A^2-MAE: A Spatial–Temporal–Spectral Unified Remote Sensing Pretraining Method Based on Anchor-Aware Masked Autoencoder

Lixian Zhang, *Member, IEEE*, Yi Zhao, Runmin Dong, *Member, IEEE*, Jinxiao Zhang, Shuai Yuan, *Graduate Student Member, IEEE*, Shilei Cao, Mengxuan Chen, Juepeng Zheng, *Member, IEEE*, Weijia Li, *Member, IEEE*, Wayne Zhang, Wei Liu, Litong Feng, Jianxi Huang, *Senior Member, IEEE*, and Haohuan Fu, *Fellow, IEEE*

***Abstract*—Vast amounts of remote sensing (RS) data provide Earth observations across multiple dimensions, encompassing critical spatial, temporal, and spectral information which is essential for addressing global-scale challenges such as land-use monitoring, disaster prevention, and environmental change mitigation. Despite various pretraining methods tailored to the characteristics of RS data, a key limitation persists: the inability to effectively integrate spatial, temporal, and spectral information within a single unified model. To unlock the potential of RS data, we construct a spatial–temporal–spectral structured dataset (STSSD) characterized by the incorporation of multiple RS sources, diverse coverage, unified locations within image sets, and heterogeneity within images. Building upon this structured dataset, we propose an anchor-aware masked autoencoder (A^2-MAE) method, leveraging intrinsic complementary information from the different kinds of images (featuring different resolutions, spectral compositions, and acquisition times) and geo-information to reconstruct the masked patches during the pretraining phase. Moreover, A^2-MAE integrates an anchor-aware masking (AAM) strategy and a geographic encoding module (GEM) to comprehensively exploit the properties of RS images. Specifically, the proposed AAM strategy dynamically adapts the masking process based on the meta-information of a preselected anchor image, thereby facilitating the training on images captured by diverse types of RS sources within one model. Furthermore, we propose a geographic encoding method to leverage accurate spatial patterns, enhancing the model's generalization capabilities for downstream applications that are generally location-related. Extensive experiments demonstrate our method achieves comprehensive improvements across various downstream tasks compared with existing RS pretraining methods, including image classification, semantic segmentation, and change detection tasks. The dataset and pretraining model are available at https://github.com/ZhaoYi1222/AAMAE**

***Index Terms*—Deep learning, diverse spatial–temporal–spectral information, foundation model, masked autoencoder (MAE), transformer.**

Received 13 May 2025; revised 25 August 2025, 20 November 2025, and 26 December 2025; accepted 10 February 2026. Date of publication 13 February 2026; date of current version 5 March 2026. This work was supported in part by the National Natural Science Foundation of China under Grant T2125006, Grant 42501588, and Grant 42301390; in part by Guangdong Science and Technology Program under Project 2024B0101040005 and Project 2025B0101080001; in part by Shenzhen Science and Technology Program under Grant KCXFZ20240903093759004 and Grant ZDCYKCX20250901094203005; and in part by Tsinghua SIGS KA Cooperation Fund. *(Lixian Zhang, Yi Zhao, and Runmin Dong contributed equally to this work.) (Corresponding authors: Runmin Dong; Juepeng Zheng; Haohuan Fu.)*

Lixian Zhang is with the High-Performance Computing Department, National Supercomputing Center in Shenzhen, Shenzhen 518055, China.

Yi Zhao, Jinxiao Zhang, and Mengxuan Chen are with the Department of Earth System Science, Tsinghua University, Beijing 100086, China.

Runmin Dong, Shilei Cao, and Juepeng Zheng are with the School of Artificial Intelligence, Sun Yat-sen University, Zhuhai 519082, China (e-mail: dongrunmin95@163.com; zhengjp8@mail.sysu.edu.cn).

Shuai Yuan is with the Department of Geography, The University of Hong Kong, Hong Kong, China.

Weijia Li is with Tsinghua Shenzhen International Graduate School, Tsinghua University, Beijing 100086, China.

Wayne Zhang, Wei Liu, and Litong Feng are with SenseTime Group, Shanghai 200030, China.

Jianxi Huang is with the Faculty of Geosciences and Engineering, Southwest Jiaotong University, Chengdu 611756, China, also with the College of Land Science and Technology, China Agricultural University, Beijing 100083, China, and also with the Key Laboratory of Remote Sensing for Agri-Hazards, Ministry of Agriculture and Rural Affairs, Beijing 100083, China.

Haohuan Fu is with the Department of Earth System Science, Tsinghua University, Beijing 100086, China, and also with Tsinghua Shenzhen International Graduate School and the Department of Earth System Science, Tsinghua University, Beijing 100086, China (e-mail: haohuan@tsinghua.edu.cn).

Digital Object Identifier 10.1109/TGRS.2026.3664662

I. Introduction

Earth observations through remote sensing (RS) constitute a fundamental tool for monitoring the evolution of global-scale phenomena, including the urbanization process [1], [2], land-use change [3], [4], and biodiversity loss [5], [6]. Over the past half-century, there has been a substantial increase in the volume of RS data, resulting in spatial, temporal, and spectral diversities within extensive RS image archives. The spatial–temporal–spectral diversities inherent in RS images offer critical and complementary information for comprehensive analysis and recognition of objects and scenes. Consequently, RS plays a pivotal role in various operational and complex research domains within the field of geoscience. To address the challenges posed by the scarcity of expensive annotations in downstream RS applications [7], [8], [9], self-supervised learning (SSL) has emerged as a promising paradigm [10], [11]. SSL derives robust feature representations from vast, unlabeled satellite image archives, which can then be fine-tuned with limited labeled data for specific tasks. The acquired feature representations can subsequently be fine-tuned with limited labeled data for specific downstream applications. Despite considerable efforts in constructing large pretrained models through methods like masked autoencoders (MAEs) or contrastive learning (CL), most of the existing RS SSL methodologies have been custom-tailored for specific scenarios, such as temporal SatMAE [12], multispectral SatMAE

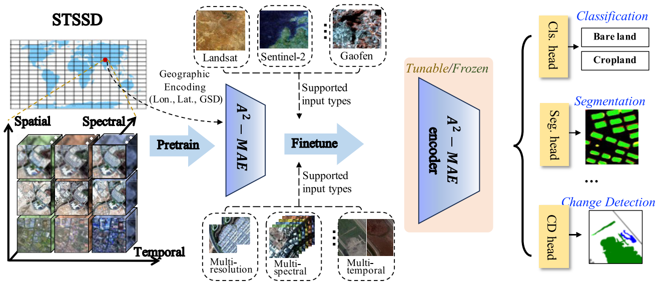

Fig. 1. Illustrative overview of the proposed global-scale dataset STSSD and pretraining method A^2-MAE. STSSD is a comprehensive RS dataset structured and characterized by the inclusion of diverse spatial, temporal, and spectral coverage. A^2-MAE facilitates the efficient utilization of the intrinsic complementarity information in STSSD within one unified spatial–temporal–spectral model. Lon., Lat., and GSD indicate longitude, latitude, and ground spatial distance information, respectively.

[12], multiresolution ScaleMAE [13], and spatiotemporal foundation models [14]. These methods enhance performance in specific downstream tasks but fall short in achieving comprehensive improvements across various downstream tasks. Besides, the existing SSL methods underutilize geographical information, which is a powerful prior for leveraging spatial patterns. In the work, we address the pivotal question: *How can we present a single spatial–temporal–spectral unified RS pretraining method to effectively leverage a diverse collection of RS images?*

The key to this question lies in two aspects. The first aspect involves the construction of a location-unified and extensive RS dataset encompassing images with varying temporal, spatial resolutions, and spectral compositions. Presently available RS datasets are typically derived from one or two satellite sources, offering limited spatial–temporal–spectral coverage. For instance, the Million-AID dataset exclusively covers optical RS images with RGB bands [15], [16]. SEN12MS [17] and SSL4EO-S12 [18] exhibit constrained temporal coverage. The SeCo [19] and CACo [20] datasets are pretrained exclusively on Sentinel-2 images. However, real-world RS data exhibits significant variations in spatial resolution, temporal coverage, and spectral composition. The SSL models trained on homogeneous RS data struggle to provide effective representation for fine-tuning of downstream tasks involving different RS sources.

The second aspect entails mining the intrinsic relevance from the images with different spatial, temporal, and spectral characteristics through SSL techniques. A straightforward approach is to design separate backbones for different types of sources and align the representations of different types [21]. However, this method leads to a linear escalation in model parameters and computational overhead with the expansion of source type count. As there are a large number of types of RS sources, such as Landsat-8 with 7 bands, Sentinel-2 with 13 bands, Gaofen-2 with 4 bands, and WorldView-2 with 8 bands, it is difficult to simultaneously model the relationship across different types of sources.

To address these challenges, we introduce STSSD (see Fig. 1), a global-scale RS dataset containing half a million sampling locations with 2.5 million spatial–temporal–spectral structured image sets collected from multiple multispectral sources. Each image set is meticulously crafted to exhibit different spatial resolutions, temporal and spectral compositions for the same location. We utilize a data pruning method to preserve heterogeneity within images and diversity across images for SSL. To harness the rich and varied representation features within STSSD effectively, we propose an Anchor-Aware MAE method (A^2-MAE), including an anchor-aware masking (AAM) strategy and geographic encoding module (GEM) (see Fig. 1). The proposed AAM strategy enables training on images captured by diverse sources within one unified spatial–temporal–spectral model. Besides, the proposed geographic encoding method allows the model to leverage accurate spatial patterns, unleashing the potential of geolocation priors for downstream tasks. Experiments verify that our method achieves comprehensive improvements across various downstream tasks compared with state-of-the-art RS SSL methods. Taking DynamicEarthNet as an example, the performance can be further enhanced by over 8.4% on mIoU through the introduction of geographic information during the fine-tuning process (refer to Section V-C).

In summary, our contributions are as follows.

1) We build the STSSD, a globally spatial–temporal–spectral structured RS dataset featuring high diversity of data sources, unification of location, and content heterogeneity. STSSD is meticulously curated to encompass diverse land-use types spatially, capture landscape changes temporally, and incorporate various band compositions spectrally.
2) We propose a pretraining method, A^2-MAE, designed to accommodate various types of RS sources within a unified backbone architecture. A^2-MAE leverages spatial–temporal–spectral relationships and geographical information to improve model representation and generalization capabilities.
3) Experiments verify the effectiveness and advantages of A^2-MAE compared to existing RS pretraining models with similar complexities across image classification, semantic segmentation, and change detection tasks.

II. Related Work

A. Large-Scale Datasets for RS Imagery Pretraining

Inspired by the achievements of computer vision (CV) datasets [22], [23], [24], [25], researchers have introduced several large-scale RS datasets [18], [26], [27], [28], [29], [30].

These datasets exhibit a gradual expansion in the volume of data, starting with the fMoW [26] encompassing 1 million images, progressing to BigEarthNet-MM [28] with 1.2 million images, and further expanding to SSL4EO-S12 [18] comprising 3 million images. In addition, there has been a progression in the diversity of spectral sources in datasets, transitioning from datasets like BigEarthNet [27] solely from Sentinel-1, to BigEarthNet-MM [28] combining Sentinel-1/2 pairs, to SatlasPretrain [30], which incorporates data from Sentinel-1/2 and NAIP, and then to DynamicEarthNet [29] containing diverse spatial–temporal–spectral images with constrained sampling locations.

Therefore, there is an urgent need to construct a large-scale spatial–temporal–spectral structured RS dataset encompassing more multispectral sources and diverse coverage. In this work, we introduce STSSD as a versatile, large-scale resource designed specifically for advancing spatial–temporal–spectral unified learning in RS.

B. SSL for Satellite Imagery

SSL primarily focuses on generating supervisory signals from unlabeled data, through the design of various pretext tasks such as masked patches reconstruction [11], [31], [32], [33], [34] and contrasting semantically similar inputs [10], [35], [36], [37], [38]. Furthermore, SSL enables the acquisition of semantic information without human annotation. Therefore, SSL plays a vital role in the RS domain [9], where annotation demands specialized expertise and incurs high costs. Existing RS pretraining methods leverage different properties of RS images or specific RS tasks [9]. For instance, Ayush et al. [39] leverage spatially aligned but temporally separated images as positive pairs to learn feature representation for 10m multispectral images. Similarly, Mall et al. [20] propose a new SSL loss for CL to distinguish between short-term and long-term changes in multispectral images. Cong et al. [12] introduce SatMAE to leverage temporal or multispectral information in data through positional encoding. Reed et al. [13] present scale-MAE to reconstruct both low and high-frequency images to learn robust multiscale representations for RS imagery. Nevertheless, these studies are customized for specific types of RS images and cannot simultaneously utilize RS images from different kinds of multispectral sources in one unified model.

Given the vast quantity and varied characteristics of RS images [40], [41], [42], it is crucial to efficiently exploit the intrinsic relevance of images with diverse spatial, temporal, and spectral attributes. Previous studies have predominantly focused on addressing specific facets of this diversity, such as multispectral [43] and multiresolution images [13], or by achieving separate pretraining models [12] and learning diverse angles [44]. Consequently, achieving comprehensive and generalizable improvements across downstream tasks that span spatial, temporal, and spectral dimensions remains challenging. In response, a contemporaneous study, Skysense [21], designs separate backbones for three types of sources in a larger model with 2.06 billion parameters, which is trained on 80 A100 GPUs. However, this approach faces difficulties in scaling model parameters and computational overhead to accommodate expanding RS sources, which is evidently unsustainable.

To fill this gap, we propose an AAM strategy to leverage intrinsic complementarity information from an image set, which can be easily extended to various multispectral sources with much less computational consumption.

C. Geography-Aware Learning

RS images offer essential metadata records containing geographic information, such as geographic location and ground sample distance (GSD) [45], [46]. This prior information enables a robust linkage between fine-tuning data and models pretrained globally [47], [48]. Consequently, it is anticipated to bolster the representational capacity of the pretrained model [49], [50]. While a few studies have leveraged recorded geographic data [39], [51], they are constrained in efficiently utilizing such information on a large scale [52]. One-hot geo-encoding [26] offers limited encoding outcomes, while GSD scaling encoding [13] cannot be jointly integrated with geo-location data. Another alternative (i.e., geo-context prototype learning [21]) demands additional computational resources while yielding encoding outcomes unsuitable for varying spatial resolutions. To bridge this gap, we introduce a GEM in A^2-MAE, providing more accurate geographical priors (i.e., latitude, longitude, and GSD) without additional computation overhead, thereby improving the generalization of applications on a global scale.

III. Dataset

A. Overview

We introduce STSSD, a large-scale RS dataset designed for spatial–temporal–spectral unified SSL. This dataset is meticulously curated through data pruning from an initial pool of 4.2 million original images collected at 1045 thousands of sampling locations across four satellite sources. The resulting STSSD comprises 510 thousands of image sets, each containing up to six images collected from two sources with different resolutions and spectral compositions. As shown in Fig. 2, STSSD consists of four kinds of image sets, featuring diverse sources, spatial resolutions, coverage, and acquisition times (details reported in Table I). It is characterized by the following four key attributes.

1) *Diversity:* STSSD offers comprehensive source diversity, incorporating four satellite sources with three distinct band compositions. It also features spectral diversity through varying band sets and resolutions (0.8–30 m/pixel) from multiple sources.
2) *Coverage:* STSSD captures dynamic geographical features across over 12 000 urban centers and 10 000 nature reserves worldwide, enhancing model performance on downstream tasks with varied geographical characteristics.
3) *Unification:* By integrating spatial, temporal, and spectral contexts, STSSD creates image sets from diverse sources and acquisition times. These unified sets provide richer multidimensional information than single-temporal or single-source data, bolstering the robustness of RS foundation models.
4) *Heterogeneity:* Using a clustering-based pruning strategy to remove redundant (e.g., desert) and low-quality images, STSSD balances interimage diversity and intraimage heterogeneity. This heterogeneity heightens the challenge of SSL image reconstruction, enabling the model to learn more powerful feature representations.

B. STSSD Construction

1) Data Sources: In the pursuit of constructing a unified RS dataset characterized by diverse sources, we strategically

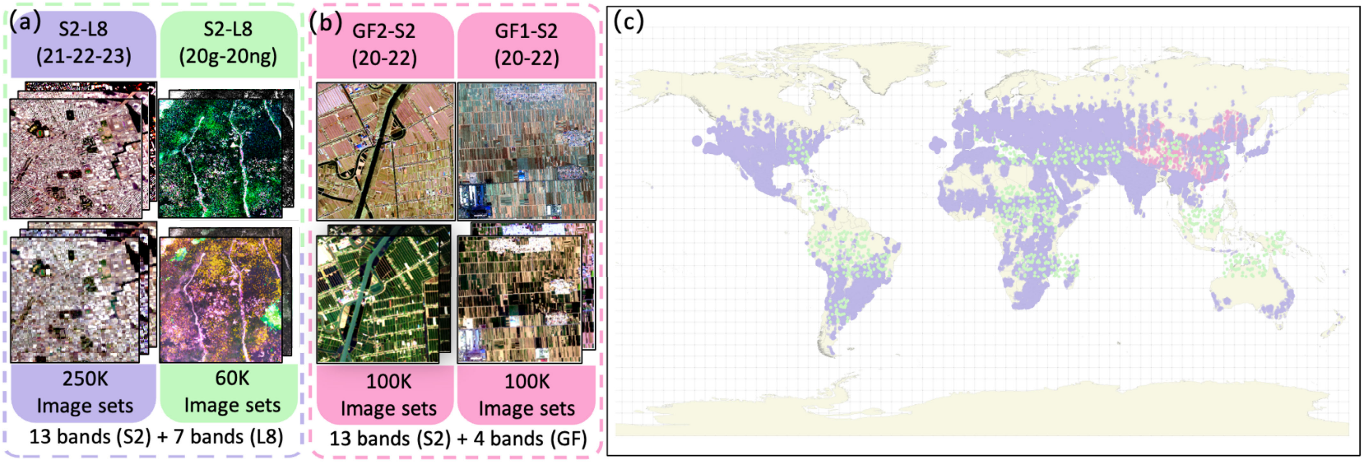

Fig. 2. Compositions of image sets and the global sampling location distribution. (a) S2-L8 image sets. "ng" denotes the nongrowth period, and "g" denotes the growth period within one year. (b) GF-S2 image sets. (c) Sampling location distribution (i.e., purple circles for S2-L8 in urban areas, green circles for S2-L8 in nature reserves, and pink circles for GF-S2).

TABLE I
CHARACTERISTICS OF SATELLITE SOURCES INCLUDED IN STSSD

Satellite Source	Number of Bands	Primary GSD	Temporal Spanning
Gaofen-1 (GF-1)	4	2 m	2020-2022
Gaofen-2 (GF-2)	4	0.8 m	2020-2022
Sentinel-2	13	10-60 m	2020-2023
Landsat-8	7	30 m	2021-2023

incorporate imagery from Gaofen-1 (4 bands, 2 m/pixel GSD), Gaofen-2 (4 bands, 0.8 m/pixel GSD), Sentinel-2 (13 bands, 10-60 m/pixel GSD), and Landsat-8 (7 bands, primarily 30 m/pixel GSD). This selection maximizes global coverage and leverages sensors with varying accessibility, resolution, and spectral characteristics. The combination of these diverse sources facilitates a high degree of potential input variability for model training. Specifically, when utilized with a training methodology involving the sampling of sensor combinations and spectral bands, the number of potential unique input combinations can reach up to 37.5 times the number of base sensor types (derived from 150 calculated combinations across four sensors) (detailed in Table I). This significantly enriches the training data diversity.

2) Global Coverage: Building upon the texture-rich images of urban areas as demonstrated by previous work [20], we further expand our dataset to include nature reserves [5], [6]. This expansion aims to enhance the model's understanding and capabilities in recognizing a more diversified and dynamic planet. Consequently, we meticulously select over the original 1045 thousands of sampling locations, spanning nature reserves (depicted in green) and main cities (depicted in purple), to collect Sentinel-2 and Landsat-8 image sets (S2-L8), as illustrated in Fig. 2(c). To capture the dynamic nature of geographical features, a time series of images are provided for each sampling location, ranging from the year 2020 to 2023, with periodic seasonal revisits. Furthermore, we utilize the locations of the available Gaofen images to gather the corresponding Sentinel-2 images, subsequently forming Sentinel-2 and Gaofen image sets (GF-S2) to enhance the representation ability for higher resolution data (depicted in pink in Fig. 2).

3) Motivations of the STSSD Structure Design: The structuring of these image sets is designed to ensure optimal resolution and band gaps for effective model learning. Specifically, there are two kinds of image sets: S2-L8 image sets and GF-S2 image sets. For S2-L8 [see Fig. 2(a)], the image sets collected from main cities comprise six images, involving 3 Sentinel-2 and 3 Landsat-8 images annually from 2021 to 2023, to capture the temporal changes in land cover. As for nature reserves, we conduct the image sets comprising four images, including 2 Sentinel-2 and 2 Landsat-8 images during both the growth and nongrowth periods in 2020, to showcase the phenological characteristics. For GF-S2 [see Fig. 2(b)], each image set integrates three images, including one Gaofen (GF-1 or GF-2) image and two Sentinel-2 images captured at different time points. Note that each image set comprises two distinct data sources with different sources and temporal snapshots. The integration of diverse image sources, characterized by varying spatial resolutions, within a single image set enables multiscale observation of the same geographical areas. This simultaneous consideration of fine-grained details and broader contextual views facilitates a more comprehensive feature representation, consequently enhancing performance across diverse downstream tasks such as building extraction and land cover mapping. The design of the STSSD dataset incorporates further deliberate choices regarding sample distribution and temporal sampling strategies to optimize it for SSL across diverse RS tasks. The sample allocation between urban and nature reserve patches is primarily driven by the higher spatial–temporal complexity of urban environments and data availability, balanced against the need for efficient yet representative coverage of natural landscapes. Temporal sampling strategies are tailored: urban areas utilize three annual timestamps to capture both seasonal variations and interannual changes relevant to urban dynamics while maintaining computational feasibility, whereas nature reserves focus on contrasting periods, specifically growing and nongrowing seasons. This focus maximizes the capture

TABLE II
COMPARISON OF STSSD AGAINST EXISTING PUBLIC DATASETS IN RS (K = THOUSANDS, M = MILLIONS)

Dataset	Spatial cover	Temporal spanning	GSD (m)	# Sources	# Patches	# Timestamps per Location
fMoW [26]	Global	2002 – 2017	0.31 – 1.60	1	132K	1 - 41
BigEarthNet [27]	Europe	2017 – 2018	10	1	590K	1
BigEarthNet-MM [28]	Europe	2017 – 2018	10	2	590K	4
SSL4EO-S12 [18]	Global	2021	10	2	250K	5 – 15
SatlasPretrain [30]	Global	2011 – 2022	1 – 10	3	3M	1
STSSD (Ours)	Global	2020 – 2023	0.8 – 30	4	510K	2 for cities, 3 for natural reserves

of ecologically significant phenological shifts, aligning with practical data availability and reliability from satellite archives. Moreover, the incorporation of multitemporal information ensures accessibility to temporal dynamics, thereby fortifying the robustness of temporal variations for downstream tasks such as change detection.

4) Data Pruning Strategy: Since the original STSSD contains observations from areas with high homogeneities, such as deserts which do not contribute significantly to the diversity and complexity of the dataset due to their uniform nature, we employ a data pruning strategy to remove redundant contents and filter out low-quality images, resulting in a more refined and curated collection of data. This process ensures that the images in STSSD are high-quality and heterogeneous. After pruning, the final STSSD owns over 510K sampling locations with 2.5 million curated images.

C. Comparison With Public RS Dataset

Since researchers have introduced several large-scale RS datasets [18], [26], [27], [28], [29], [30], we provide a comprehensive comparison of STSSD against existing public datasets in RS in Table II.

Our STSSD dataset offers several notable contributions and strengths compared to existing public datasets in the RS domain. First, STSSD boasts a comprehensive spatial coverage, spanning globally, and a relatively recent temporal range from 2020 to 2023. This temporal range surpasses many other datasets, such as fMoW [26]. Moreover, STSSD provides a diverse range of ground sampling distances (GSD) ranging from 0.8 to 30 m, offering finer spatial resolutions than most datasets, such as SSL4EO-S12 [18]. In addition, STSSD incorporates data from four distinct sources, surpassing the number of sources in other datasets like BigEarthNet and BigEarthNet-MM [27]. Consequently, STSSD stands out for its extensive coverage, recent temporal span, multiple spatial resolutions, and diverse data sources, collectively contributing to its significance as a valuable resource for RS applications.

D. Implementation Details

1) Sample Selection Details: We employ a sampling strategy similar to that used for city sampling [19], [20], with larger scales and finer filters. To construct a globally representative dataset of urban environments, we implemented a city selection strategy based on population size. We selected the top 12 000 cities ranked by population, representing over 92% of the world's recognized urban settlements, thereby encompassing significant diversity in geographic location, economic development, and land-use patterns, while balancing representativeness with computational feasibility for subsequent analysis. Subsequently, a Gaussian sampling approach is employed within a region centered at the geographical coordinates of each selected city, with a radius of 50 km. Upon selecting a candidate point (i.e., longitude, latitude) within this region, additional checks are conducted to ensure the spatial diversity (no marine areas, no overlapping areas), and cloud cover ($\leq$ 10% covering). For nature reserves, a similar approach is adopted, focusing on regions designated as protected natural areas. Geographic datasets containing information about nature reserves [53], [54] are used to identify suitable sampling locations. The same sampling methodology described for cities is applied within the boundaries of these nature reserves, ensuring spatial coverage across diverse ecological habitats. Similarly, checks for cloud cover and proximity to previously sampled locations are conducted to ensure the integrity and representativeness of the sampled data. The final curated dataset combines samples from both global urban areas and designated nature reserves. It comprises 250 000 patches from urban areas and 60 000 patches from nature reserves, resulting in an approximate 4:1 ratio of urban to natural samples. This sample distribution is deliberately chosen to prioritize the representation of highly heterogeneous and complex urban landscapes, crucial for robust model training, while ensuring efficient yet comprehensive coverage of diverse natural environments, thereby balancing data representativeness with computational tractability for large-scale pretraining.

2) Data Pruning Details: Due to the extensive scale of our dataset, challenging examples contribute significantly to its diversity and complexity, thus enhancing our model's feature representation capability. To address the homogeneous images derived from nature reserves, we exclusively apply the data pruning strategy to the observation data from nature reserves. Initially, we partition the collected RS dataset from nature reserves into 341 nonoverlapping subregions based on geolocations, each delineated by a $1^\circ \times 1^\circ$ grid. As geographical representative features within each subregion vary significantly, we employ individual data pruning models based on k-means clustering to effectively preserve the geographical diversity within nature reserves. Specifically, we gauge the difficulty of each data point by measuring its Euclidean distance to the nearest cluster centroid [55]. Simple examples are those most prototypical, while hard examples deviate furthest from the norm. The simplest examples provide limited information, while the hard examples provide diverse information about earth observation. Specifically, adopting the common land

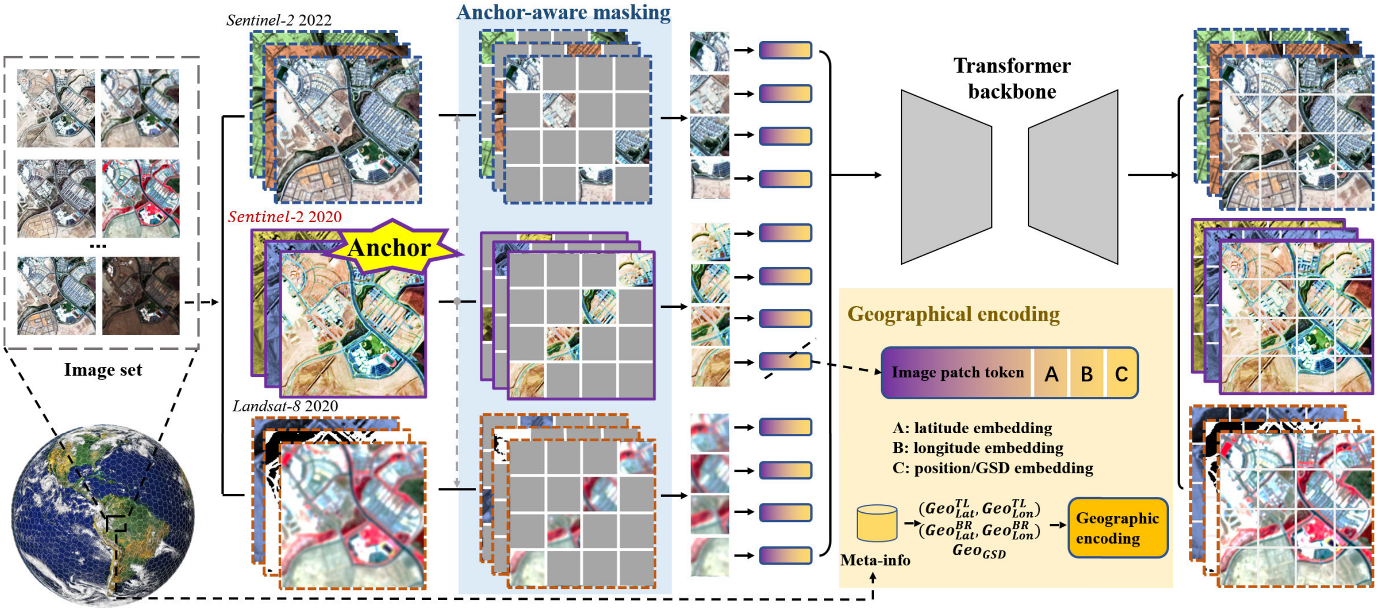

Fig. 3. Overall framework of A^2-MAE. A^2-MAE incorporates an AAM strategy and a GEM, allowing for the efficient utilization of spatial-, temporal-, and spectral-variant information in large-scale RS imagery. This figure illustrates an example of the image set P_i, consisting of $\{I_i^{2020,Sen2}$, $I_i^{2020,Lan8}$, and $I_i^{2022,Sen2}\}$. Taking the middle image $I_i^{2020,Sen2}$ as the referenced anchor, we utilize consistent masking strategy, random masking strategy, and mutually exclusive masking strategy for the $I_i^{2020,Sen2}$, $I_i^{2020,Lan8}$, and $I_i^{2022,Sen2}$, respectively.

cover classification system (IGBP), we set k, the number of clusters, to 20. After the k-means clustering, we retain the most difficult 10% images of nature reserves in the final STSSD. To provide deeper insights into example difficulty across various metrics, we visually represent selected images and discarded images based on our self-supervised prototype metric. Qualitatively, simple examples typically exhibit high similarity and redundancy, while hard examples tend to be highly heterogeneous and complexity.

3) Construction Details: To enhance the quality of the constructed STSSD, targeted preprocessing steps are undertaken for each data source. Initially, all data undergoes processing for atmosphere and radiation correction. Subsequently, the Gaofen series images are pan-sharpened to achieve higher resolution. Afterward, the STSSD is organized based on geographical locations, with each location containing 3–6 pairs of images, contingent upon the data source availability in the respective geographical area. To ensure alignment across locations and address variations in image resolutions within each pair, the images are cropped to sizes ranging from 256 × 256 to 3200 × 3200. Specifically, for nature reserves, Landsat-8 images are resized to 256 × 256, while Sentinel-2 images are resized to 768 × 768. In the case of the main cities, Landsat-8 images are adjusted to 256 × 256, and Sentinel-2 images to 480 × 480. Moreover, for the geographic region of China, Sentinel-2 images are standardized to 256 × 256, Gaofen-1 images to 1280 × 1280, and Gaofen-2 images to 3200 × 3200.

IV. Methodology

A. Overall Architecture

As illustrated in Fig. 3, A^2-MAE is a self-supervised pretraining method based on the MAE [11], which makes two key contributions to the MAE framework to unlock representative potentials in STSSD. First, A^2-MAE presents an AAM strategy to utilize the image sets collected from different sources. The AAM dynamically adjusts the masking strategy according to the meta-information (i.e., the spatial resolutions, temporal, and spectral information) of a preselected anchor image for each training iteration. This adaptation learning allows the model to leverage the intrinsic complementarity of spatial–temporal–spectral information to reconstruct the masked patches, thereby improving the model's representation ability. In addition, A^2-MAE introduces a GEM to obtain the geo-embedding of the given image set to provide accurate geographical priors for A^2-MAE, improving the model's generalization ability.

B. Setup

Since the image sets in the STSSD are gathered by geo-locations, we denote $P_i = \{I_i^{1,1}, \ldots, I_i^{t,s}\}$ represents the image set at the *ith* location. $I_i^{t,s} \in \mathbb{R}^{H\times W\times C}$ represents an RS image with H height, W width, and C channels, which is captured by source s at time t. Note that images from different sources s have different representative features, including different spectral compositions and GSDs. Three images of P_i are randomly selected as input I_{in} for A^2-MAE, ensuring a minimum of two different s and two different t to capture sufficient diversity in spatiotemporal–spectral relationships while balancing computational cost. The A^2-MAE then patchifies the selected I_{in} into three sets of sequence Seq of independent patches. After randomly removing a fraction of the obtained patches, the A^2-MAE reconstructs the removed patches by leveraging the complementarity information within the remaining patches from Seq. Unlike the traditional MAE architecture, the A^2-MAE includes an AAM to encourage the A^2-MAE to implicitly leverage the intrinsic complementarity information within I_{in} and a GEM to introduce the geographic-related information.

C. AAM Strategy

Existing RS SSL methods are often tailored for specific scenarios, limiting the capabilities to leverage symbiotic features among images and increasing the costs for transferring to other scenarios. In contrast, our method works toward a unified pretraining method that potentially benefits various representative features between images with different spatial resolutions, temporal, and spectral compositions in P_i. However, this symbiotic and diverse complementarity information within the image sets also poses challenges for obtaining robust and generalized RS representative features due to the complexity of spatial–temporal–spectral relationships.

To jointly utilize the spatial–temporal–spectral information within the image sets, a straightforward method is to apply the random masking strategy to different spectral combinations of the input image set. However, when input image sets have diverse combinations of spatial resolutions and temporal compositions, the random masking strategy may lead to feature leakage from the remaining high-resolution patches during the reconstruction of the removed low-resolution patches at the same position, resulting in shortcut learning during model pretraining. To this end, we propose the AAM to dynamically adjust the masking strategy of images for each input I_{in}, enabling training with images from diverse sources while preventing feature leakage. Specifically, we adopt a consistent masking strategy for images from different sources s at the same retrieval time t, a mutually exclusive masking strategy for images from the same s at different t, and a random masking strategy for the other circumstances. This strategy is designed to optimize feature learning for the unique spatial, temporal, and spectral diversity inherent in RS data, preventing the model from resorting to trivial reconstruction tasks. Adopting a single, uniform masking strategy, such as consistent masking across all scenarios, would be suboptimal as it fails to account for the multidimensional heterogeneity of RS data. While effective for leveraging spectral differences in same-time imagery, consistent masking applied across different timestamps would undermine the learning of temporal dynamics by allowing the exploitation of static spatial overlap. Besides, another key challenge in training on multisource RS data is the significant variation in the number of spectral bands across different sensors (e.g., 4 for Gaofen, 13 for Sentinel-2, 7–11 for Landsat-8). To handle this diversity while maintaining a unified model architecture, we employ a dynamic band selection strategy during pretraining. Specifically, for each training iteration involving an anchor image, we randomly select three spectral bands from the set of bands available in that anchor image. These three bands form the input channels for the model in that iteration. This approach serves to train a band-agnostic feature encoder. Instead of learning representations specific to a fixed band composition or requiring sensor-specific input layers, the model is compeled to learn robust relationships between spectral channels regardless of which particular triplet is presented as input. Over the course of extensive pretraining across numerous iterations, the model is exposed to a vast multitude of possible 3-band combinations available within the full spectral range of the anchor datasets. This dynamic sampling strategy offers significant benefits: it ensures architectural uniformity and simplicity, making A^2-MAE inherently adaptable to imagery from any multispectral sensor without needing sensor-specific parameters or complex band standardization. For a quantitative ablation study on AAM, please refer to Section V-C

In the example depicted in Fig. 3, three images is randomly sampled from an image set P_i, specifically, $I_{in} = \{I_i^{2020,Sen2}, I_i^{2020,Lan8}, I_i^{2022,Sen2}\}$. Taking the middle image $I_i^{2020,Sen2}$ as the referenced anchor, A^2-MAE explicitly aware the meta-information (i.e., source s and time t) to opt specific masking strategy for removing patches from the other two images (i.e., $I_i^{2020,Lan8}$ and $I_i^{2022,Sen2}$). We first randomly select three bands of the anchor image $I_i^{2020,Sen2}$ to encompass a substantial diversity of band compositions while balancing the computational costs. If an image in I_{in} differs in s but the same in t as the anchor image (i.e., the bottom image $I_i^{2020,Lan8}$), a consistent masking strategy is employed, obtaining a patch sequence $Seq_i^{2020,Lan8}$ where the patches are removed from the same position as those in $Seq_i^{2020,Sen2}$. This ensures that remaining patches maintain their coarsest version for the same source s, preventing the A^2-MAE from feature leakage during pretraining. To address temporal disparities, if an image in I_{in} has a different time t from the same s (i.e., the upper image $I_i^{2022,Sen2}$), a mutually exclusive masking strategy is adopted to ensure the removed patch retains positional uniqueness with $Seq_i^{2020,Sen2}$, enhancing A^2-MAE's capacity to leverage multitemporal symbiotic features. In addition, the incorporation of I_{in} offers sufficient diversity in spatiotemporal–spectral relationships, encouraging A^2-MAE to effectively leverage multiscale symbiotic features for patch reconstruction.

D. Geographic Encoding Module

The metadata stored with the RS images contains geographic information, including the latitude, longitude, and GSD. The latitude and longitude information indicate the absolute location of the retrieved image on the Earth, which is of significance in leveraging the geographic pattern in the pretraining on worldwide RS datasets. The GSD indicates the ground scale of the RS image, which is critical to understanding the spatial ranges and frequency specificity of the image. For example, an image with a low GSD has more details in high frequency than a high GSD image does. Therefore, we propose the GEM to explicitly incorporate essential geographic priors into the MAE model, thereby enhancing its generalization capabilities for downstream applications.

As illustrated in Fig. 4, given an RS image, the corresponding metadata records one GSD Geo_{GSD} and four sets of latitude and longitude (i.e., the four corners $(Geo^c_{Lat}, Geo^c_{Lon}), c \in \{TL, TR, BL, BR\}$). Instead of directly utilizing the decimal geographic information, the GEM views the RS image as a group of squared grids and encodes it to achieve better representative geo-encoding features. Let $\mathbb{G}$ be the set of grids

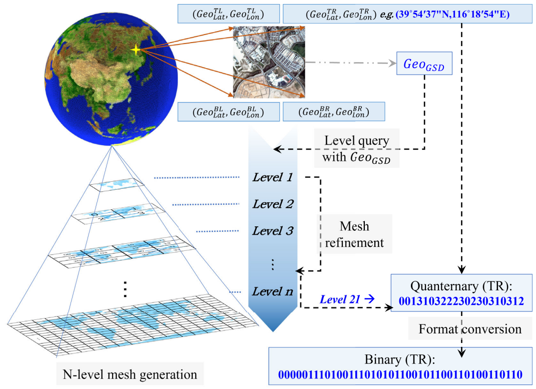

Fig. 4. Proposed GEM in A^2-MAE. By efficiently encoding the geographic metadata (i.e., a Geo_{GSD} and four sets of $(Geo^c_{Lat}, Geo^c_{Lon})$ in RS images, the GEM encourages A^2-MAE to be explicitly aware of this crucial geographic prior for varying geographic information in images.

formed with latitudes and longitudes. The $\mathbb{G}$ contains several levels of mesh in a pyramid way, which are composed of equally subdivided grids with integer coding. Let *Level 0* mesh $G_0 = \{g_0\} \in \mathbb{G}$ be a 512° × 512° grid which covers the whole global area. *Level 1* mesh $G_1 = \{g_1, \ldots, g_{n1}\} \in \mathbb{G}$ is defined as equally subdivided four grids, each of which has 256° × 256° in the height and width. Following this logic, *Level k* mesh G_k is obtained by a quadtree division of *Level k-1* mesh G_{k-1}. The higher level mesh represents finer resolution with low GSD. In all, given the metadata of an RS image, we first query the closest *Level k* according to its Geo_{GSD}, then four sets of latitude and longitude are embedded as sequences of binary arrays. For example, the Geo_{GSD} of Landsat-8 images is 30 m, which can be referred to grids in *Level 21* (32 m of the size of each grid in the equator). Considering that the Geo_{GSD} is utilized as an approximate version in this encoding strategy, we further encode the precise Geo_{GSD} by replacing the positional embedded vector with the ground-scaled positional encoding vector (inspired by [13]), which can be embedded as follows:

$$v_{gsd,x}(pos, 2i) = sin\frac{\widehat{Geo_{GSD}}}{Geo_{GSD}}\frac{pos}{10\,000^{\frac{2i}{D}}} \tag{1}$$

$$v_{gsd,y}(pos, 2i+1) = cos\frac{\widehat{Geo_{GSD}}}{Geo_{GSD}}\frac{pos}{10\,000^{\frac{2i}{D}}} \tag{2}$$

where *pos* is the position of the embedded patch along the given axis, *i* is the patch index, and *D* is the number of embedded dimensions, exactly as introduced in [56]. $\widehat{Geo_{GSD}}$ is the reference GSD (nominally set to 1 m).

As a result, the proposed GEM efficiently embeds the geographic metadata, providing unique embedded features for images with specific locations and GSDs.

V. Experiments

A. Implementation Details and Baselines

We adopt the ViT-Large architecture as the backbone for the proposed A^2-MAE, and pretrain the A^2-MAE using the constructed STSSD. We employ a progressive training strategy [43] by starting with S2-L8 data and then progressively transitioning to GF-S2 in STSSD. The patch size is fixed to 16 × 16 pixels. A^2-MAE is pretrained for 130 epochs with a batch size of 1024 on 8 NVIDIA A800 GPUs. AdamW optimizer [57] is utilized with an initial learning rate of 0.0001, coupled with a half-cycle cosine decay schedule. According to existing works [11], [12], we adopt a masking ratio of 75%, balancing training efficiency and pretext task difficulty. Eight SSL methods with the officially released pretrained weights are selected as competing methods in this study, including 2 ResNet-50-based methods (SeCo [19], CACo [20]) and 6 ViT-Large-based methods (vanilla MAE pretrained on ImageNet-1k [11], SatMAE (the version for spectral data) [12], and ScaleMAE [13]), Prithvi 2.0 [58], DOFA [59], and MA3E [44]. Full fine-tuning is employed in the downstream tasks for all methods, of which the first layer is modified to fit the data structure of specific datasets.

To ensure rigorous evaluation of model adaptability and feature quality, we conduct experiments using both full fine-tuning and linear probing protocols across downstream tasks. In full fine-tuning, all model parameters undergo optimization with modified input layers to align with specific dataset structures, enabling holistic adaptation while maintaining architectural consistency across compared methods. Conversely, linear probing evaluates the intrinsic pretrained features by freezing the backbone and training task-specific linear classifiers. All competing methods adopt the same head for each downstream dataset (specified in Table III).

As presented in Table IV, this study utilizes a range of downstream datasets and implements various tasks to evaluate the effectiveness of the proposed approach. To ensure a fair comparison, the competing methods adopt the officially released pretrained weights for each downstream task. The utilized datasets include AID (featuring diverse scene categories in a single year), BigEarthNet (spanning ten countries over two years), Sen1Floods11 (representing 14 biomes in a single year), CropSeg (focusing on the contiguous United States for a single year), LEVIR-CD (covering 20 districts over 5–14 years), and OSCD (encompassing 24 urbanized regions over three years). It's important to note that all of the utilized downstream datasets have a large-scale spatial coverage, diverse in spectral composition, and various temporal spanning, contributing to a comprehensive evaluation of the generalization capability of the competing methods in addressing the diversities in spatial, temporal, and spectral coverage. All the competing methods are trained and evaluated on the same partition of training and test materials for each downstream task.

For the implementation details, as summarized in Table III, each dataset is associated with specific implementation details, including the input size, input channel, optimizer (AdamW), learning rate, learning rate schedule (multistep), weight decay, batch size, maximum iterations/epoch, warmup strategy (linear), and task-specific head and loss function.

The proposed A^2-MAE approach exhibits versatility in accommodating various downstream tasks, as depicted in Fig. 5. By leveraging its flexibility, A^2-MAE can effectively address tasks such as scene classification, semantic segmentation, and change detection, each with distinct input combinations in terms of spectral, temporal, and spatial

TABLE III

IMPLEMENTATION DETAILS IN THE DOWNSTREAM TASK FINE-TUNING. LR. INDICATES THE LEARNING RATE. SCH. INDICATES THE SCHEDULER. ITER. INDICATES THE ITERATION

Dataset	AID	BigEarthNet	Sen1Floods11	CropSeg	LEVIR-CD	OSCD
Optimizer	AdamW	AdamW	AdamW	AdamW	AdamW	AdamW
Input size	600×600	128×128	512×512	512×512	256×256	96×96
Input channel	RGB	13 bands	6 bands	6 bands	RGB	RGB
Base LR.	1e-3	1e-3	1e-4	1e-4	3e-4	3e-4
LR. sch.	multi-step	multi-step	multi-step	multi-step	multi-step	multi-step
Weight decay	0.01	0.01	0.01	0.01	0.01	0.01
Batch size	64	256	32	16	96	96
Max epoch	100 epoch	100 epoch	100 epoch	100 epoch	200 epoch	100 epoch
Warmup	linear	linear	linear	linear	linear	linear
Head	Linear cls.	Linear cls.	UperNet	UperNet	ChangeFormer [60]	U-Net
Loss function	CrossEntropy	Soft-margin	BCE	BCE	BCE	BCE

TABLE IV

DETAILS OF THE DATASETS UTILIZED IN DOWNSTREAM TASKS

Dataset	Spatial coverage	Temporal coverage	Spectral coverage	Resolution
AID	mainly 7 countries	single year	RGB	0.5 to 8 m
BigEarthNet	10 countries	2 years	13 bands of Sentinel-2	10 m
Sen1Floods11	14 biomes	single year	13 bands of Sentinel-2	10 m
CropSeg	Contiguous United States	single year	7 bands of HLS	30 m
LEVIR-CD	20 districts	5-14 years	RGB	0.5 m
OSCD	24 urbanized regions	3 years	13 bands of Sentinel-2	10 m

TABLE V

COMPARISON RESULTS (%) OF CLASSIFICATION, SEGMENTATION, AND CHANGE DETECTION. THE VANILLA MAE MODEL WAS PRETRAINED ON IMAGENET-1K

Methods	Classification		Segmentation		Change detection		
	AID	BigEarthNet	Sen1Floods11	CropSeg	LEVIR-CD	OSCD	DynamicEarthNet
	Accuracy	mAP	mIoU	mIoU	mIoU	F1	mIoU
SeCo[19]	86.67	82.6	79.10	32.78	82.37	46.78	43.1
CACo[20]	86.76	82.1	84.62	32.83	82.90	51.74	41.2
Vanilla MAE[11]	86.96	80.1	85.42	44.51	80.41	34.74	39.8
SatMAE[12]	72.76	82.1	88.77	44.21	82.11	47.63	44.7
ScaleMAE[13]	86.90	82.2	85.05	44.04	83.07	48.70	44.3
Prithvi 2.0[58]	85.2	77.6	85.24	44.71	82.57	49.07	42.6
DOFA[59]	77.32	78.0	83.90	**44.94**	83.18	49.25	43.4
MA3E[44]	77.54	76.9	85.34	40.97	82.99	52.72	42.8
A^2-MAE	**87.08**	**83.0**	**88.87**	44.81	**84.32**	**53.97**	**46.0**

resolutions. This flexibility allows A^2-MAE to adapt to the specific requirements of each task, thereby enhancing its applicability across a wide range of RS applications. Whether the task demands spectral information, temporal sequences, or high-resolution spatial details, A^2-MAE can seamlessly integrate these input modalities to deliver accurate and robust results. Consequently, the proposed approach presents a versatile solution for RS practitioners, enabling them to tackle diverse challenges with a unified framework.

B. Comparison Results

We conduct experiments on seven datasets with diverse distributions in spatial, temporal, and spectral coverage, encompassing various data sources. This ensures a comprehensive assessment of the capabilities in efficiently utilizing spatial-, temporal-, and spectral-variant features, involving different downstream tasks, including classification, segmentation, and change detection. We employ the encoder of the competing pretrained models for all downstream tasks, and details of the training setups for fine-tuning of each task are included in Table III. As shown in Table V, A^2-MAE achieves comprehensive improvements across downstream tasks, indicating the effectiveness of A^2-MAE in exploiting the properties of RS images.

1) Land Cover Classification: We perform the scene classification task on AID [61] and the multilabel classification task on BigEarthNet [27]. AID is an aerial image dataset featuring diverse scene categories with higher spatial resolution (0.5–8 m) and RGB channels. BigEarthNet encompasses 590K Sentinel-2 images with 13 bands collected from ten countries. As presented in Table V, despite that several competing methods (i.e., SeCo, CACo, and SatMAE) are custom-tailored and pretrained on the Sentinel-2 image dataset, A^2-MAE still outperforms all competing methods in both AID and BigEarthNet datasets, highlighting the proposed A^2-MAE's effectiveness in leveraging diverse RS information within one unified model.

2) Semantic Segmentation: We perform experiments on Sen1Floods11 [62] and CropSeg [63]. Sen1Floods11 is

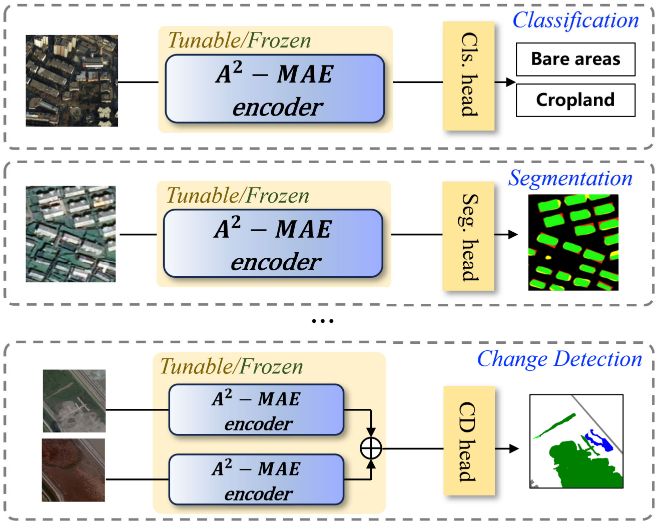

Fig. 5. Illustration of the proposed A^2-MAE to accommodate different downstream tasks. A^2-MAE is flexible in addressing downstream tasks with different input combinations.

a surface water segmentation dataset including 4831 Sentinel-2 imagery with 13 bands covering 120 406 km^2 and spans 14 biomes and six continents of the world across 11 flood events. CropSeg is a cropland segmentation dataset containing 3854 Harmonized Landsat-Sentinel imagery with 7 bands at 30-m resolution across the Contiguous United States. Given that SeCo and CACo are pretrained on datasets covering only urban regions, A^2-MAE, which is pretrained on STSSD covering diverse land cover types, achieves significant improvements by 9.77%/12.03% against SeCo and 4.25%/11.98% against CACo, highlighting its superior generalization ability when pretrained on STSSD with diverse coverage and geographical characteristics.

3) Change Detection: We conduct experiments on the LEVIR-CD [64], OSCD [65], and DynamicEarthNet datasets [29]. LEVIR-CD comprises 637 image pairs with a resolution of 0.5 m and a time span ranging from 5 to 14 years. The OSCD dataset comprises Sentinel-2 images with 13 bands collected from 24 urbanized regions worldwide. DynamicEarthNet provides daily images from Planet with 4 bands at 3-m resolution and monthly images from Sentinel-2 with 13 bands across approximately 75 areas of interest worldwide. Table V presents the quantitative evaluation results of baselines and A^2-MAE. A^2-MAE outperforms competing methods across various backbones and self-supervised architectures by 1.25% on mIoU/2.23% on F1/1.3% on mIoU against the second-best results in LEVIR-CD, OSCD, and DynamicEarthNet, respectively. These improvements demonstrate A^2-MAE's effectiveness in leveraging multitemporal RS images within a unified model.

We present cases of visualization comparisons in the downstream task, i.e., OSCD change detection. The visual comparisons of the downstream task in Fig. 6 clearly demonstrate the superior performance of the proposed A^2-MAE method compared to competing methods. A^2-MAE exhibits superior performance in accurately detecting changes between two images, as indicated by its results being more closely aligned with the ground truth labels compared to CACo, SatMAE, ScaleMAE, and SeCo. These visual comparisons underscore the effectiveness of A^2-MAE in producing high-quality outputs, highlighting its potential for advancing RS applications.

4) Comparison With Vision Foundation Model DINOv2: The proposed A^2-MAE approach demonstrates superior performance compared to competing vision foundation models across various RS tasks, as illustrated in Table VI. In classification tasks on AID and BigEarthNet datasets, A^2-MAE achieves the highest accuracy, outperforming the Vanilla MAE model pretrained on ImageNet-1k. Similarly, in segmentation tasks on Sen1Floods11 and CropSeg datasets, A^2-MAE achieves the highest mIoU scores, indicating superior segmentation quality compared to both Vanilla MAE [11] and DINOv2 [66]. Moreover, in change detection tasks on LEVIR-CD and OSCD datasets, A^2-MAE consistently outperforms competing models in terms of mIoU and F1 scores, demonstrating its effectiveness in detecting changes accurately and comprehensively. These comprehensive evaluation results underscore the superior performance of the proposed A^2-MAE approach across multiple RS tasks, highlighting its potential as a robust solution for various RS applications. On BigEarthNet, DINOv2 scores 81.2% mAP with 3-band inputs, trailing SeCo's ResNet-50 (82.6%) due to domain shift on pretraining datasets (natural versus RS imagery) and limited spectral sensitiveness. A^2-MAE, pretrained on STSSD with GEM and full-band flexibility, reaches 74.3% mAP (12-band), surpassing SeCo's 72.6% (3-band), highlighting its RS-specific design advantages.

C. Ablation Study

To efficiently and fairly investigate the key contributions of A^2-MAE, namely, AAM and the GEM, we conduct comprehensive ablation experiments. For these ablations, we utilize DynamicEarthNet as the pretraining corpus, training different configurations of A^2-MAE from scratch (same for the SatMAE and Van. MAE in the ablation experiments) on its designated training splits (55 locations), ensuring computational feasibility for rapidly iterating and testing numerous design variations and allowing to isolate the impact of our proposed modules from the sheer scale advantage of the full STSSD, thereby demonstrating their generality and robustness. The pretraining and fine-tuning phases for these ablation variants are conducted on the DynamicEarthNet training and testing locations, respectively. We also conduct comparisons in terms of masking strategy (i.e., random masking and tube masking strategies [12]) and geographic encoding method (i.e., one-hot geographic encoding [26] and scale encoding [13]). We further encode the geographic information using the proposed GEM during fine-tuning on DynamicEarthNet. This configuration (denoted as A^2-MAE$^+$) gives us a promising glance at the potential of fully utilizing the GEM when the geographic metadata of RS images is available in downstream RS tasks.

As shown in Table VII, A^2-MAE outperforms the SatMAE by 1.6/0.8 of Pix. Acc./mIoU, decoupling and showcasing the contributions of the proposed pretraining method. Besides, both the AAM (+6.6% on mIoU) and GEM (+0.9% on mIoU)

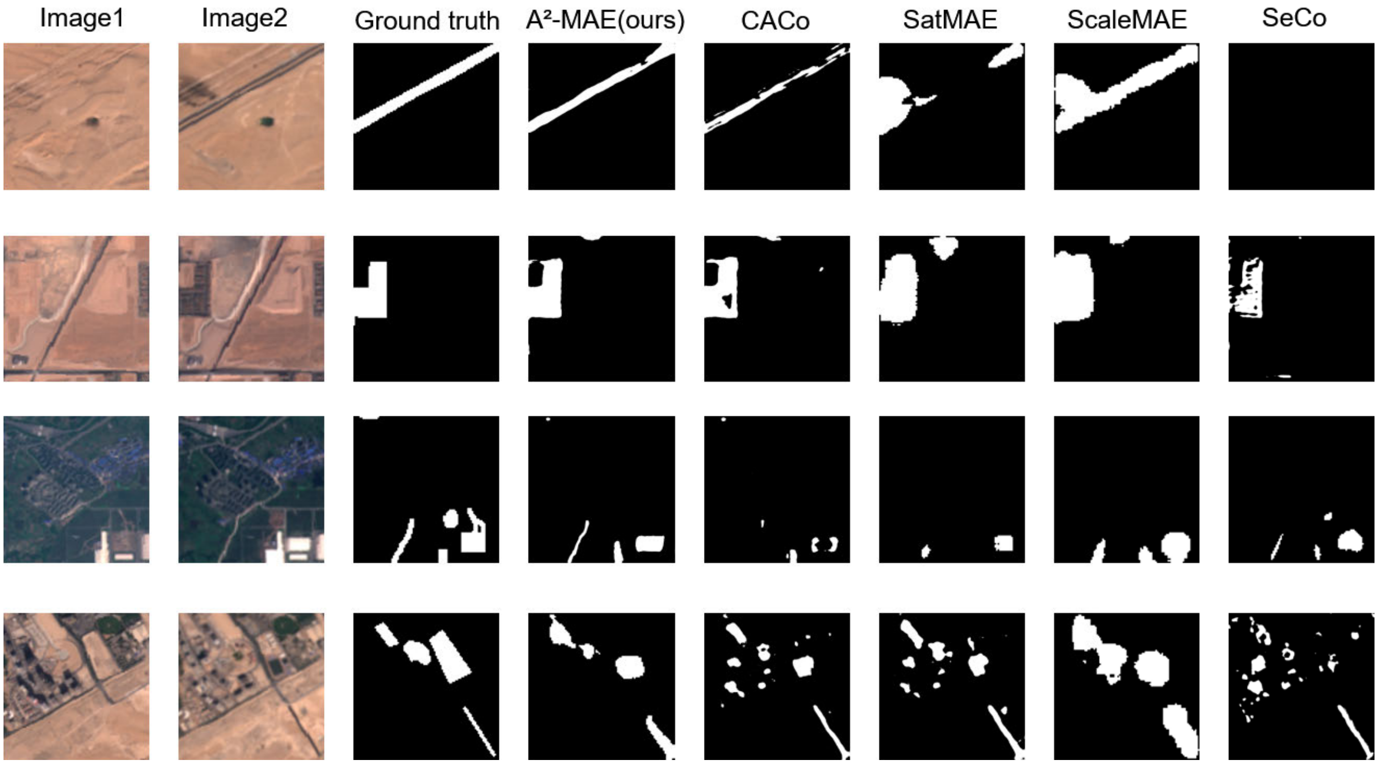

Fig. 6. Visualization of change detection results on OSCD change detection. The columns from left to right are input images (Image 1 and Image 2), ground truth labels, and results of A^2-MAE (ours), CACo [20], SatMAE [12], ScaleMAE [13], and SeCo [19], respectively.

TABLE VI

Comparison Results (%) in Classification, Segmentation, and Change Detection Tasks. The Vanilla MAE Model Is Pretrained on ImageNet-1k

Methods	Classification		Segmentation		Change detection		
	AID	BigEarthNet	Sen1Floods11	CropSeg	LEVIR-CD	OSCD	DynamicEarthNet
	Accuracy	mAP	mIoU	mIoU	mIoU	F1	mIoU
Vanilla MAE [11]	86.96	80.1	85.42	44.51	80.41	34.74	39.8
DINOv2 [66]	82.04	81.2	88.02	42.07	82.11	53.61	41.7
A^2-MAE (ours)	**87.08**	**83.0**	**88.87**	**44.81**	**84.32**	**53.97**	**46.0**

TABLE VII

Performance Comparison (%) Between Different Subcomponents of A^2-MAE on DynamicEarthNet Change Detection. All Models Utilize a Scratch Version of Themselves Pretrained on DynamicEarthNet and Further Fine-Tuned on DynamicEarthNet. A^2-MAE$^+$ Denotes Introducing the Geographic Encoding During Fine-Tuning

Pre-training	Pix. Acc.	mIoU
SatMAE [12]	72.2	45.9
Van. MAE (random masking)	69.2	40.2
+ tube masking [12]	67.4	38.6
+ AAM	72.7	46.8
+ One-hot geo-encoding[26]	71.9	47.1
+ scale encoding [13]	72.8	47.3
+ full GEM (A^2-MAE)	73.8	47.7
A^2-MAE$^+$	76.4	56.1

contribute to the significant performance improvements of A^2-MAE.

Specifically, for masking strategies, the proposed AAM outperforms random masking and tube masking strategies by 6.6% on mIoU, indicating the effectiveness of AAM. As a cornerstone of A^2-MAE, AAM is designed to tackle the challenges posed by heterogeneous RS data. Unlike conventional random masking, which risks feature leakage when applied to co-registered images from different sources (e.g., Sentinel-2 and Landsat-8 at the same location), AAM leverages meta-information from a preselected anchor image—such as spatial resolution or spectral bands—to dynamically tailor masking patterns. Our ablation study (see Table VII) quantifies AAM's impact. Specifically, it boosts the mIoU by +6.6% compared to random masking and +8.2% over tube masking on the DynamicEarthNet dataset. These gains highlight AAM's ability to harness cross-source complementarity effectively. By mitigating feature leakage and encouraging the model to learn robust spatiotemporal and spectral relationships, AAM significantly strengthens feature representation. For example, the model can use detailed Gaofen-2 patches to infer masked regions in Sentinel-2 images, enhancing reconstruction quality without increasing computational overhead.

For geographic encoding methods, comparisons against one-hot geographic encoding [26] show improvements by 1.9%/0.6% on Acc./mIoU. Further ablation studies reveal enhancements by 0.1%/0.5% on Acc./mIoU for GSD embedding and 1.0%/0.4% on Acc./mIoU for Lat/Lon embedding.

TABLE VIII

Performance Comparison (%) Between the Variants of A^2-MAE When Removing Progressive Training Strategy or Clustering-Based Pruning Technique. A^2-MAE-mix_pretraining Denotes an A^2-MAE Model Pretrained on a Variant of the STSSD Dataset, Where Data From All Sources (Gaofen-1, Gaofen-2, Sentinel-2, and Landsat-8) and Timestamps Were Combined and Mixed Together From the Beginning, Without Employing the Progressive Approach. A^2-MAE-no_pruning Denotes an A^2-MAE Model Pretrained Using the Full, Unpruned Collection of Nature Reserve Patches Available, in Addition to the Urban Patches, Without Applying the Diversity-Based Selection Strategy. The Vanilla MAE Model Is Pretrained on ImageNet-1k

Methods	Classification		Segmentation		Change detection		
	AID	BigEarthNet	Sen1Floods11	CropSeg	LEVIR-CD	OSCD	DynamicEarthNet
	Accuracy	mAP	mIoU	mIoU	mIoU	F1	mIoU
Vanilla MAE	86.96	80.1	85.42	44.51	80.41	34.74	39.8
A^2-MAE-mix_pretraining	86.96	77.8	83.62	44.24	82.86	50.74	43.1
A^2-MAE-no_pruning	**87.40**	78.7	81.48	43.97	83.95	50.01	44.2
A^2-MAE (full)	87.08	**83.0**	**88.87**	**44.81**	**84.32**	**53.97**	**46.0**

GEM complements AAM by embedding geographic metadata—latitude, longitude, and GSD—into learnable representations, enabling A^2-MAE to align its features with real-world spatial contexts. This is critical for RS tasks, where spatial variability is pronounced. For example, GSD embeddings allow the model to differentiate between fine-grained urban features (e.g., 0.8 m) and coarser agricultural landscapes (e.g., 30 m), while latitude and longitude embeddings capture biome-specific spectral signatures, such as those distinguishing tropical rainforests from boreal forests. AAM drives the model to reconstruct temporally varying patches—such as newly constructed buildings—by leveraging multisource inputs, while GEM enforces spatial consistency, reducing false positives in stable regions like forests or water bodies. This synergy manifests in A^2-MAE's superior performance across multiple downstream tasks, rigorously confirming the complementary roles of the proposed AAM and GEM.

Furthermore, by introducing the geographic information via the GEM in the fine-tuning phase, A^2-MAE$^+$ achieves a notable improvement of 8.4% on mIoU against A^2-MAE, indicating the promising potential of the GEM. It reveals that for downstream tasks that provide raw geographic metadata, introducing the GEM during fine-tuning can improve the results by a large margin. AAM facilitates efficient multisource learning by preventing feature leakage and maximizing cross-source information, while GEM enhances spatial awareness through geographic embeddings. Together, they enable a unified approach to spatial–temporal–spectral modeling, unlocking the pretrained A^2-MAE's full potential in RS downstream tasks.

To comprehensively investigate the necessity of our progressive training strategy and clustering-based pruning technique, we conduct two ablation experiments, pretrained on STSSD (denoted as A^2-MAE-mix_pretraining and A^2-MAE-no_pruning) and then fine-tuned on all the utilized downstream datasets. In the first experiment (denoted as A^2-MAE-mix_pretraining), we pretrain A^2-MAE on a variant of the STSSD dataset where data from all sources (i.e., Gaofen-1, Gaofen-2, Sentinel-2, and Landsat-8) and timestamps are combined and mixed together from the beginning, without employing the progressive approach, to quantify the difficulty of directly mixing highly heterogeneous data. In the second experiment (denoted as A^2-MAE-no_pruning), A^2-MAE is pretrained using the STSSD with a full, unpruned collection of nature reserve patches available, in addition to the urban patches, without applying the diversity-based selection strategy.

As reported in Table VIII, the A^2-MAE-mix variant performs noticeably worse than our proposed A^2-MAE, which uses the progressive strategy across most downstream tasks and datasets. For example, it shows significant performance drops on BigEarthNet mAP (77.8% versus 83.0%), Sen1Floods11 mIoU (83.62% versus 88.87%), OSCD F1 (50.74% versus 53.97%), and DynamicEarthNet mIoU (43.1% versus 46.0%), and slightly lower performance on others. The lower probing scores for models like Prithvi 2.0 highlight their task-specific strengths (e.g., temporal modeling on specific bands), which shine under fine-tuning but may not yield linearly separable features for tasks with unseen spectral compositions. These results suggest that directly training on the full, highly heterogeneous STSSD dataset from scratch poses significant challenges, likely due to the model struggling to simultaneously learn unified representations across vastly different spatial resolutions, spectral bands, and temporal dynamics present when all data is mixed directly. This empirical evidence supports the necessity of our progressive training strategy, which gradually introduces complexity, allowing the model to build foundational representations from more homogeneous subsets before adapting to the full heterogeneity of STSSD, thereby mitigating the issues observed with direct mixing and leading to superior downstream performance.

Comparing A^2-MAE-no_pruning to the full A^2-MAE (which utilizes the pruned 60 000 nature reserve patches), A^2-MAE (with pruning) generally achieves superior performance across most downstream tasks. Notably, pruning leads to improved results on BigEarthNet mAP (83.0% versus 78.7%), segmentation on Sen1Floods11 mIoU (88.87% versus 81.48%) and CropSeg mIoU (44.81% versus 43.97%), and change detection on LEVIR-CD mIoU (84.32% versus 83.95%), OSCD F1 (53.97% versus 50.01%), and DynamicEarthNet mIoU (46.0% versus 44.2%). While A^2-MAE-no_pruning shows a slightly higher accuracy on AID classification (87.40% versus 87.08%), the overall trend indicates that pretraining on the pruned subset of 60 000 diverse nature reserve samples is more effective for learning represen-

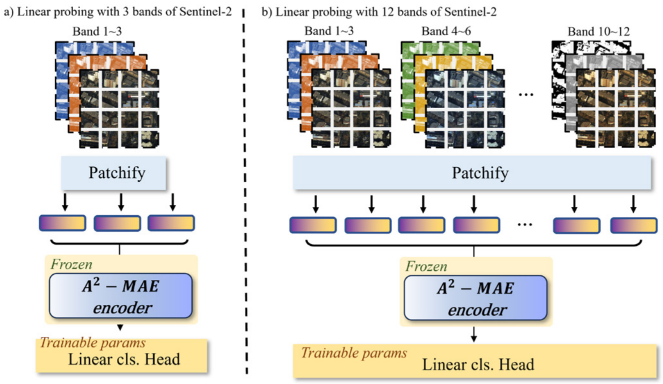

Fig. 7. Illustration of the usage of A^2-MAE in (a) 3 bands of Sentinel-2 and (b) 12 bands of Sentinel-2 images with the frozen backbone.

tations that transfer well to a wider range of downstream tasks compared to using the larger, unpruned set. This validates our clustering-based pruning strategy, confirming that focusing on a diverse subset, rather than simply maximizing sample count in natural landscapes, enhances the quality and transferability of the learned pretrained features.

D. Results of Linear Probing in Downstream Tasks

Experiments on the BigEarthNet dataset demonstrate the effectiveness of the A^2-MAE approach in leveraging pretrained features with frozen backbones for downstream tasks involving multispectral images. Specifically, A^2-MAE, pretrained on datasets encompassing diverse spectral compositions, exhibits its adaptability to both 3-band (RGB) and multiple-band inputs. To comprehensively evaluate A^2-MAE's representative capabilities on BigEarthNet, we provide two versions of linear probing results.

First, we evaluate A^2-MAE utilizing only three bands (RGB) of Sentinel-2 images [see Fig. 7(a)], enabling a fair comparison against competing methods trained solely on 3-band imagery. Second, we explore A^2-MAE's potential by incorporating all 12 bands of Sentinel-2 [see Fig. 7(b)], showcasing its ability to fully unlock pretrained representative features. In the latter case, we partition the input multispectral image into subimages, each comprising three bands, and concatenate the tokens generated from these patches. This innovative process allows the frozen model not only to handle the standard 3-band Sentinel-2 images but also the full 12-band of spectral information.

TABLE IX

Multilabel Classification Results (%) on BigEarthNet. The Mark of * Indicates the Version of the Frozen Model That Adopts All 12-Band of Sentinel-2 Images

Model	Input channels	Backbone	mAP
CACo [20]	B2,B3,B4 (RGB)	ResNet-50	72.6
SatMAE [12]	B2,B3,B4 (RGB)	ViT-Large	66.2
ScaleMAE [13]	B2,B3,B4 (RGB)	ViT-Large	73.2
A^2-MAE (ours)	B2,B3,B4 (RGB)	ViT-Large	73.5
SatMAE* [12]	B1-8,8A,9,11-12	ViT-Large	66.5
DOFA*	B1-8,8A,9,11-12	ViT-Large	57.5
MA3E*	B1-8,8A,9,11-12	ViT-Large	64.6
Prithvi 2.0*	B1-8,8A,9,11-12	ViT-Large	55.8
A^2-MAE* (ours)	B1-8,8A,9,11-12	ViT-Large	**74.3**

The comparison results presented in Table IX underscore A^2-MAE's superior performance in multilabel classification tasks using images with 3-band of images on BigEarthNet, achieving a mean average precision (mAP) of 73.5%. A^2-MAE outperforms competing models tailored for the standard 3-band images, including CACo [20], SatMAE [12], and ScaleMAE [13]. Notably, A^2-MAE's utilization of frozen backbone features extends beyond the conventional RGB channels to encompass a broader spectrum (all 12-band of spectral information) with a mAP of 74.3%, achieving 0.8% higher than the 3-band version. The gained improvement is higher than that of SatMAE (+0.3% of mAP when utilizing all 12-band of Sentinel-2 images). This capability enhances its adaptability to multispectral imagery, emphasizing the robustness and versatility of A^2-MAE in effectively harnessing images with diverse spectral combinations using the pretrained features.

Beyond full fine-tuning, A^2-MAE excels as a frozen foundation model. Linear probing with a frozen backbone yields 74.3% mAP on BigEarthNet (12-band input), surpassing SatMAE (66.2%) and ScaleMAE (73.2%), and achieves competitive mIoU/F1 scores on Sen1Floods11 (88.87%) and OSCD (53.97%) against DINOv2. These results affirm its readiness for efficient, adaptation-free deployment in RS applications.

Furthermore, by exposing the model to diverse spectral subsets, it encourages the learning of features that generalize well to downstream tasks, which may utilize different band combinations or the full spectrum of a sensor. As demonstrated by our strong performance on various downstream benchmarks, including those utilizing more than 3 bands (e.g., 12 bands for BigEarthNet, 13 for Sen1Floods11), demonstrating that A^2-MAE can effectively leverage the full spectral richness of the data to learn highly transferable representations.

As reported in Table X, full linear probing using 3-band (RGB) on downstream tasks reveals A^2-MAE's robust features (e.g., highest frozen mIoU on Sen1Floods11 at 88.04%), outperforming competitors by leveraging STSSD's diversity. Fine-tuning gaps (e.g., +4.9% mIoU for A^2-MAE on DynamicEarthNet) highlight adaptation benefits, with A^2-MAE consistently leading. Compared to the supervised baseline (we utilize a ResNet-50 as backbone with the same heads as Table III from random initialization on each dataset's training split, using identical hyperparameters for fairness.), A^2-MAE's fine-tuning and linear probing results demonstrate SSL's advantages in leveraging a small number of labeled data for robust RS representations. For datasets with rich spatial–temporal–spectral information, such as Sen1Floods11, CropSeg, and DynamicEarthNet, the features learned by A^2-MAE are highly effective, allowing the linear probing result of the proposed A^2-MAE to surpass the supervised baseline. This demonstrates the power of pretraining on complex, multidimensional data. Conversely, for datasets that are comparatively simpler in their spatiotemporal and spectral dimensions, such as AID and LEVIR-CD, the fully supervised ResNet-50 baseline now shows stronger performance than all linear probing results. It reveals that a smaller, supervised

TABLE X

LINEAR PROBING (FROZEN BACKBONE) RESULTS AND SUPERVISED LEARNING BASELINE (%) ACROSS CLASSIFICATION, SEGMENTATION, AND CHANGE DETECTION TASKS

	Classification		Segmentation		Change detection		
Methods	AID	BigEarthNet	Sen1Floods11	CropSeg	LEVIR-CD	OSCD	DynamicEarthNet
	Accuracy	mAP	mIoU	mIoU	mIoU	F1	mIoU
SeCo[19]	40.78	**74.5**	65.54	25.51	47.47	37.38	38.2
CACo[20]	36.70	72.6	65.94	25.89	47.45	34.95	37.5
Vanilla MAE[11]	37.74	74.4	79.19	38.61	39.01	40.47	35.1
SatMAE[12]	38.72	66.5	77.70	35.42	52.39	36.00	40.3
ScaleMAE[13]	47.58	73.2	80.49	38.65	60.51	40.63	39.8
Prithvi 2.0[58]	50.54	55.8	80.34	**43.25**	60.77	25.72	38.4
DOFA[59]	35.70	57.5	82.73	38.48	**62.59**	26.18	39.1
MA3E[44]	44.68	64.6	79.72	42.23	58.51	30.73	37.9
A^2-MAE	**55.16**	74.3	**88.04**	39.33	62.47	**41.49**	**41.1**
Supervised learning baseline	83.04	79.2	81.44	33.62	81.65	46.20	40.2
A^2-MAE (full-fintune)	87.08	83.0	88.87	44.81	84.32	53.97	46.0

model can become highly specialized and fully optimized for a single, less complex task, whereas linear probing only evaluates the preexisting linear separability of general-purpose features from a large foundation model without fine-tuning the backbone. This highlights a known tradeoff between the powerful, general representations of a foundation model and the task-specific optimization of a smaller supervised model. While distinct model paradigms prioritize different objectives and thus exhibit varying advantages under linear evaluation, A^2-MAE demonstrates consistently superior or competitive performance across the entire benchmark.

VI. CONCLUSION

In this study, we introduce STSSD, a spatial–temporal–spectral structured dataset comprising 510 000 of sampling locations with 2.5 million structured images collected from multiple RS sources. To exploit different kinds of multispectral sources in one unified backbone, we propose an AAM strategy to harness the intrinsic complementary information from different kinds of images, thus achieving more powerful feature representations. Furthermore, we propose the GEM to leverage geographic information, thereby improving the model's generalization ability. Experiments verify the effectiveness and advantages of our method compared to existing RS pretraining models with the same parameter amount across image classification, semantic segmentation, and change detection tasks. In future work, we will expand the diversity of modalities such as synthetic aperture radar and hyperspectral images in STSSD and A^2-MAE.

REFERENCES

[1] C. Small, F. Pozzi, and C. Elvidge, "Spatial analysis of global urban extent from DMSP-OLS night lights," *Remote Sens. Environ.*, vol. 96, nos. 3–4, pp. 277–291, Jun. 2005.

[2] Y. Zhou et al., "A global map of urban extent from nightlights," *Environ. Res. Lett.*, vol. 10, no. 5, May 2015, Art. no. 054011.

[3] X. Huang, A. Schneider, and M. A. Friedl, "Mapping sub-pixel urban expansion in China using MODIS and DMSP/OLS nighttime lights," *Remote Sens. Environ.*, vol. 175, pp. 92–108, Mar. 2016.

[4] M. Imhoff, "Using nighttime DMSP/OLS images of city lights to estimate the impact of urban land use on soil resources in the United States," *Remote Sens. Environ.*, vol. 59, no. 1, pp. 105–117, Jan. 1997.

[5] W. Turner, "Sensing biodiversity," *Science*, vol. 346, no. 6207, pp. 301–302, 2014.

[6] R. Wang and J. A. Gamon, "Remote sensing of terrestrial plant biodiversity," *Remote Sens. Environ.*, vol. 231, Sep. 2019, Art. no. 111218.

[7] Q. Li, L. Mou, Y. Hua, Y. Shi, and X. X. Zhu, "Building footprint generation through convolutional neural networks with attraction field representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5609017.

[8] R. Dong et al., "Large-scale land cover mapping with fine-grained classes via class-aware semi-supervised semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 16737–16747.

[9] Y. Wang, C. M. Albrecht, N. A. A. Braham, L. Mou, and X. X. Zhu, "Self-supervised learning in remote sensing: A review," 2022, *arXiv:2206.13188*.

[10] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9729–9738.

[11] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16000–16009.

[12] Y. Cong et al., "SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 197–211.

[13] C. J. Reed et al., "Scale-MAE: A scale-aware masked autoencoder for multiscale geospatial representation learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 4088–4099.

[14] F. Yao et al., "RingMo-sense: Remote sensing foundation model for spatiotemporal prediction via spatiotemporal evolution disentangling," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5620821.

[15] Y. Long et al., "On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and million-AID," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4205–4230, 2021.

[16] Y. Long, G.-S. Xia, L. Zhang, G. Cheng, and D. Li, "Aerial scene parsing: From tile-level scene classification to pixel-wise semantic labeling," 2022, *arXiv:2201.01953*.

[17] M. Schmitt, L. H. Hughes, C. Qiu, and X. X. Zhu, "SEN12MS—A curated dataset of georeferenced multi-spectral Sentinel-1/2 imagery for deep learning and data fusion," 2019, *arXiv:1906.07789*.

[18] Y. Wang, N. A. A. Braham, Z. Xiong, C. Liu, C. M Albrecht, and X. Xiang Zhu, "SSL4EO-s12: A large-scale multi-modal, multi-temporal dataset for self-supervised learning in Earth observation," 2022, *arXiv:2211.07044*.

[19] O. Manas, A. Lacoste, X. Giro-i-Nieto, D. Vazquez, and P. Rodriguez, "Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9414–9423.

[20] U. Mall, B. Hariharan, and K. Bala, "Change-aware sampling and contrastive learning for satellite images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 5261–5270.

[21] X. Guo et al., "SkySense: A multi-modal remote sensing foundation model towards universal interpretation for Earth observation imagery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 27662–27673.

[22] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[23] T. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[24] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, Jan. 2015.

[25] S. Abu-El-Haija et al., "YouTube-8M: A large-scale video classification benchmark," 2016, *arXiv:1609.08675*.

[26] G. Christie, N. Fendley, J. Wilson, and R. Mukherjee, "Functional map of the world," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6172–6180.

[27] G. Sumbul, M. Charfuelan, B. Demir, and V. Markl, "BigearthNet: A large-scale benchmark archive for remote sensing image understanding," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2019, pp. 5901–5904.

[28] G. Sumbul et al., "BigEarthNet-MM: A large-scale, multimodal, multilabel benchmark archive for remote sensing image classification and retrieval [software and data sets]," *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 3, pp. 174–180, Sep. 2021.

[29] A. Toker et al., "DynamicEarthNet: Daily multi-spectral satellite dataset for semantic change segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 21158–21167.

[30] F. Bastani, P. Wolters, R. Gupta, J. Ferdinando, and A. Kembhavi, "SatlasPretrain: A large-scale dataset for remote sensing image understanding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 16772–16782.

[31] Z. Tong, Y. Song, J. Wang, and L. Wang, "VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 10078–10093.

[32] S. Zhang, H. Chen, H. Yang, X. Sun, P. S. Yu, and G. Xu, "Graph masked autoencoders with transformers," 2022, *arXiv:2202.08391*.

[33] R. Bachmann, "Multimae: Multi-modal multi-task masked autoencoders," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2022, pp. 348–367.

[34] S. Woo et al., "ConvNeXt v2: Co-designing and scaling ConvNets with masked autoencoders," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 16133–16142.

[35] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[36] J.-B. Grill et al., "Bootstrap your own latent-a new approach to self-supervised learning," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 21271–21284.

[37] X. Chen and K. He, "Exploring simple Siamese representation learning," in *Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognit.*, Jun. 2021, pp. 15750–15758.

[38] M. Caron et al., "Emerging properties in self-supervised vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9650–9660.

[39] K. Ayush et al., "Geography-aware self-supervised learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10181–10190.

[40] L. Zhang, Z. Ren, B. Chen, P. Gong, B. Xu, and H. Fu, "A prolonged artificial nighttime-light dataset of China (1984–2020)," *Sci. Data*, vol. 11, no. 1, p. 414, Apr. 2024.

[41] Q. Li, L. Mou, Y. Shi, and X. X. Zhu, "BANet: A bilateral attention network for extracting changed buildings between remote sensing imagery and cadastral maps," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 139, May 2025, Art. no. 104486.

[42] R. Dong et al., "Building bridges across spatial and temporal resolutions: Reference-based super-resolution via change priors and conditional diffusion model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 27684–27694.

[43] D. Hong et al., "SpectralGPT: Spectral remote sensing foundation model," 2024, *arXiv:2311.07113*.

[44] Z. Li et al., "Masked angle-aware autoencoder for remote sensing images," in *Proc. Eur. Conf. Comput. Vis.*Cham, Switzerland: Springer, 2025, pp. 260–278.

[45] X. Wu, Z. Shi, and Z. Zou, "A geographic information-driven method and a new large scale dataset for remote sensing cloud/snow detection," *ISPRS J. Photogramm. Remote Sens.*, vol. 174, pp. 87–104, Apr. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271621000290

[46] W. Li, K. Chen, H. Chen, and Z. Shi, "Geographical knowledge-driven representation learning for remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5405516.

[47] D. Zhao, Q. Wang, J. Zhang, and C. Bai, "Mine diversified contents of multispectral cloud images along with geographical information for multilabel classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023.

[48] G. Chu et al., "Geo-aware networks for fine-grained recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 247–254.

[49] K. Klemmer, E. Rolf, C. Robinson, L. Mackey, and M. Rußwurm, "SatCLIP: Global, general-purpose location embeddings with satellite imagery," in *Proc. AAAI Conf. Artif. Intell.*, 2023, pp. 4347–4355.

[50] L. Bai et al., "Geographic mapping with unsupervised multi-modal representation learning from VHR images and POIs," *ISPRS J. Photogramm. Remote Sens.*, vol. 201, pp. 193–208, Jul. 2023.

[51] L. Scheibenreif, J. Hanna, M. Mommert, and D. Borth, "Self-supervised vision transformers for land-cover segmentation and classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 1421–1430.

[52] M. Rußwurm, K. Klemmer, E. Rolf, R. Zbinden, and D. Tuia, "Geographic location encoding with spherical harmonics and sinusoidal representation networks," 2023, *arXiv:2310.06743*.

[53] H. C. Bingham et al., "Sixty years of tracking conservation progress using the world database on protected areas," *Nature Ecol. Evol.*, vol. 3, no. 5, pp. 737–743, Apr. 2019.

[54] P. R. Elsen, W. B. Monahan, and A. M. Merenlender, "Reply to you et al.: The world database on protected areas is an invaluable resource for global conservation assessments and planning," *Proc. Nat. Acad. Sci. USA*, vol. 115, no. 39, pp. E9029–E9030, Sep. 2018.

[55] B. Sorscher, R. Geirhos, S. Shekhar, S. Ganguli, and A. Morcos, "Beyond neural scaling laws: Beating power law scaling via data pruning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 19523–19536.

[56] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. neural Inf. Process. Syst.*, vol. 30, 2025, pp. 5998–6008. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/11296896/

[57] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[58] D. Szwarcman et al., "Prithvi-EO-2.0: A versatile multi-temporal foundation model for Earth observation applications," 2024, *arXiv:2412.02732*.

[59] Z. Xiong et al., "Neural plasticity-inspired multimodal foundation model for Earth observation," 2024, *arXiv:2403.15356*.

[60] W. G. C. Bandara and V. M. Patel, "A transformer-based Siamese network for change detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2022, pp. 207–210.

[61] G.-S. Xia et al., "AID: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, Jul. 2017.

[62] D. Bonafilia, B. Tellman, T. Anderson, and E. Issenberg, "Sen1Floods11: A georeferenced dataset to train and test deep learning flood algorithms for Sentinel-1," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 835–845.

[63] I. N. Geospatial. *IBM-NASA-Geospatial/Multi-Temporal-Crop-Classification. Datasets At Hugging Face*. Accessed: Nov. 7, 2023. [Online]. Available: https://huggingface.co/datasets/ibm-nasa-geospatial/multi-temporal-crop-classification

[64] H. Chen and Z. Shi, "A spatial–temporal attention-based method and a new dataset for remote sensing image change detection," *Remote Sens.*, vol. 12, no. 10, p. 1662, May 2020.

[65] R. C. Daudt, B. Le Saux, A. Boulch, and Y. Gousseau, "Urban change detection for multispectral Earth observation using convolutional neural networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2018, pp. 2115–2118.

[66] M. Oquab et al., "DINOv2: Learning robust visual features without supervision," 2023, *arXiv:2304.07193*.