

# CSMoE: An Efficient Remote Sensing Foundation Model with Soft Mixture-of-Experts

Leonard Hackel<sup>✉</sup>, *Graduate Student Member, IEEE*, Tom Burgert<sup>✉</sup>, *Member, IEEE*,  
and Begüm Demir<sup>✉</sup>, *Senior Member, IEEE*

**Abstract**—Self-supervised learning (SSL) through masked autoencoders (MAEs) has recently attracted great attention for remote sensing (RS) foundation model (FM) development, enabling improved representation learning across diverse sensors and downstream tasks. However, existing RS FMs often either suffer from substantial computational complexity during both training and inference or exhibit limited representational capacity. These issues restrict their practical applicability in RS. To address this limitation, we propose an adaptation for enhancing the efficiency of RS FMs by integrating the Soft mixture-of-experts (MoE) mechanism into the FM. The integration of Soft MoEs into the FM allows modality-specific expert specialization alongside shared cross-sensor representation learning. To demonstrate the effectiveness of our adaptation, we apply it on the Cross-Sensor Masked Autoencoder (CSMAE) model, resulting in the Cross-Sensor Mixture-of-Experts (CSMoE) model. In addition, we introduce a thematic-climatic descriptor-driven sampling strategy for the construction of a representative and diverse training set to train our CSMoE model. Extensive experiments on scene classification, semantic segmentation, and content-based image retrieval (CBIR) demonstrate that our adaptation yields a reduction in computational requirements while maintaining or improving representational performance. Compared to state-of-the-art RS FMs, CSMoE achieves a superior trade-off between representational capacity, accuracy, and computational efficiency. On average, CSMoE achieves more than twice the computational efficiency of existing RS FMs, while maintaining competitive performance across all experiments. These results highlight the effectiveness of the proposed adaptation for creating scalable and computationally efficient RS FMs. The associated code for the model and the training set creation, as well as the pretrained model weights, will be available at <https://git.tu-berlin.de/rsim/csmoe>.

**Index Terms**—Foundation models, self-supervised learning, mixture of experts, data subsampling, cross-modal retrieval, scene classification, semantic segmentation.

## I. INTRODUCTION

WITH the advances in self-supervised learning (SSL) and the increasing availability of large-scale earth observation (EO) data, the development of foundation models (FMs) has attracted great attention in the remote sensing (RS) community for representation learning problems [1], [2], [3], [4], [5], [6], [7], [8], [9]. FMs aim to learn general-purpose, task-agnostic representations that can process data from diverse sensors and solve downstream tasks with minimal fine-tuning. Unlike conventional deep learning (DL) models in RS, that are often tailored to specific tasks and data modalities, RS FMs aim for broader generalization. This is achieved by relying

on SSL-based learning objectives such as reconstruction-based (e.g., masked image modeling (MIM) [10]) or contrastive (e.g., MOCO [11] and DINOv2 [12]) learning for large-scale pretraining using a large amount of unlabeled data. Recently, the design and development of FMs has mainly evolved along three primary axes: i) scaling up model size and training data to increase representational capacity [1], [2], [13]; ii) introducing architectural innovations to accurately represent the complex content of RS images. [4], [5], [6]; and iii) integrating multiple EO modalities to enhance multi-modal and cross-sensor characteristics [7], [8].

There are several FMs developed in RS, focusing on scaling model and dataset sizes to improve generalization of learned representations. As an example, Prithvi [1] scales to more than 600 million parameters and is trained on global time-series data from Sentinel-2 (S2) and Landsat-8/9 satellites, enabling improved performance on tasks such as disaster response and ecosystem monitoring. Satlas [2] introduces a multi-task dataset with over 300 million annotations. SSL4EO [3] complements these efforts with a global, seasonally diverse pretraining dataset, while Major TOM [13] contributes an extensible framework based on a geographical indexing system, and introduces a multi-modal dataset (called Major TOM Core) with up to 23 TB per image modality. Beyond scale, architectural innovations have emerged to better model the complex content of RS images. As an example, RingMo [4] introduces a generative masking strategy for masked autoencoders (MAEs) tailored to extracting fine-grained features from RS images. As another example, SatMAE [5] incorporates temporal and spectral encodings, while ScaleMAE [6] leverages resolution-aware positional embeddings for improved cross-scale performance. Progress has also been made in developing modality-agnostic FMs. For example, DOFA [7] employs a wavelength-conditioned dynamic patch embedding layer to accommodate different channel configurations, thereby supporting the change of the image modality at the time of inference without retraining. TerraMind [8], on the other hand, enables cross-modal generation through a dual-scale architecture and modality-conditioned decoding. All these developments reflect a growing trend toward scalable, general, and multi-modal FMs in RS. For a comprehensive overview of FMs in RS, we refer the reader to [14], [15], [16].

However, this progress comes at a significant computational cost. While FMs benefit from a high representational capacity, reflected in a large number of parameters that enable them to model complex functions, this capacity is often accompanied by substantial computational complexity, commonly measured

The authors are with the Berlin Institute for the Foundations of Learning and Data (BIFOLD) and Technische Universität Berlin, 10623 Berlin, Germany (emails: l.hackel@tu-berlin.de (corresponding author), t.burgert@tu-berlin.de, demir@tu-berlin.de).

TABLE I: Comparison of RS FMs regarding their model sizes as number of parameters (#P), computational complexity in FLOPs, resulting  $C_2C$  ratio, and number of pixels in the pretraining dataset (PT DS). The FLOP-calculation is based on a forward pass of a single S2-image (at  $224 \times 224$  pixels with the subset of bands supported by the respective model) for feature extraction. For our CSMoE model a patch size of 16 was used. Evaluation based on reference implementations in TerraTorch [23]. M = Million, B = Billion, T = Trillion.

Model	multi-modal	#P $\uparrow$	FLOPs $\downarrow$	$C_2C$ Ratio $\uparrow$	PT DS # pixels $\downarrow$
Prithvi V2-300 [1]	$\times$	304M	59.85B	5.08	210.7B
Prithvi V2-600 [1]	$\times$	<b>631M</b>	162.18B	3.89	210.7B
Satlas [2]	$\times^1$	88M	17.12B	5.14	14.6T
DOFA [7]	$\checkmark$	111M	17.47B	6.35	20.6B
TerraMind [8]	$\checkmark$	87M	17.84B	4.88	451.6B
CSMAE [9]	$\checkmark$	87M	<b>5.64B</b>	15.43	<b>3.9B</b>
CSMoE (ours)	$\checkmark$	271M	10.11B	<b>26.81</b>	14.5B

in floating-point operations (FLOPs), needed for inference of an image. The importance of computational efficiency has already been recognized in various RS tasks, such as scene classification [17], [18], [19], semantic segmentation [19], [20], and visual question answering (VQA) [21], [22]. By contrast, such efficiency-oriented approaches remain largely unexplored in the context of FMs. Table I shows the computational complexity and representational capacity of different FMs as well as their ability to process multiple data modalities and the size of their pretraining datasets. To illustrate the trade-off between computational complexity and representational capacity, in the table we also report the ratio of parameters (in millions) to FLOPs (in billions), which we refer to as the capacity-to-compute ( $C_2C$ ) ratio:

$$C_2C = \frac{\# \text{ Params (in millions)}}{\text{FLOPs (in billions)}}. \quad (1)$$

Higher ratios indicate more efficient FMs in terms of representational capacity per unit of computation. By analyzing the table, one can see that some FMs focus on maximizing representational capacity (e.g., Prithvi2-600 with over 630 million parameters). However, their computational complexity during inference exceeds 160 billion FLOPs per image, resulting in a low  $C_2C$  ratio. Others, such as Satlas, operate at a lower computational budget but offer limited representational capacity with only 88 million parameters. Models like DOFA and TerraMind introduce valuable multi-modal learning capabilities, yet their per-sample inference cost remains high (approximately 17.5 billion FLOPs), indicating that improvements in generalization often come with significant computational resource demands.

Although inference efficiency is essential for model deployment, pretraining efficiency is equally critical given the scale of recent RS FMs. Reducing the dataset size by eliminating redundant samples is therefore essential to improve training efficiency and scalability [24]. Training on datasets comprising

hundreds of billions of pixels introduces substantial engineering and computational challenges [25]. Efficient storage, loading, and throughput must be ensured to prevent data pipeline bottlenecks, which often require specialized infrastructure. In addition to these computational challenges, training datasets can contain redundant samples, such as images from the same land-use/land-cover (LULC) class under the same climate zone (e.g., extensive deserts or forests) [1], [3], [26], [27]. When included in the training data, these additional samples may not contribute significant new information [28]. Therefore, when the considered models are computationally intensive, a high amount of redundancies may increase the training costs without improving the capability of the learned representations, resulting in minor improvements relative to the increasing computational demands. Recent studies have begun to address similar challenges in general machine learning through automatic data selection and curation strategies (e.g., graph-based [29] and clustering-based [24] approaches). In the context of large-scale RS FMs, however, such strategies remain underexplored and are not yet widely established. In practice, the prevailing paradigm still assumes that increasing the size of the pretraining dataset uniformly translates into better performance, often without explicitly accounting for redundancy. The above-mentioned issues result in: i) the computational inefficiency of current RS FMs during pretraining and inference; and ii) the absence of efficient and widely established data selection strategies to mitigate redundancies in pretraining data for RS FMs.

To address these issues and achieve efficiency in training and inference, we propose an approach that aims to inject a mixture-of-experts (MoE) into a RS FM. Although prior studies have explored the use of MoEs in RS model development [30], [31], [32], [33], these works primarily employ MoEs as a means to scale model capacity, rather than to explicitly target computational efficiency. In contrast, our focus is on leveraging MoEs to enhance efficiency, for which we adopt the Soft MoE [34], a variant designed to combine high representational capacity with reduced computational cost. When Soft MoEs are applied in transformers [35], [36], each input token is softly routed to multiple expert branches, and their outputs are combined using the learned routing weights. The mechanism is particularly well-suited for efficient RS FMs, as it combines a lightweight routing mechanism with soft token mixing, thereby increasing representational capacity while reducing computational complexity. We inject the Soft MoE into the Cross-Sensor Masked Autoencoder (CSMAE) [9] FM to create our FM that we call Cross-Sensor Mixture-of-Experts (CSMoE). In addition, we introduce an efficient thematic-climatic descriptor-driven sampling strategy that selects a representative training set from a large-scale image archive, while retaining the full geographic and thematic-climatic diversity of the original archive. To this end, our proposed thematic-climatic descriptor-driven sampling strategy consists of two stages: The first stage aims at preserving thematic-climatic diversity while reducing the sample size. This is achieved by assigning each sample a combination of climatic and thematic descriptors (e.g., a climate zone and a set of LULC classes). Then, a fixed number of samples is drawn from

<sup>1</sup>Different versions of the model exist for individual modalities, but no unified model for multiple modalities.

each combination of thematic-climatic descriptors to build an initial subsampled dataset. In the second stage, the aim is to enhance the spatial diversity of the subset. This is achieved by applying a genetic algorithm [37] that maximizes the geodesic distance within the selected samples. Through these two stages, the strategy reduces redundancy by discarding semantically similar samples (e.g., multiple samples from the same combination of climate and thematic product descriptors) while maintaining the geographic and thematic-climatic diversity of the original archive. This helps reduce the training time and environmental impact of FM training in RS. We train our CSMoE model using a subset of MajorTOM Core [13], where we select a training set using our thematic-climatic descriptor-driven sampling strategy. Through extensive experiments, we demonstrate that our CSMoE model archives performance compared to state-of-the-art FMs, while significantly reducing inference cost. These results underscore CSMoE's efficiency and strong generalization capability across diverse remote sensing tasks.

The main contributions of this work are summarized as follows:

- We propose the first computationally efficient FM in RS, CSMoE, that injects a Soft MoE into CSMAE. As shown in Table I, this integration enhances the model's representational capacity while significantly reducing computational complexity during both training and inference compared to models of similar size.
- We introduce a new thematic-climatic descriptor-driven sampling strategy for sampling from large-scale, unlabeled RS archives. The strategy selects the optimal training set based on the user-defined sample count.
- We conduct extensive experiments on a suite of single- and multi-label scene classification and semantic segmentation (pixelwise classification) benchmarks, as well as both uni-modal and cross-modal content-based image retrieval (CBIR) tasks in RS and compare our CSMoE model with state-of-the-art FMs in RS. We demonstrate that CSMoE has a comparable or superior downstream performance while requiring fewer computational resources than competing models.

The remainder of this paper is organized as follows: In Section II, we introduce the proposed CSMoE model and our efficient thematic-climatic descriptor-driven sampling strategy. Section III describes the experimental setup, including the construction of the pretraining dataset via the proposed sampling strategy, as well as the downstream datasets and implementation details. Section IV presents a comprehensive evaluation of the CSMoE model on scene classification, semantic segmentation, and image retrieval tasks, along with a sensitivity analysis. Finally, Section V concludes the paper.

## II. PROPOSED EFFICIENT REMOTE SENSING FOUNDATION MODEL WITH SOFT MIXTURE-OF-EXPERTS

Let  $\mathcal{X}$  and  $\mathcal{Y}$  denote two co-registered multi-modal remote sensing image archives associated with different image modalities (i.e., acquired by different sensors). The archives  $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^N$  and  $\mathcal{Y} = \{\mathbf{y}_i\}_{i=1}^N$  each include  $N$  images,

where  $\mathbf{x}_i$  and  $\mathbf{y}_i$  are the  $i$ th RS images in the respective archives and  $(\mathbf{x}_i, \mathbf{y}_i)$  is the  $i$ th multi-modal image pair that includes two images acquired by different sensors on the same geographical area. We assume that an unlabeled training set  $\mathcal{T} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$  is available for representation learning. Although our approach can be extended to multiple modalities by applying the proposed processing steps in a pairwise combination, we focus on the dual-modality case to simplify the mathematical treatment in this paper.

The computational complexity of current RS FMs during both training and inference poses significant challenges for operational deployment. To address this limitation, we propose to integrate the MoE mechanism into existing MAE-based FM architectures to achieve improved computational efficiency while maintaining or increasing representational capacity. For representation learning from the training set  $\mathcal{T}$ , MIM (i.e., SSL through MAEs) has recently emerged as one of the most successful approaches in RS. Among various MAE variants, the CSMAE model [9] has shown promising results in both unimodal and cross-modal representation learning. However, like all existing multi-modal MAE-based FMs, CSMAE suffers from a high computational complexity during training and inference with limited representational capacity. As a first time in RS, we explore the effectiveness of integrating the Soft MoE mechanism into MAEs to create computationally efficient FMs. To this end, we select CSMAE as our base model due to its proven cross-modal capabilities and apply our MoE injection adaptation to create CSMoE. In the following subsections, we first provide background information on the CSMAE model, and then present our adaptation that injects the Soft MoE into CSMAE and our efficient thematic-climatic descriptor-driven sampling strategy.

### A. Basics on Cross-Sensor Masked Autoencoder

Given an image  $\mathbf{x} \in \mathcal{X}$  or  $\mathbf{y} \in \mathcal{Y}$ , MAEs operate by dividing the image into  $P$  non-overlapping image patches of size  $\rho \times \rho$ , forming a token set  $\mathcal{P} = \{\mathbf{p}_n\}_{n=1}^P$ . A random subset  $\mathcal{M} \subset \{1, \dots, P\}$  is masked, while the remaining tokens  $\mathcal{U} = \{1, \dots, P\} \setminus \mathcal{M}$  are passed to a transformer encoder  $e$  to produce latent representations  $\mathcal{Z}_{\mathcal{U}}$  as follows:

$$\mathcal{Z}_{\mathcal{U}} = e(\{\mathbf{p}_n\}_{n \in \mathcal{U}}). \quad (2)$$

These representations, along with a learned mask token  $\mathbf{z}_m$ , are used by a decoder  $d$  to get the reconstructed tokens  $\{\hat{\mathbf{p}}_n\}$  as follows:

$$\{\hat{\mathbf{p}}_n\} = d(\mathcal{Z}_{\mathcal{U}}, \mathbf{z}_m). \quad (3)$$

The learning objective is computed over the masked positions on the reconstructed tokens:

$$\mathcal{L}_{\text{UMR}} = \text{RecL}(\{\hat{\mathbf{p}}_n\}_{n \in \mathcal{M}}, \{\mathbf{p}_n\}_{n \in \mathcal{M}}), \quad (4)$$

where RecL is a reconstruction loss (e.g., mean average error or mean square error).

The CSMAE model [9] extends this formulation by introducing two adaptations of MAEs to enable cross-modal representation learning: i) extending the encoder into a multi-sensor encoder and a cross-sensor encoder; and ii) extending

the learning objective with latent similarity preservation and cross-modal reconstruction. To accomplish this, first two independent masks  $\mathcal{M}^{\mathcal{X}}$  and  $\mathcal{M}^{\mathcal{Y}}$  and their inverse  $\mathcal{U}^{\mathcal{X}}$  and  $\mathcal{U}^{\mathcal{Y}}$  are created. These are used on the input images, as described above, to get a set of tokens for each image modality. Then, the multi-sensor encoder, which uses one sensor-specific vision transformer (ViT) encoder per modality, is applied to the unmasked patches. Here, each sensor-specific encoder ( $e_{MS}^{\mathcal{X}}$  or  $e_{MS}^{\mathcal{Y}}$ ) has its own set of parameters, enabling the accurate modeling of sensor-specific image characteristics. Then the cross-sensor encoder  $e_{CS}$  (a ViT encoder that uses shared weights among all modalities) is applied on the output of the multi-sensor encoder. It aligns the latent representations of the different modalities into a shared embedding space by processing the output of the multi-sensor encoder as follows:

$$\mathcal{Z}_{\mathcal{U}}^{\mathcal{X}} = e_{CS}(e_{MS}^{\mathcal{X}}(\{\mathbf{p}_n^{\mathcal{X}}\}_{n \in \mathcal{U}^{\mathcal{X}})}), \quad (5)$$

$$\mathcal{Z}_{\mathcal{U}}^{\mathcal{Y}} = e_{CS}(e_{MS}^{\mathcal{Y}}(\{\mathbf{p}_n^{\mathcal{Y}}\}_{n \in \mathcal{U}^{\mathcal{Y}})}). \quad (6)$$

To enable cross-modal alignment, the CSMAE model is trained using two cross-modal learning objectives: i) cross-modal reconstruction; and ii) latent similarity preservation. For cross-modal reconstruction, two sensor-specific decoders  $d^{\mathcal{X}}$  and  $d^{\mathcal{Y}}$  are employed. Each decoder takes as input the features of one modality ( $\mathcal{Z}_{\mathcal{U}}^{\mathcal{X}}$  or  $\mathcal{Z}_{\mathcal{U}}^{\mathcal{Y}}$ ) as well as a modality-specific mask token ( $\mathbf{z}_m^{\mathcal{X}}$  or  $\mathbf{z}_m^{\mathcal{Y}}$ ) to reconstruct the masked patches of the other modality as follows:

$$\{\hat{\mathbf{p}}_n^{\mathcal{Y}}\} = d(\mathcal{Z}_{\mathcal{U}}^{\mathcal{X}}, \mathbf{z}_m^{\mathcal{Y}}), \quad (7)$$

$$\{\hat{\mathbf{p}}_n^{\mathcal{X}}\} = d(\mathcal{Z}_{\mathcal{U}}^{\mathcal{Y}}, \mathbf{z}_m^{\mathcal{X}}). \quad (8)$$

The cross-modal reconstruction loss  $\mathcal{L}_{\text{CMR}}$  is then calculated using the same reconstruction loss as the single-modal reconstruction:

$$\begin{aligned} \mathcal{L}_{\text{CMR}} = & \text{ReL}(\{\hat{\mathbf{p}}_n^{\mathcal{Y}}\}_{n \in \mathcal{M}^{\mathcal{X}}}, \{\mathbf{p}_n^{\mathcal{Y}}\}_{n \in \mathcal{M}^{\mathcal{X}}}) \\ & + \text{ReL}(\{\hat{\mathbf{p}}_n^{\mathcal{X}}\}_{n \in \mathcal{M}^{\mathcal{Y}}}, \{\mathbf{p}_n^{\mathcal{X}}\}_{n \in \mathcal{M}^{\mathcal{Y}}}). \end{aligned} \quad (9)$$

Note that here, for the loss calculation of one modality, the mask of the other modality is used, as this mask is also used to calculate the latent features.

For preservation of latent similarity, the mutual information loss  $\mathcal{L}_{\text{MI}}$  [38], which is based on contrastive learning with normalized temperature-scaled cross-entropy [39], is used. For a batch of  $|\mathcal{B}|$  image pairs, let  $\mathbf{c}_i^{\mathcal{X}}, \mathbf{c}_i^{\mathcal{Y}}$  be projected representations (e.g., projected [CLS] tokens) that should be aligned in the latent space. Then, using a similarity function  $\text{sim}(\cdot, \cdot)$  (e.g., cosine similarity), a temperature parameter  $\tau$  and the indicator function  $\mathbb{1}$ ,  $\mathcal{L}_{\text{MI}}$  can be defined as:

$$\ell^i(\mathcal{X}, \mathcal{Y}) = -\log \left( \frac{\exp(\text{sim}(\mathbf{c}_i^{\mathcal{X}}, \mathbf{c}_i^{\mathcal{Y}})/\tau)}{\sum_{q \in \mathcal{B}} \mathbb{1}_{[q \neq i]} \exp(\text{sim}(\mathbf{c}_i^{\mathcal{X}}, \mathbf{c}_q^{\mathcal{Y}})/\tau)} \right), \quad (10)$$

$$\mathcal{L}_{\text{MI}}(\mathcal{B}) = \frac{1}{2|\mathcal{B}|} \sum_{i \in \mathcal{B}} (\ell^i(\mathcal{X}, \mathcal{Y}) + \ell^i(\mathcal{Y}, \mathcal{X})). \quad (11)$$

Additionally, the same learning objective  $\mathcal{L}_{\text{UMR}}$  as in single-modal MAEs is applied on both modalities individually.

## B. The proposed CSMoE Model

To simultaneously address the computational complexity limitations of masked autoencoder-based models while preserving cross-modal representation learning, we introduce a general adaptation for efficient multi-modal processing. Although the proposed adaptation can be applied to any transformer-based FM, in this paper we apply the adaptation to CSMAE, resulting in the Cross-Sensor Mixture-of-Experts (CSMoE) model, the first compute-efficient multi-modal FM for RS. The proposed CSMoE model extends CSMAE by integrating Soft MoE mechanisms into the encoder components, enabling selective expert activation while maintaining both intra-modal and inter-modal characteristics. To achieve this, we adapt CSMAE by: i) integrating expert routing; ii) adapting the encoder architecture; and iii) including regularizing training objectives. Fig. 1 shows an illustration of the CSMoE model, while our adaptations are explained in detail in the following.

The proposed CSMoE model leverages Soft MoE to reduce computational complexity while maintaining model capacity through selective expert activation. Unlike sparse MoEs [40] that route individual tokens to top- $k$  out of  $R$  experts, Soft MoE employs a two-stage routing mechanism based on  $S$  intermediate representations called slots  $\{\mathbf{s}_{\vartheta}\}_{\vartheta=1}^S$ . As shown in Fig. 2, each slot summarizes a subset of input features through weighted aggregation, where slot-feature similarities are converted into dispatch weights  $\alpha_{\vartheta, n}$  and combine weights  $\hat{\alpha}_{\vartheta, n}$ :

$$\alpha_{\vartheta, n} = \text{softmax}_n \left( \frac{\hat{\mathbf{s}}_{\vartheta, n}}{\tau} \right), \quad \hat{\alpha}_{\vartheta, n} = \text{softmax}_{\vartheta}(\hat{\mathbf{s}}_{\vartheta, n}), \quad (12)$$

where  $\tau$  is a temperature parameter,  $\text{softmax}_n$  and  $\text{softmax}_{\vartheta}$  are softmax operations along the token and slot dimensions, respectively and  $\{\hat{\mathbf{s}}_{\vartheta, n}\}_{n=1}^P$  are slot embeddings (learned projections of the input features  $\{\mathbf{z}_n\}_{n=1}^P$ ) for expert  $\vartheta$ . Each slot  $\mathbf{s}_{\vartheta}$  is formed as a weighted average of the input features as:

$$\mathbf{s}_{\vartheta} = \sum_{n=1}^P \alpha_{\vartheta, n} \cdot \mathbf{z}_n. \quad (13)$$

The slots are then processed by one assigned expert each and the final outputs are reconstructed through weighted aggregation. This approach reduces expert calls from  $k \cdot P$  to  $S$ , enabling efficient processing while preserving representational capacity.

The CSMoE model incorporates Soft MoE into both encoder components of CSMAE. The multi-sensor encoder employs modality-specific MoE layers, where separate expert networks with different parameters are utilized for different image modalities, allowing accurate modeling of sensor-specific characteristics. Additionally, the cross-sensor encoder employs shared MoE layers, where the same expert networks process features from all modalities to facilitate inter-modal pattern learning. Based on this, we define the CSMoE model as follows:

1) *Modality-Specific Encoding*: Each image modality is processed through its dedicated encoder consisting of  $L_{\text{MS}}^E$  Soft MoE transformer encoder layers. For example, for modality  $\mathcal{X}$ ,



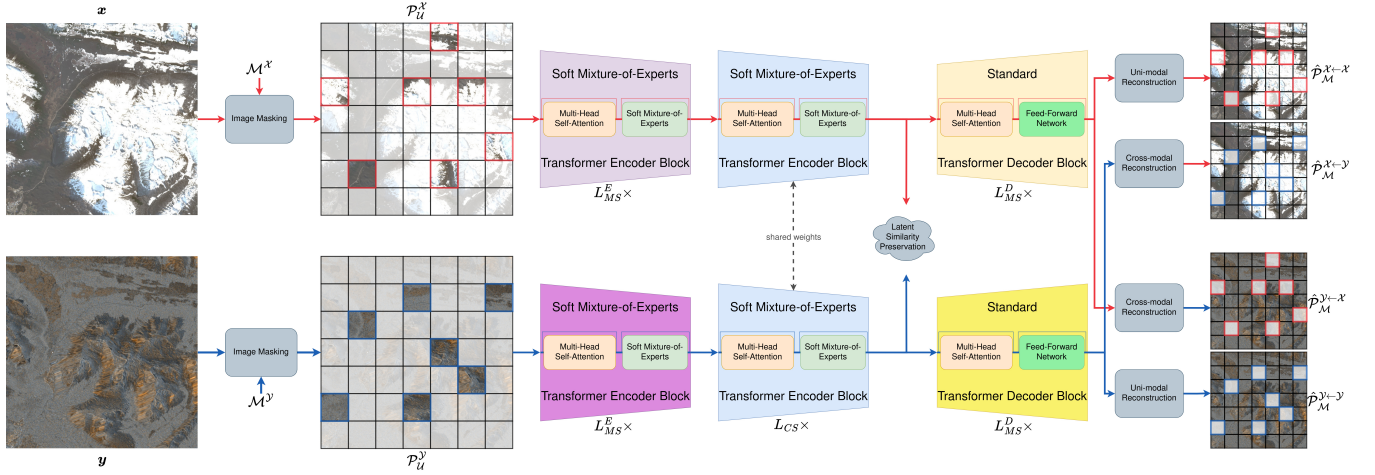


Fig. 1: An illustration of the CSMoE model for two image modalities. Each modality has one modality-specific encoder consisting of  $L_{MS}^E$  Transformer-MoE layers and one modality-specific decoder consisting of  $L_{MS}^D$  Transformer layers. Additionally, all modalities share a cross-sensor encoder consisting of  $L_{CS}^E$  Transformer-MoE layers. Different colors indicate the processing pathways for each RS image modality.

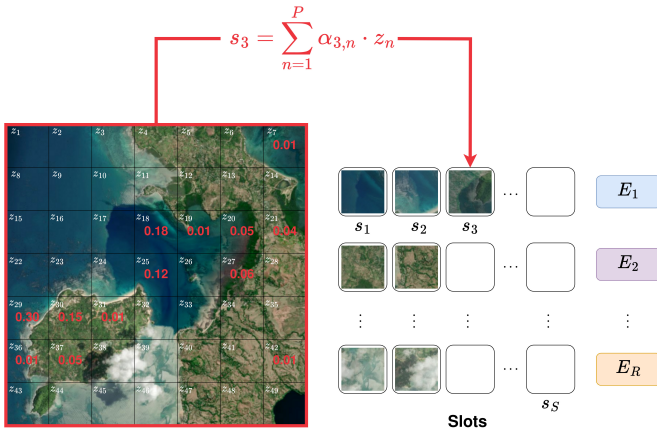


Fig. 2: Qualitative example of the dispatch process. Slots ( $s_1, \dots, s_S$ ) are formed as a weighted combination of the input features ( $z_1, \dots, z_{49}$  for  $P = 49$  patches) and assigned to experts ( $E_1, \dots, E_R$ ). The dispatch weights  $\alpha_{3,n}$  for slot 3 are shown in red, features without annotation have a zero-weight. Note, that in the model this process happens in the embedding space but for visualization it is shown in the image space.

the unmasked tokens  $\mathcal{P}_U^x$  are processed using its modality-specific encoder  $e_{MS}^x$  as:

$$\mathcal{Z}_{MS}^x = e_{MS}^x(\mathcal{P}_U^x), \quad (14)$$

where each encoder layer applies self-attention followed by Soft MoE processing. For modality  $\mathcal{Y}$ , the processing is identical using its respective modality-specific encoder  $e_{MS}^y$ . These modality-specific encoders enable expert specialization for different semantic patterns within each image modality.

2) *Cross-Sensor Encoding*: The outputs from both modality-specific encoders are processed through a shared

cross-sensor encoder  $e_{CS}$  consisting of  $L_{CS}^E$  Soft MoE transformer encoder layers:

$$\mathcal{Z}_{CS}^x = e_{CS}(\mathcal{Z}_{MS}^x) \quad (15)$$

$$\mathcal{Z}_{CS}^y = e_{CS}(\mathcal{Z}_{MS}^y). \quad (16)$$

The shared Soft MoE transformer encoder layers learn to route cross-modal patterns to appropriate experts, facilitating inter-modal relationship modeling while maintaining computational efficiency.

3) *Modality-Specific Decoding*: For reconstruction, separate decoders  $d_{MS}^j$  consisting of  $L_{MS}^D$  standard transformer layers are employed. The decoders process cross-sensor representations to reconstruct both unimodal and cross-modal targets:

$$\hat{\mathcal{P}}_{\mathcal{M}}^{j \leftarrow j'} = \text{Linear}_{j \leftarrow j'}(d_{MS}^j(\mathcal{Z}_{CS}^{j'}, \mathcal{Z}_{\mathcal{M}}^{j'})), \quad (17)$$

where  $j, j' \in \{\mathcal{X}, \mathcal{Y}\}$  and  $j \leftarrow j'$  denotes the reconstruction of the image modality  $j$  using features of the image modality  $j'$ . The decoders do not include MoE components as they are removed after pretraining.

4) *Training Objective Extensions*: To regularize expert routing and ensure stable training, we extend the CSMAE training objectives with two additional loss functions. The slot repulsion loss  $\mathcal{L}_{\text{REP}}$  encourages reduced correlation between slot embeddings:

$$\mathcal{L}_{\text{REP}} = -\frac{1}{S^2} \sum_{\vartheta=1}^S \sum_{\vartheta'=1}^S (\langle \tilde{s}_{\vartheta}, \tilde{s}_{\vartheta'} \rangle)^2, \quad (18)$$

while the entropy loss  $\mathcal{L}_{\text{ENT}}$  promotes balanced expert usage:

$$\mathcal{L}_{\text{ENT}} = -\frac{1}{SP} \sum_{\vartheta=1}^S \sum_{n=1}^P \alpha_{\vartheta,n} \log(\alpha_{\vartheta,n} + \varepsilon), \quad (19)$$

where  $\varepsilon$  is a small term for numerical stability.

The overall training objective combines the CSMAE losses with regularization terms:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{UMR}} + \mathcal{L}_{\text{CMR}} + \mathcal{L}_{\text{MI}} + \lambda \cdot \mathcal{L}_{\text{REP}} + \gamma \cdot \mathcal{L}_{\text{ENT}}, \quad (20)$$

where  $\lambda$  and  $\gamma$  are scaling parameters controlling the regularization influence. In this way, the characterization of cross-modal RS image representations is achieved by learning to reconstruct both intra-modal and inter-modal image content while maintaining computational efficiency through selective expert activation and promoting stable expert usage through regularized routing.

### C. A Thematic-Climatic Descriptor-Driven Sampling Strategy

The efficiency of FM-training in RS depends not only on model design choices, but also on training data selection. The use of global datasets often introduces substantial redundancy, where vast amounts of data contribute little additional information. For instance, when training on global coverage, more than 70% of the images may depict oceans, which are irrelevant for most land-focused applications. Even within land areas, extended homogeneous regions, such as the Amazon rainforest, the Siberian taiga, or the Sahara desert, dominate the dataset but provide limited additional information.

To address this issue, we propose a thematic-climatic descriptor-driven sampling strategy that samples a reduced but representative training set from a large-scale RS data archive. This strategy efficiently samples a training set with a wide range of geographic and thematic-climatic variability. The strategy consists of two stages: i) auxiliary descriptor generation, where each image is associated with climatic and thematic descriptors derived from global climate zone and thematic product datasets; and ii) entropy-maximizing stratified sampling, where a genetic algorithm selects spatially dispersed and thematically-climatically balanced samples.

1) *Auxiliary Descriptor Generation*: Let  $\mathcal{D} = \{\varphi_i\}_{i=1}^N$  denote the set of bounding boxes (the set of coordinates that define the spatial extend) of the images within a large-scale RS data archive  $\mathcal{T}$  that consists of  $N$  images or co-registered image tuples, where  $\varphi_i$  contains the longitude and latitude coordinates of the bounding box of the  $i$ th image of  $\mathcal{T}$ . We assume that the images are unlabeled, i.e., no thematic (e.g., LULC) or climatic information is directly available. In the first stage, each image is associated with a thematic and a climatic descriptor derived from global raster layers. Specifically, each bounding box  $\varphi_i$  is linked to: i) a climate zone raster  $M_{\text{climate}}$  (e.g., Köppen-Geiger [41]), which provides coarse climatic information; and ii) a thematic land-cover product  $M_{\text{TP}}$  (e.g., ESA WorldCover [42]), which captures land-use properties at finer scale. This combination allows us to incorporate both climate and local land-cover information, two complementary dimensions of variability, into the sampling process.

In detail, each bounding box  $\varphi_i$  is associated with a climate stratum  $u_i$  and a thematic stratum  $v_i$ , depending on the overlay with the reference rasters. The resulting set of tuples is therefore represented as  $\mathcal{D}' = \{(\varphi_i, u_i, v_i)\}_{i=1}^{N'} \mid N' \leq N$ , where only images covered by both rasters are retained. We refer to these thematic-climatic strata as descriptors since they are not

manually labeled but derived from external data sources. These descriptors are used for the subsequent stratified sampling stage by embedding thematic and climatic variability directly into the sampling process. The algorithm of the descriptor generation stage is given in Algorithm 1.

---

#### Algorithm 1 Assign Thematic and Climatic Descriptors to Coordinates

---

**Require:** Set of coordinates  $\mathcal{D} = \{\varphi_i\}_{i=1}^N$   
**Require:** Climate map  $M_{\text{climate}}$  with bounds  $M_{\text{climate}}^B$   
**Require:** Thematic product map  $M_{\text{TP}}$  with bounds  $M_{\text{TP}}^B$   
**Ensure:** Descriptors for each coordinate: climate stratum  $u_i$  and thematic product stratum  $v_i$

- 1: Initialize empty annotated coordinates  $\mathcal{D}' = \{\}$
- 2: **for** each coordinate  $\varphi_i \in \mathcal{D}$  **do**
- 3:   **if**  $\varphi_i \notin M_{\text{climate}}^B$  or  $\varphi_i \notin M_{\text{TP}}^B$  **then**
- 4:     Continue to next coordinate
- 5:   **end if**
- 6:    $u_i \leftarrow M_{\text{climate}}[\varphi_i]$
- 7:    $v_i \leftarrow M_{\text{TP}}[\varphi_i]$
- 8:   Add  $(\varphi_i, u_i, v_i)$  to  $\mathcal{D}'$
- 9: **end for**
- 10: **return** Set of pseudo-annotated tuples  $\mathcal{D}'$

---

2) *Entropy-Maximizing Stratified Sampling*: Given the set of tuples  $\mathcal{D}'$  containing locations and descriptors, the entropy-maximizing stratified sampling aims to construct a reduced training set  $\mathcal{D}^*$  that preserves the spatial and thematic-climatic variability of  $\mathcal{T}$  while minimizing redundancies among the selected samples. To this end, we stratify  $\mathcal{D}'$  by joined thematic and climatic descriptors and then sample within each joined stratum using an entropy-driven sampling process. Specifically, for each pair  $(u, v)$  of climatic and thematic strata, we define:

$$\mathcal{S}_{u,v} = \{(\varphi_i, u_i, v_i) \in \mathcal{D}' \mid u_i = u, v_i = v\}. \quad (21)$$

While some joined strata are small and can be fully retained (i.e., the size of the joined stratum is smaller than a user-defined target sample count per stratum  $N_s$ ), others are extensive and can be sampled from to reduce redundancy. In contrast to random sampling, we use the spatial dispersion (distance) of selected samples within each joined stratum as a criterion for sampling. To this end, we employ a genetic algorithm that iteratively optimizes a subset of samples within  $\mathcal{S}_{u,v}$ . Each candidate solution  $\{\mathcal{S}_{u,v} \mid \mathbf{b}\}$  is a set of samples, where  $\mathbf{b}$  is a binary mask indicating which samples of the joined stratum are selected for the candidate solution. The fitness of the candidate solution is measured by an entropy-based score that uses the pairwise Haversine distances. This encourages the selection of samples that are representative and spatially diverse. Standard evolutionary operators are used to optimize candidate solutions. Additionally, to maintain the target sample counts per joined stratum  $N_s$ , we use an adaptive constraint mechanism: if a candidate solution exceeds 110% of the target sample count, it is randomly pruned; if it falls below 90% it is randomly augmented.

After convergence, or if a given compute budget is reached, the candidate solution with the highest fitness is selected. The

process is repeated independently across all joined strata, and the union of candidate solutions with the highest respective fitness defines the training set  $\mathcal{D}^*$ . This strategy ensures that the resulting training set is spatially well-dispersed, thematically and climatically balanced, and substantially less redundant than the original archive. The algorithm of the optimization stage is given in Algorithm 2.

---

**Algorithm 2** Entropy-Based Subsampling via Genetic Optimization

---

**Require:** Set of location-descriptor tuples  $\mathcal{D}'$   
**Require:** Number of samples per joined stratum  $N_s$ , number of iterations  $T$ , population size  $N_p$ , climate zones  $C_u$ , thematic product classes  $C_v$   
**Require:** Mutation rate  $r_m$ , crossover rate  $r_c$   
**Ensure:** Optimized stratified sample subset  $\mathcal{D}^*$  with high spatial entropy

- 1: Initialize empty sample subset  $\mathcal{D}^* = \{\}$
- 2: **for all** climatic strata  $u \in C_u$  **do**
- 3:   **for all** thematic strata  $v \in C_v$  **do**
- 4:     Extract subset  $\mathcal{S}_{u,v} \subseteq \mathcal{D}'$  where  $u_i = u, v_i = v$
- 5:      $n_g \leftarrow |\mathcal{S}_{u,v}|$
- 6:     **if**  $n_g \leq N_s$  **then**
- 7:       Add all elements of  $\mathcal{S}_{u,v}$  to  $\mathcal{D}^*$
- 8:     **else**
- 9:       Define objective function
- 10:        $f_g(\mathbf{b}) \leftarrow \text{Entropy}(\{\mathcal{S}_{u,v} \mid \mathbf{b}\})$
- 11:       Run genetic algorithm
- 12:        $\mathbf{b}^* \leftarrow \text{GA}(f_g, n_g, T, N_p, N_s, r_m, r_c)$
- 13:        $\mathcal{S}_{u,v}^* \leftarrow \{\mathcal{S}_{u,v} \mid \mathbf{b}^*\}$
- 14:        $\mathcal{D}^* \leftarrow \mathcal{D}^* \cup \mathcal{S}_{u,v}^*$
- 15:     **end if**
- 16:   **end for**
- 17: **end for**
- 18: **return**  $\mathcal{D}^*$

---

### III. DATASETS AND EXPERIMENTAL SETUP

#### A. Efficient Sampling of Large-Scale Pretraining Data

We pretrain on a dataset that we construct based on Major TOM Core (MTom) [13]. We select MTC as our large-scale RS data archive since it is publicly available, contains globally distributed samples, and has already been adopted in recent studies, indicating its potential as an emerging benchmark for global-scale representation learning. This makes it a suitable and reproducible foundation for our sampling strategy. To construct our pretraining dataset, which we refer to as Minor TOM Core (MTom $_{\mu}$ ), we use our proposed thematic-climatic descriptor-driven sampling strategy on MTom. Additionally, to evaluate the effect of using the entropy-maximizing stratified sampling, we create a second dataset, denoted as Minor TOM Core (random) (MTom $_r$ ), which is created using the same thematic-climatic stratified sampling but without applying the genetic algorithm to increase the spatial diversity. Both datasets include only tiles with matching Sentinel-1 (S1) synthetic aperture radar (SAR) and S2 multispectral image pairs. For auxiliary descriptor generation, we utilize

TABLE II: Statistics on the full MTom dataset, a stratified subset MTom $_r$  and our entropy-based stratified subset MTom $_{\mu}$ .

Statistic	MTom	MTom $_r$	MTom $_{\mu}$
Number of tiles	1 302 691	18 785	18 846
Tiles per climate zone	43 423	626	628
Tiles per LULC class	118 426	1 707	1 713
Tiles per combination ( $N_s > 90$ )	7 434 $\sigma$ 16 395	98 $\sigma$ 4	100 $\sigma$ 2
Average Distance	6 737 km	8 192 km	8 648 km
Min. distance	10 km	32 km	36 km
Distance within $\mathcal{S}_{u,v}$	2 748 km	1 811 km	5 106 km

the Köppen-Geiger climate classification map [41] for climate zone information and ESA WorldCover [42] as the thematic product. We associate each tile in MTom with a Köppen-Geiger class and ESA WorldCover class. For computational reasons, we only use the center location of each tile for this. Based on the association, we use our entropy-maximizing stratified sampling for MTom $_{\mu}$  and a random selection for MTom $_r$ . Both sets contain  $N_s \approx 100$  samples per joined stratum. For the sampling of MTom $_{\mu}$ , the genetic algorithm is run for  $T = 2\,500$  iterations with population size  $N_p = 10$ , crossover rate  $r_c = 0.5$ , and mutation rate  $r_m = \frac{N_s}{n_g \cdot 25}$ , where  $n_g$  is the total number of samples in the joined stratum.

Table II shows that both datasets have similar class balance and distance between tiles, but MTom $_{\mu}$  contains a significantly greater spatial diversity within each class. Fig. 3 visualizes the distribution of samples with respect to their location. As one can see from the figure, large homogeneous regions (e.g., the Amazon forest, the Sahara desert, or Siberia) and small islands exhibit significantly improved spatial diversity when entropy-maximizing stratified sampling is applied to the dataset. Following [43], we split each tile of MTom $_{\mu}$  into patches of size  $120 \times 120$  pixels, discarding those that are too small (due to non-integer multiples of tile-patch size relations) or contain invalid pixels (see Fig. 4). This yields a total of 1 055 080 valid training patches, of which we randomly select 5% to track for validation loss during the training.

#### B. Description of Downstream Tasks

We evaluate our pretrained CSMoE model on four scene classification and two semantic segmentation tasks from the geobench benchmark collection [44] as well as on unimodal and cross-modal CBIR on the BigEarthNet-v2 benchmark dataset [43]. The datasets for scene classification are the following:

- **m-bigeearthnet** is a multi-label LULC classification dataset covering ten European countries. It contains 22 000 S2 images, each labeled with one or more of the 43 CLC2018 [45] classes.
- **m-brick-kiln** is a binary scene classification dataset for brick kiln detection in Bangladesh. It contains 17 061 S2 images that are labeled with brick-kiln or no-brick-kiln.
- **m-so2sat** is a multi-class climate zone classification dataset. It contains 21 964 S1 and S2 images from 42 globally distributed cities and is associated with one of 17 local climate zones.

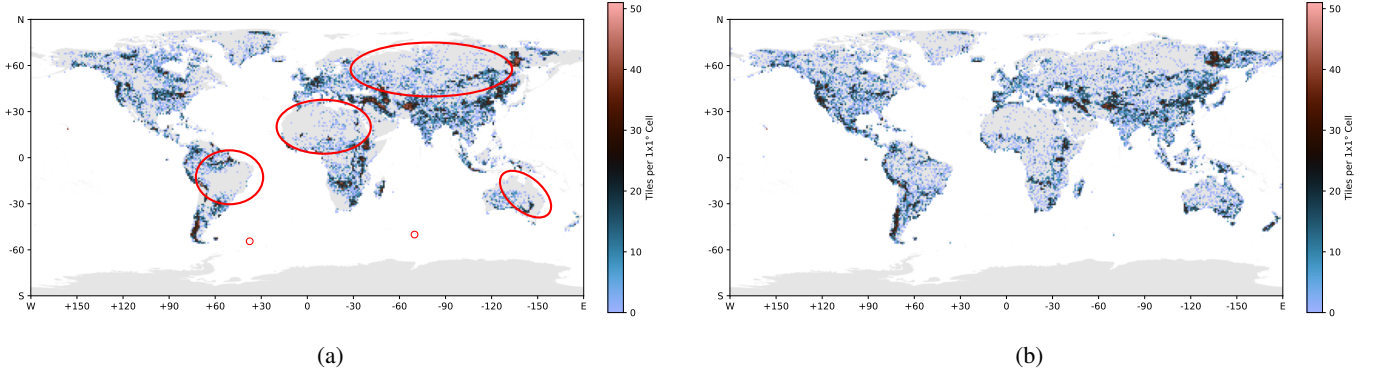


Fig. 3: Tile density per  $1^\circ \times 1^\circ$  square (a) in random sampling ( $\text{MTom}_r$ ); and (b) using our entropy-maximizing stratified sampling ( $\text{MTom}_\mu$ ). Red circles mark regions that have a significantly lower spatial diversity of samples in  $\text{MTom}_r$  compared to  $\text{MTom}_\mu$ , e.g., the Amazon forest, the Sahara desert, or Siberia.

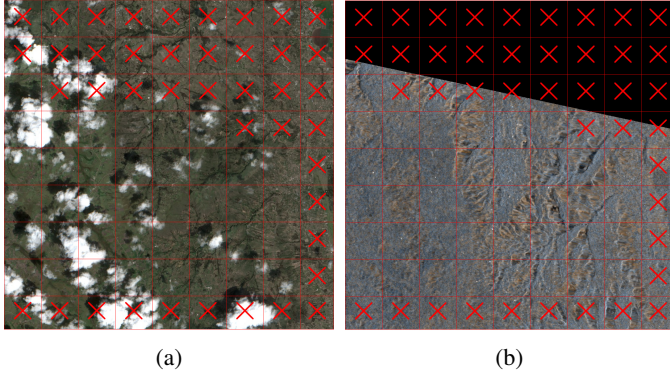


Fig. 4: Result of splitting and filtering a  $\text{MTom}_\mu$  tile into patches: (a) True color representation of the Sentinel-2 tile; (b) False colour composites of the Sentinel-1 tile in decibel-scale with red channel as VV, green channel as VV+VH, and blue channel as VH. Patches marked with a red  $\times$  are dropped due to invalid pixels or incompatible size.

- **m-eurosat** is a multi-class LULC classification dataset covering Europe. It contains 4000 S2 images, each labeled with one of 13 classes.

For semantic segmentation, we use the following datasets:

- **m-cashew-plant** is a cashew plantation segmentation dataset from 1800 S2 images obtained over Benin. Each pixel is labeled with one of seven classes.
- **m-SA-crop-type** is a crop-type segmentation dataset from S2 images. It contains 5000 images from Brandenburg, Germany, and Cape Town, South Africa. Each pixel is associated with one of ten crop type labels.

For CBIR, we conducted experiments on the BigEarthNet-v2 benchmark dataset [43]. We use the following sets of image pairs:

- **BENv2-14k** comprises of 13683 BigEarthNet-v2 image pairs acquired over Serbia during the summer months. Each S1-S2 pair is labeled with one or more classes from the 19-class nomenclature introduced in [46].
- **BENv2-243k** comprises of 243130 BigEarthNet-v2 image pairs acquired over the ten European countries during

the summer and autumn months. Each S1-S2 pair is labeled with one or more classes from the same 19-class nomenclature.

We separately stacked: i) the VV and VH bands of S1 images (if applicable); and ii) the S2 bands associated with 10m and 20m spatial resolution, while nearest-neighbor interpolation was applied to the 20m bands. For the datasets in the geobench benchmark collection, we follow the proposed train/validation/test-split from [44]. For the CBIR experiments, the validation split as proposed in [43] of the respective set of images was used to select query images, while images were retrieved from the test split. We evaluate two scenarios for CBIR: i) unimodal CBIR, where query images and retrieved images belong to the same image modality; and ii) cross-modal CBIR, where query images are selected from one image modality and retrieved from the other image modality.

### C. Experimental Setup

We trained our CSMoE model in four variants with different patch sizes  $\rho \in \{32, 28, 16, 14\}$ . If not noted differently, the CSMoE model was trained for 150 epochs on  $\text{MTom}_\mu$  with a mini-batch size of 256–512, depending on the memory requirements of the model, and an image size of  $224 \times 224$  pixels. All model variants use four modality-specific encoder layers ( $L_{MS}^E = 4$ ), two cross-sensor encoder layers ( $L_{CS}^E = 2$ ), four modality-specific decoder layers ( $L_{MS}^D = 4$ ), and eight experts per layer ( $S = 8$ ). The embedding dimension was set to 768 for the encoder and 256 for the decoder. During pretraining, we follow [9] and set both the masking ratio and the temperature of the  $\mathcal{L}_{MI}$  to 0.5. The AdamW optimizer with learning rate  $10^{-4}$  and a cosine annealing learning rate schedule with linear warm-up was utilized. All the experiments were conducted on NVIDIA 4×A100 or 4×H200 GPUs.

We carried out three different kinds of experiments to: i) perform a sensitivity analysis with respect to different variants of the CSMoE model; ii) compare the CSMoE model variants with other FMs in terms of their scene classification and semantic segmentation performance relative to their computational complexity; and iii) compare the CSMoE model variants with the baseline CSMAE [9] model in terms of

TABLE III: The hyperparameters selected for evaluation on the geobench benchmark collection for the comparison of different FMs.

Task	Dataset	CBIR	Probing	Metric
		$K$	LR	
Classification	m-bigearthnet	-	$1 \times 10^{-2}$	$\text{mAP}_\mu$
	m-brick-kiln	-	$3 \times 10^{-3}$	AA
	m-so2sat	-	$1 \times 10^{-2}$	AA
	m-eurosat	-	$5 \times 10^{-2}$	AA
Segmentation	m-cashew-plant	-	$3 \times 10^{-4}$	IoU
	m-SA-crop-type	-	$1 \times 10^{-2}$	IoU
CBIR	BENv2-14k	10	-	$F_1$
	BENv2-243k	10	-	$F_1$

their uni-modal and cross-modal CBIR performance. For the sensitivity analysis, we vary the: 1) patch size  $\rho$ ; 2) strategies of constructing the classification token; and 3) the number of pretraining epochs. For the comparison with other RS FMs, we compare the CSMoE variants with: Prithvi V2 [1] in the 300 million and 600 million parameter versions; Satlas [2] in the swin-base configuration; DOFA [7] in the base configuration; TerraMind [8] in the base configuration; and CSMAE [9] in the SECD configuration. For all models except for CSMAE, we use the implementations and checkpoints provided in TerraTorch [23]. For CSMAE, we use the implementation provided in [9] and train one model on BigEarthNet-V2 [43] and one on MTom $_\mu$ . We use the model trained on BigEarthNet-V2 for the scene classification and semantic segmentation following [9], and the model trained on MTom $_\mu$  for unimodal and cross-modal CBIR to avoid evaluating based on training data bias.

For the scene classification and semantic segmentation tasks, we consider probing as our downstream scenario, where we freeze the backbone of the FM and only train a linear layer and a UPerNet [47] decoder for scene classification and semantic segmentation, respectively, for 50 epochs. We performed a hyperparameter search and fixed the hyperparameters as shown in Table III. We report the results in terms of their micro-mean average precision ( $\text{mAP}_\mu$ ) for multi-label scene classification, average accuracy (AA) for multi-class scene classification, mean intersection-over-union (IoU) for semantic segmentation, and  $F_1$ -score for uni- and cross-modal retrieval. All scores are reported in %. For each CBIR task, the task is denoted as  $\langle Q \rangle \rightarrow \langle R \rangle$ , where  $\langle Q \rangle$  denotes the image modality of the query images and  $\langle R \rangle$  denotes the image modality of the retrieved images.

#### IV. EXPERIMENTAL RESULTS

##### A. Sensitivity Analysis

In this subsection, we investigate the impact of three key design factors of the proposed CSMoE model, which influence its computation cost and the effectiveness of using its capacity (which can be seen by measuring the performance on downstream tasks): 1) patch size  $\rho$ ; 2) classification token construction strategy; and 3) training duration. In addition, we evaluate these design choices to quantify their trade-offs in terms of computational efficiency.

1) *Patch Size*: To analyze the effect of patch size  $\rho$  on the performance and computational cost of the CSMoE model, we train four CSMoE variants with four different patch sizes  $\rho \in \{14, 16, 28, 32\}$ . Table IV shows the downstream performance on the scene classification and semantic segmentation datasets as well as the computational requirements in terms of number of parameters, FLOPs and C<sub>2</sub>C ratio for different patch sizes  $\rho$ .

From the table, one can see that CSMoE variants trained with smaller patch sizes ( $\rho \in \{14, 16\}$ ) outperform those using larger patch sizes ( $\rho \in \{28, 32\}$ ) across all classification and segmentation tasks, with particularly large gains on the segmentation datasets m-cashew-plant and m-SA-crop-type. For example, CSMoE with patch size  $\rho = 14$  achieves 59.4% IoU and 39.8% IoU, compared to 46.0% and 35.8% for CSMoE with patch size  $\rho = 32$ , on the m-cashew-plant and m-SA-crop-type datasets, respectively. A possible explanation is that smaller patches retain finer spatial structures in the embeddings, which may be particularly beneficial for tasks requiring precise delineation of heterogeneous land-cover types. However, this improvement comes with a substantial increase in FLOPs, from 2.92B to 13.40B, and a corresponding drop in the C<sub>2</sub>C ratio from 94.86 to 20.22. We observe that patch sizes of 14 or 16 offer the best trade-off between performance and compute, while larger patch sizes, though more efficient, lead to under-utilization of the model’s capacity.

2) *Classification Token Construction Strategy*: To assess how different strategies for constructing the classification token affect the downstream performance, we use the pretrained CSMoE model with patch size  $\rho = 16$  and train a linear classifier, where we vary the strategy for extracting the input token to the classifier. As shown in Table V, directly using the [CLS] token yields the best overall performance across all classification tasks, with 62.6%  $\text{mAP}_\mu$  on m-bigearthnet and 84.9% AA on m-eurosat, outperforming alternatives such as averaging all tokens (60.1%  $\text{mAP}_\mu$ , 82.3% AA) or averaging all tokens excluding the [CLS] token (60.2%  $\text{mAP}_\mu$ , 82.3% AA). It is worth noting that reusing the normalization applied during contrastive pretraining, with or without the projection head, leads to significantly degraded results. For example, using the normalized (denoted as “norm. [CLS]”) or the normalized and projected [CLS] token (denoted as “norm. & proj. [CLS]”) on m-so2sat yields only 21.9% AA, which is less than half of the un-normalized [CLS] token (denoted as “only [CLS]”), suggesting a misalignment between the pretraining and probing objectives. We would also like to note that the strategy for constructing the classification token does not affect the results on segmentation, as the classification token is not used as a feature for segmentation. We conclude that while simple averaging strategies offer reasonable performance, using the raw [CLS] token remains the most effective and robust choice for classification probing. However, it is critical for achieving good classification performance that the [CLS] token is used without normalization, although it was trained during pretraining with normalization.

3) *Training Length*: To assess the influence of training duration on downstream performance and convergence stability and identify an optimal trade-off between final loss and down-



TABLE IV: Comparison of different patch sizes  $\rho$  using linear probing. Performance is reported as  $\text{mAP}_\mu$  (%) for m-bigearthnet, AA (%) for m-brick-kiln, m-so2sat, and m-eurosat, and IoU (%) for m-cashew-plant and m-SA-crop-type.

Model	$\rho$	Classification				Segmentation		# Params $\uparrow$	FLOPs $\downarrow$	$\text{C}_2\text{C}$ Ratio $\uparrow$
		m-bigearthnet	m-brick-kiln	m-so2sat	m-eurosat	m-cashew-plant	m-SA-crop-type			
CSMoE	32	62.6	93.2	44.1	84.9	46.0	35.8	<b>277M</b>	<b>2.92B</b>	<b>94.86</b>
	28	65.1	93.8	46.3	84.9	48.3	36.7	275M	3.67B	74.93
	16	<b>66.5</b>	94.3	48.0	86.2	55.7	38.6	271M	10.11B	26.81
	14	66.0	<b>94.4</b>	<b>49.6</b>	<b>88.3</b>	<b>59.4</b>	<b>39.8</b>	271M	13.40B	20.22

TABLE V: Comparison of different strategies of constructing the classification token using linear probing on CSMoE ( $\rho = 32$ ). Performance is reported as  $\text{mAP}_\mu$  (%) for m-bigearthnet and AA (%) for m-brick-kiln, m-so2sat, and m-eurosat.

Classification Token Construction Strategy	m-bigearthnet	m-brick-kiln	m-so2sat	m-eurosat
avg. w/o. [CLS]	60.2	89.4	40.1	82.3
avg. all tokens	60.1	88.9	40.2	82.3
only [CLS]	<b>62.6</b>	<b>93.2</b>	<b>44.1</b>	<b>84.9</b>
norm. [CLS]	40.8	82.4	21.9	69.8
norm. & proj. [CLS]	40.8	82.4	21.9	69.8

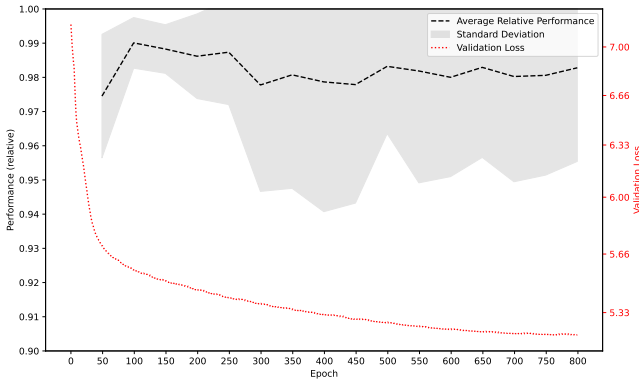


Fig. 5: Normalized average performance (%) and validation loss obtained by CSMoE with patch size  $\rho = 16$  when pretrained for different numbers of epochs on MTom $_\mu$  and evaluated via linear and segmentation probing.

stream performance consistency, we evaluate CSMoE with patch size  $\rho = 16$  trained on MTom $_\mu$  for up to 800 epochs, where we save one weight checkpoint every 50 epochs. For each checkpoint, we calculate the normalized performance per dataset as the ratio of the current score to the best score achieved across all epochs for that dataset. This normalization allows for averaging across heterogeneous task metrics (e.g., AA,  $\text{mAP}_\mu$ , and IoU). All checkpoints are evaluated under the same linear probing protocol on the geobench benchmark collection. Fig. 5 shows the normalized downstream performance as well as the validation loss over the number of epochs trained. As one can see from the figure, the resulting average normalized performance increases from 97.5% at epoch 50 to a peak of 99.0% at epoch 100, while the standard deviation across datasets simultaneously decreases, indicating improved stability. After epoch 150 ( $98.8\% \pm 0.71\%$ ), performance plateaus, while inter-dataset variance gradually increases, with the standard deviation exceeding 3% from epoch 300 onward.

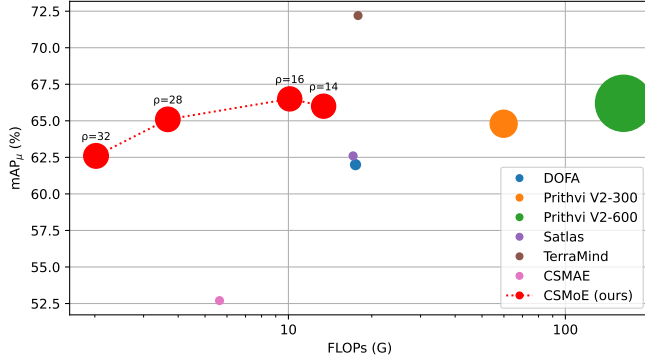
The validation loss, however, decreases for the full duration of the training, albeit at a diminishing rate (5.71 at epoch 50, 5.21 at epoch 799). We adopt 150 epochs as the default pretraining duration, as it yields near-optimal average performance with low inter-dataset variance, representing the best trade-off between training efficiency and generalization stability.

### B. Comparison with other Foundation Models

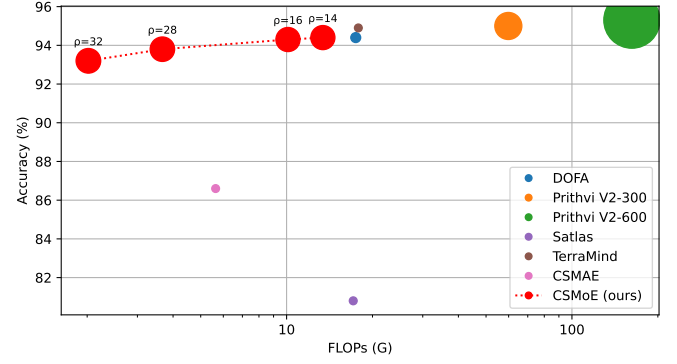
In this subsection, we analyze the effectiveness of the proposed CSMoE model for scene classification by comparing it with state-of-the-art FMs across four scene classification datasets in the geobench benchmark collection. Fig. 6 shows the corresponding scene classification results in terms of  $\text{mAP}_\mu$  for m-bigearthnet and AA for the remaining datasets with respect to the required number of FLOPs. One can see that the CSMoE model variants consistently achieve high classification performance across all datasets while requiring significantly fewer FLOPs compared to most existing approaches. As an example, on the m-bigearthnet dataset, all CSMoE model variants reach an  $\text{mAP}_\mu$  scores comparable to Prithvi V2-600 while requiring an order of magnitude fewer FLOPs (see Fig. 6a). On m-eurosat (Fig. 6c) and m-brick-kiln (Fig. 6b), the CSMoE model variants yield an AA close to or even surpassing FMs such as the CSMAE and Satlas models, while maintaining a much lower computational budget. In particular, on the m-brick-kiln, all of our model variants achieve an AA comparable to much larger models such as Prithvi V2-300 and TerraMind, despite operating with considerably fewer FLOPs. On the more challenging m-so2sat dataset (Fig. 6d), even the most lightweight CSMoE model variant that uses a patch size of  $\rho = 32$  outperforms the baseline CSMAE model, whereas the CSMoE model variants with smaller patch sizes narrow the gap to the best-performing FMs while maintaining high computational efficiency. We would like to note that the FLOPs reported in the figures are shown in logarithmic scale to better visualize the trade-off between computational complexity and performance across models. As a result, differences in FLOPs may appear visually less pronounced. In general, our model variants with a smaller patch size ( $\rho \in \{16, 14\}$ ) tend to achieve better performance across datasets, except on the m-bigearthnet dataset, where the variant using a patch size of  $\rho = 16$  yields a slightly higher  $\text{mAP}_\mu$  than the variant using a patch size of  $\rho = 14$ .

We observe similar trends in semantic segmentation, as shown in Fig. 7, which presents results on m-cashew-plant (Fig. 7a) and m-SA-crop-type (Fig. 7b). All of our CSMoE model variants demonstrate strong performance while operating at significantly reduced computational cost. For instance,

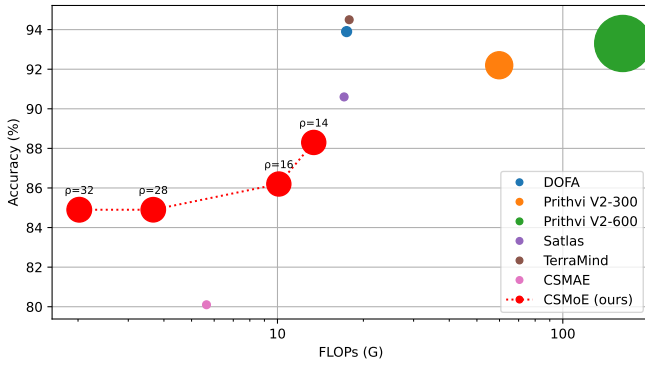




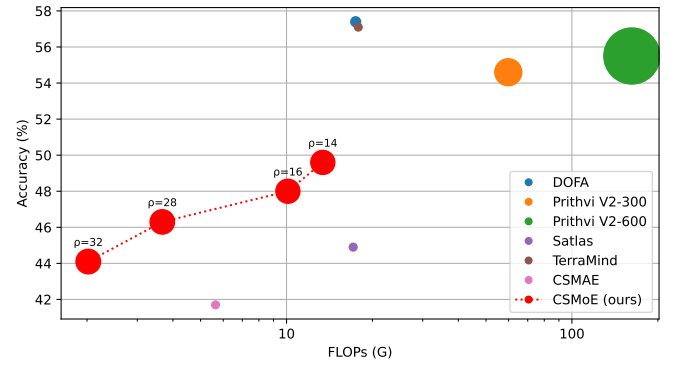
(a)



(b)

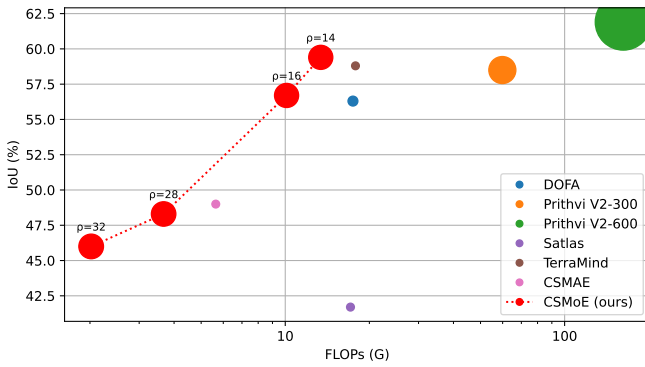


(c)

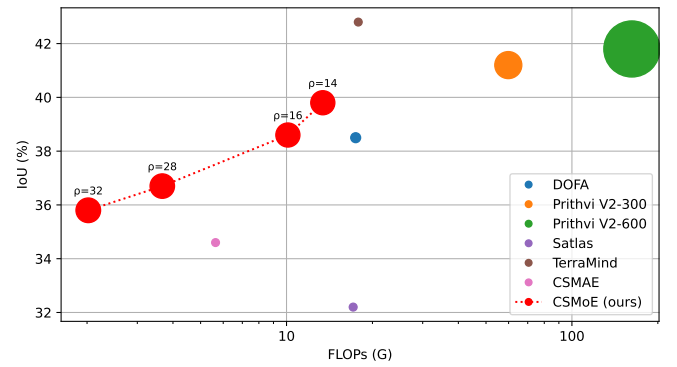


(d)

Fig. 6: Scene classification results versus the number of FLOPs on: (a) m-bigearthnet; (b) m-brick-kiln; (c) m-eurosat; and (d) m-so2sat.



(a)



(b)

Fig. 7: Semantic segmentation results versus the number of FLOPs on: (a) m-cashew-plant; and (b) m-SA-crop-type.

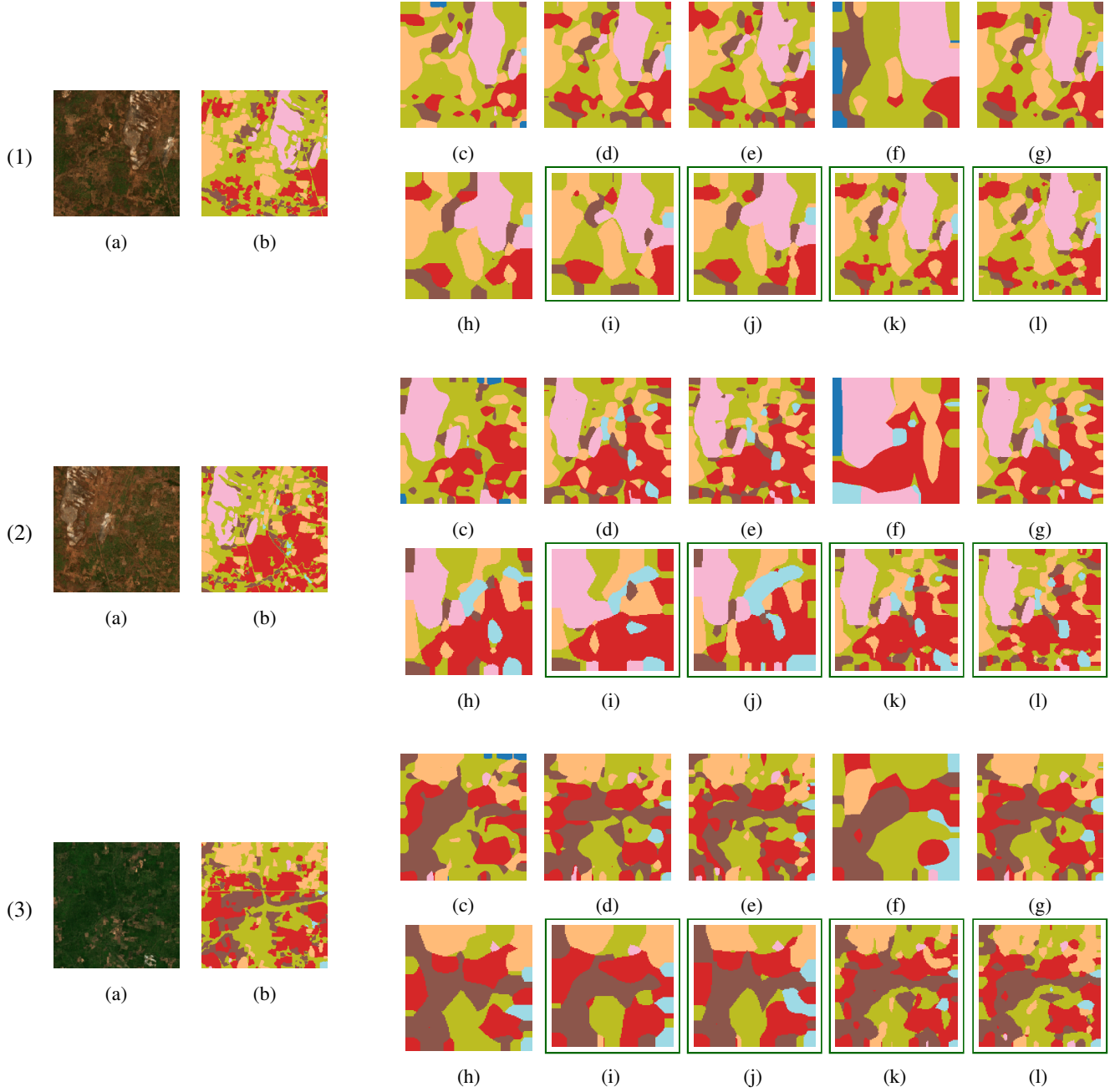


Fig. 8: Qualitative comparison on three samples from segmentation probing on the m-cashew-plant dataset: (a) input image (RGB); (b) reference maps; and (c) – (l) obtained segmentation maps from: (c) DOFA; (d) Prithvi V2-300; (e) Prithvi V2-600; (f) Satlas; (g) TerraMind; (h) CSMAE; and (i-l) CSMoE (ours) with patch sizes  $\rho = 32, 28, 16$  and  $14$ , respectively. Our results are highlighted with a green border.

on m-cashew-plant, the CSMoE model variant with patch size  $\rho = 14$  achieves an IoU of 59.4%, outperforming TerraMind (58.8%) and Prithvi V2-300 (58.5%) and closely approaching the performance of Prithvi V2-600 (61.9%), while requiring substantially fewer FLOPs (13.4G vs. 17.8G, 59.9G, and 162.2G, respectively). On the m-SA-crop-type dataset, the CSMoE model variant with patch size  $\rho = 14$  similarly outperforms or closely matches the performance of mid-sized FMs like CSMAE, Satlas, and DOFA. In line with our

results on scene classification, the CSMoE model variants using smaller patch sizes generally yield stronger results, with the variant that uses a patch size of  $\rho \in \{14, 16\}$  consistently outperforming their counterparts with larger patch sizes. These trends are further illustrated in Fig. 8, which compares model outputs across three scenes from the m-cashew-plant dataset. From top to bottom, each example shows the true color representation of the input image (a), the ground truth mask (b), predictions from other FMs (c–h), followed by

TABLE VI:  $F_1$ -scores (%) obtained by CSMAE and of CSMoE with different patch sizes  $\rho$  on uni-modal and cross-modal CBIR when the image sets of BENv2-14k and BENv2-243k are considered.

Model	patch size $\rho$	Training Set								# Params $\uparrow$	FLOPs $\downarrow$	$C_2C$ Ratio $\uparrow$
		BENv2-14k				BENv2-243k						
		Uni-Modal CBIR		Cross-Modal CBIR		Uni-Modal CBIR		Cross-Modal CBIR				
		$S_I \rightarrow S_I$	$S_2 \rightarrow S_2$	$S_I \rightarrow S_2$	$S_2 \rightarrow S_I$	$S_I \rightarrow S_I$	$S_2 \rightarrow S_2$	$S_I \rightarrow S_2$	$S_2 \rightarrow S_I$			
CSMAE [9]	16	66.61	72.29	30.86	42.63	63.59	70.78	<b>33.04</b>	<b>37.83</b>	87M	5.64B	15.43
CSMoE (ours)	32	62.84	69.13	37.48	45.86	60.06	67.36	32.26	32.63	<b>277M</b>	<b>2.92B</b>	<b>94.86</b>
	28	63.82	71.41	40.57	38.65	61.55	69.48	32.93	27.02	275M	3.67B	74.93
	16	65.95	71.73	36.14	<b>49.49</b>	63.16	70.25	24.59	32.16	271M	10.11B	26.81
	14	<b>66.71</b>	<b>72.37</b>	<b>43.01</b>	45.43	<b>64.14</b>	<b>70.89</b>	32.30	31.04	271M	13.40B	20.22

our CSMoE model variants with patch sizes  $\rho = 32$ ,  $\rho = 28$ ,  $\rho = 16$  and  $\rho = 14$  (i-l). The results reveal that especially our model variants with smaller patch sizes (k, l) produce cleaner and more coherent segmentation maps, often recovering fine structures and class boundaries more accurately than the less efficient baseline FMs.

For CBIR, we evaluate the CSMAE and the proposed CSMoE model on the BENv2-14k and BENv2-243k sets of images from BigEarthNet-v2. Table VI shows the corresponding  $F_1$ -scores, the required number of model parameters and the FLOPs when the two training sets are considered for unimodal and cross-modal CBIR. By assessing the table, one can see that in the unimodal CBIR scenario all of our CSMoE model variants achieve CBIR performance comparable to or surpassing that of CSMAE in both sets of images. For instance, on BENv2-14k, the CSMoE model variant using a patch size of  $\rho = 14$  yields an  $F_1$ -score of 66.71% on the  $S1 \rightarrow S1$  task and 72.37% on the  $S2 \rightarrow S2$  task, slightly outperforming CSMAE (66.61% and 72.29%, respectively). On the BENv2-243k set of images, the same CSMoE model variant again yields the highest  $F_1$ -score of the evaluated model in the  $S1 \rightarrow S1$  and  $S2 \rightarrow S2$  tasks with 64.14% and 70.89% compared to 63.59% and 70.89% for the CSMAE model. In the more challenging cross-modal CBIR scenario, we observe a general drop in performance across all models. Here, on BENv2-243k, the CSMAE model achieves the best results with 33.04%  $F_1$  on the  $S1 \rightarrow S2$  task and 37.83% on  $S2 \rightarrow S1$ , followed closely by our CSMoE model with 32.93% when the variant with patch size  $\rho = 28$  on the  $S1 \rightarrow S2$  task is considered, and 32.63% for the variant that is using a patch size of  $\rho = 32$  on the  $S2 \rightarrow S1$  task. It is worth noting that these two CSMoE model variants operate with 35% and 48% fewer FLOPs compared to CSMAE while achieving only slightly lower  $F_1$ -scores. Additionally, on BENv2-14k, the CSMoE model variants with smaller patch sizes ( $\rho = 14$  and  $\rho = 16$ ) again achieve the highest  $F_1$  scores with 43.01 % and 49.49% on  $S1 \rightarrow S2$  and  $S2 \rightarrow S1$ , respectively, significantly outperforming CSMAE (30.86% on  $S1 \rightarrow S2$  and 42.63% on  $S2 \rightarrow S1$ ). Furthermore, we would like to note that CSMAE was explicitly optimized for CBIR, whereas CSMoE was trained more generally. Overall, CSMoE achieves comparable or superior performance in most settings, demonstrating strong adaptability across unimodal and cross-modal retrieval tasks.

Overall, for both sets of images, we find that smaller

patch sizes lead to stronger CBIR performance, in agreement with observations from the scene classification and semantic segmentation tasks. However, this trend is less evident than for scene classification and semantic segmentation, and smaller patch sizes are generally more beneficial in the unimodal case, while differences diminish in cross-modal settings. We theorize that this is likely due to the increased complexity and domain gap, which benefits more from overall strong encoders than from specific patch sizes. In particular, the performance of our most efficient model variant in terms of FLOPs, CSMoE with patch size  $\rho = 32$ , remains competitive despite its reduced computational budget, underscoring the efficiency of the proposed model. In summary, these results confirm the success of the proposed CSMoE model in CBIR tasks under restricted computational constraints.

## V. CONCLUSION

In this paper, for the first time in RS, we have investigated the effectiveness of injecting Soft MoEs into FMs to decrease their computational requirements while retaining their representational capacity. To this end, we introduce a general adaptation for efficient multi-modal processing, which injects a Soft MoE into CSMAE to address computational complexity limitations while preserving cross-modal representation learning capabilities to create our CSMoE model. Based on our adaptation that includes integration of expert routing, encoder architecture modifications, and regularization of training objectives, we have created the first compute-efficient multi-modal FM in RS, which we call CSMoE. Additionally, we introduced a novel thematic-climatic descriptor-driven sampling strategy that leverages climate zones and thematic products to ensure thematic-climatic diversity while promoting spatial diversity through genetic optimization. We trained our FM using a training set that we created using our thematic-climatic descriptor-driven sampling strategy. To evaluate our CSMoE model, we carried out extensive experiments on scene classification, semantic segmentation, and image retrieval tasks, while comparing the CSMoE model with other state-of-the-art FMs. Experimental results demonstrate the effectiveness of CSMoE, achieving performance comparable or superior to existing FMs while requiring significantly fewer FLOPs. The success of CSMoE relies on our two contributions: i) an effective integration of Soft MoE mechanisms into both modality-specific and cross-sensor encoders; and ii) an efficient data

sampling strategy that preserves thematic-climatic diversity while reducing computational overhead.

In our experiments, we also investigated the effects of patch size, strategies for constructing the classification token, and the number of pretraining epochs on the required computational resources and the resulting downstream performance of our CSMoE model. From this sensitivity analysis, we have derived guidelines to select the appropriate CSMoE model variant under different computational requirements in RS as follows:

- If computational resources are severely constraint, the CSMoE model variant with patch size  $\rho = 32$  can be selected due to its exceptional C<sub>2</sub>C ratio of 94.86 while maintaining competitive performance.
- For the best trade-off between efficiency and performance, the CSMoE model variant with patch size  $\rho = 16$  can be selected for its high performance, especially in segmentation tasks.
- For the highest downstream performance, the CSMoE model variant that is using a patch size of  $\rho = 14$  can be selected, achieving the highest performance of the CSMoE model variants while maintaining a lower computational complexity than the existing FMs.

We would like to note that, in this paper, we realized our Soft MoE adaptation in the context of one specific RS FM, namely CSMAE. However, the proposed adaptation can be applied to most existing FMs in RS to improve both their representational capacity and computational complexity. Additionally, they are not limited to models using masked image modeling as the learning objective, but can also be used for models using contrastive learning or any other self-supervised learning objective. As future works, we plan to: i) extend our model to be able to process other modalities such as video and text; and ii) increase the efficiency of the proposed model by including task-specific expert pruning or expert creation at runtime.

#### REFERENCES

- [1] D. Szwarcman, S. Roy, P. Fraccaro, P. E. Gíslason, B. Blumenstiel, R. Ghosal, P. H. de Oliveira, J. L. d. S. Almeida, R. Sedona, Y. Kang, et al., “Prithvi-2.0: A versatile multi-temporal foundation model for earth observation applications,” *arXiv preprint arXiv:2412.02732*, 2024.
- [2] F. Bastani, P. Wolters, R. Gupta, J. Ferdinando, and A. Kembhavi, “Satlaspretrain: A large-scale dataset for remote sensing image understanding,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 16 772–16 782.
- [3] Y. Wang, N. A. A. Braham, Z. Xiong, C. Liu, C. M. Albrecht, and X. X. Zhu, “Ssl4eo-s12: A large-scale multimodal, multitemporal dataset for self-supervised learning in earth observation [software and data sets],” *IEEE Geoscience and Remote Sensing Magazine*, vol. 11, no. 3, pp. 98–106, 2023.
- [4] 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*, vol. 61, pp. 1–22, 2022.
- [5] 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.
- [6] C. J. Reed, R. Gupta, S. Li, S. Brockman, C. Funk, B. Clipp, K. Keutzer, S. Candido, M. Uyttendaele, and T. Darrell, “Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4088–4099.
- [7] Z. Xiong, Y. Wang, F. Zhang, A. J. Stewart, J. Hanna, D. Borth, I. Papoutsis, B. L. Saux, G. Camps-Valls, and X. X. Zhu, “Neural plasticity-inspired multimodal foundation model for earth observation,” *arXiv preprint arXiv:2403.15356*, 2024.
- [8] J. Jakubik, F. Yang, B. Blumenstiel, E. Scheurer, R. Sedona, S. Maurogiovanni, J. Bosmans, N. Dionelis, V. Marsocci, N. Kopp, et al., “Terramind: Large-scale generative multimodality for earth observation,” *arXiv preprint arXiv:2504.11171*, 2025.
- [9] J. Hackstein, G. Sumbul, K. Norman Clasen, and B. Demir, “Exploring masked autoencoders for sensor-agnostic image retrieval in remote sensing,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–14, 2025.
- [10] 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*, 2022, pp. 15 979–15 988.
- [11] 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*, 2020, pp. 9729–9738.
- [12] M. Oquab et al., “DINOv2: Learning robust visual features without supervision,” *Transactions on Machine Learning Research*, 2024.
- [13] A. Francis and M. Czerkawski, “Major tom: Expandable datasets for earth observation,” in *IEEE International Geoscience and Remote Sensing Symposium*, 2024, pp. 2935–2940.
- [14] L. Jiao, Z. Huang, X. Lu, X. Liu, Y. Yang, J. Zhao, J. Zhang, B. Hou, S. Yang, F. Liu, et al., “Brain-inspired remote sensing foundation models and open problems: A comprehensive survey,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 10 084–10 120, 2023.
- [15] A. Xiao, W. Xuan, J. Wang, J. Huang, D. Tao, S. Lu, and N. Yokoya, “Foundation models for remote sensing and earth observation: A survey,” *IEEE Geoscience and Remote Sensing Magazine*, pp. 2–29, 2025.

- [16] S. Lu, J. Guo, J. R. Zimmer-Dauphinee, J. M. Nieuwsma, X. Wang, P. vanValkenburgh, S. A. Wernke, and Y. Huo, "Vision foundation models in remote sensing: A survey," *IEEE Geoscience and Remote Sensing Magazine*, pp. 2–27, 2025.
- [17] Z. Chen, J. Yang, Z. Feng, and L. Chen, "Rscnet: An efficient remote sensing scene classification model based on lightweight convolution neural networks," *Electronics*, vol. 11, no. 22, 2022.
- [18] G. Cheng, X. Xie, J. Han, L. Guo, and G.-S. Xia, "Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 3735–3756, 2020.
- [19] Z. Xu, W. Zhang, T. Zhang, Z. Yang, and J. Li, "Efficient transformer for remote sensing image segmentation," *Remote Sensing*, vol. 13, no. 18, 2021.
- [20] L. Wang, R. Li, C. Zhang, S. Fang, C. Duan, X. Meng, and P. M. Atkinson, "Unetformer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 190, pp. 196–214, 2022.
- [21] L. Hackel, K. N. Clasen, M. Ravanbakhsh, and B. Demir, "Lit-4-rsvqa: Lightweight transformer-based visual question answering in remote sensing," in *IEEE International Geoscience and Remote Sensing Symposium*, 2023, pp. 2231–2234.
- [22] L. Hackel, K. N. Clasen, and B. Demir, "Configilm: A general purpose configurable library for combining image and language models for visual question answering," *SoftwareX*, vol. 26, p. 101731, 2024.
- [23] C. Gomes, B. Blumenstiel, J. L. d. S. Almeida, P. H. de Oliveira, P. Fraccaro, F. M. Escofet, D. Szwarcman, N. Simumba, R. Kienzler, and B. Zadrozny, "Terra-torch: The geospatial foundation models toolkit," *arXiv preprint arXiv:2503.20563*, 2025.
- [24] H. V. Vo et al., "Automatic data curation for self-supervised learning: A clustering-based approach," *Transactions on Machine Learning Research*, 2024.
- [25] P. Dias, A. Tsaris, J. Bowman, A. Potnis, J. Arndt, H. L. Yang, and D. Lunga, "Oreole-fm: Successes and challenges toward billion-parameter foundation models for high-resolution satellite imagery," in *Proceedings of the 32nd ACM International Conference on Advances in Geographic Information Systems*, 2024, pp. 597–600.
- [26] R. Roscher, M. Russwurm, C. Gevaert, M. Kampffmeyer, J. A. Dos Santos, M. Vakalopoulou, R. Hänsch, S. Hansen, K. Nogueira, J. Prexl, et al., "Better, not just more: Data-centric machine learning for earth observation," *IEEE Geoscience and Remote Sensing Magazine*, vol. 12, no. 4, pp. 335–355, 2024.
- [27] R. Stanimirova, K. Tarrio, K. Turlej, K. McAvoy, S. Stonebrook, K.-T. Hu, P. Arévalo, E. L. Bullock, Y. Zhang, C. E. Woodcock, et al., "A global land cover training dataset from 1984 to 2020," *Scientific Data*, vol. 10, no. 1, p. 879, 2023.
- [28] T. Kerdreux, A. Tuel, Q. Febvre, A. Mouche, and B. Chapron, "Efficient self-supervised learning for earth observation via dynamic dataset curation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2025, pp. 3017–3027.
- [29] H. Van Assel and R. Balestrierio, "A graph matching approach to balanced data sub-sampling for self-supervised learning," in *Advances in Neural Information Processing Systems Workshop: Self-Supervised Learning-Theory and Practice*, 2024.
- [30] H. Lin, D. Hong, S. Ge, C. Luo, K. Jiang, H. Jin, and C. Wen, "Rs-moe: A vision-language model with mixture of experts for remote sensing image captioning and visual question answering," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–18, 2025.
- [31] H. Bi, Y. Feng, B. Tong, M. Wang, H. Yu, Y. Mao, H. Chang, W. Diao, P. Wang, Y. Yu, et al., "Ring-moe: Mixture-of-modality-experts multi-modal foundation models for universal remote sensing image interpretation," *arXiv preprint arXiv:2504.03166*, 2025.
- [32] X. Liu and Z. Lian, "Rsuniqlm: A unified vision language model for remote sensing via granularity-oriented mixture of experts," *arXiv preprint arXiv:2412.05679*, 2024.
- [33] X. Chen, S. Yan, J. Zhu, C. Chen, Y. Liu, and M. Zhang, "Generalizable multispectral land cover classification via frequency-aware mixture of low-rank token experts," *arXiv preprint arXiv:2505.14088*, 2025.
- [34] J. Puigcerver, C. R. Ruiz, B. Mustafa, and N. Houlsby, "From sparse to soft mixtures of experts," in *International Conference on Learning Representations*, 2024.
- [35] 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.
- [36] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.
- [37] J. H. Holland, "Genetic algorithms," *Scientific american*, vol. 267, no. 1, pp. 66–73, 1992.
- [38] G. Sumbul, M. Müller, and B. Demir, "A novel self-supervised cross-modal image retrieval method in remote sensing," in *IEEE International Conference on Image Processing*, 2022, pp. 2426–2430.
- [39] P. Bachman, H. R. Devon, and W. Buchwalter, "Learning representations by maximizing mutual information across views," *Advances in Neural Information Processing Systems*, vol. 32, pp. 15535–15545, 2019.
- [40] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," in *International Conference on Learning Representations*, 2017.
- [41] W. P. Köppen, R. Geiger, et al., "Köppen-geiger, klima erde= climate of the earth," *Klett-Perthes, Gotha*, 1961.

- [42] D. Zanaga, R. Van De Kerchove, D. Daems, W. De Keersmaecker, C. Brockmann, G. Kirches, J. Wevers, O. Cartus, M. Santoro, S. Fritz, et al., “Esa worldcover 10 m 2021 v200,” *Zenodo*, 2022.
- [43] K. N. Clasen, L. Hackel, T. Burgert, G. Sumbul, B. Demir, and V. Markl, “reBEN: Refined bigearthnet dataset for remote sensing image analysis,” in *IEEE International Geoscience and Remote Sensing Symposium*, 2025.
- [44] A. Lacoste, N. Lehmann, P. Rodriguez, E. Sherwin, H. Kerner, B. Lütjens, J. Irvin, D. Dao, H. Alemohammad, A. Drouin, et al., “Geo-bench: Toward foundation models for earth monitoring,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 51 080–51 093, 2023.
- [45] J. Feranec, T. Soukup, G. Hazeu, and G. Jaffrain, *European Landscape Dynamics: CORINE Land Cover Data*, 1st. CRC Press, Inc., 2016.
- [46] G. Sumbul, A. De Wall, T. Kreuziger, F. Marcelino, H. Costa, P. Benevides, M. Caetano, B. Demir, and V. Markl, “Bigearthnet-mm: A large-scale, multimodal, multilabel benchmark archive for remote sensing image classification and retrieval [software and data sets],” *IEEE Geoscience and Remote Sensing Magazine*, vol. 9, no. 3, pp. 174–180, 2021.
- [47] 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*, 2018, pp. 418–434.