# MAPEX: Modality-Aware Pruning of Experts for Remote Sensing Foundation Models

Joëlle Hanna*[1], *Student Member, IEEE,* Linus Scheibenreif*[2,1], *Student Member, IEEE,* and Damian Borth[1], *Member, IEEE*

***Abstract*—Remote sensing data is commonly used in a wide range of tasks, such as natural disaster monitoring and land-use studies. For each task, scientists carefully choose appropriate modalities or leverage data from purpose-built instruments. Recent work on remote sensing foundation models pre-trains computer vision models on large amounts of remote sensing data to learn general-purpose representations. However, this progress comes at the cost of very large models, which are particularly challenging to deploy on edge devices due to their high inference costs. Moreover, downstream applications often rely on only a subset of modalities and operate under strict resource constraints, creating a mismatch between the large, multi-purpose models produced by pre-training and the lightweight, task-specific models needed in practice. We address this mismatch with MAPEX, a remote sensing foundation model based on mixture-of-modality experts. MAPEX is pre-trained on multi-modal remote sensing data using a novel modality-conditioned token routing mechanism that naturally encourages the emergence of modality-specialized experts. To apply the model on a specific task, we propose a modality-aware pruning technique that retains only the experts relevant to the task's modalities, resulting in lightweight, task-specific models that can be directly extracted from the pre-trained foundation model, at no additional cost. Our approach yields efficient modality-specific models while simplifying fine-tuning and deployment for the modalities of interest. We experimentally validate MAPEX on diverse remote sensing datasets and show strong performance compared to fully supervised training and state-of-the-art remote sensing foundation models. Code will be available at github.com/HSG-AIML/MAPEX.**

## I. Introduction

COMPUTER vision methods are widely applied in the remote sensing domain for tasks such as classification, segmentation, or detection in imagery obtained from satellites or aerial platforms [1]. Increasingly, the visual foundation models tailored to characteristics of remote sensing imagery, such as top-down viewpoint, varying resolution, or multi-spectral data, are adopted for these tasks [2], [3].

Advanced foundation models pre-trained on unlabeled data with self-supervised objectives present a great opportunity in the remote sensing domain, where labeled datasets are small and expensive, but unlabeled data is widely available in huge quantities [4]. This is particularly important for tasks that rely on scarce annotations, where labels are costly to obtain and only weak supervision is often available [5].

However, existing foundation models for remote sensing still face two fundamental mismatches with real-world applications. The first is a **modality mismatch**: most remote sensing foundation models are pre-trained on individual modalities such as optical RGB [6] or multi-spectral [7], [8]. This becomes challenging when the downstream task requires a different modality, for example applying a model trained on multispectral data to radar-based applications [8], [6]. In contrast, multi-modal foundation models attempt to overcome this by representing all modalities jointly [9], [10], [11], but this creates a **size mismatch**: such models are typically very large, which is advantageous during pre-training but makes them costly to store, fine-tune, and deploy on edge devices, in particular when only a subset of modalities is needed downstream [12], [11], [13].

Models that exclusively work with multiple modalities can be restrictive in remote sensing, where instruments and satellite missions are often designed to solve specific tasks, such that multi-modal data might not be necessary downstream. For instance, synthetic aperture radar (SAR) data is commonly used to map flooding because it is sensitive to water cover and able to penetrate clouds, which are common around flood events [14]. Pairing SAR with optical data that cannot penetrate cloud cover for the sake of maintaining the multi-modal pre-training setup for downstream fine-tuning might not be useful in such cases. Similar arguments could be made for a series of other remote sensing tasks in areas ranging

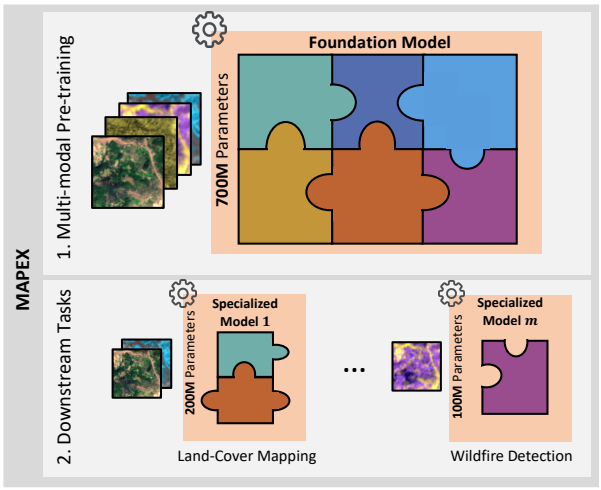


Fig. 1. Overview of MAPEX. A multi-modal foundation model (top) is pretrained using a Mixture-of-Experts architecture with modality-aware routing. During fine-tuning (bottom), irrelevant experts are pruned, yielding smaller, specialized models tailored to specific downstream tasks (e.g., land-cover mapping, wildfire detection, etc.), making them lighter and easier to fine-tune.

* Both authors contributed equally to this work

The authors are with the (1) AIML Lab at the School of Computer Science, University of St.Gallen, Switzerland (email: firstname.lastname@unisg.ch) and (2) ETH Zürich, Switzerland (email: lscheibenrei@ethz.ch)

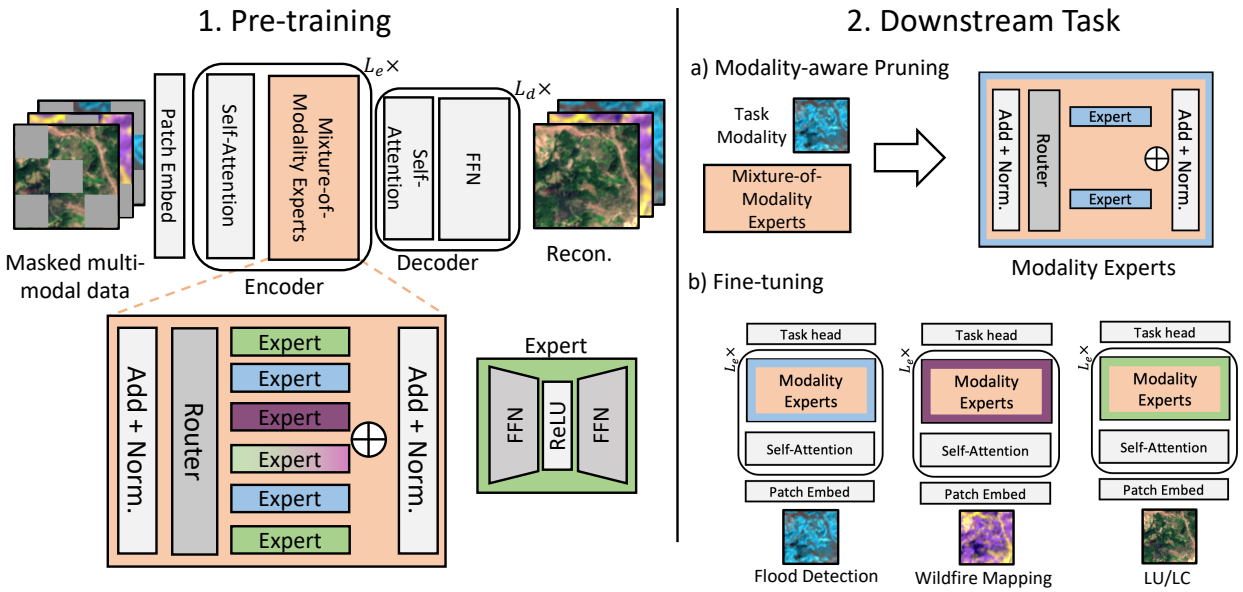


Fig. 2. Overview of the pre-training and downstream application process. **1.** The model is trained with all modalities on a masked reconstruction task. Modalities are processed by specialized experts in the mixture-of-modality experts module. **2.a)** Modality experts are pruned for a modality of interest. **2.b)** The resulting modality experts model is fine-tuned on a downstream task of the corresponding modality.

from wildfire detection [15] to environmental monitoring or agricultural applications [16].

While multi-modal pre-training gives computer vision models a more complete picture of the world and allows them to leverage valuable synergies between instruments [17], there remains a clear gap between the large foundation models produced by pre-training and the lightweight, task-specific models needed in practice. We **address** these mismatches by introducing a novel modality-aware pre-training method and successive expert pruning step (see Figure 2). First, we propose a novel pre-training method for multi-modal remote sensing data based on a mixture-of-experts architecture with modality-conditioned token routing. Unlike standard MoE approaches that route tokens individually based on feature similarity, MAPEX performs modality-level routing, where all tokens from the same modality share the same expert selection determined by learned modality embeddings. This method facilitates the use of multi-modal pre-training data while enforcing modality-specific structure in large parts of the pre-trained model. In the second step, we leverage this modality-specific structure with **M**odality-**A**ware **P**runing of **E**xperts (**MAPEX**) for the modality of the downstream task (see Figure 1). This yields a significantly smaller, more focused model that is specialized for one (or multiple) modalities of interest and can be easily fine-tuned on that modality. By doing so, MAPEX can adapt to any set of available modalities during inference (solving the modality mismatch) while producing much smaller, task-specific models compared to the original foundation network (solving the size mismatch). This approach allows us to combine the best of both worlds: leverage large-scale multi-modal data for self-supervised pre-training of large foundation models, but also provide small, specialized models that are easy to use and pre-trained for a wide variety of modalities.

In summary, our **contributions** are as follows:

- A new multi-modal pre-training approach for a remote sensing mixture-of-experts model, enabling a single large foundation model to learn modality-specialized experts.
- A novel modality-conditioned expert routing method to create modality specific experts within the multi-modal model.
- A modality-aware pruning technique that transforms this large foundation model into multiple lightweight, modality-specific models tailored to different downstream tasks.

## II. Related Work

*a) Foundation Models:* Foundation models are based on the idea of training big models on large, unlabeled data with self-supervised learning techniques before adapting them to specific downstream tasks [18]. This approach has produced strong, general models for natural language [19], computer vision [20], and more specialized variants for domains such as remote sensing [21]. Increasingly, research in foundation models has focused on multi-modal data, particularly the combination of language and visual data [22]. Multi-modal foundation models benefit from training on datasets of multiple paired modalities, producing models that perform well on individual modalities and enabling novel applications with multi-modal data (see [23] for a review). Large scale pre-training improves performance on downstream applications, but also requires big models for optimally compute-efficient training [24]. This increases latency as well as deployment and model serving costs, limiting foundation model use in time or cost sensitive settings. A growing body of literature addresses the foundation model efficiency issues with techniques like pruning [25], distillation [26] or parameter-efficient fine-tuning methods [27].

*b) Mixture-of-Experts:* A complementary area of research focuses on conditional computation as a way of increasing model size efficiently. Mixture-of-experts (MoE) models are commonly used in large foundation models [28], [29], [30]. These models consist of multiple *expert* modules and a *router* that assigns tokens to each expert during training and inference. In standard MoE models, the router $r$ produces logits from tokens $t$ as $r(t) = W_r \cdot t$, which are turned into routing probabilities with a softmax to determine the routing of $t$. In transformer models, MoEs replace feed-forward network layers (FFN) with multiple experts, each implemented as a FFN itself. This approach has been used to train performant large-scale language [31] and vision models [29] and has recently gained traction in remote sensing [32], [33], [34]. Sparse MoE models conditionally route tokens to a subset of expert modules, making inference more efficient [31], while large-scale distributed training benefits from conditional computation by facilitating parallelism across GPUs [28].

In the multi-modal context, MoEs have been applied to language-image models [35], incorporating modality-specific experts with token routing based on each token's modality. Furthermore, DAMEX [36] introduced a dataset-aware MoE model for object detection where experts specialize in specific datasets rather than modalities. In DAMEX, MoE layers process data from different object detection datasets with experts specifically dedicated to each dataset. In contrast to our work, the model is trained in supervised fashion for object detection with mostly optical RGB data and without explicit pruning step for inference or fine-tuning on individual datasets.

*c) Remote Sensing Foundation Models:* Remote sensing enables Earth observation through active and passive sensors mounted on satellites or aerial platforms, capturing vast amounts of heterogeneous data [37], [38]. Unlike traditional RGB imagery, satellite data varies significantly across sensor types, making model transferability challenging. To address this, foundation models for remote sensing leverage large-scale self-supervised pretraining on unlabeled data. Most foundation models target specific modalities, such as optical RGB [8], [6], multispectral [21] images which sample the electromagnetic spectrum with a few broad bands, hyperspectral [39] data, which provides almost continuous coverage of the electromagnetic spectrum with many narrow bands, or SAR [40] images that are acquired through active sensing with radar waves. More recent works, such as MMEarth [41], DOFA [9], RingMo-SAM [11], SkySense [13], and SSL-based multi-modal models [42], explore large-scale multi-modal pretraining by integrating diverse sensor data to improve generalization. However, these models typically remain monolithic, processing all available modalities. In contrast, MAPEX introduces a selective Mixture-of-Experts approach that activates only the relevant experts per input, enabling the extraction of smaller, task-specific models for downstream applications. This improves computational efficiency while preserving strong multi-modal performance.

## III. METHODS

Figure 2 presents a high-level overview of our method.

### A. Problem Setup

We consider a multi-modal dataset $\mathcal{D}_n^m$ consisting of $n$ co-located and temporally matched samples from $m$ different modalities. Our model is based on the vision transformer (ViT) base architecture [43]. Input samples $\mathbf{x}_i^j \sim \mathcal{D}_n^m$ of shape $\mathbf{x} \in \mathbb{R}^{(\sum_j^m C^j) \times H \times W}$ are split into patches and linearly embedded per modality with learnable patch embedding transforms $\{f^j\}_{j=1,\dots,m}$. This yields a set of $m$ $D$-dimensional tokens for each sample patch. Sinusoidal positional embeddings to encode spatial information are added to each token [44]. We investigate sinusoidal and learnable *modality*-tokens to encode the modality of tokens across the channel dimension. To mark modality transitions in the token sequence, learnable end-of-modality tokens are injected before attention operations.

### B. Masked Image Modeling for Multi-modal Pretraining

We utilize masked image modeling as self-supervised pretraining technique [45]. After patch-embedding, a binary mask is applied to the input sequence of tokens, independently masking $75\%$ of the tokens corresponding to each modality as in [45]. The remaining tokens are processed by a ViT encoder. Once encoded, the original sequence length is restored by introducing special, learnable, `[MASK]` tokens at the positions of dropped input tokens. A ViT decoder then aims to reconstruct the original (normalized) pixel values of the masked patches. We use a mean-squared error (MSE) objective between pixel values of the original masked patches and their reconstruction for training [45].

*a) Modality-wise Decoding:* To keep the reconstruction task difficult in the presence of multiple independently masked modalities (a spatial patch might be masked in one modality but visible in another), we reconstruct all modalities independently. All tokens of a modality are processed by the decoder in a distinct forward pass, incurring a total of $m$ decoder forward passes, while also reducing the sequence length in each pass by a factor of $\frac{1}{m}$.

### C. Mixture-of-Modality Experts

Mixture-of-experts (MoE) models replace the feed-forward layers in each ViT block with a routing module and multiple feed-forward *expert* modules [28]. The router consists of a linear layer that produces logits for each token, from which a softmax distribution over the $e$ experts is constructed. The expert outputs for each token are then weighted by the routing softmax weights. Unlike token-based routing, MAPEX performs modality-level routing, where all tokens from the same modality are routed to the same experts. We adapt the MoE framework to enforce expert$\leftrightarrow$modality relationships in the model. Starting from a fixed assignment of experts to modalities, we ultimately design a learnable routing mechanism based on modality information that consistently routes tokens from a modality $M$ to the same subset of experts. Top-$k$ selection of the $k$ highest ranked experts is used to achieve sparse conditional computation for each modality. Our **deterministic routing** setup defines a fixed one-to-one relationship between modalities and experts. Every modality

is assigned to exactly one expert a priori, and no experts are shared between modalities. **Positional embedding routing** utilizes sinusoidal positional embeddings [44] across the stack of modalities as input to the routing module. While the routing weights are learned, the input embeddings remain fixed and unique for each modality. This approach supports an arbitrary number of experts but introduces an arbitrary ordering among modalities. **Modality routing** introduces learnable [MODALITY] embeddings that represent each modality in lieu of channel-wise positional embeddings. Unlike tokens based on modalities' relative position in the input data, learnable embeddings do not introduce an implicit order between the modalities. These tokens are added to each modality and used as input to the routing module. Explicitly, for token $t^m$ of modality $m$, the routing weights are computed by passing the associated modality embedding to the routing module $r$: $w_r^m = r([\texttt{MODALITY}]^m)$. Unlike standard MoE implementations, this routing strategy guarantees that tokens of a modality will always be routed to the same experts, even across samples.

*a) Load Balancing Loss:* Imbalanced expert utilization is a common problem with learned routing mechanisms [28]. We observe that some experts randomly receive a lower shares of tokens at the beginning of training, resulting in fewer gradient updates and continued low utilization of these experts throughout training. We address this issue by introducing a differentiable load-balancing loss that encourages a more uniform distribution of routing probabilities between experts:

$$\mathcal{L}_{\text{load}} = \frac{1}{e}\sum_{i=1}^{e}(U_i - \frac{1}{e})^2, \tag{1}$$

with $e$ the total number of experts and $U_i$ the empirical utilization fraction of the $i^{\text{th}}$ expert (i.e., the fraction of all tokens that expert $i$ processes). This additional loss term encourages more uniform expert utilization. During pre-training, the load balancing loss is combined with the reconstruction loss to create the training objective: $\mathcal{L} = \mathcal{L}_{\text{rec}} + \alpha \cdot \mathcal{L}_{\text{load}}$ where the weighting factor $\alpha$ scales the load balancing loss.

*b) Shared Expert:* We hypothesize that some features are shared between modalities and introduce an additional shared expert that gets access to tokens from all modalities. The shared expert can process common features (such as edges, corners, or colors) that are shared between modalities, allowing the modality experts to focus on information that is specific to individual modalities. We limit the number of shared experts to one to maintain lightweight modality-specific models after pruning (see Sec. V-A0d for a discussion on the trade-off between model size and number of experts)

*c) Modality Dropout:* In order to prime the model for downstream applications where the full set of pre-training modalities is not available, we utilize modality dropout during the pre-training stage. In each batch, the tokens corresponding to a pre-defined fraction of modalities are set to zero.

### D. Downstream Modality-Aware Expert Pruning

After pre-training the mixture-of-modality experts model with masked autoencoding on unlabeled multi-modal data, we transfer the model encoder to the downstream task of interest. Often only a subset of the pre-training modalities will be available for the downstream task. Accordingly, we prune the mixture-of-modality experts to suit the available data modalities. For a model pre-trained with experts $E^p$ on $\mathcal{M}^p$ pre-training modalities and $k$ experts per modality, all experts and patch-embedding projections corresponding to modalities that are not in the set of downstream modalities are dropped. The downstream experts $E^d$ are extracted from each pre-trained model layer as:

$$E^d = \{E_i | i \in \arg\operatorname{top} k(w_r^m) \forall m \in \mathcal{M}^d\}_{i=1,\dots,|E^p|}, \tag{2}$$

with $w_r^m$ the pre-training routing probabilities for modality $m$. In the case of a single downstream modality, this yields a model with $k$ downstream experts. To utilize the pruned model on a supervised downstream task, we perform standard fine-tuning on labeled data. This step adapts the purely unsupervised multi-modal model trained for reconstruction to modality-specific supervised tasks.

## IV. Datasets

*a) BigEarthNet-GeoEnvironment (ben-ge):* The *ben-ge* dataset [46] is an extension to the BigEarthNet-MM [47] dataset. It consists of 590 326 colocated and temporally matched samples from Sentinel-1 (SAR), Sentinel-2 (multi-spectral) as well as elevation and environmental data. We utilize the *ben-ge* dataset for multi-modal pre-training of the mixture-of-modality experts model. We group the data into six distinct modalities: *RGB* (Sentinel-2 B04, B03, B02), *Red Edge* (Sentinel-2 B05, B06, B07), *SWIR* (short-wave infrared, Sentinel-2 B11, B12, B8A), *NIR* (near-infrared, Sentinel-2 B08), *SAR* (Sentinel-1 VV and VH polarizations), and *ELE* (elevation map). **ben-ge-8k** is a subset of *ben-ge* consisting of ≈8 000 samples that we use for downstream evaluation on multi-label land-use/land-cover classification with different modalities.

*b) SEN12-FLOOD:* This dataset consists of Sentinel-1 and Sentinel-2 imagery acquired in areas prone to flooding and corresponding binary flood labels [48] across the world. It contains 366 time-series of varying length, with 512×512 pixels per tile. The dataset is split by locations and tiles within a time-series are considered independently. We use the SAR data from SEN12-FLOOD for downstream evaluation of the pre-trained model on a single modality.

*c) California Wildfire:* We construct a dataset specifically for wildfire detection by collecting fire perimeter data from the California Fire Perimeters dataset [49] and recent CAL FIRE incidents records [50]. These records include detailed spatial and temporal information on wildfire events, such as locations, dates, and sizes. For each fire perimeter, we acquire corresponding Sentinel-2 satellite imagery [51]. To create a binary classification dataset, we crop small image patches (224×224 pixels) and label them as either `wildfire` or `no-wildfire` based on overlap with the wildfire perimeter. We utilize Sentinel-2 short-wave infrared (SWIR) bands (B8A, B11, and B12), which have proven most effective in detecting burned areas [52], [15]. The resulting dataset

TABLE I
COMPARISON OF ROUTING MECHANISMS USING $k$-NN ACCURACY FOR LAND-COVER CLASSIFICATION ON DIFFERENT *ben-ge-8k* [46] MODALITIES.

| Modality | Deterministic | Pos. Embed | Modality-based |
|---|---|---|---|
| RGB | 63.5 | 63.1 | **67.2** |
| Red Edge | 63.2 | 64.9 | **66.7** |
| SWIR | 68.7 | 67.9 | **72.4** |
| NIR | 51.6 | 55.9 | **56.8** |
| SAR | 62.2 | 65.3 | **66.1** |
| ELE | 46.0 | **48.3** | 48.2 |

contains 20 000 samples, with 11 000 labeled as `wildfire` and 9 000 as `no-wildfire`.

## V. EXPERIMENTS AND RESULTS

### A. Self-supervised Pre-training

MAPEX models are pre-trained with masked autoencoding on multi-modal remote sensing data with balanced modality shares. Representations generated by the mixture-of-modality experts encoder (6 experts, 360M parameters) are mapped back to pixel-space for each modality with an 8-block transformer decoder (see Section III-B). We pre-train for 50 epochs with AdamW [53] on the *ben-ge* dataset. The load balancing loss is weighted by $\alpha$=0.01 which yields near uniform expert utilization. Pre-training results in good reconstruction performance (see Fig. 3) for most modalities. Reconstruction is most difficult for the SAR backscatter data, where reconstruction MSE remains $\approx 3\times$ higher than on the other modalities.

*a) Routing Mechanism:* In contrast to standard MoE approaches, this work aims to elicit strong expert $\leftrightarrow$ modality relationships during pre-training, which facilitates expert-pruning for downstream tasks. In standard token-based routing, routing probabilities for each token $t$ are computed by the router as $w_r$=$r(t)$. Initial experiments with token-based routing on multi-modal remote sensing data failed to produce a clear specialization of experts for individual modalities, and suffered from unbalanced expert utilization. We address and evaluate these shortcomings with three alternative routing mechanisms (see Section III-C): deterministic routing, positional embedding routing, and modality-based routing. We find that pre-training with modality-based routing consistently outperforms the two other routing mechanisms across downstream tasks of different modalities (see Table I). This confirms our hypothesis that the routing parameters should be learnable (unlike the deterministic routing), and that learnable modality tokens are a suitable input for the routing module.

*b) Multi-modal Pre-training:* We assess the effect of multi-modal pre-training on the *ben-ge* dataset. We compare uni-modal pre-training (only on RGB) against multi-modal pre-training (using all available modalities). After pre-training, all models are pruned to a single expert. Table III (first two rows) reports the $k$-NN accuracy on the RGB modality. The results show that multi-modal pre-training leads to better performance on the RGB downstream task compared to pre-training exclusively on the RGB modality. This suggests that the model benefits from shared representations learned across multiple modalities, even when only RGB is available at test time.

*c) Expert Modules:* To evaluate the benefit of using multiple expert modules in multi-modal data, we compare multi-modal MAPEX variants with a single and multiple experts (see Table III; last two rows). After pre-training, both models are pruned to one expert and evaluated on the RGB modality. The model pre-trained with multiple experts significantly outperforms the model with one expert. Please note that both models are the same size after the pruning stage.

*d) Shared Expert:* We study the impact of using a shared expert on MAPEX performance across multiple modalities. The shared expert is hypothesized to capture common cross-modal information, enabling modality-specific experts to specialize in unique features. We fine-tune the shared expert along with the retained modality-specific experts. Table IV indicates that, while the shared expert generally enhances performance across most modalities, it also introduces a significant increase in model parameters. This raises a trade-off between performance improvement and computational efficiency, suggesting that the shared expert may not always provide a cost-effective benefit, particularly in resource-constrained applications.

*e) Modality Dropout:* We investigate the impact of varying modality dropout values (0%, 10%, and 50%) during pre-training, on downstream model performance. Table V shows that a dropout value of 50% outperforms both 10% and 0% by around 1.5% (absolute) on average. Intuitively, a higher dropout value encourages the model to become more robust by learning representations that do not rely on the presence of all modalities. By setting a modality dropout value of 50% during pre-training, the model learns to handle partial data input, which aligns better with scenarios where some modalities may be unavailable. This approach is particularly beneficial as a significant portion of the model is shared during pre-training, with only certain parts pruned for downstream tasks. As a result, the shared component gains robustness from the dropout, enhancing the model's adaptability across different downstream specialized tasks.

### B. Downstream Evaluation

We evaluate pre-trained MAPEX across several downstream tasks and compare to other MAE-based remote sensing foundation models. MAPEX is based on the ViT-base architecture and retains two experts after pruning for one modality. SatMAE [8] is a ViT-base model trained on multispectral Sentinel-2 data, while Scale-MAE [6] is trained on the RGB bands of Sentinel-2. Since Scale-MAE only provides pre-trained weights for a ViT-large architecture, we use that configuration in our comparisons. Our setup is as follows: after pre-training, we retain only the top two FFN experts for each modality, pruning the others. This results in smaller, modality-specialized models (*i.e.* RGB experts for RGB input, SAR experts for SAR input, *etc.*). These specialized models

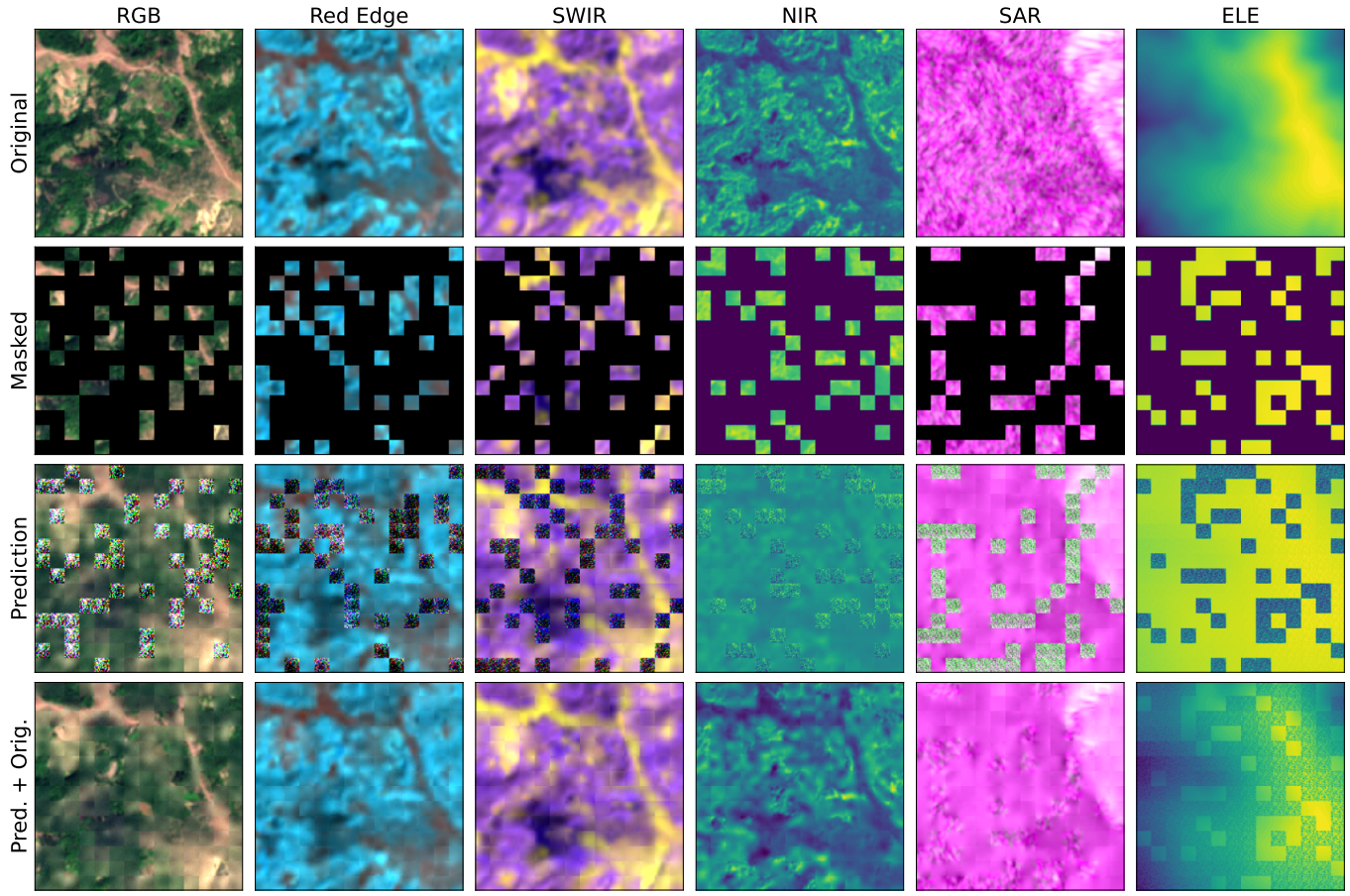


Fig. 3. Reconstruction examples from mixture-of-modality experts model on multi-modal data. From left to right: optical RGB, red edge, short-wave infrared, near infrared, synthetic aperture radar, and elevation from *ben-ge* [46]. The same decoder is used for all modalities.

TABLE II
$k$-NN AND FINE-TUNING (FT) PERFORMANCE (ACCURACY IN %) FOR LAND-COVER CLASSIFICATION ON DIFFERENT MODALITIES OF THE *ben-ge-8k* DATASET [46]. FOR EACH MODALITY, THE BEST PERFORMING MODEL IS HIGHLIGHTED IN **BOLD**.

| Modality | Supervised FS | SatMAE [8] | | Scale-MAE [6] | | MAPEX (Ours) | |
|---|---|---|---|---|---|---|---|
| | | $k$-NN | FT | $k$-NN | FT | $k$-NN | FT |
| Architecture<br># params | *MAPEX*<br>130M | *vit_base*<br>90M | | *vit_large*<br>310M | | *MAPEX*<br>130M | |
| RGB | 71.2 | 64.5 | 70.8 | 75.1 | **79.9** | 67.2 | 75.6 |
| Red Edge | 73.3 | 65.5 | 69.7 | 74.7 | **80.3** | 66.7 | 77.8 |
| SWIR | 76.5 | 68.3 | 74.5 | 76.2 | **83.7** | 72.4 | 79.8 |
| NIR | 56.1 | 50.0 | 56.9 | 49.9 | 57.2 | 56.8 | **59.4** |
| SAR | 69.7 | 62.2 | 67.2 | 65.3 | 73.1 | 66.1 | **73.6** |
| ELE | 51.8 | 46.5 | 47.1 | 46.6 | 51.4 | 48.2 | **58.8** |

TABLE III
EFFECT OF UNI VS. MULTI-MODAL PRE-TRAINING AND NUMBER OF PRE-TRAINING EXPERTS ON *ben-ge-8k* [46]. $k$-NN ACCURACY ON THE RGB MODALITY AFTER PRUNING ALL MODELS TO 1 EXPERT.

| Pre-training Modalities | Pre-training Experts | $k$-NN Acc. |
|---|---|---|
| 1 (RGB) | 1 | 62.1 |
| 6 (All) | 1 | 63.2 |
| 6 (All) | 6 | 65.1 |

are also compared against the same model architecture and size (2 experts) trained fully supervised from scratch (FS).

*a) Specialized Models:* Table II shows that fine-tuning our specialized models on their respective input modalities outperforms other methods for NIR, SAR, and ELE inputs. This pattern is consistent whether we perform $k$-NN classification on features extracted from the specialized models or fine-tune the entire models. On modalities that are closer to Scale-MAE pre-training data, this model outperforms our approach. This is partly explained by the fact that Scale-MAE is 2.3× larger

TABLE IV
$k$-NN PERFORMANCE FOR LAND-COVER CLASSIFICATION ON DIFFERENT MODALITIES OF THE BEN-GE-8K DATASET [46], WITH AND WITHOUT THE USE OF A SHARED EXPERT.

| Modality | 2 Experts | 2 Experts + 1 shared |
|---|---|---|
| # params | 130M | 200M |
| RGB | 67.2 | **67.6** |
| Red Edge | 66.7 | **67.1** |
| SWIR | 72.4 | **72.9** |
| NIR | **56.8** | 56.0 |
| SAR | 66.1 | **66.7** |
| ELE | **48.2** | 47.8 |

TABLE V
COMPARISON OF DIFFERENT MODALITY DROPOUT VALUES. $k$-NN ACCURACY FOR LAND-COVER CLASSIFICATION ON DIFFERENT MODALITIES OF THE BEN-GE-8K DATASET [46] IS REPORTED.

| Modality | Dropout = 0 | Dropout = 0.1 | Dropout = 0.5 |
|---|---|---|---|
| RGB | 65.8 | 65.8 | **67.2** |
| Red Edge | 66.1 | 66.2 | **66.7** |
| SWIR | 71.6 | 71.9 | **72.4** |
| NIR | 54.2 | 54.3 | **56.8** |
| SAR | 65.7 | 65.8 | **66.1** |
| ELE | 46.6 | 47.1 | **48.2** |

than our specialized models. More importantly, our method outperforms the bigger Scale-MAE on those modalities that differ the most from the pre-training data used for Scale-MAE.

Besides land-cover classification, we test MAPEX on two other specific tasks: flood and wildfire detection. These are important applications that are commonly tackled with specific remote sensing modalities (see Figure 4).

For example, SAR (Sentinel-1) data is more informative than RGB data for detecting water in cloudy weather, while the SWIR bands of Sentinel-2 provide better information than RGB data for detecting burned areas. Since most foundation models are commonly trained on either the full spectrum of Sentinel-2 or only on RGB data, we demonstrate in Table VI that our smaller, specialized models outperform general foundation models. Specifically, for flood detection, MAPEX outperforms SatMAE by 3% (absolute) and ScaleMAE by 1%, while for wildfire detection, it surpasses SatMAE by 2% and performs comparably to ScaleMAE despite being considerably smaller. Beyond classification and detection, we also evaluate MAPEX on segmentation tasks to assess its ability to handle dense predictions. To this end, we add the same simple segmentation head to both the RGB-specific pruned model (with MAPEX) and SatMAE, and we compare the results to a U-Net baseline. Table VII shows that MAPEX achieves the best performance, demonstrating its versatility across multiple remote sensing tasks.

*b) Number of Experts :* We explore how the number of selected experts (top-$k$) influences the downstream performance of specialized models across the different modalities. We train models with up to 6 experts per modality and adaptively drop experts for downstream evaluation. We face a trade-off between expert number and model size: pruned models with 1 expert have 90M parametes, while 6 experts

TABLE VI
$k$-NN AND FINE-TUNING PERFORMANCE (ACCURACY IN %) FOR BINARY FLOOD DETECTION USING SENTINEL 1 DATA [48] AND WILDFIRE DETECTION [49] USING SHORT-WAVE INFRARED (SWIR) DATA (SENTINEL-2 B8A, B11, B12 BANDS).

| Modality | FS | SatMAE | | Scale-MAE | | MAPEX | |
|---|---|---|---|---|---|---|---|
| | | $k$-NN | FT | $k$-NN | FT | $k$-NN | FT |
| Architecture<br># params | *MAPEX*<br>130M | *vit_base*<br>90M | | *vit_large*<br>310M | | *MAPEX*<br>130M | |
| Flood Detection | 78.1 | 77.4 | 80.3 | 83.0 | 82.6 | 80.1 | **83.2** |
| Wildfire Detection | 84.7 | 84.2 | 88.6 | 86.6 | **90.8** | 87.3 | 90.5 |

TABLE VII
MIOU (IN %) FOR SEMANTIC SEGMENTATION ON THE *ben-ge-8k* DATASET [46]

| Modality | U-Net [54] | SatMAE (*vit_base*) | MAPEX |
|---|---|---|---|
| RGB | 56.3 | 53.1 | **56.9** |

inflate the model to 370M. Figure 5 shows that most modality-specific models achieve their highest accuracy when using the top 2 or 3 experts. Higher top-$k$ values often reduce accuracy likely because the model becomes too large, with too much capacity for a small dataset. Given that, we choose the top 2 as it yields a smaller, more efficient model with minimal loss in accuracy.

*c) Specialization of Experts:* This work focuses on transforming a general, modality-agnostic foundation model into smaller modality-specific models by selectively keeping experts specialized to each type of data. To test the effectiveness of these specialized models, we create modality-specific versions and evaluate them across all available data types. Results are shown in Figure 6. This involves, for example, retaining only the RGB experts to build an RGB-specific model, then testing it with various input modalities (represented in the first row of Figure 6). We repeat this process for each modality. The heatmap reveals that each specialized model performs best when tested on its intended data type (e.g., RGB experts on RGB data, NIR experts on NIR data). These results indicate that MAPEX expert routing successfully instills expert↔modality relationships into the model during pre-training.

*d) Balancing Expert Size and Number:* To find the ideal number of experts and sizes, we conduct an experiment with fixed downstream model size (90M, ViT-Base) while varying the number and size of experts per modality. To ensure a fair comparison without inflating model sizes when using more experts, we fix the number of parameters added to the base model across expert configurations. Models are pretrained with different numbers of experts and pruned to a fixed number per modality: 1, 2, 5, or 10 experts are retained per modality. This corresponds to expert sizes of 4.7M, 2.4M, 0.9M, and 0.5M parameters, respectively. Figure 7 shows that optimal expert configurations vary by modality, but the majority of modalities achieve the highest accuracy with 2 experts per modality. For RGB, accuracy peaks with 5 smaller experts (0.9M each),

TABLE VIII
$k$-NN AND FINE-TUNING (FT) PERFORMANCE (ACCURACY IN %) FOR LAND-COVER CLASSIFICATION ON DIFFERENT COMBINATION OF MODALITIES OF THE *ben-ge-8k* DATASET [46].

| Modality | Supervised FS | SatMAE [8] | | ScaleMAE [6] | | MAPEX (Ours) | |
|---|---|---|---|---|---|---|---|
| | | $k$-NN | FT | $k$-NN | FT | $k$-NN | FT |
| Architecture | *MAPEX* | *vit_base* | | *vit_large* | | *MAPEX* | |
| # params | 130M | 90M | | 310M | | 130M | |
| RGB | 71.2 | 64.5 | 70.8 | 75.1 | 79.9 | 67.2 | 75.6 |
| SAR | 69.7 | 62.2 | 67.2 | 65.3 | 73.1 | 66.1 | 73.6 |
| RGB + SAR | 74.8 | 70.8 | 74.7 | 67.2 | 80.2 | 72.9 | 79.9 |

TABLE IX
FEW-SHOT RESULTS FOR FINE-TUNING (ACCURACY IN %) ON LAND-COVER CLASSIFICATION USING THE RGB MODALITY OF *ben-ge-8k* [46].

| | Supervised FS | SatMAE [8] | ScaleMAE [6] | MAPEX (Ours) |
|---|---|---|---|---|
| Architecture | *MAPEX* | *vit_base* | *vit_large* | *MAPEX* |
| # params | 130M | 90M | 310M | 130M |
| 10-shot | 32.1 | 34.5 | 38.9 | 38.6 |
| 100-shot | 48.1 | 53.0 | 57.9 | 56.1 |

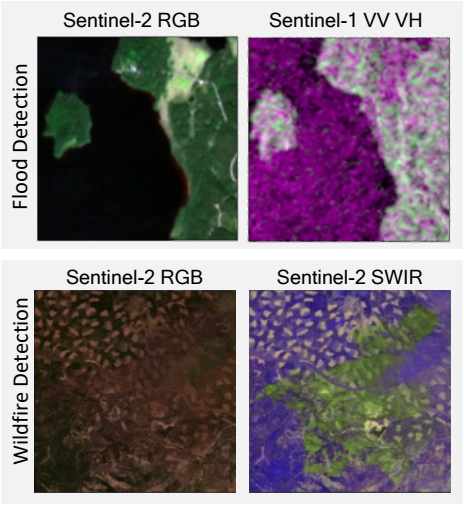


Fig. 4. Task specific modalities: Sentinel-1 bands provide valuable information for detecting water bodies in the presence of clouds. Similarly, SWIR bands are more effective than RGB for identifying burned areas.

suggesting that moderate specialization enhances learning. In contrast, elevation (ELE) performs best with fewer, larger experts, indicating a less diverse feature space that benefits from stronger individual experts. These findings suggest that a fixed number of experts per modality is suboptimal, but 2 experts per modality is a good compromise across most cases.

*e) Multimodal Downstream Evaluation:* Since combining multiple modalities is often beneficial for some downstream tasks, we evaluate the multimodal capabilities of MAPEX by testing its performance on different spatially co-registered data modalities: RGB, SAR, and their combination (RGB + SAR) from the *ben-ge-8k* dataset. For unimodal inputs, we use the top-2 experts per modality, while for the combined RGB + SAR input, we use the top-1 expert per modality to maintain the same model size during evaluation. Note that for the multimodal setup, we merge the data channel-wise before feeding it to the model. Table VIII shows that MAPEX outperforms SatMAE and the supervised baseline for both RGB and SAR inputs while closely matching ScaleMAE, despite its significantly smaller architecture. These results show that MAPEX can easily adapt and compete with other methods, depending on the available modalities at test time.

*f) Few-shot Learning:* We perform few-shot experiments to evaluate the label efficiency of our method and compare its performance with other approaches. In each k-shot experiment, we randomly select k samples for every class from the training set. We keep the same experimental setup as described in Section V-B. Table IX reports the performance of MAPEX compared to SatMAE [8] and ScaleMAE [6] on the RGB modality of the *ben-ge-8k* dataset. The results, presented for both 10-shot and 100-shot scenarios, show that MAPEX achieves competitive performance, outperforming SatMAE in both scenarios and nearing the performance of ScaleMAE, despite ScaleMAE having a significantly larger architecture. This demonstrates that MAPEX can handle limited labeled data effectively, making it a practical and efficient alternative to larger models like ScaleMAE.

### C. Computational Efficiency Analysis

Furthermore, we evaluate MAPEX from a computational efficiency perspective. Table X reports inference latency, FLOPs, and peak memory usage for MAPEX before and after pruning, as well as for SatMAE and ScaleMAE. Modality-aware pruning reduces computational cost from 866.7 G to 33.9 G FLOPs and decreases peak memory from 14.8 GB to 0.96 GB, while improving latency by over an order of magnitude. These results demonstrate that MAPEX not only yields lightweight, modality-specialized expert models but also significantly improves runtime efficiency.

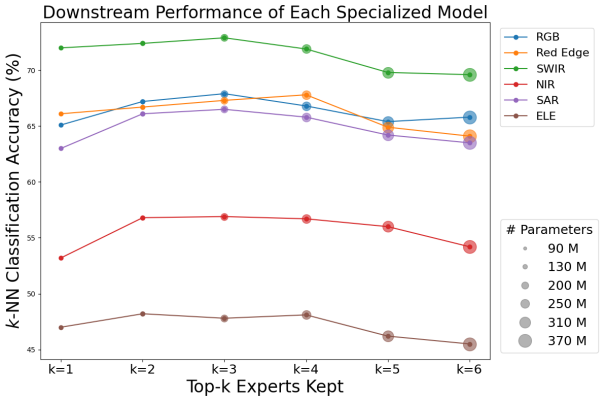


Fig. 5. Downstream accuracy with different numbers of retained modality experts $k$. Most models achieve peak accuracy with top-$k\in\{2,3\}$, while more experts tend to reduce accuracy.

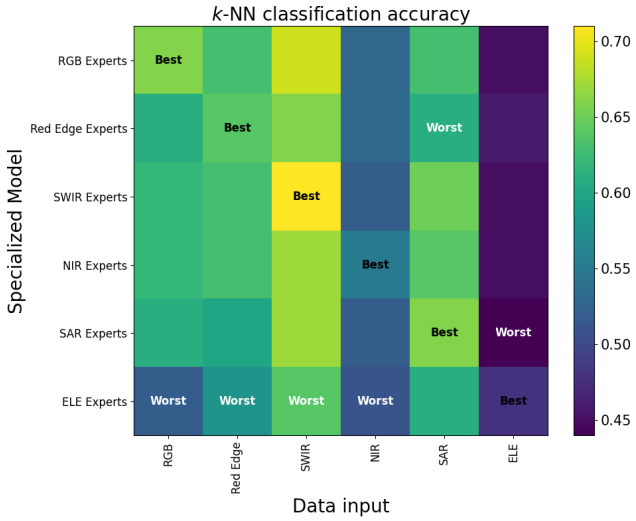


Fig. 6. Expert↔modality specialization on the *ben-ge-8k* dataset [46]. Each row corresponds to the performance of MAPEX pruned for a specific modality and applied across different modalities. The best and worst specialized models for each input modality are highlighted.

TABLE X
COMPUTATIONAL EFFICIENCY COMPARISON OF MAPEX (BEFORE PRUNING (B.P.) AND AFTER PRUNING (A.P.)) AGAINST BASELINE MODELS. LATENCY IS MEASURED WITH BATCH SIZE 1 ON AN NVIDIA H100

| Model | Params | FLOPs | Latency | Peak Memory |
|---|---|---|---|---|
| MAPEX (b.p.) | 713M | 866.7 G | 187.5 ms | 14.8 GB |
| MAPEX (a.p.) | 130M | 33.9 G | 14.7 ms | 0.96 GB |
| SatMAE | 88M | 58.2 G | 6.4 ms | 0.82 GB |
| ScaleMAE | 303M | 84 G | 10.3 ms | 1.80 GB |

### D. Implementation Details

To evaluate a pre-trained model with $k$-NN classification, we extract and average-pool image-level features for each sample. Then, a $k$-NN classifier with $k$=5 is fit to the feature representations. For fine-tuning on classification tasks, we append an additional linear classification layer to the pre-trained model. For pre-training and downstream evaluation, images are resized to $224\times224$ pixels and normalized channel-wise. We fine-tune the models for 30 epochs, using AdamW [53] optimizer and a learning rate scheduler that reduces the rate when the validation loss ceases to improve.

## VI. Limitations & Future Work

This work introduced a new method to efficiently leverage large pre-trained remote sensing foundation models for modality-specific downstream applications. Although many remote sensing tasks can still be addressed using data from individual, purpose-built sensors, the increasing availability of multi-modal observations will drive demand for downstream models capable of jointly exploiting heterogeneous data sources. While the MAPEX approach can be applied in multi-modal settings, the benefits of the proposed model-pruning strategy diminish as the number of modalities grows. This limitation suggests a need for new techniques that can more effectively manage multi-modal information during pre-training, fine-tuning, and inference. Beyond multi-modality, integrating multi-temporal remote sensing data into this framework remains an important and open challenge for future work.

## VII. Discussion & Conclusion

This work proposes a mixture-of-modality experts module suitable for large-scale self-supervised pre-training on multi-modal data. To improve downstream performance on individual modalities, we present a pruning mechanism to derive efficient downstream models specialized for individual modalities from the pre-trained model. The MAPEX model is successfully pre-trained on multi-modal remote sensing data, achieving good reconstruction performance across modalities. Our experiments show that the modality-based expert routing mechanism elicits specialization among expert modules, resulting in sub-networks that can be mapped to individual modalities. When pruned for a specific modality, the large-scale pre-trained model turns into a much smaller specialized model that performs well on downstream tasks without any fine-tuning, as $k$-NN evaluation experiments show. Fine-tuned, our modality-specific models perform competitively with existing fine-tuned remote sensing foundation models.

## References

[1] 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.

[2] 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," *arXiv preprint arXiv:2410.16602*, 2024.

[3] S. Lu, J. Guo, J. R. Zimmer-Dauphinee, J. M. Nieusma, X. Wang, P. VanValkenburgh, S. A. Wernke, and Y. Huo, "AI Foundation Models in Remote Sensing: A Survey," *arXiv preprint arXiv:2408.03464*, 2024.

[4] M. Chi, A. Plaza, J. A. Benediktsson, Z. Sun, J. Shen, and Y. Zhu, "Big Data for Remote Sensing: Challenges and Opportunities," *Proceedings of the IEEE*, vol. 104, no. 11, pp. 2207–2219, 2016.

[5] J. Hanna, M. Mommert, and D. Borth, "Sparse multimodal vision transformer for weakly supervised semantic segmentation," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2145–2154, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:260911594

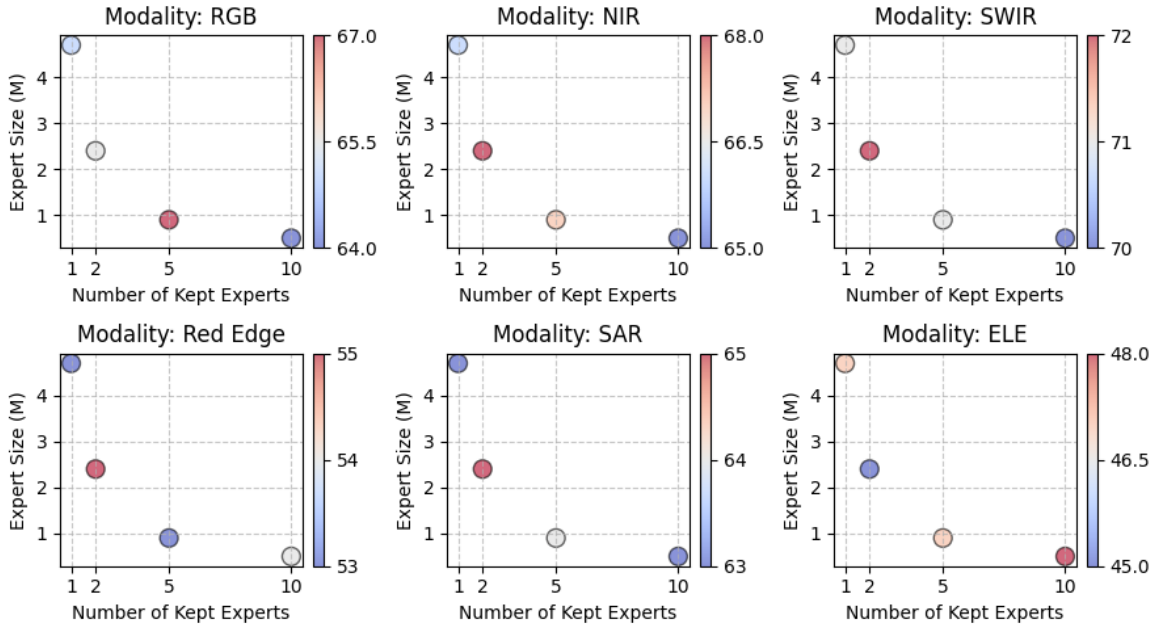


Fig. 7. Downstream accuracy (%) for different expert numbers and sizes per modality. Total model size is fixed at 90M.


[6] C. Reed, R. Gupta, S. Li, S. Brockman, C. Funk, B. Clipp, S. Candido, M. Uyttendaele, and T. Darrell, "Scale-MAE: A Scale-Aware Masked Autoencoder for Multiscale Geospatial Representation Learning," *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4065–4076, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:255340927

[7] J. Jakubik, S. Roy, C. Phillips, P. Fraccaro, D. Godwin, B. Zadrozny, D. Szwarcman, C. Gomes, G. Nyirjesy, B. Edwards, D. Kimura, N. Simumba, L. Chu, S. K. Mukkavilli, D. Lambhate, K. Das, R. Bangalore, D. Oliveira, M. Muszynski, K. Ankur, M. Ramasubramanian, I. Gurung, S. Khallaghi, H. Li, M. Cecil, M. Ahmadi, F. Kordi, H. Alemohammad, M. Maskey, R. K. Ganti, K. Weldemariam, and R. Ramachandran, "Foundation models for generalist geospatial artificial intelligence," *ArXiv*, vol. abs/2310.18660, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:264590307

[8] 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," *ArXiv*, vol. abs/2207.08051, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:250626671

[9] 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 foundation model for observing the earth crossing modalities," *ArXiv*, vol. abs/2403.15356, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:271708675

[10] A. Fuller, K. Millard, and J. Green, "Croma: Remote sensing representations with contrastive radar-optical masked autoencoders," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[11] Z. Yan, J. Li, X. Li, R. Zhou, W. Zhang, Y. Feng, W. Diao, K. Fu, and X. Sun, "RingMo-SAM: A Foundation Model for Segment Anything in Multimodal Remote-sensing Images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–16, 2023.

[12] V. Nedungadi, A. Kariryaa, S. Oehmcke, S. Belongie, C. Igel, and N. Lang, "MMEarth: Exploring Multi-Modal Pretext Tasks for Geospatial Representation Learning," *arXiv preprint arXiv:2405.02771*, 2024.

[13] X. Guo, J. Lao, B. Dang, Y. Zhang, L. Yu, L. Ru, L. Zhong, Z. Huang, K. Wu, D. Hu *et al.*, "Skysense: A Multi-modal Remote Sensing Foundation Model Towards Universal Interpretation for Earth Observation Imagery," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 27 672–27 683.

[14] D. Amitrano, G. Di Martino, A. Di Simone, and P. Imperatore, "Flood Detection with SAR: A review of Techniques and Datasets," *Remote Sensing*, vol. 16, no. 4, p. 656, 2024.

[15] A. E. Cocke, P. Z. Fulé, and J. E. Crouse, "Comparison of Burn Severity Assessments using Differenced Normalized Burn Ratio and Ground Data," *International Journal of Wildland Fire*, vol. 14, no. 2, pp. 189–198, 2005.

[16] R. P. Sishodia, R. L. Ray, and S. K. Singh, "Applications of Remote Sensing in Precision Agriculture: A Review," *Remote sensing*, vol. 12, no. 19, p. 3136, 2020.

[17] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[18] 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.

[19] T. B. Brown, "Language Models are Few-shot Learners," *arXiv preprint arXiv:2005.14165*, 2020.

[20] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment Anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.

[21] D. Hong, B. Zhang, X. Li, Y. Li, C. Li, J. Yao, N. Yokoya, H. Li, P. Ghamisi, X. Jia *et al.*, "SpectralGPT: Spectral Remote Sensing Foundation Model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[22] 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*. PMLR, 2021, pp. 8748–8763.

[23] C. Li, Z. Gan, Z. Yang, J. Yang, L. Li, L. Wang, J. Gao *et al.*, "Multimodal Foundation Models: From Specialists to General-purpose Assistants," *Foundations and Trends® in Computer Graphics and Vision*, vol. 16, no. 1-2, pp. 1–214, 2024.

[24] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling Laws for Neural Language Models," *arXiv preprint arXiv:2001.08361*, 2020.

[25] Sietsma and Dow, "Neural Net Pruning-Why and How," in *IEEE 1988 international conference on neural networks*. IEEE, 1988, pp. 325–333.

[26] G. Hinton, "Distilling the Knowledge in a Neural Network," *arXiv preprint arXiv:1503.02531*, 2015.

[27] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank Adaptation of Large Language Models," *arXiv preprint arXiv:2106.09685*, 2021.

[28] 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," *arXiv preprint arXiv:1701.06538*, 2017.

[29] C. Riquelme, J. Puigcerver, B. Mustafa, M. Neumann, R. Jenatton, A. Susano Pinto, D. Keysers, and N. Houlsby, "Scaling Vision with Sparse Mixture of Experts," *Advances in Neural Information Processing Systems*, vol. 34, pp. 8583–8595, 2021.

[30] Y. Zhou, T. Lei, H. Liu, N. Du, Y. Huang, V. Zhao, A. M. Dai, Q. V. Le, J. Laudon *et al.*, "Mixture-of-experts with Expert Choice Routing," *Advances in Neural Information Processing Systems*, vol. 35, pp. 7103–7114, 2022.

[31] W. Fedus, B. Zoph, and N. Shazeer, "Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity," *Journal of Machine Learning Research*, vol. 23, no. 120, pp. 1–39, 2022.

[32] B. Han, S. Zhang, X. Shi, and M. Reichstein, "Multisensory Geospatial Models via Cross-Sensor Pretraining."

[33] J. Cao, Y. You, and J. Liu, "MV-MOE: A Visual Mixture-of-Experts Model for Optical-SAR Image Matching," in *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2024, pp. 9676–9679.

[34] H. Lin, D. Hong, S. Ge, C. Luo, K. Jiang, H. Jin, and C. Wen, "RS-MoE: Mixture of Experts for Remote Sensing Image Captioning and Visual Question Answering," *arXiv preprint arXiv:2411.01595*, 2024.

[35] X. V. Lin, A. Shrivastava, L. Luo, S. Iyer, M. Lewis, G. Gosh, L. Zettlemoyer, and A. Aghajanyan, "MoMa: Efficient Early-fusion Pre-training with Mixture of Modality-aware Experts," *arXiv preprint arXiv:2407.21770*, 2024.

[36] Y. Jain, H. S. Behl, Z. Kira, and V. Vineet, "Damex: Dataset-aware mixture-of-experts for visual understanding of mixture-of-datasets," *ArXiv*, vol. abs/2311.04894, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:265050925

[37] T. Lillesand, R. W. Kiefer, and J. Chipman, *Remote Sensing and Image Interpretation*. John Wiley & Sons, 2015.

[38] C. Toth and G. Jóźków, "Remote Sensing Platforms and Sensors: A Survey," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 115, pp. 22–36, 2016.

[39] D. Wang, M. Hu, Y. Jin, Y. Miao, J. Yang, Y. Xu, X. Qin, J. Ma, L. Sun, C. Li *et al.*, "HyperSIGMA: Hyperspectral Intelligence Comprehension Foundation Model," *arXiv preprint arXiv:2406.11519*, 2024.

[40] Y. Glaser, J. E. Stopa, L. M. Wolniewicz, R. Foster, D. Vandemark, A. Mouche, B. Chapron, and P. Sadowski, "WV-Net: A Foundation Model for SAR WV-mode Satellite Imagery Trained using Contrastive Self-supervised Learning on 10 Million Images," *arXiv preprint arXiv:2406.18765*, 2024.

[41] V. Nedungadi, A. Kariryaa, S. Oehmcke, S. J. Belongie, C. Igel, and N. Lang, "Mmearth: Exploring multi-modal pretext tasks for geospatial representation learning," *ArXiv*, vol. abs/2405.02771, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:269605353

[42] L. Scheibenreif, J. Hanna, M. Mommert, and D. Borth, "Self-supervised Vision Transformers for Land-cover Segmentation and Classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1422–1431.

[43] A. Dosovitskiy, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *arXiv preprint arXiv:2010.11929*, 2020.

[44] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[45] 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. 16 000–16 009.

[46] M. Mommert, N. Kesseli, J. Hanna, L. Scheibenreif, D. Borth, and B. Demir, "Ben-ge: Extending BigEarthNet with geographical and environmental data," in *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2023, pp. 1016–1019.

[47] 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," *IEEE Geoscience and Remote Sensing Magazine*, vol. 9, no. 3, pp. 174–180, 2021.

[48] C. Rambour, N. Audebert, E. Koeniguer, B. Le Saux, M. Crucianu, and M. Datcu, "Sen12-flood: a SAR and Multispectral Dataset for Flood Detection," *IEEE: Piscataway, NJ, USA*, 2020.

[49] California Department of Forestry and Fire Protection, "California Fire Perimeters (all)," https://catalog.data.gov/dataset/california-fire-perimeters-all-b3436, accessed: 2024-11.

[50] ——, "CAL FIRE Incidents," https://www.fire.ca.gov/incidents, accessed: 2024-11.

[51] M. Drusch, U. Del Bello, S. Carlier, O. Colin, V. Fernandez, F. Gascon, B. Hoersch, C. Isola, P. Laberinti, P. Martimort *et al.*, "Sentinel-2: ESA's optical high-resolution mission for GMES operational services," *Remote sensing of Environment*, vol. 120, pp. 25–36, 2012.

[52] Geospatial Data Platform, "Normalized burn ratio (NBR)," https://docs.up42.com/help/spectral-indexes/nbr, accessed: 2024-11.

[53] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *International Conference on Learning Representations*, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:53592270

[54] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *ArXiv*, vol. abs/1505.04597, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:3719281