, Bitterfeld, Langen, Neu Isenburg, Papenburg, Königs Wusterhausen, Sankt Ingbert, Mörfelden-Walldorf, Seelze, Barsinghausen, Viernheim, Dietzenbach, Radebeul, Bad Vilbel, Wedel, Ahrensburg, Wernigerode, Lampertheim, Bad Nauheim, Hoyerswerda, Fürstenwalde, Achim, Georgsmarienhütte, Bramsche, Einbeck, Schönebeck, Burgdorf, Geesthacht, Riesa, Taunusstein, Andernach, Schwedt (Oder), Friedberg, Mīt Ghamr, Ziftá, Ra’s Ghārib, Vigo, Gijón, Badalona, Cartagena, Sabadell, Jerez de la Frontera, Mataró, Marbella, Algeciras, Lorca, El Puerto de Santa María, Mijas, Avilés, Rubí, Gandía, Benidorm, Benalmádena, Villanueva y Geltrú, La Línea de la Concepción, Arrecife, Granollers, Linares, Motril, Torrelavega]

B. Implementation Details

B.1. Pre-training Details

Following MAE [11], we use a mask ratio of 0.75 to mask partial tokens along both the spatial and temporal dimensions. We calculate the loss between the reconstructed and original pixels using normalized settings, as in MAE. Our experiments are carried out on NVIDIA RTX 4090 GPUs, with an effective batch size of 128, a base learning rate of 1.5×10^{-4} , and a weight decay of 0.05. We pre-train the model for 100 epochs, including 10 warmup epochs, and we adopt the same learning rate scheduler and weight decay parameters as described in [6].

B.2. Deforestation Monitoring

We fine-tune TiMo and all comparative models on the MultiEarth dataset [58] for 30 epochs with a base learning rate of 0.0001 and a batch size of 2. For the decoder, we use UperNet to generate multi-temporal segmentation images for each timestamp. We fine-tune all models using four Nvidia RTX 4090 GPUs for 30 epochs with a base learning rate of 0.0001. We also employ the AdamW optimizer

with a weight decay of 0.01.

B.3. Multi-class Land Cover Segmentation

For fine-tuning on the MultiSenGE dataset [2], the input consists of multi-temporal images, while the output is a single land-cover segmentation map. Since the encoder generates multi-temporal features, we use UperNet to recover the spatial resolution of each image. We then apply a Max Pooling layer on the temporal axis to generate a mono-temporal feature. We fine-tune all models using Nvidia RTX 4090 GPUs for 30 epochs with a base learning rate of 0.0001, a batch size of 2, and the AdamW optimizer with default settings.

B.4. Crop Type Classification

Similar to the MultiSenGE dataset [2], the MTLCC dataset contains multi-temporal images and mono-temporal, pixel-level crop type annotations. We fine-tuned all models for 50 epochs using a base learning rate of 0.01, a batch size of 4, and the AdamW optimizer. The same decoder structure used for the MultiSenGE dataset was also employed for fine-tuning on the MTLCC dataset.

B.5. Flood Disaster Assessment

For the Sen12Flood dataset [60], we train the models for 30 epochs with a base learning rate of 0.005 on the Sen12Flood dataset, using the AdamW optimizer with a weight decay of 0.05. We adopt a cosine decay learning rate schedule. Since this task aims to generate image-level prediction results, we apply average pooling to the spatial dimension, following the approach of [11].

For the KuroSiwo dataset [61], we train the models for 30 epochs with a base learning rate of 0.001. The same decoder structure used for the MultiSenGE dataset is also employed for fine-tuning on the KuroSiwo dataset.

C. More Qualitative Results

Figures 8-11 present the prediction outcomes of our model across various applications, including deforestation monitoring, land cover segmentation, crop type classification, and flood disaster assessment. The results illustrate that our approach produces highly accurate segmentation maps from multi-temporal imagery, showcasing its exceptional ability to capture the spatiotemporal dynamics inherent in satellite image time series.

D. Datasheet for MillionST

D.1. Motivation

1. For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

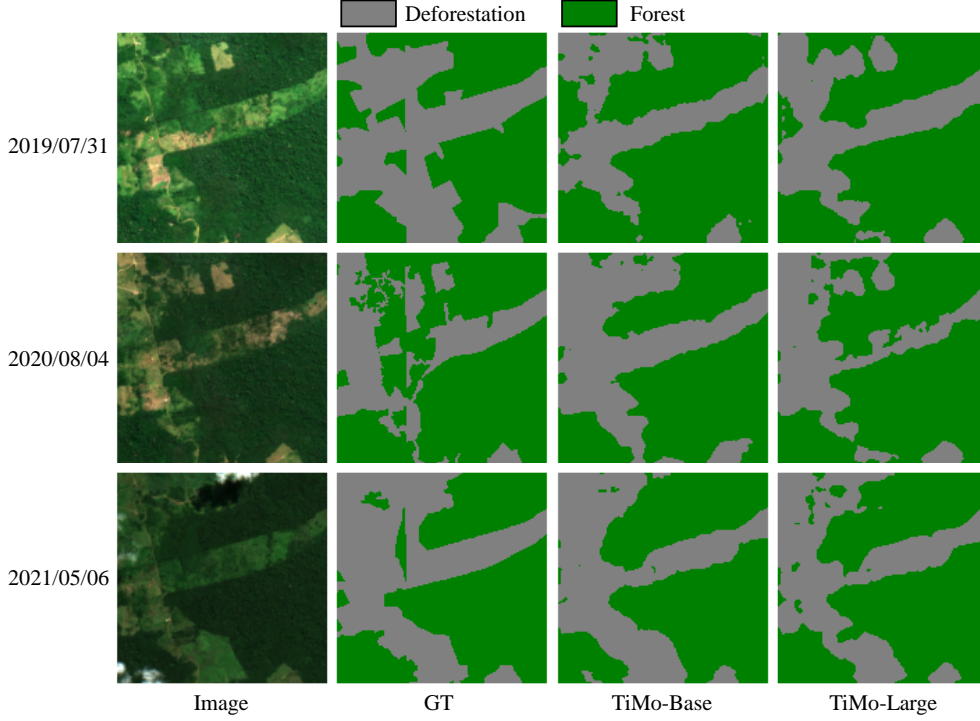


Figure 8. Visualization of the prediction results on MultiEarth dataset. GT denotes ground truth.

A1: MillionST is designed to enable unsupervised pre-training of deep learning models using unlabeled satellite image time series (SITS) data. To address the limitations of many existing pre-training datasets, which often suffer from restricted temporal coverage, MillionST was created as a large-scale SITS dataset specifically tailored for spatiotemporal pre-training. The dataset comprises data collected from 100,000 geographic locations, with each location providing images from 10 timestamps spanning five years. Extensive spatiotemporal diversity in MillionST will facilitate the development of multi-temporal remote sensing community.

2. Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

A2: This dataset is created by the authors of this paper.

3. Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

A3: N/A.

D.2. Composition

1. What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

A1: MillionST is comprised of image subsets captured by the Sentinel-2 satellite. Each subset consists of multi-spectral images from 10 timestamps and each image covers 2.65×2.65 km, consisting of 12 spectral bands at 10m, 20m, and 60m.

2. How many instances are there in total (of each type, if appropriate)?

A2: MillionST has 100,000 instances, totally including 1,000,000 image patches.

3. Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

A3: MillionST is a sample of instances from a larger set. The larger set includes all Sentinel-2 images covering all timestamps around the world since the launch of the Sentinel-2 mission. Since the images in MillionST are mainly collected from Europe, North Africa, and West Asia, it cannot represent the larger set. Nevertheless, it contains images of 10 temporal phrases across five years, encompassing abundant spatiotemporal variations.

4. What data does each instance consist of? “Raw” data

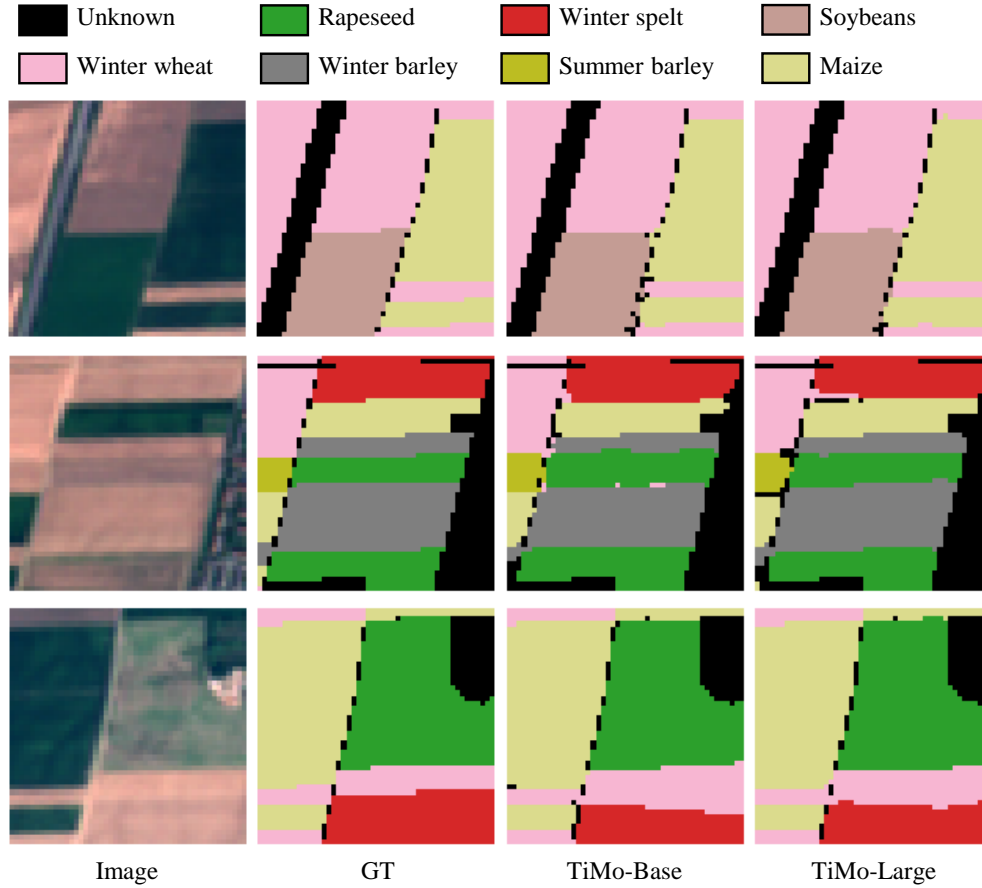


Figure 9. Visualization of the prediction results on MTLCC dataset. GT denotes ground truth.

(e.g., unprocessed text or images) or features? In either case, please provide a description.

A4: Each instance consists of raw Sentinel-2 multispectral image patches from 10 temporal phrases.

5. Is there a label or target associated with each instance? If so, please provide a description.

A5: No, since this dataset is intended for spatiotemporal self-supervised learning.

6. Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

A6: No.

7. Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

A7: Yes. The relationships between different instances are shown in the folder name and `metadata.json`.

8. Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a

description of these splits, explaining the rationale behind them.

A8: Yes, we recommend utilizing the whole dataset for spatiotemporal self-supervised pre-training.

9. Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

A9: No.

10. Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

A10: The dataset is self-contained since the samples are satellite imagery and can be downloaded from official web-

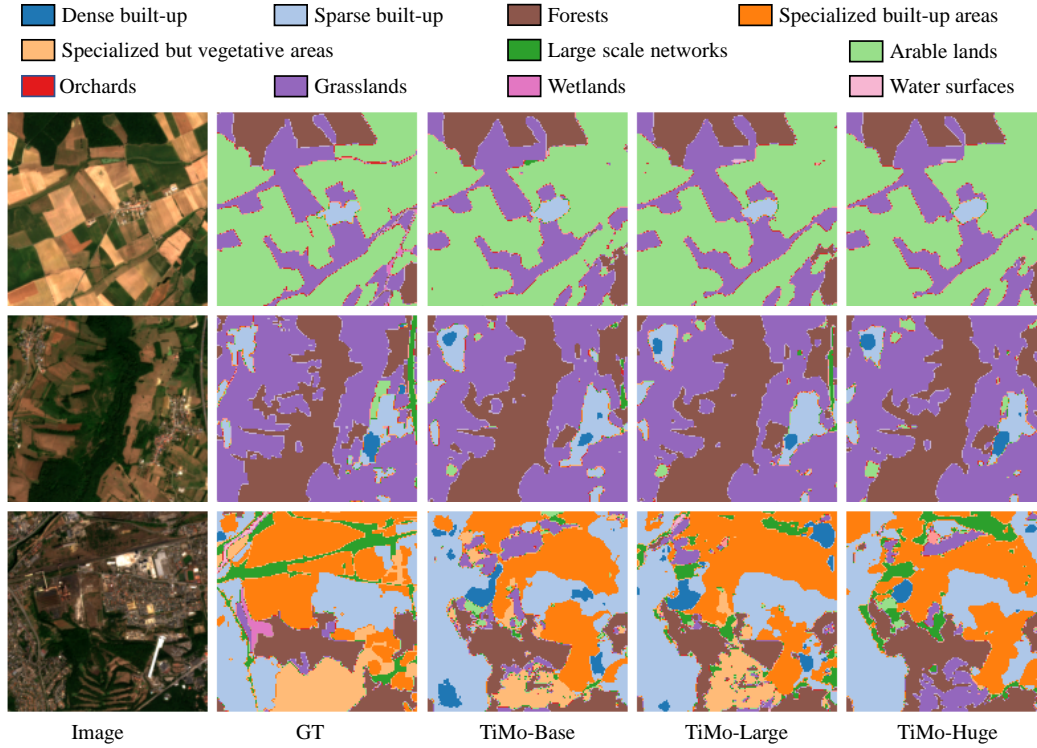


Figure 10. Visualization of the prediction results on MultiSenGE dataset. GT denotes ground truth.

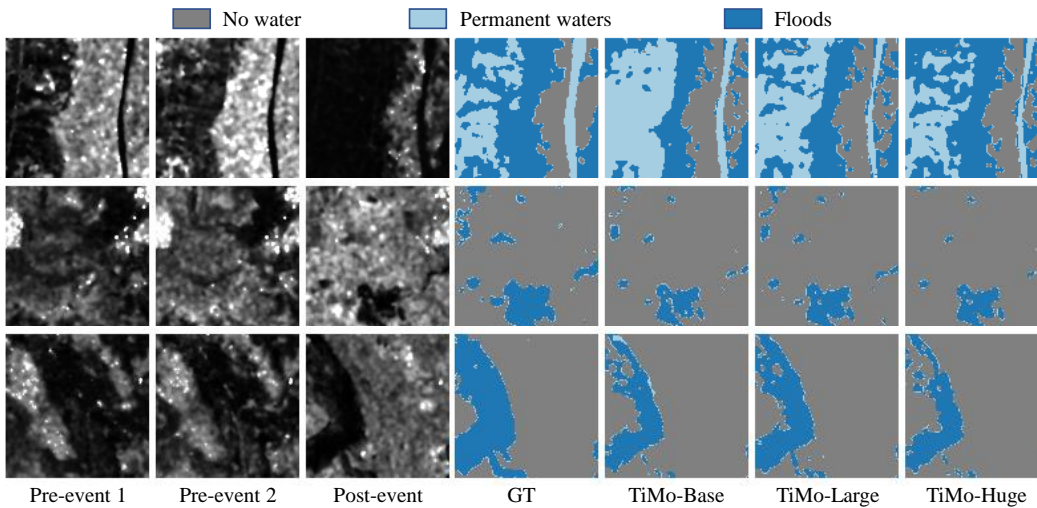


Figure 11. Visualization of the prediction results on KuroSiwo dataset. GT denotes ground truth.

sites. Users should refer to the official websites for information on any possible restrictions.

11. Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor/patient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.

A11: No.

12. Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

A12: No.

D.3. Collection Process

1. How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

A1: The data associated with each instance are directly observable, as they are stored in the GeoTIFF format and can be accessed via [Rasterio](#).

2. What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

A2: The data after radiometric calibration and atmospheric correction are collected from Google Earth Engine through cloud removal, filtering of low cloud content images, temporal and spatial sampling, and cropping. All operations are controlled by Python scripts to operate the Google Earth Engine. The correctness of Python scripts is validated.

3. If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

A3: The images in the dataset are probabilistically sampled.

Spatially, for each sample, the process begins with uniform sampling from 1,317 cities in Europe, North Africa, and West Asia. These cities are selected from the top 10,000 most populous cities globally. Subsequently, around each chosen city, Gaussian distribution sampling is performed, with the city center as the mean and a standard deviation of 50 kilometers.

Temporally, starting from a random time in the second half of 2017, timestamps are generated at six-month intervals until the first half of 2022, resulting in 10 timestamps. For each sample, a random offset is applied to these timestamps. Sampling is then conducted within a one-month window, covering 15 days before and after each offset timestamp. As a result, each location is associated with images corresponding to 10 timestamps.

4. Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

A4: The authors of this paper.

5. Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe

in which the data associated with the instances was created.

A5: Since all operations are performed online, downloading is seriously affected by the network connection status, and collecting data costs about 1 month.

D.4. Preprocessing/cleaning/labeling

1. Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

A1: No.

2. Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the “raw” data.

A2: N/A.

3. Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

A3: N/A.

D.5. Uses

1. Has the dataset been used for any tasks already? If so, please provide a description.

A1: No.

2. Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

A2: N/A.

3. What (other) tasks could the dataset be used for?

A3: It can be used for the research of RS self-supervised learning, especially for spatiotemporal pre-training.

4. Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

A4: No.

5. Are there tasks for which the dataset should not be used? If so, please provide a description.

A5: No.

D.6. Distribution

1. Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on

behalf of which the dataset was created? If so, please provide a description.

A1: Yes. The dataset will be publicly available.

2. How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

A2: It will be publicly available on the project website.

3. When will the dataset be distributed?

A3: The dataset will be distributed once the paper is accepted after peer review.

4. Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

A4: It will be distributed under the [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 License](#).

5. Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

A5: No.

6. Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

A6: No.

D.7. Maintenance

1. Who will be supporting/hosting/maintaining the dataset?

A1: The authors.

2. How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

A2: They can be contacted via email available on the project website.

3. Is there an erratum? If so, please provide a link or other access point.

A3: No.

4. Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

A4: No.

5. Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If

not, please describe how its obsolescence will be communicated to users.

A5: N/A.

6. If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

A6: N/A.