SpectralGPT: Spectral Foundation Model

Danfeng Hong, *Senior Member, IEEE,* Bing Zhang, *Fellow, IEEE,* Xuyang Li, Yuxuan Li, Chenyu Li, Jing Yao, *Member, IEEE,* Naoto Yokoya, *Member, IEEE,* Hao Li, Pedram Ghamisi, *Senior Member, IEEE,* Xiuping Jia, *Fellow, IEEE,* Antonio Plaza, *Fellow, IEEE,* Gamba Paolo, *Fellow, IEEE,* Jon Atli Benediktsson, *Fellow, IEEE,* Jocelyn Chanussot, *Fellow, IEEE*

***Abstract*—The foundation model has recently garnered significant attention due to its potential to revolutionize the field of visual representation learning in a self-supervised manner. While most foundation models are tailored to effectively process RGB images for various visual tasks, there is a noticeable gap in research focused on spectral data, which offers valuable information for scene understanding, especially in remote sensing (RS) applications. To fill this gap, we created for the first time a universal RS foundation model, named SpectralGPT, which is purpose-built to handle spectral RS images using a novel 3D generative pretrained transformer (GPT). Compared to existing foundation models, SpectralGPT 1) accommodates input images with varying sizes, resolutions, time series, and regions in a progressive training fashion, enabling full utilization of extensive RS big data; 2) leverages 3D token generation for spatial-spectral coupling; 3) captures spectrally sequential patterns via multi-target reconstruction; 4) trains on one million spectral RS images, yielding models with over 600 million parameters. Our evaluation highlights significant performance improvements with pretrained SpectralGPT models, signifying substantial potential in advancing spectral RS big data applications within the field of geoscience across four downstream tasks: single/multi-label scene classification, semantic segmentation, and change detection.**

***Index Terms*—Artificial intelligence, deep learning, downstream, foundation model, tensor masked modeling, progressive, remote sensing, spectral data, transformer.**

This work was supported by the National Key Research and Development Program of China under Grant 2022YFB3903401, the National Natural Science Foundation of China under Grant 42241109, Grant 42271350, and Grant 62201553, the MIAI@Grenoble Alpes (ANR-19-P3IA-0003), and the AXA Research Fund. (*Corresponding author: Bing Zhang*)

D. Hong, X. Li and Y. Li are with the Aerospace Information Research Institute, Chinese Academy of Sciences, 100094 Beijing, China, and also with the School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, 100094 Beijing, China. (e-mail: hongdf@aircas.ac.cn)

B. Zhang is with the Aerospace Information Research Institute, Chinese Academy of Sciences, 100094 Beijing, China, and with the College of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100049, China. (e-mail: zb@radi.ac.cn)

C. Li is with the School of Mathematics and Statistics, Southeast University, 211189 Nanjing, China, and also with the Aerospace Information Research Institute, Chinese Academy of Sciences, 100094 Beijing, China. (e-mail: lichenyu@seu.edu.cn)

J. Yao is with the Aerospace Information Research Institute, Chinese Academy of Sciences, 100094 Beijing, China. (e-mail: yaojing@aircas.ac.cn)

N. Yokoya is with the Department of Complexity Science and Engineering, Graduate School of Frontier Sciences, the University of Tokyo, Chiba 277–8561, Japan. (e-mail: yokoya@k.u-tokyo.ac.jp)

H. Li is with the Big Geospatial Data Management, Technical University of Munich, Munich 85521, Germany. (e-mail: hao_bgd.li@tum.de)

P. Ghamisi is with the Helmholtz-Zentrum Dresden-Rossendorf (HZDR), Helmholtz Institute Freiberg for Resource Technology, 09599 Freiberg, Germany, and is also with the Institute of Advanced Research in Artificial Intelligence (IARAI), 1030 Vienna, Austria (e-mail: p.ghamisi@gmail.com).

X. Jia is with the School of Engineering and Information Technology, University of New South Wales, Canberra, ACT 2612, Australia.

A. Plaza is with the Hyperspectral Computing Laboratory, Department of Technology of Computers and Communications, Escuela Politécnica, University of Extremadura, 10003 Cáceres, Spain. (e-mail: aplaza@unex.es)

P. Gamba is with the Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Pavia 27100, Italy. (e-mail: paolo.gamba@unipv.it)

J. Benediktsson is with the Faculty of Electrical and Computer Engineering, University of Iceland, Reykjavik 101, Iceland. (e-mail: benedikt@hi.is)

J. Chanussot is with the Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-Lab, 38000 Grenoble, France, and also with the Aerospace Information Research Institute, Chinese Academy of Sciences, 100094 Beijing, China. (e-mail: jocelyn@hi.is)

I. Introduction

SPECTRAL imaging is capable of capturing a vast array of spectral information, thereby enabling highly accurate analysis and recognition of objects and scenes beyond what is possible with RGB data alone [1]–[3]. This has made multi/hyper-spectral (MS/HS) remote sensing (RS) data the preferred tool of choice and a key component in a wide range of Earth Observation (EO) applications [4]–[6], including land use/land cover mapping, ecosystem monitoring, weather forecasting, energy resource development, biodiversity conservation, and geological exploration.

The rapid expansion in the availability and accessibility of spectral data from RS satellite missions, such as Landsat-8/9, Sentinel-2, Gaofen-1/2/6, and others, has further opened up opportunities for new discoveries and advancements in fields related to EO [7], [8]. Nevertheless, this growth also gives rise to two challenging difficulties that require prompt attention and effective solutions.

- **Limited information extraction and mining capabilities from massive spectral data.** The existing expert-centric and data-driven models have reached their limits and are insufficient for effectively learning visual representations from such vast amounts of spectral RS data. There is an urgent need to create new-generation models to improve the intelligent processing and analysis capabilities of spectral RS big data to a level that matches its volume.
- **Limited prediction and interpretation abilities for downstream EO tasks on a few label and label-free cases.** Compared to the availability of spectral RS data, there is a scarcity of corresponding labels at both the pixel and image levels. This shortage of labeled data hinders the application of full supervision in deep learning and AI models for practical EO tasks. Urgent action is needed to create RS foundation models embodied with spectral knowledge.

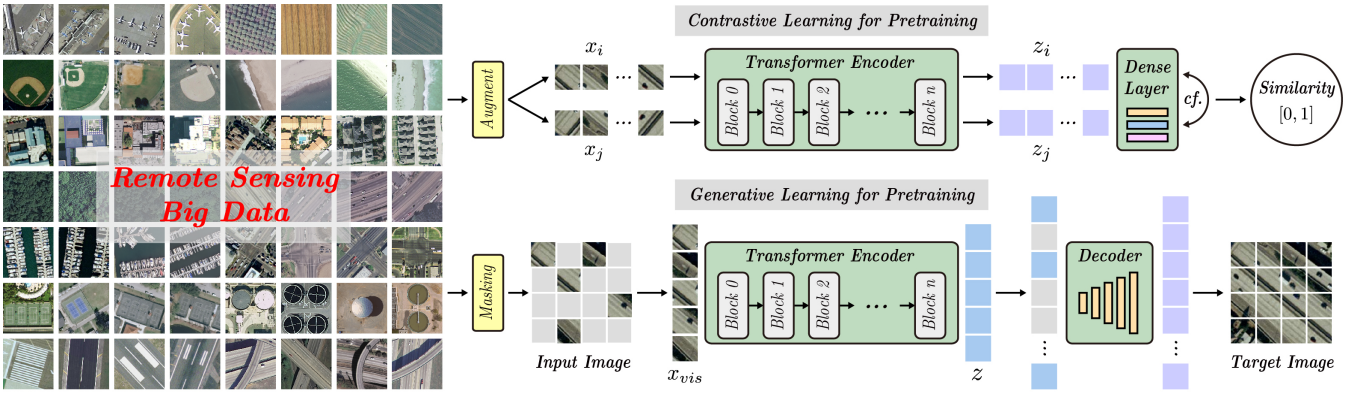

Fig. 1. An illustration to clarify the differences between contrastive learning and generative learning for pretraining in terms of the RS foundation model.

The development of pretraining techniques based on foundational models [9] is currently experiencing a remarkable surge, driven by advancements in self-supervised learning techniques [10] and the transformative capabilities of transformer-based methods [11]. Notably, this surge is particularly evident in the domains of natural language processing and computer vision. Pretraining agent tasks are usually divided into contrastive learning [12] and generative learning [13]. As the name suggests, the former aims to teach the model to differentiate between similar and dissimilar examples, while the latter focuses on training a model to generate new data or recover complete data from partial observations. Their differences are illustrated in Fig. 1. Two representative frameworks in contrastive learning are momentum contrast (MoCo) [14] and simple contrastive learning (SimCLR) [15]. MoCo introduces momentum updates to improve the contrastive learning process, while SimCLR leverages data augmentations to enhance the variety and complexity of the image pairs used for contrastive learning. There have been numerous variants of the MoCo and SimCLR frameworks developed since their initial proposals. These variants aim to address specific challenges or limitations of the original frameworks or to further improve their performance. For example, some variants of SimCLR have incorporated new types of data augmentations or improved the training objective [16]–[18], while some variants of MoCo have explored different momentum update strategies or used additional losses to improve contrastive learning [19]–[21]. Accompanied by the rise of vision transformers (ViT) [22], there has been significant progress in generative learning based on masked image modeling (MIM) for visual pretraining tasks. Bidirectional encoder representation from image transformers (BEiT), as presented in [23], is a prominent example of a MIM architecture built on top of the ViT. MIM allows for the input of all image patches, which provides flexibility for adapting to various network architectures. Yet the high computational cost associated with MIM can to some extent limit its practical use in certain applications. He *et al.* [24] proposed masked autoencoders (MAE) as a particular alternative to MIM. In MAE, unmasked patches or pixels are used to reconstruct those that are masked. This approach is computationally more efficient and also enhances the inference ability of the pretrained models, thereby making it more practical for various applications.

However, these advanced models have been relatively underexplored in RS. Wang *et al.* [25] trained a plain vision transformer with 100 million parameters on RGB images towards RS task design and developed a new rotated varied-size window attention mechanism for fine-tuning the model on downstream tasks. Unlike MAE-based methods that rely on only a few visible image patches to infer the entire image, Sun *et al.* [26] considered all image patches, whether masked or unmasked, by implementing a MIM strategy in their RS pretraining models. Despite the increased computational cost and the reduction in inference efficiency, MIM allows for the flexible use of various deep architectures as network backbones, such as ViT and Swin transformers [27]. The success of these two initial studies indicates the significant potential of pretrained models for applications in RS. The rapid progress in imaging spectroscopy has solidified the significance of spectral RS in EO. This prominence arises from its unique ability to effectively utilize the wealth of spectral information available. However, existing RS foundation models encounter challenges when applied to spectral data due to their limited capacity to model multi-band data. The specific gaps between spectral data and existing foundation models can be summarized as follows.

- **Gap 1 (*cf.* existing RS foundation models):** They often struggle to capture spatial-spectral representations inherent in 3D tensor data. A majority of these models are predominantly designed for processing data resembling RGB imagery, which constrains their capability to fully capture and characterize spectral information. Consequently, their applicability to handle such data types remains constrained.
- **Gap 2 (*cf.* foundation models for video data):** There have been foundation models designed for video data in computer vision [28], [29], yet there are significant differences between video data and spectral data. The main distinctions lie in the varying content between continuous frames in videos and the redundancy that often exists between all frames. As a result, the network designs pretrained for video data are often not well-suited for spectral data.
- **Gap 3 (*cf.* foundation models for spectral data):** Indeed, recent research about foundation models for

spectral data has been relatively scarce. Only one conference paper, namely SatMAE [30], has delved into the utilization of pretrained transformers, e.g., MAE, for spectral satellite images. SatMAE's central design approach involves grouping adjacent spectral bands, akin to RGB bands. However, this practice inadvertently disrupts spectral continuity, leading to a suboptimal capture of 3D spatial-spectral coupling traits and spectrally sequential data. Moreover, constraints pertaining to the number of pretraining samples and effective training strategies have further impeded performance enhancements in this context.

To fill these gaps, we devise SpectralGPT, a groundbreaking RS foundation model meticulously tailored for spectral data. SpectralGPT features pioneering elements, such as a 3D masking strategy, an encoder for learning visual representations from spatial-spectral mixed tokens, and a decoder with multi-target reconstruction for preserving spectrally sequential characteristics. These innovations significantly enhance SpectralGPT's ability to learn intrinsic knowledge representations from spectral data, providing valuable insights for scene understanding in various RS downstream applications. Fig. 2 illustrates a visual overview of SpectralGPT's pretraining and its versatile application in diverse downstream tasks, underscoring its profound contributions.

- **Customized foundation model for spectral data:** SpectralGPT is the first purpose-built foundation model designed explicitly for spectral RS data. SpectralGPT considers unique characteristics of spectral data, i.e., spatial-spectral coupling and spectral sequentiality, in the MAE framework with a simple yet effective 3D GPT network.
- **Large-scale training data:** SpectralGPT is trained on an extensive dataset derived from the Sentinel-2 satellite with over one million spectral images. This effort culminates in the creation of three distinct model iterations–Base, Large, and Huge–comprising approximately 100 million, 300 million, and 600 million parameters, respectively.
- **Flexibility in pretraining:** SpectralGPT employs a progressive training strategy, enabling it to process input images with varying sizes, resolutions, time series, and geographical regions. This innovative design exposes the model's encoder to a diverse array of information, ultimately enhancing its capability to represent a wide range of features effectively.
- **Advanced 3D masking and reconstruction:** SpectralGPT leverages a 3D tensor-shaped spatial-spectral mask with a masking rate of at least 90% on spectral RS data. Additionally, it employs a groundbreaking multi-target reconstruction strategy to comprehensively capture locally spatial-spectral characteristics and spectrally sequential information. These innovations substantially improve the model's learning capabilities through inference.
- **Superior performance across downstream tasks:** SpectralGPT's impact extends to downstream RS models, where it outperforms existing state-of-the-art (SOTA) competitors across various tasks, including single/multi-label scene classification, semantic segmentation, and change detection.
- **New benchmark dataset:** We curate a new benchmark dataset, named SegMunich, which focuses on urban areas and their adjacent neighborhoods within Munich City, Germany. This dataset is designed to cater to the requirements of semantic segmentation tasks with 13 classes to facilitate downstream analysis.

II. The Proposed SpectralGPT

A. A Brief Recall of MAE

MAE is a simple autoencoding method [31] that enables the reconstruction of the original signal. Like all autoencoders, MAE includes an encoder that maps the observed signal to a potential representation and a decoder that reconstructs the original signal from the potential representation. However, in contrast to classical autoencoders, MAE uses an asymmetric design that enables the encoder to operate only on partial and observed signals (without mask tokens). Additionally, MAE employs a lightweight decoder to reconstruct the complete signal from potential representations and mask tokens.

In detail, the implementation process of the MAE can be broken down into the following steps:

Step 1. Given the input image $\boldsymbol{x}$ with $H \times W$ pixels by C dimensions, the strategy in ViT is adapted to divide it into regular, non-overlapping patches with the size of $p \times p \times C$, denoted as $\boldsymbol{x} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_i, ..., \boldsymbol{x}_{\frac{H}{p} \times \frac{W}{p}}\}$.

Step 2. Next, a masking operation is performed on these patches to identify visible (or unmasked) and masked patches, i.e., $\boldsymbol{x}_{vis} = \{\boldsymbol{x}_i | i \in vis\}$. Only the visible patches are sent into the to-be-learned encoder.

Step 3. The encoder f_{en} is implemented using ViT, where each visible patch is first linearly projected by a shared matrix $\boldsymbol{E}_s$, combined with positional embeddings $\boldsymbol{E}_{pos}$, and then processed through a series of transformer blocks. Thus, the encoder output in the i-th patch can be expressed as $\boldsymbol{z}_i = f_{en}(\boldsymbol{E}_s \boldsymbol{x}_i + \boldsymbol{E}_{pos})$.

Step 4. The input to the MAE decoder, denoted by g_{de}, is a complete set of tokens that includes the encoded visible patches and mask tokens (e.g., $\boldsymbol{z}_m$). The encoded features, which are the latent representations from the encoder, and the mask tokens are used as inputs and combined with positional embeddings to the lightweight ViT decoder. The final layer of the decoder is a linear projection (e.g., $\boldsymbol{W}$) that outputs several channels equal to the number of pixels in a patch. The output is then reshaped to reconstruct the image as $\hat{\boldsymbol{x}} = g_{de}(\boldsymbol{W}([\boldsymbol{z}_{vis}, \boldsymbol{z}_m]) + pos)$, where $\boldsymbol{z}_{vis}$ is the encoded representations of visible patches.

Step 5. The loss function used in MAE is the mean squared error (MSE), and it is calculated for the visible and masked patches (similar to BERT [32]), i.e., $\mathcal{L} = \frac{1}{vis+mask} \sum_{i \in vis \bigcup mask} (\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i)^2$.

It is worth noting that a normalization approach is performed, i.e., the mean and standard deviation of pixel values in each patch are calculated, and the patch is normalized accordingly, i.e., $\boldsymbol{x}_{norm} = \{\frac{\boldsymbol{x}_i - u_i}{\sigma_i} | i \in vis\}$. In this case, the encoder reconstruction task changes to reconstruct normalized pixel values.

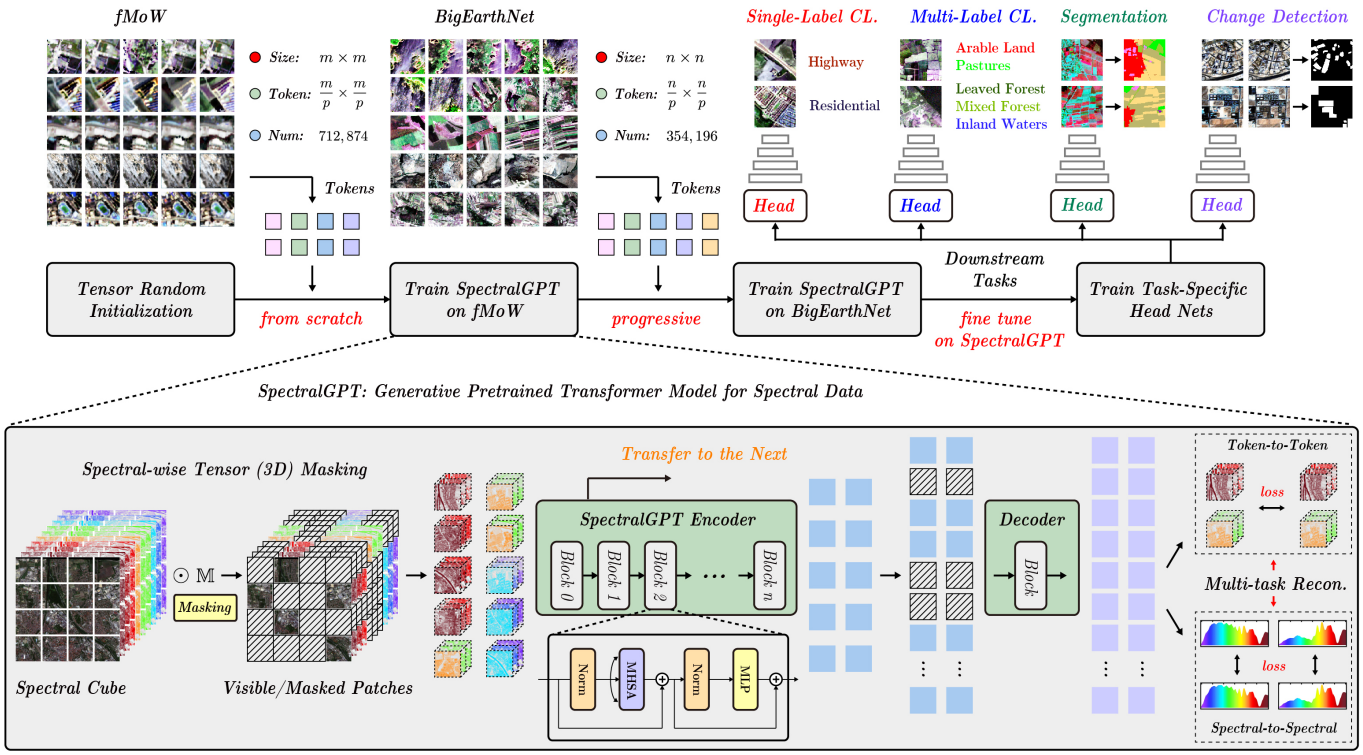

Fig. 2. An illustrative workflow of the proposed SpectralGPT foundation model and its adaptation to downstream tasks. In the pretraining phase, SpectralGPT starts to train the model from scratch on one dataset (e.g., fMoW-S2, with 712,874 images) with a (3D) tensor-based random weight initialization. Subsequently, the model undergoes progressive training on more datasets (e.g., BigEarthNet-S2, with 354,196 images) with varying image sizes, time series information, and geographic regions. SpectralGPT is constructed following the MAE architecture [24] and incorporates spectral-wise tensor (3D) masking, where 90% of the tokens are masked out. For downstream tasks, such as classification, segmentation, and change detection, the pretrained SpectralGPT is connected with task-specific Head networks to be trained and then performs fine-tuning.

B. Methodological Overview of SpectralGPT

Our SpectralGPT model is structured with three key components: 3D masking for processing spectral data, an encoder to learn spectrally visual representations, and a decoder for multi-target reconstruction. What sets our approach apart is a progressive training manner, where the model is trained using diverse types of spectral data. This strategy enhances the proposed SpectralGPT foundation model, endowing it with greater flexibility, robustness, and generalization capabilities. Fig. 2 provides an illustrative workflow of the proposed SpectralGPT with various downstream tasks.

C. 3D Masking on Spectral Data

Inspired by the spacetime-agnostic sampling in video-like data using the MAE-based framework [29], we model multi-band spectral images as 3D tensor data. To achieve this, we implement a 3D cube masking strategy that enables efficient processing of tensor-like spectral data. Our approach leverages a 90% masking rate to capture spatially and spectrally visual representations in an effective way, leading to more accurate and diverse knowledge extraction from input spectral data.

Given a 3D cube-shaped spectral image $\boldsymbol{x} \in \mathbb{R}^{H \times W \times D}$, we partition it into non-overlapping 3D tensor tokens along both the spatial and spectral dimensions. Each token has a size of $p \times p \times k$, where p and k are the token sizes in spatial and spectral dimensions, respectively. Using these settings, we then have $\frac{H}{p} \times \frac{W}{p} \times \frac{D}{k}$ tokens, denoted as $\boldsymbol{x} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_i, ..., \boldsymbol{x}_{\frac{H}{p} \times \frac{W}{p} \times \frac{D}{k}}\}$. This results in the following visible and masked representations (e.g., $\boldsymbol{x}_{vis}$ and $\boldsymbol{x}_{mask}$) on the spectral data.

$$[\boldsymbol{x}_{vis}, \boldsymbol{x}_{mask}] = \mathbb{M} \odot \boldsymbol{x}, \tag{1}$$

where $\mathbb{M} \in \{0,1\}^{\frac{H}{p} \times \frac{W}{p} \times \frac{D}{k}}$ is a token-wise binary mask indicating which tokens should be masked, i.e., all pixels in the token are set to zero.

D. Encoder for Visible Tokens

Similar to the encoder in MAE, all visible tokens $\{\boldsymbol{x}_i | i \in vis\}$ spatial-spectral mixed representations are first transformed into feature embeddings using a shared linear projection $\boldsymbol{E}_s$. The learned representations via the encoder f_θ with respect to the variable $\boldsymbol{\theta}$ are denoted as $f_{\boldsymbol{\theta}}(\boldsymbol{E}_s \boldsymbol{x}_i + \boldsymbol{E}_{pos})$, where $\boldsymbol{E}_{pos}$ represents the positional encoding. The encoder f_θ comprises several stacking self-attention (SA) transformer blocks. The SA module used in the encoder can be constructed as follows.

- The input embeddings $\boldsymbol{z}_i$ are linearly transformed to *query* $\boldsymbol{Q}_i$, *key* $\boldsymbol{K}_i$, and *value* $\boldsymbol{V}_i$ embeddings using learnable projection matrices $\boldsymbol{W}_Q$, $\boldsymbol{W}_K$, and $\boldsymbol{W}_V$, respectively.
- The attention scores $\boldsymbol{S}_i$ between *query* and *key* embeddings are computed as a dot product scaled by $\frac{1}{\sqrt{d}}$ and

passed through a softmax function. The resulting scores are then used to weight the *value* embeddings, which are summed to produce the final output embeddings, i.e., $\boldsymbol{z}_i = \boldsymbol{S}_i\boldsymbol{V}_i$. The formula of the original SA's complete process can be written as

$$\begin{aligned} &\boldsymbol{Q}_i = \boldsymbol{x}_i\boldsymbol{W}_Q,\ \boldsymbol{K}_i = \boldsymbol{x}_i\boldsymbol{W}_K,\ \boldsymbol{V}_i = \boldsymbol{x}_i\boldsymbol{W}_V, \\ &\boldsymbol{S}_i = \text{softmax}(\frac{\boldsymbol{Q}_i\boldsymbol{K}_i^\top}{\sqrt{d}}), \\ &\boldsymbol{z}_i = \text{Attention}(\boldsymbol{Q}_i, \boldsymbol{K}_i, \boldsymbol{V}_i) = \boldsymbol{S}_i\boldsymbol{V}_i, \end{aligned} \tag{2}$$

- Finally, the output features $\boldsymbol{z}_i$ have the same dimension as $\boldsymbol{x}_i$ and can be further processed by subsequent encoders.

E. Lightweight Decoder with Multi-Target Reconstruction

Given the encoder output features $\boldsymbol{z}$, we simultaneously train a lightweight decoder with a multi-target reconstruction strategy with respect to the variable $\boldsymbol{\phi}$ on $\boldsymbol{z}$ to recover the original image tokens from potential embeddings of visible and masked image tokens. Mathematically, the reconstructed image tokens $\hat{\boldsymbol{x}}$ can be formulated as $\hat{\boldsymbol{x}} = g_{\boldsymbol{\phi}}(f_{\boldsymbol{\theta}}(\mathbb{M} \odot \boldsymbol{x}))$. The decoder $g_{\boldsymbol{\phi}}$ is typically narrower and shallower than the encoder, and usually consists of a few transformer blocks and a linear reconstruction layer. The proposed SpectralGPT trains the encoder $f_{\boldsymbol{\theta}}$ and decoder $g_{\boldsymbol{\phi}}$ in an end-to-end manner to minimize the reconstruction loss between the reconstructed image tokens $\hat{\boldsymbol{x}}$ and the original image tokens $\boldsymbol{x}$. In our approach, the reconstruction loss consists of two components: token-to-token and spectral-to-spectral. This multi-target reconstruction allows the learned representations to capture spatial-spectral coupling characteristics and spectrally sequential information effectively. This overall loss $\mathcal{L}$ is quantified mathematically using MSE in the pixel space, defined as follows:

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_{token} + \lambda\mathcal{L}_{spectral} \\ &= \frac{1}{m}\sum_{i\in vis}(\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i)^2 + \frac{1}{n}\sum_{j=1}^{n}(\boldsymbol{x}_j - \hat{\boldsymbol{x}}_j)^2 \\ &= \frac{1}{m}\sum_{i\in vis}(\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i)^2 \\ &\quad + \frac{1}{n}\sum_{j=1}^{n}([\boldsymbol{x}_{r,c,1}, ..., \boldsymbol{x}_{r,c,\frac{D}{k}}]_j - [\hat{\boldsymbol{x}}_{r,c,1}, ..., \hat{\boldsymbol{x}}_{r,c,\frac{D}{k}}]_j)^2, \end{aligned} \tag{3}$$

where m and n represent the number of visible tokens ($\frac{H}{p} \times \frac{W}{p}$) and spectral tokens (each spectral token consists of $\frac{D}{k}$ standard tokens along with spectral dimension), and (r, c) denotes the standard token in the i-th row and the j-th column of the spectral data.

F. Progressive Pretraining

The proposed SpectralGPT model has the advantage of being highly adaptable to different input image sizes, which is especially useful for processing large datasets with images of varying size, resolution, temporal variability, and geographical coverage. This is achieved by dividing the input image into fixed-sized 3D tokens (e.g., $8 \times 8 \times 3$), which are then independently processed through the encoder-decoder pipeline.

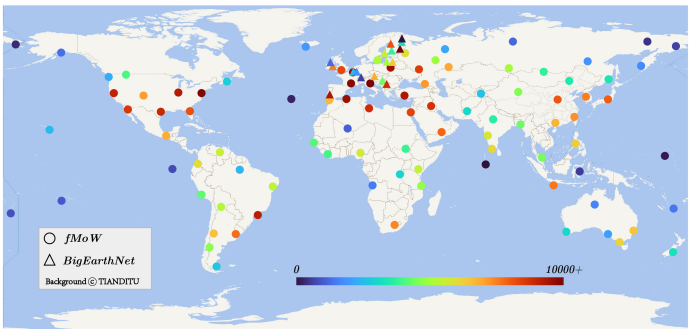

Fig. 3. Sample density distribution collected from Sentinel-2 sources (i.e., fMoW, BigEarthNet) over the Earth's inhabited areas, amounting to a total of 1,473,105 images.

The resulting tokens are then stitched back together to form the final output image. This approach ensures that the model can handle images of arbitrary dimensions in theory, without requiring any changes to the architecture or hyperparameters. With this characteristic, the proposed model allows for feeding varying-sized images into the encoder networks and also might enable the input images with different sensors, resolutions, time series, and modalities, as long as the 3D tokens are cropped to a fixed size.

It is worth emphasizing that the progressive feeding of different types of input images into the networks is not only useful for enabling greater flexibility in the type and size of input images but also for improving the model's ability to extract valuable knowledge from diverse data sources, thereby enhancing model generalization. This can be achieved, for instance, by first inputting images with a size of 96×96 pixels and then progressively feeding in the images with the size of 128×128 pixes, or by starting with Sentinel-2 data and then transitioning to Landsat-8 or Gaofen-2 data. More broadly, the ability to process varying types and sizes of input images can lead to more robust and generalizable features that are not limited to a specific input image type or size, thus improving model generalization and performance on previously unseen data. Moreover, this flexibility in input image size and type is particularly beneficial in practical applications, where input images may come from different sources or have varying resolutions, allowing the model to be more adaptable to real-world scenarios where input images can be unpredictable.

G. Pretrained Dataset

Our foundation model is trained on a comprehensive dataset comprising over one million spectral images from the Sentinel-2 satellite. This dataset encompasses 12 spectral bands and draws from two primary sources: fMoW-S2 [30], a globally diverse collection labeled with 62 categories based on the Functional Map of the World (fMoW) [33], and BigEarthNet [34], a regional dataset originating from over ten European countries. To provide an overview of the dataset, Fig. 3 illustrates the distribution of image samples over Earth's inhabited areas, amounting to a total of 1,473,105 images.

Researchers at Stanford University meticulously curated a dataset by leveraging geo-coordinates and timestamps from the fMoW dataset. This process aims to construct a time

series of Sentinel-2 images. To ensure data quality, locations with exclusively pre-Sentinel-2 fMoW images were excluded. For locations with partially preceding fMoW images, selective curation is performed, involving the exclusion of these specific images and the introduction of supplementary captures at 6-month intervals to enrich the temporal sequence. This approach culminates in the creation of the fMoW Sentinel-2 dataset, denoted as fMoW-S2. This dataset predominantly covers fMoW locations and preserves labels mirroring those in the original fMoW dataset.

The fMoW-S2 dataset comprises Sentinel-2 spectral images (B1-12 and B8A) and is partitioned into three subsets: 712,874 training images, 84,939 validation images, and 84,966 test images, totaling 882,779 images. Each image has an average dimension of approximately 45 pixels in height and 60 pixels in width. For additional details about the fMoW-S2 dataset, interested parties can refer to the dedicated website[1].

Furthermore, the study incorporates the BigEarthNet dataset, specifically the BigEarthNet-S2 variant[2], comprising 590,326 distinct, non-overlapping Sentinel-2 spectral image tokens. The pretraining of different variant models using the proposed method involves the use of 712,874 fMoW-S2 images and 354,196 BigEarthNet-S2 images. Notably, only 10% of the BigEarthNet-S2 images with labels, amounting to 35,420 images, are utilized for fine-tuning in downstream tasks.

H. Implementation Details and Experimental Setup

Following established conventions, we acknowledge that Sentinel-2 images comprise 13 spectral bands. However, to harmonize datasets across pretrained and downstream tasks in terms of channel composition, we have opted to retain the 12 dominant bands, excluding band B10, in all fMoW dataset images. To ensure data consistency, we normalize the spectral images band by band, scaling their values to a standardized range of 0 to 1. Subsequently, we follow established methodologies [30] for preprocessing. This involves random image cropping within the range of 0.2x to 1.0x of the original size, resizing them to 96×96 pixels, and applying horizontal flips. These meticulous steps collectively contribute to the robustness and compatibility of our spectral foundation model.

We employ the vanilla ViT-Base architecture as the network backbone. To adapt the model to spectral data, we employ a token size of $8 \times 8 \times 3$ pixels, effectively partitioning the images. For instance, an image with a size of $96 \times 96 \times 12$ pixels is segmented into $12 \times 12 \times 4$ tokens. Drawing inspiration from a prior work [29], our approach incorporates two learnable positional embeddings. One of these embeddings is dedicated to spatial information, while the other is tailored to capture variations across spectral channels. This augmentation further refines the model's ability to extract meaningful features from the spectral input.

Our pretraining closely adheres to the approach outlined in a prior study [30]. Utilizing the computational power of 8 NVIDIA GeForce RTX 4090 GPUs and AMD EPYC 7Y83 CPU, we implement the AdamW optimizer [35] with a foundational learning rate of 10^{-4}, coupled with a half-cycle cosine decay schedule. To ensure robustness, we adopt a 3D masking ratio of 90%, facilitating effective training. The model undergoes a comprehensive pretraining regimen spanning 200 epochs on the fMoW-S2 dataset. After this phase, the model's training continues on the BigEarthNet-S2 dataset for 100 epochs. While this phase necessitates a modification in input dimensions to $128 \times 128 \times 12$, the other settings remain consistent. To distinguish between different stages, the model pretrained solely on the fMoW dataset is denoted as SpectralGPT, whereas the model pretrained on both datasets in a progressive way is represented as SpectralGPT$^+$. This meticulous strategy effectively enhances the model's adaptability and performance across diverse datasets.

[1]https://purl.stanford.edu/vg497cb6002
[2]https://bigearth.net

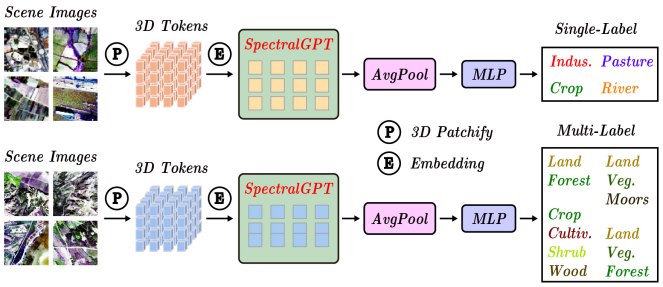

Fig. 4. Network architecture for downstream tasks in terms of single-label (Top) RS scene classification and multi-label (Bottom) RS scene classification by leveraging our pretrained SpectralGPT model.

III. Experiments

In this section, we rigorously evaluate the performance of our SpectralGPT model by benchmarking it against several SOTA foundation models: ResNet50 [36], SeCo [37], ViT [22], and SatMAE [30]. Further, we assess its capabilities across four downstream EO tasks, including single-label scene classification, multi-label scene classification, semantic segmentation, and change detection, as well as extensive ablation studies.

We quantitatively assess the performance of pretrained foundation models across four downstream tasks in terms of recognition accuracy for the single-label RS scene classification task, macro and micro mean average precision (mAP), i.e., macro-mAP (micro-mAP), for the multi-label RS scene classification task, overall accuracy (OA) and mean intersection over union (mIoU) for the semantic segmentation task, and precision, recall, and F1 score for the change detection. Additionally, we conduct insightful ablation studies, exploring critical factors such as masking ratio, decoder depth, model size, patch size, and training epochs. Utilizing the computational power of 4 NVIDIA GeForce RTX 4090 GPUs, we meticulously fine-tune pretrained foundation models for both downstream tasks and ablation studies, thereby offering comprehensive insights into SpectralGPT's capabilities and adaptability within the RS domain.

A. Single-Label RS Scene Classification on EuroSAT

For the downstream single-label RS scene classification task, we employ the EuroSAT dataset [38]. This dataset

consists of 27,000 Sentinel-2 satellite images collected from 34 European countries. These images are classified into 10 land use classes, each containing between 2,000 to 3,000 labeled images. Each image in this dataset has a resolution of 64×64 pixels and encompasses 13 spectral bands. It's worth noting that, for consistency with prior data processing, band B10 has been excluded from all images. Additionally, we follow the train/validation splits as recommended in [39].

On the EuroSAT dataset, these pretrained models undergo fine-tuning, spanning 150 epochs with a batch size of 512. This fine-tuning process employs the AdamW optimizer with a base learning rate of 2×10^{-4}, and it incorporates data augmentations in line with prior work [24], including weight decay (0.05), drop path (0.1), reprob (0.25), mixup (0.8), and cutmix (1.0). The foundational encoder of the pretrained model is utilized, and its output is passed through an average pooling layer to generate predictions. The training objective is to minimize the cross-entropy loss. Fig. 4 illustrates the network architecture for the downstream single-label scene classification task.

The pretrained model's encoder serves as the foundational backbone, and its output is subject to an average pooling layer to generate predictions. The training objective involves minimizing the cross-entropy loss. In Table I, we present a comparative analysis of our proposed method against alternative pretraining models, reporting the highest Top1 accuracy on the validation set. The obtained results highlight the efficacy of the proposed approach, achieving an impressive accuracy of 99.15%. Furthermore, when the model undergoes pretraining on both the fMoW-S2 and BigEarthNet datasets, a noteworthy performance boost is observed, culminating in a remarkable accuracy of 99.21%. This underscores the advantage of leveraging diverse data sources for improved model performance.

TABLE I
QUANTITATIVE RESULTS OF SOTA PRETRAINED FOUNDATION MODELS FOR THE DOWNSTREAM SINGLE-LABEL RS SCENE CLASSIFICATION TASK IN TERMS OF RECOGNITION ACCURACY ON THE EUROSAT DATASET. THE BEST RESULT IS SHOWN IN BOLD.

Method	Pretrained Dataset	Acc. (%)
ResNet50 [36]	ImageNet-1k	96.72
SeCo [37]	SeCo	97.23
ViT [22]	Random Init.	98.73
ViT-22k [22]	ImageNet-22k	98.91
SatMAE [30]	fMoW-S2	99.09
SpectralGPT	fMoW-S2	99.15
SpectralGPT^{+}	fMoW-S2+BigEarthNet	**99.21**

B. Multi-Label RS Scene Classification on BigEarthNet

For the multi-label RS scene classification task, we utilize the BigEarthNet-S2 dataset [34]. This extensive dataset consists of 125 Sentinel-2 tiles and comprises 590,326 12-band images that span 19 classes for multi-label classification. The images encompass resolutions ranging from 10 to 60 meters, with 12% of low-quality images being excluded. The training and validation sets align with prior research [39], with 354,196 training samples and 118,065 validation samples. To prepare the images for model training, those of varying resolutions are standardized to a uniform size of 128×128 pixels using bilinear interpolation.

On the BigEarthNet-S2 dataset, these foundation models are fine-tuned using a 10% subset of the training data, following similar settings to those applied in the EuroSAT fine-tuning experiments, except for an increased learning rate of 2×10^{-4}, which is in line with findings from previous research [30], [37]. Most existing methods, including those that use pretrained foundation models, typically utilize all available images for training in the BigEarthNet-S2 dataset. In contrast, our proposed SpectralGPT achieves higher classification performance, even when utilizing only 10% of the training samples. Given the multi-label classification nature of this task, our training objective involves the multi-label soft margin loss, and performance evaluation is based on the mAP metric. Notably, we calculate mAP using both macro and micro mAP measurements. This approach is particularly relevant for the BigEarthNet-S2 dataset, which exhibits class imbalance. The multi-label classification framework is shown in Fig. 4.

Table II presents a comparative analysis of our pretrained model against other proposed pretrained models and models trained from scratch, showcasing the exceptional performance of the proposed approach. In particular, when compared to ViT pretrained on ImageNet-22k and SatMAE, our SpectralGPT model outperforms them by 0.84% (0.82%) and 0.71% (0.68%) in terms of macro-mAP (micro-mAP), respectively. Notably, the introduction of additional pretraining data (BigEarthNet), i.e., SpectralGPT^{+}, leads to a significant performance boost, with the model achieving an impressive 88.22% (87.50%) macro-mAP (micro-mAP), surpassing the model solely trained on fMoW-S2 by 2.19% (1.86%). This substantial improvement can be attributed to two key factors. Firstly, the model's initial pretraining on BigEarthNet (even without labels) equips it with a strong grasp of the dataset's distribution, accelerating convergence during fine-tuning and enhancing mAP. Secondly, the adoption of the MIM method as a pretraining pretext task, coupled with a substantial data scale, necessitates alignment with the training strategy, emphasizing the significance of the random masking framework and a 90% masking ratio to facilitate more robust representation learning. Furthermore, as our evaluation focuses on a multi-label classification task and employs only 10% of the training data, the results underscore the superior generalization and few-shot learning capabilities of our proposed model in tackling challenging downstream tasks.

C. RS Semantic Segmentation on SegMunich

For the semantic segmentation task, we create a new SegMunich dataset, which is derived from the Sentinel-2 spectral satellite [41]. This dataset consists of a 10-band best-pixel composite with dimensions of $3,847 \times 2,958$ pixels and a spatial resolution of 10 meters. It captures Munich's urban landscape over a span of three years up to April 2020 and includes a segmentation mask that meticulously delineates 13 Land Use and Land Cover (LULC) classes. The data for this mask is sourced from various places, including OpenStreetMap for street network data and the OSMLULC

TABLE II
QUANTITATIVE RESULTS OF SOTA PRETRAINED FOUNDATION MODELS FOR THE DOWNSTREAM MULTI-LABEL RS SCENE CLASSIFICATION TASK IN TERMS OF MEAN AVERAGE PRECISION (MAP) ON THE BIGEARTHNET DATASET. THE BEST RESULT IS SHOWN IN BOLD.

Method	Pretrained Dataset	macro-mAP	micro-mAP
ResNet50 [36]	ImageNet-1k	80.76	80.06
SeCo [37]	SeCo	83.16	82.82
ViT [22]	Random Init.	81.57	80.15
ViT-22k [22]	ImageNet-22k	85.08	84.67
SatMAE [30]	fMoW-S2	85.21	84.93
SpectralGPT	fMoW-S2	86.03	85.61
SpectralGPT$^+$	fMoW-S2+BigEarthNet	**88.22**	**87.50**

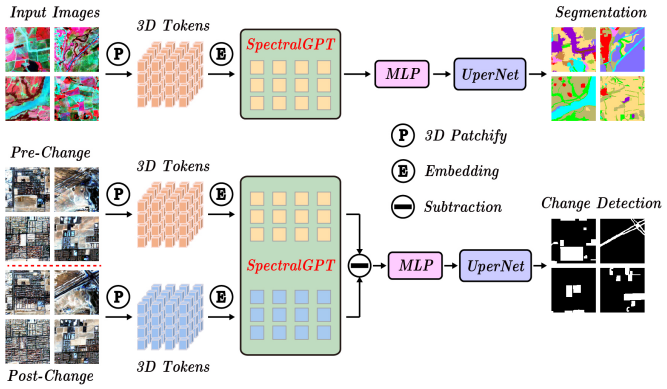

Fig. 5. Network architecture for downstream tasks in terms of semantic segmentation (Top) and change detection (Bottom) by leveraging our pretrained SpectralGPT model and training a follow-up UperNet [40] Head.

platform[3] for the remaining 12 classes, all obtained at the same 10-meter spatial resolution. To create a comprehensive feature representation for semantic segmentation, the dataset combines the 10-meter spectral bands (B1, B2, B3, and B4) with resampled 20-meter spectral bands (B5, B6, B7, B8A, B11, B12), which have been upsampled to match the 10-meter resolution. This amalgamation of spectral bands ensures that the dataset provides rich and informative data for the semantic segmentation task.

On the SegMunich dataset, we employ the UperNet framework [40] in conjunction with the pretrained foundation models, initially consolidating the four tokens per pixel from the encoder's final layer into a single token. Image data is divided into 128×128-pixel tokens with a 50% overlap. The dataset is then split into a training-validation ratio of 8:2 and subjected to data augmentation techniques, including random flips and rotations. During fine-tuning on this dataset, we use a batch size of 96 and set the base learning rate to 5×10^{-4}. The optimization and loss functions remain consistent with those employed in the EuroSAT experiment, ensuring a coherent and uniform approach to model training and evaluation. The segmentation architecture is detailed in Fig. 5.

Table III lists quantitative results in terms of OA and mIoU for the semantic segmentation task. Our SpectralGPT (SpectralGPT$^+$) outperforms all others, exhibiting a significant lead with a 1.1% (2.3%) higher mIoU than the second-best result (i.e., SatMAE). Fig. 6(a) offers a visual depiction of the Munich area under study for the segmentation task, along with the proportions of the 13 classes. The qualitative comparison in several ROIs, as shown in Fig. 6(b), highlights our model's superior ability to recognize a wider range of land use categories well compared to the competing models in most instances. Furthermore, when considering ViT-22k as the baseline for performance comparison, our model consistently excels across all segmentation classes, as evident in Fig. 6(c), particularly for categories, such as *crops*, *pastures*, *open spaces*, *vegetations*, and others. By amalgamating category statistics with class-wise IoU outcomes, it becomes evident that our SpectralGPT model excels in mitigating the challenges posed by category-imbalanced classification. This results in a substantial enhancement in performance when compared to other foundation models.

D. RS Change Detection on OSCD

For the change detection task, we utilize the OSCD dataset [42]. Fig. 7(a) shows several examples at the city scale. This dataset comprises 24 cities of Sentinel-2 images, with 14 images used for training and 10 images for evaluation. These images were captured between 2015 and 2018 and encompassed 13 spectral bands with resolutions of 10m, 20m, and 60m. The dataset is annotated at the pixel level to indicate changes, specifically focusing on urban developments.

On the OSCD dataset, we perform image cropping to create patches of size 128×128 pixels with a 50% overlap rate, and we apply random flips and rotations as data augmentation techniques. For each pair of images, both are simultaneously processed through a shared encoder, and the difference between their features is computed and then passed to a UperNet. Each feature pixel consists of 4 tokens, similar to the segmentation approach, and we use a linear layer to consolidate these 4 tokens into 1 token. The model is trained for 60 epochs with a batch size of 64 and a learning rate set to 1×10^{-3}, using negative log-likelihood loss as the training objective. The entire framework for leveraging the pretrained SpectralGPT model in the change detection task is depicted in Fig. 5.

Model performance is assessed through precision, recall, and F1 score, with quantitative outcomes shown in Table IV on the OSCD dataset, where our proposed model achieves the highest F1 score, surpassing the second-best model (i.e., SatMAE) by a substantial margin of 0.75% (1.53%). However, it is worth noting that our model excels in F1 score and recall but exhibits relatively lower precision compared to other models. This phenomenon can be attributed to two main factors. Firstly, extreme imbalance of the inherent data within the change detection task (see Fig.7(b)), where the number of positive and negative samples varies significantly, may lead the model to classify negative cases as positive to improve recall at the cost of precision. Secondly, the complexity of the ViT architecture demands a substantial amount of data to mitigate overfitting. The model may struggle with overfitting and become less adaptable to out-of-domain data. Addressing this challenge could involve providing additional fine-tuning data

[3]https://osmlanduse.org/

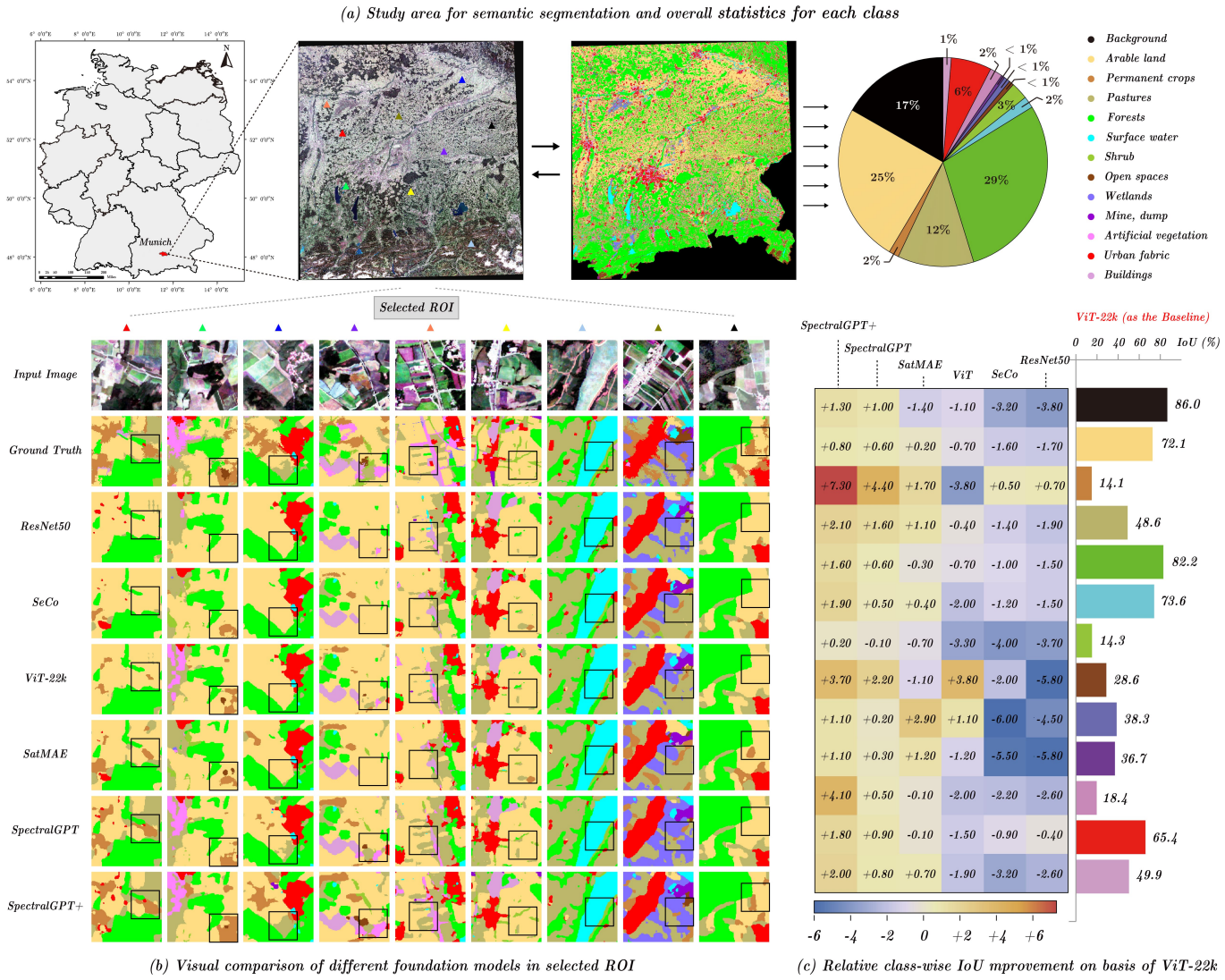

Fig. 6. Qualitative and quantitative semantic segmentation results of different pretrained foundation models on the SegMunich dataset. (a) Study area and overall statistics for each class in the semantic segmentation task. (b) Segmentation visualization maps of different foundation models in selected ROIs. (c) Relative IoU performance improvement for each class on the basis of ViT-22k.

TABLE III
QUANTITATIVE RESULTS OF SOTA PRETRAINED FOUNDATION MODELS THAT ARE FINELY TUNED USING UPERNET FOR THE DOWNSTREAM RS SEMANTIC SEGMENTATION TASK IN TERMS OF OVERALL ACCURACY (OA) AND MEAN INTERSECTION OVER UNION (MIOU) ON THE SEGMUNICH DATASET. THE BEST RESULT IS SHOWN IN BOLD.

Method	Pretrained Dataset	OA	*Background*	*Arable land*	*Perm. Crops*	*Pastures*	*Forests*	*Surface water*	*Shrub*	*Open spaces*	*Wetlands*	*Mine, dump*	*Artificial veg.*	*Urban fabric*	*Buildings*	mIoU
ResNet50 [36]	ImageNet-1k	80.1	82.2	70.4	14.8	46.7	80.7	72.1	10.6	22.8	33.8	30.9	15.8	65.0	47.3	45.6
SeCo [37]	SeCo	80.3	82.8	70.5	14.6	47.2	81.2	72.4	10.3	26.6	32.3	31.2	16.2	64.5	46.7	45.9
ViT [22]	Random Init	81.0	84.9	71.4	10.3	48.2	81.5	71.6	11.0	**32.4**	39.4	35.5	16.4	63.9	48.0	47.3
ViT-22k [22]	ImageNet-22k	81.7	86.0	72.1	14.1	48.6	82.2	73.6	14.3	28.6	38.3	36.7	18.4	65.4	49.9	48.3
SatMAE [30]	fMoW-S2	81.5	84.6	72.3	15.8	49.7	81.9	74.0	13.6	27.5	**41.2**	**37.9**	18.3	65.3	50.6	48.7
SpectralGPT	fMoW-S2	82.5	87.6	73.1	16.3	50.6	83.6	74.7	14.2	32.5	39.2	37.7	19.4	67.0	51.7	49.8
SpectralGPT$^+$	fMoW-S2+BigEarthNet	**82.7**	**88.0**	**73.1**	**22.5**	**51.0**	**84.1**	**76.0**	**14.5**	**33.7**	39.6	**38.7**	**22.5**	**67.4**	**52.0**	**51.0**

or reducing the model's rank. In terms of qualitative results, our model excels in predicting change pixels with fewer false negatives in selected ROIs of Fig. 7(d). Significantly, Fig. 7(c) accentuates the exceptional performance of SpectralGPT, with our model achieving the top results in half of the testing cities. In addition, there is a consistent performance trend among the compared foundation models across 10 different testing cities, with *Lasvegas* and *Montpellier* consistently achieving

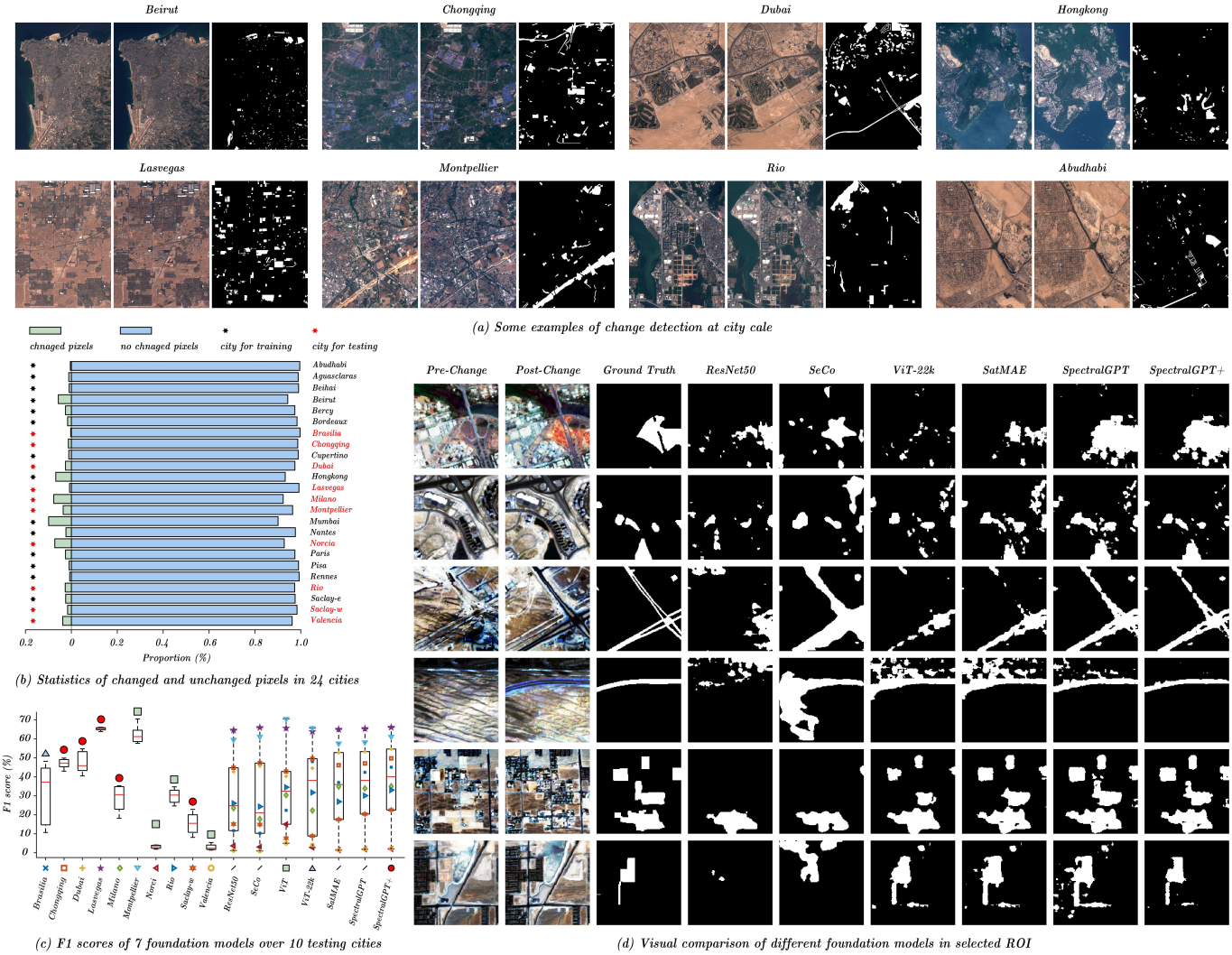

Fig. 7. Qualitative and quantitative change detection results of different pretrained foundation models on the OSCD dataset. (a) Some studied examples of change detection at the city scale. (b) Statistics of changed and unchanged pixels in 24 studied cities, including training and testing scenes. (c) F1 scores of different foundation models over 10 testing cities. (d) Change detection visualization maps of different foundation models in selected ROIs.

TABLE IV
QUANTITATIVE RESULTS OF SOTA PRETRAINED FOUNDATION MODELS THAT ARE FINELY TUNED USING UPERNET FOR THE DOWNSTREAM RS CHANGE DETECTION TASK IN TERMS OF PRECISION, RECALL, AND F1 SCORE ON THE OSCD DATASET. THE BEST RESULT IS SHOWN IN BOLD.

Method	Pretrained Dataset	Precision	Recall	F1
ResNet50 [36]	ImageNet-1k	**65.42**	38.86	48.10
SeCo [37]	SeCo	57.71	49.23	49.82
ViT [22]	Random Init.	56.71	47.52	51.71
ViT-22k [22]	ImageNet-22k	52.09	52.37	52.23
SatMAE [30]	fMoW-S2	55.18	50.54	52.76
SpectralGPT	fMoW-S2	51.65	56.15	53.51
SpectralGPT$^+$	fMoW-S2+BigEarthNet	52.39	**57.20**	**54.29**

the highest and second-highest F1 scores, respectively.

E. Ablation Studies

During the pretraining stage, we conduct a comprehensive study of various factors that may impact downstream task performance. These factors encompass masking ratio, ViT patch size, data scale, reconstruction target, decoder depth, and model size. To provide a more rigorous assessment of pretrained models, we subject all ablation models to fine-tuning on the BigEarthNet multi-label classification dataset with only a 10% subset of the train set, which poses a more formidable challenge, evaluated using the mAP measurement. Our choice of ViT-B as the backbone model ensures consistency across experiments. Except for ablations involving data scale and training schedule length, all models undergo pretraining on the fMoW-S2 dataset for a duration of 200 epochs. This comprehensive evaluation framework enables us to gain deeper insights into the impact of these factors on model performance.

1) Token Size: Table V(a) Fig. 8(a) offers crucial insights into the impact of token size on model performance, consistently demonstrating that larger patch sizes lead to reduced model performance, aligning with prior research findings [30]. This phenomenon can be attributed to the intrinsic characteristics of ViT architectures. With larger token sizes, such as 16×16, each image contains fewer tokens, resulting in

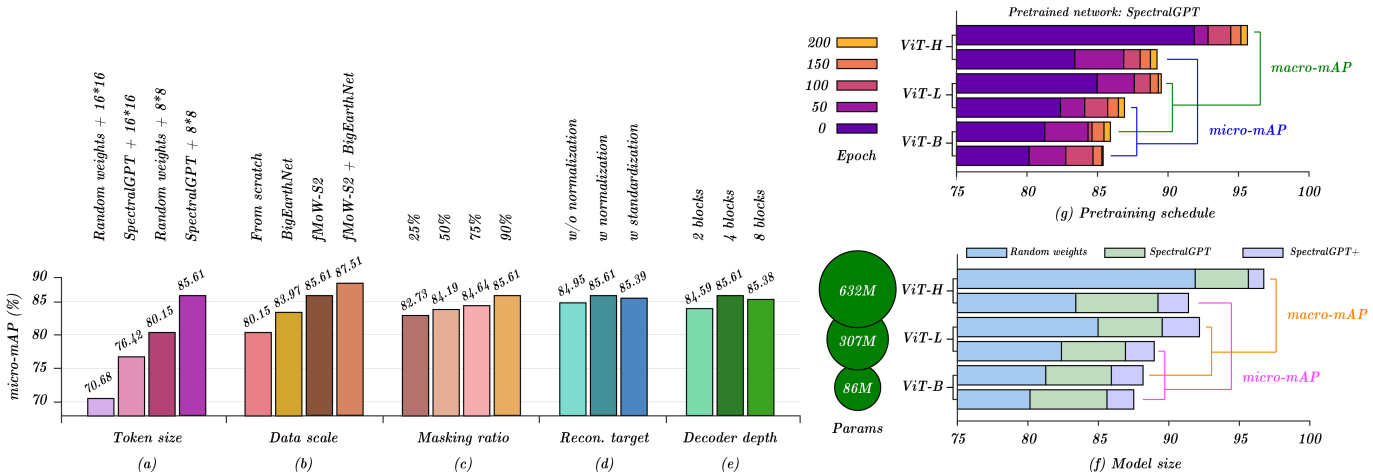

Fig. 8. An illustration for ablation analysis of the proposed SpectralGPT foundation model on BigEarthNet-S2 dataset in terms of (a) token size, (b) data scale, (c) masking ratio, (d) reconstruction target, and (e) decoder depth, as well as (f) model size (i.e., ViT-Base, ViT-Large, and ViT-Huge) and their (g) pretraining length/epoch.

TABLE V
ABLATION ANALYSIS OF THE PROPOSED SPECTRALGPT FOUNDATION MODEL IN TERMS OF MASKING RATIO, PATCH SIZE, DATA SCALE, RECONSTRUCTION TARGET, AND DECODER DEPTH, RESPECTIVELY. THE BEST RESULT IS SHOWN IN BOLD.

(a) Token Size

Init. Weights	Patch Size	mAP
Random	16	70.68
SpectralGPT	16	76.42
Random	8	80.15
SpectralGPT	**8**	**85.61**

(b) Data Scale

Pretrained Dataset	mAP
From scratch	80.15
BigEarthNet	83.97
fMoW-S2	85.61
fMoW-S2+BigEarthNet	**87.51**

(c) Masking Ratio

Ratio	mAP
25%	82.73
50%	84.19
75%	84.64
90%	**85.61**

(d) Reconstruction Target

Case	mAP
Without norm	84.95
Normalization	**85.61**
Standardization	85.39

(e) Decoder Depth

Blocks	mAP
2	84.59
4	**85.61**
8	85.38

a diminished level of fine-grained spatial information as the model progresses through its deeper layers. Consequently, this reduction in spatial detail negatively affects the model's overall performance. However, it is noteworthy that the pretrained model consistently enhances mAP regardless of the token size settings, emphasizing its capacity to improve performance across various configurations. Significantly, the recognition performance with a token size of 8×8 is notably superior to that with 16×16, despite the input images being 96×96 or 128×128 in size, underscoring the versatility and efficacy of the pretrained model.

2) Data Scale: Table V(b) and Fig. 8(b) present a comprehensive analysis focusing on the impact of pretraining data in our research. We conducted pretraining using two datasets (i.e., fMoW-S2, BigEarthNet) while maintaining a standardized input image size of 96×96. To delve deeper into this comparison, we initially pretrained models exclusively on fMoW-S2, followed by a seamless continuation of pretraining on BigEarthNet without any intermediate fine-tuning steps. Our pretraining datasets consisted of the extensive train set of fMoW-S2, which encompasses an impressive 712,874 images from around the world, and the BigEarthNet training set, which comprises 351,496 images within the European region after excluding those affected by snow, clouds, or cloud shadows.

This analysis in Table V(b) underscores the substantial influence of both data scale and distribution on model pretraining. Models pretrained on the same dataset as the downstream task consistently demonstrate superior performance, highlighting the crucial role of dataset coherence in effective transfer learning. Furthermore, fMoW-S2 outperforms BigEarthNet in pretraining, primarily due to its larger dataset and broader geographic coverage. Interestingly, the concept of continual pretraining, which combines both datasets, results in models with higher mAP scores. This improvement can be attributed in part to the transition from 96×96 images during fMoW-S2 pretraining to 128×128 images during BigEarthNet pretraining, underscoring the beneficial impact of increasing image size on overall model efficacy.

3) Masking Ratio: Table V(c) and Fig. 8(c) shed light on the impact of the masking ratio, revealing a noteworthy trend where higher masking ratios correspond to improved model performance. Unlike the conventional masking ratio of 75% often applied to natural RGB images, we find that the optimal masking ratio for multi-spectral images is 90%. This observation aligns with the hypothesis presented in [29] that the masking ratio in MIM methods is intricately linked to the information redundancy within the data. Multi-spectral images inherently exhibit greater information redundancy, with strong correlations among their spectral bands. Consequently, a higher masking ratio is essential for the model to effectively learn meaningful representations from these images. Moreover, a 90% masking ratio significantly enhances the efficiency of the pretraining stage, reducing memory complexity and

expediting training times, offering a practical advantage in model development.

4) Reconstruction Target: Table V(d) and Fig. 8(d) conduct an insightful analysis of the influence of reconstruction targets on normalized, standardized data and raw data without normalization or standardization in the context of multi-spectral images. Normalization, which scales all data to the [0, 1] range, and standardization, which transforms data to have a mean of 0 and a standard deviation of 1, are the two examined targets. Remarkably, the results show minimal disparity in model performance between normalization and standardization reconstruction targets, primarily because both targets pertain to pixel-level data transformations. However, the model pretrained on raw data performs much worse than the models with normalized reconstruction targets. We attribute this phenomenon to the characteristics of multi-spectral images. The spectral values are usually numerically large and vary from band to band, thereby the model pretraining on the raw data may need a longer pretaining schedule to converge and show the same performance compared with those models pretrained on normalized and standardized data. Our perspective suggests that employing a more semantically meaningful target in a specific representation space could potentially yield improved model performance.

5) Decoder Depth: Table V(e) and Fig. 8(e) examine the impact of decoder depth on model performance, following the principles of MIM methods where the pretrained encoder serves as the backbone for downstream tasks while discarding the decoder component. Notably, the results reveal that a shallow decoder configuration is ill-suited for spectral model pretraining. This observation aligns with the hypothesis that spectral images, characterized by high dimensionality and complexity, require a decoder with enhanced capacity, consistent with prior findings in the field [29].

6) Model Size: Table VI and Fig. 8(f) give a comparative analysis between fine-tuning results of ViT-B and ViT-L quantitatively and qualitatively, revealing compelling insights. The macro average precision and micro average precision are listed to comprehensively evaluate the performance of models. ViT-B, equipped with 12 transformer layers and 86 million parameters, exhibits promising performance gains when employing the proposed method, achieving a mAP(micro) of 85.41, surpassing the ViT-B trained from scratch by 5.26. On the other hand, ViT-L, featuring 24 layers and 307 million parameters, notably outperforms ViT-B, with a mAP(micro) of 86.92, surpassing the model trained from scratch by a significant margin of 4.44. Besides, ViT-H, concluding 32 layers and 632 million parameters, highly enhances the performance of the neural network on BigEarthNet with the mAP(micro) of 89.23. Notably, though our models are only fine-tuned with 10% downstream training data, the ViT-H model employing SpectralGPT$^+$ pretrained weights beats all the models even trained with the whole train set, with a SOTA mAP(micro) of 91.39. These results underscore the pivotal role of an appropriate pretraining strategy and indicate that larger ViT models are capable of learning more intricate image representations, rendering them highly suitable for tasks demanding superior accuracy.

TABLE VI
PERFORMANCE COMPARISON USING DIFFERENT PRETRAINED MODELS ACROSS THREE VIT-BASED NETWORK SCALES (I.E., BASE, LARGE, HUGE) ON THE BIGEARTHNET DATASET. THE BEST RESULT IS SHOWN IN BOLD.

Network Scale	Params	Pretained Network	macro-mAP	micro-mAP
ViT-Base	86M	Random Init.	81.27	80.15
		SpectralGPT	85.92	85.61
		SpectralGPT$^+$	**88.17**	**87.50**
ViT-Large	307M	Random Init.	84.98	82.38
		SpectralGPT	89.53	86.92
		SpectralGPT$^+$	**92.17**	**88.96**
ViT-Huge	632M	Random Init.	91.87	83.40
		SpectralGPT	95.64	89.23
		SpectralGPT$^+$	**96.73**	**91.39**

7) Pretraining Schedule: In Fig. 8(g), we present the fine-tuning results for models trained with varying pre-training epochs, evaluated using the macro-mAP and micro-mAP metrics, respectively. Notably, the models pretrained for just 50 epochs exhibit significant performance gains compared to those trained from scratch. The observed trend in the figure indicates that the models continue to benefit from longer pretraining epochs, suggesting that extended training can further enhance performance. Moreover, the results in Table VI reinforce this finding, as ViT-L and ViT-H consistently achieve higher mAP compared to ViT-B, highlighting the effectiveness of both extended pretraining and larger model architectures.

F. Visual Comparison and Geo-characteristic Recoverability

With varying masking ratios (i.e., 50%, 75%, 90%, and 95%) as the input, Fig. 9 visually illustrates the image reconstruction results obtained using SatMAE and our SpectralGPT. Not unexpectedly, as the masking ratio increases, the reconstructed images deviate more from the originals. It is worth highlighting, however, that the proposed SpectralGPT outperforms SatMAE significantly in terms of spectral image reconstruction performance, particularly in preserving visual structures and textural details. To be specific, when utilizing 50% visible patches, the reconstructed results with SatMAE are comparable to those with SpectralGPT, albeit with some slight blurriness in certain fine details in the SatMAE results. As the percentage of masked patches increases (e.g., from a 75% mask to 90% and further to 95%), the reconstruction performance of SatMAE experiences a substantial decline. In contrast, our SpectralGPT exhibits a superior reconstruction capability (*cf.* SatMAE). Even with a masking rate exceeding 90%, critical structures and shape components remain preserved in the visuals, demonstrating our model's robust learning, reasoning, and generalization capabilities.

In addition to the in-depth discussion and sensitivity analysis concerning the masking ratio, we have undertaken more extensive investigations into spectral-wise reconstruction capabilities by using only 10% of visible patches, with the remainder masked out. These investigations prioritize the representation of geographical characteristics, utilizing various spectral band combinations. As illustrated in Fig. 10, we present visualizations of eight different band combinations. These visualizations distinctly highlight the remarkable superiority of our proposed SpectralGPT (closer to that generated

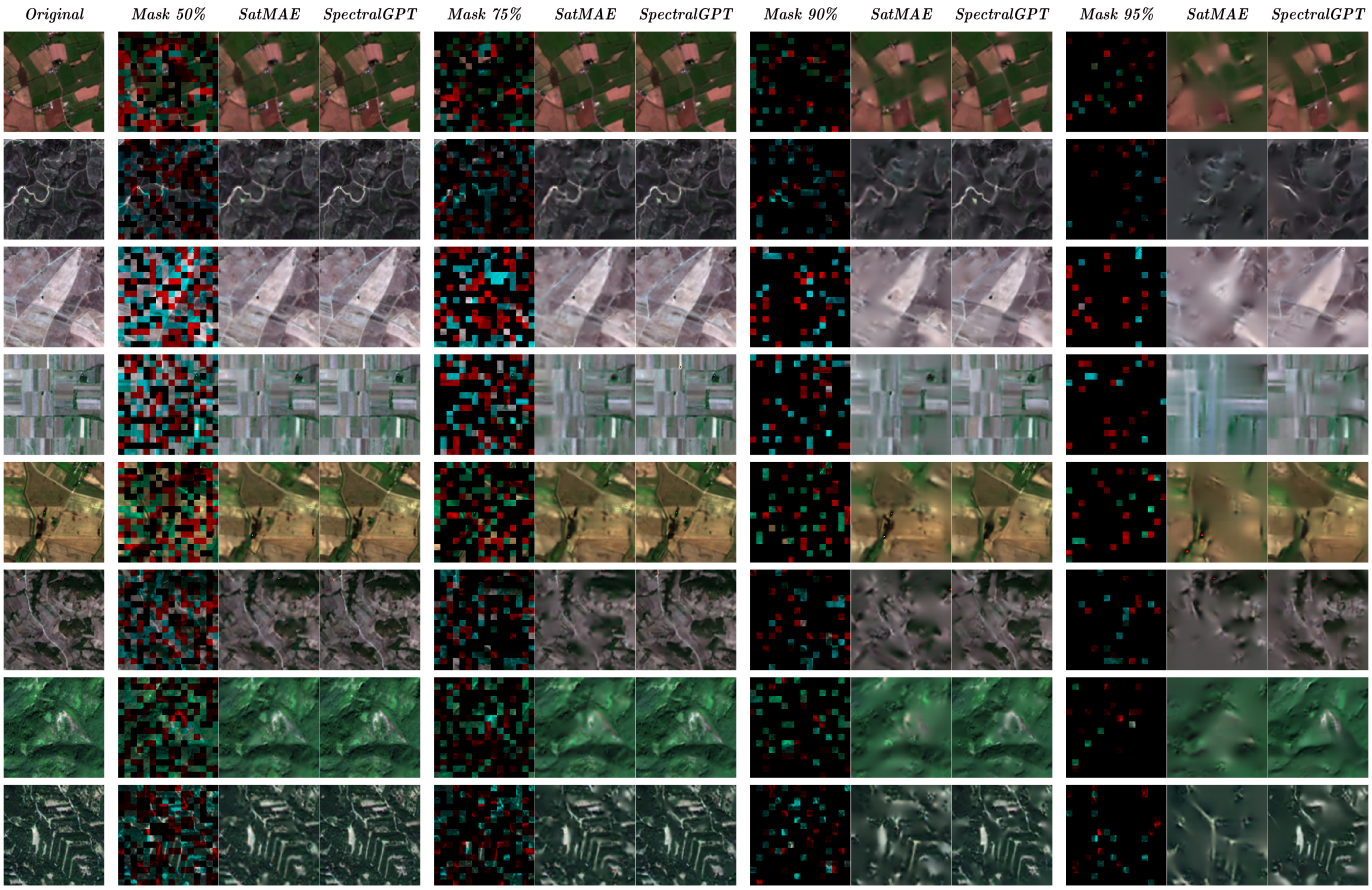

Fig. 9. Visual comparison from the nature-color (R: 4, G: 3, B: 2) image reconstruction perspective between SatMAE and SpectralGPT with varied masking ratios of 50%, 75%, 90%, and 95%, respectively. By masking out a greater number of patches, the reconstructed images exhibit noticeable differences from the originals (e.g., 50% vs. 95%), which is expected. It is worth noting, however, that SpectralGPT holds stronger reconstruction capability (*cf. SatMAE*), even if the masking rate has reached over 90%, showing its powerful learning, reasoning, and generalizing performance.

TABLE VII
LIST OF GEOGRAPHICAL CHARACTERISTICS AND VISUALIZED BAND COMBINATIONS CORRESPONDING TO OBSERVATION TARGETS IN PRACTICAL GEOSCIENCE APPLICATIONS.

Geo-characteristic	Band Combination	Observation Target
Agriculture Condition	B11, B8, B2	Crop Health
Bathymetric Survey	B4, B3, B1	Coast
Vegetation Health	B8, B4, B3	Chlorophyll
Geological Structure	B12, B11, B2	Faults, Lithology
Moisture Content	B11, B8A	Plant Water Pressure
Vegetation Density	B12, B8A, B4	Vegetation Cover, Soil, Building
Vegetation Index (NDVI)	B8, B4	Tree Crown, Urban, Waterscape
Atmospheric Penetration	B12, B11, B8A	Particle, Smoke, Haze, Thin Cloud

by original images) when compared to SatMAE, particularly in terms of band-wise spectral reconstruction capabilities and its application value in the context of EO tasks. In our study, we have identified eight geo-characteristics corresponding to observation targets in practical applications, as detailed in Table VII. Furthermore, notable visual differences are evident in the geo-characteristics obtained using SatMAE and SpectralGPT. These pronounced visual disparities can be attributed to spectral degradation stemming from the comparatively limited reconstructive and inferential capabilities of SatMAE when compared to our more powerful SpectralGPT.

IV. CONCLUSION

The explosive development of foundation models represents a significant technological revolution following the advent of deep learning. Currently, various industries are witnessing significant leaps in technology and application advancements, largely driven by the emergence of foundation models. The RS field is no exception, with numerous EO applications, reaping significant benefits.

Spectral imaging has gained recognition in EO for its ability to provide rich insights into the composition of observed objects and materials, making it a transformative technology with vast potential to address global challenges and reshape various industries. However, the ever-expanding availability of spectral data from various RS platforms undeniably presents formidable challenges. There is a pressing demand for the development of foundation models specifically designed for spectral RS data. To fully unlock and leverage the potential of spectral RS data, it is imperative to overcome and resolve several challenging obstacles. These include efficiently processing and utilizing diverse RS spectral big data from various sources, extracting meaningful knowledge representations from complex spatial-spectral mixed information, and tackling the spectral degradation of neighboring spectral relevance modeling.

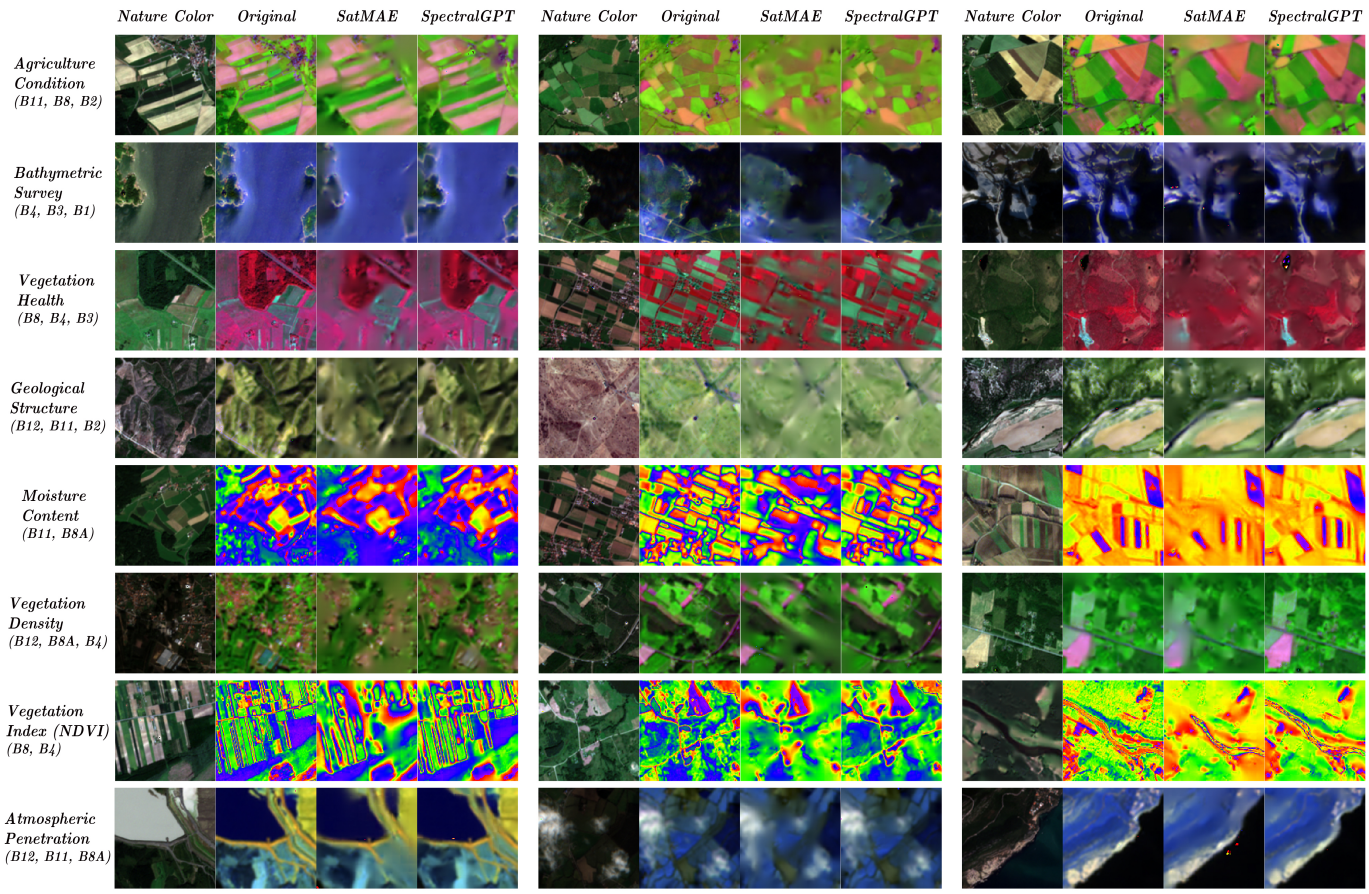

Fig. 10. Band combination visualization for geographical characteristic representations of reconstructed spectral images. These images are reconstructed and inferred by SatMAE and SpectralGPT models, respectively, only using 10% visible patches (masking 90% patches). Eight band combinations are given to highlight the characteristics, encompassing applications in agriculture, oceanography, geology, vegetation, atmosphere, and more, further serving a variety of EO applications.

In response to these challenges, we propose SpectralGPT, a customized spectral RS foundation model, featuring a novel 3D GPT architecture. With its innovative 3D GPT architecture, trained on over one million spectral images and over 600 million parameters, SpectralGPT empowers intelligent processing of spectral RS big data. SpectralGPT can flexibly handle diverse inputs in terms of size, resolution, temporal variability, and geographical coverage. This 3D masking strategy enables effective information extraction from spatial-spectral coupling tokens. Moreover, the innovative multi-target reconstruction is capable of capturing sequentially preserving spectral characteristics and meanwhile mitigating spectral degradation. Notably, our progressive training paradigm empowers the foundation model, surpassing transitional points in performance. These breakthroughs achieved by SpectralGPT democratize access to spectral RS big data, rendering it more accessible and cost-effective for large-scale EO applications.

Our study also includes a comprehensive assessment of MAE-based pretrained foundation models, with a focus on spectral reconstruction capabilities. We systematically evaluated model performance with inputs ranging from 50% to as low as 5% visible tokens. This extensive analysis allows us to gauge their proficiency in spectral-wise reconstruction and inference, especially significant in Geo-fields, such as agriculture, oceanography, geology, and vegetation. Visualizing the band combinations of reconstructed spectral images using both SatMAE and SpectralGPT demonstrates the latter's potential in practical EO tasks and Geo-field applications.

Looking ahead, our research will pursue several objectives. We plan to expand the volume and diversity of RS data used for training, encompassing various modalities, resolutions, time series, and image sizes. This enrichment will enhance the robustness of the RS foundation model. Furthermore, we aim to extend SpectralGPT's capabilities by incorporating a wider range of downstream tasks. This will transform SpectralGPT into a versatile AI model with improved generalization, well-suited for diverse EO and geoscience applications.

REFERENCES

[1] A. F. Goetz, G. Vane, J. E. Solomon, and B. N. Rock, "Imaging spectrometry for earth remote sensing," *Science*, vol. 228, no. 4704, pp. 1147–1153, 1985.

[2] B. Manifold, S. Men, R. Hu, and D. Fu, "A versatile deep learning architecture for classification and label-free prediction of hyperspectral images," *Nature Machine Intelligence*, vol. 3, no. 4, pp. 306–315, 2021.

[3] D. Hong, W. He, N. Yokoya, J. Yao, L. Gao, L. Zhang, J. Chanussot, and X. Zhu, "Interpretable hyperspectral artificial intelligence: When non-convex modeling meets hyperspectral remote sensing," *IEEE Geoscience and Remote Sensing Magazine*, vol. 9, no. 2, pp. 52–87, 2022.

[4] A. Plaza, J. A. Benediktsson, J. W. Boardman, J. Brazile, L. Bruzzone, G. Camps-Valls, J. Chanussot, M. Fauvel, P. Gamba, A. Gualtieri, *et al.*, "Recent advances in techniques for hyperspectral image processing," *Remote Sensing of Environment*, vol. 113, pp. S110–S122, 2009.

[5] M. Reichstein, G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, and f. Prabhat, "Deep learning and process understanding for data-driven earth system science," *Nature*, vol. 566, no. 7743, pp. 195–204, 2019.

[6] M. Ziatdinov, A. Ghosh, C. Y. Wong, and S. V. Kalinin, "Atomai framework for deep learning analysis of image and spectroscopy data in electron and scanning probe microscopy," *Nature Machine Intelligence*, vol. 4, no. 12, pp. 1101–1112, 2022.

[7] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4, pp. 8–36, 2017.

[8] D. Tuia, B. Kellenberger, S. Beery, B. R. Costelloe, S. Zuffi, B. Risse, A. Mathis, M. W. Mathis, F. van Langevelde, T. Burghardt, *et al.*, "Perspectives in machine learning for wildlife conservation," *Nature Communications*, vol. 13, no. 1, p. 792, 2022.

[9] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[10] J. Tian, X. Sun, Y. Du, S. Zhao, Q. Liu, K. Zhang, W. Yi, W. Huang, C. Wang, X. Wu, *et al.*, "Recent advances for quantum neural networks in generative learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[12] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, pp. 8748–8763, PMLR, 2021.

[13] X. Liu, D. Hong, J. Chanussot, B. Zhao, and P. Ghamisi, "Modality translation in remote sensing time series," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2021.

[14] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.

[15] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning*, pp. 1597–1607, PMLR, 2020.

[16] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, "Big self-supervised models are strong semi-supervised learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 22243–22255, 2020.

[17] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9912–9924, 2020.

[18] X. Chen and K. He, "Exploring simple siamese representation learning," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pp. 15750–15758, 2021.

[19] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *arXiv preprint arXiv:2003.04297*, 2020.

[20] Y. Xiong, M. Ren, and R. Urtasun, "Loco: Local contrastive representation learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 11142–11153, 2020.

[21] Z. Xie, Y. Lin, Z. Zhang, Y. Cao, S. Lin, and H. Hu, "Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16684–16693, 2021.

[22] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[23] H. Bao, L. Dong, S. Piao, and F. Wei, "Beit: Bert pre-training of image transformers," in *International Conference on Learning Representations*, 2022.

[24] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

[25] D. Wang, Q. Zhang, Y. Xu, J. Zhang, B. Du, D. Tao, and L. Zhang, "Advancing plain vision transformer towards remote sensing foundation model," *IEEE Transactions on Geoscience and Remote Sensing*, 2022.

[26] X. Sun, P. Wang, W. Lu, Z. Zhu, X. Lu, Q. He, J. Li, X. Rong, Z. Yang, H. Chang, *et al.*, "Ringmo: A remote sensing foundation model with masked image modeling," *IEEE Transactions on Geoscience and Remote Sensing*, 2022.

[27] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.

[28] Z. Tong, Y. Song, J. Wang, and L. Wang, "Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, pp. 10078–10093, 2022.

[29] C. Feichtenhofer, Y. Li, K. He, *et al.*, "Masked autoencoders as spatiotemporal learners," *Advances in Neural Information Processing Systems*, vol. 35, pp. 35946–35958, 2022.

[30] Y. Cong, S. Khanna, C. Meng, P. Liu, E. Rozi, Y. He, M. Burke, D. Lobell, and S. Ermon, "Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery," *Advances in Neural Information Processing Systems*, vol. 35, pp. 197–211, 2022.

[31] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, and L. Bottou, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion.," *Journal of Machine Learning Research*, vol. 11, no. 12, 2010.

[32] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[33] G. Christie, N. Fendley, J. Wilson, and R. Mukherjee, "Functional map of the world," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6172–6180, 2018.

[34] G. Sumbul, M. Charfuelan, B. Demir, and V. Markl, "Bigearthnet: A large-scale benchmark archive for remote sensing image understanding," in *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pp. 5901–5904, IEEE, 2019.

[35] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *The International Conference on Learning Representations (ICLR)*, 2019.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[37] O. Manas, A. Lacoste, X. Giró-i Nieto, D. Vazquez, and P. Rodriguez, "Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9414–9423, 2021.

[38] P. Helber, B. Bischke, A. Dengel, and D. Borth, "Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 7, pp. 2217–2226, 2019.

[39] M. Neumann, A. S. Pinto, X. Zhai, and N. Houlsby, "In-domain representation learning for remote sensing," in *International Conference on Learning Representations (ICLR) AI for Earth Sciences Workshop*, 2020.

[40] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 418–434, 2018.

[41] D. Hong, B. Zhang, X. Li, Y. Li, C. Li, J. Yao, N. Yokoya, H. Li, X. Jia, A. Plaza, P. Gamba, J. A. Benediktsson, and J. Chanussot, "SpectralGPT: The first remote sensing foundation model customized for spectral data," Oct. 2023.

[42] R. C. Daudt, B. Le Saux, A. Boulch, and Y. Gousseau, "Urban change detection for multispectral earth observation using convolutional neural networks," in *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pp. 2115–2118, Ieee, 2018.