REVIEW ARTICLE

Mixture-of-Experts in Remote Sensing: A Survey

Yongchuan Cui[1,2], **Peng Liu**[1,2,*] **and Lajiao Chen**[1,2]

[1] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China
[2] School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 101408, China

Abstract

Remote sensing data analysis and interpretation present unique challenges due to the diversity in sensor modalities and spatiotemporal dynamics of Earth observation data. Mixture-of-Experts (MoE) model has emerged as a powerful paradigm that addresses these challenges by dynamically routing inputs to specialized experts designed for different aspects of a task. However, despite rapid progress, the community still lacks a comprehensive review of MoE for remote sensing. This survey provides the first systematic overview of MoE applications in remote sensing, covering fundamental principles, architectural designs, and key applications across a variety of remote sensing tasks. The survey also outlines future trends to inspire further research and innovation in applying MoE to remote sensing.

Keywords: Mixture-of-Experts, remote sensing, image classification, vision-language models, object detection, change detection, multi-modal fusion, super-resolution.

Submitted: 17 December 2025
Accepted: 13 January 2026
Published: 23 March 2026

Vol. 1, **No.** 1, 2026.
10.62762/JGRS.2025.140654

***Corresponding author:**
✉ Peng Liu
liupeng202303@aircas.ac.cn

1 Introduction

Remote sensing has become an indispensable information source for observing the Earth's surface and atmosphere, supporting applications such as land-cover and land-use mapping, agriculture and forestry monitoring, urban planning, climate and environmental surveillance, and disaster management [1–3]. Modern remote sensing archives cover multiple sensor modalities (optical, Synthetic Aperture Radar (SAR), Light Detection and Ranging (LiDAR), multispectral, hyperspectral), spatial resolutions from sub-meter to kilometre scale, and dense temporal sampling from days to decades [4–6]. This diversity and scale create rich opportunities for detailed Earth observation, but they also pose significant challenges for classical Machine Learning (ML) and standard Deep Learning (DL) pipelines. In particular, the same land-cover class may exhibit very different signatures across sensors, resolutions, seasons and viewing geometries, while different classes can appear very similar in a single modality or band [1, 2]. Handling such heterogeneity, domain shifts and long-tailed label distributions remains a central problem in operational remote sensing systems, where traditional ML approaches often struggle to generalize across diverse acquisition conditions [7–10].

Over the past decade, DL has substantially improved

Citation
Cui, Y., Liu, P., & Chen, L. (2026). Mixture-of-Experts in Remote Sensing: A Survey. *Journal of Geoscience and Remote Sensing*, 1(1), 4–38.

performance in many remote sensing tasks by leveraging powerful generic architectures originally developed for Computer Vision (CV) and pattern recognition. Convolutional Neural Networks (CNNs) such as AlexNet [11] and ResNet [12], fully convolutional networks for dense prediction [13], and U-Net-style encoder–decoder models [14] have all been adapted to high-resolution aerial and satellite imagery for classification, detection and segmentation [1–3]. More recently, Transformer-based models and large-scale pre-training have further improved representation quality and cross-task transfer, following developments such as the Vision Transformer and attention-based sequence modeling [6, 15, 16]. Nevertheless, most existing remote sensing models are still trained as monolithic networks for relatively narrow settings: a fixed set of sensors, a limited range of geographic regions, or a single family of tasks. As a result, they often struggle to generalize across sensors, domains and tasks without extensive re-training or careful domain adaptation [4, 9].

feature
expert
MoE
Remote
multi
accuracy
attention
detection
module
classification
semantic
object
scale
tasks
spatial
segmentation
transformer
network
scene
resolution
spectral
input
layer
fusion
parameter
domain
dataset
mixture
Vision
Sensing
hyperspectral
modalities
modality
frequency
loss
encoder
channel
multimodal
foundation
region
neural
fine
level
map
time
stage
classes
process
limited
single
token
strategy
output
visual
land
set
target
tuning
CNN
distillation
global
knowledge
framework
noise
SAR
HSI
LiDAR
class
weight
temporal
extraction
measurement
optical
uncertainty
specific
specialized
prediction
recognition
source
structure
cross
cover
DOTA
RS
patch
local
context
Landsat
label
routing
samples
study
values
sensor
router
satellite
band
field
mIoU
LoRA
mask
mechanism
ocean
climate
convolution
LLM
Earth
deep
Net

Figure 1. Word cloud of the most frequent words appearing in MoE-related remote sensing papers.

Mixture-of-Experts (MoE) models provide an alternative way to increase model flexibility and capacity while retaining efficiency. The basic idea, introduced in the early 1990s, is to decompose a complex prediction problem into simpler sub-problems that are handled by multiple expert networks, with a gating function that assigns input-dependent weights or responsibilities to each expert [17–19]. Adaptive mixtures of local experts and hierarchical mixtures of experts formalized this *divide-and-conquer* principle in a probabilistic framework, leading to theoretical results on approximation properties, identifiability and consistency for mixtures-of-experts and related generalized linear models [20–24]. Later surveys provide unified treatments of MoE from both practical and theoretical viewpoints and emphasize that the key design choices concern the form of the experts, the gating strategy, and the way experts are regularized and trained jointly [21, 22]. Compared with a single dense network, an MoE architecture can allocate different subsets of parameters to different regions of the input space, classes, modalities or tasks, making it a natural candidate for heterogeneous remote sensing data.

In recent years, MoE has re-emerged as a central mechanism for scaling up deep neural networks in Natural Language Processing (NLP) and CV. Sparsely-gated MoE layers [25] and GShard [26] demonstrated that conditional computation can decouple parameter count from per-example computation by activating only a small subset of experts for each token. Switch Transformers [27] and GLaM [28] further refined gating and load-balancing schemes to train language models with hundreds of billions to over a trillion parameters at practical cost, while MoE via shallow embedding explored channel-wise MoE routing within CNNs [29]. For multi-task and multi-modal learning, M^{3}ViT [30] integrates MoE layers into Vision Transformers [31] to reduce interference between tasks while keeping inference efficient, Mod-Squad [32] treats experts as reusable modules that can be shared or specialized across tasks, and MoE-based semantic segmentation frameworks use expert combinations to analyze multi-modal inputs [33]. At the systems level, DeepSpeed-MoE [34] provides an end-to-end framework for training and serving very large MoE models efficiently, and recent surveys on MoE in large language models synthesize developments in routing, optimization and deployment [35]. These results indicate that MoE architectures are especially suitable for scenarios with heterogeneous data and tasks, where conditional computation and expert specialization can be exploited.

These properties make MoE particularly attractive for remote sensing, where multi-sensor fusion, long-term time series analysis and multi-task learning are often required in a single workflow. However, explicit MoE usage in remote sensing is still relatively new compared with NLP and general CV. Early work primarily relied on ensembles or multi-classifier systems without an explicit learned gating network [39, 40], and most deep remote sensing models remain dense. In the last few years, a number of studies have begun to introduce MoE explicitly into remote sensing pipelines. MixtureRS [41] replaces dense feed-forward blocks by sparse MoE layers in a cross-modality Transformer for hyperspectral–LiDAR land-use classification. Heterogeneous MoEs architectures have been proposed for remote sensing image super-resolution, with expert groups specialized to different ground-object characteristics and dual routing to adapt reconstruction to local content [42]. For multi-modal change detection, M^{2}CD [43] integrates MoE modules into the backbone to explicitly handle the distribution gap between optical and SAR images, while an uncertainty-aware MoE model has been designed to address long-tailed crop type mapping from multi-source imagery [44]. In the context of remote sensing vision–language models, RS-MoE and RSUniVLM [45, 46] adopt expert-based routing mechanisms to improve captioning, visual question answering, and multi-granularity reasoning. At an even larger scale, RingMoE [47] proposes a multi-modal MoE foundation model that jointly pre-trains modality-specific and shared experts on massive optical and SAR datasets, and mixture-of-experts networks have also been leveraged for specialized applications such as burned area mapping from multi-temporal satellite imagery [48]. The word cloud in Figure 1 summarizes the most frequent terms appearing in MoE-related remote sensing papers. The word cloud is based on analysis of 57 core MoE-related remote sensing papers published between 2016 and 2025. Text from paper titles, abstracts, and key technical sections was first tokenized into words using standard word tokenization methods. Word frequency was then computed from the tokenized text, with normalization to account for paper length variations. Standard English stop words were removed using a stop word list to focus on domain-specific technical terms. Literature selection criteria included papers that explicitly apply MoE architectures to remote sensing tasks, published in peer-reviewed journals and conferences as well as recent preprints, given that many MoE-related papers in remote sensing are very recent. The selection covers major application domains (classification, detection, change detection, multi-modal fusion, etc.). The dominance of words such as *expert*, *feature*, *multi*, *module* and *MoE* confirms that current work is primarily concerned with designing expert modules and feature-processing pipelines, rather than completely new backbone architectures. The prominence of *remote*, *sensing*, *classification*, *detection*, *segmentation*, *semantic* and *object* indicates that most MoE applications concentrate on standard high-level vision tasks (scene and land-cover classification, object detection, semantic segmentation), while terms such as *change*, *time*, and *series* appear but are noticeably smaller, reflecting that temporal and change detection problems are less explored. Frequent occurrence of *multi*, *modal*, *scale*, *spectral*, *spatial* and sensor-related words (*optical*, *SAR*, *hyperspectral*, *LiDAR*) highlights a strong emphasis on multi-modal and multi-scale fusion, where MoE is used to manage heterogeneity across modalities and resolutions. Meanwhile, architectural and training-related terms such as *attention*, *transformer*, *token*, *LoRA*, *foundation*, *uncertainty* and *router* suggest that many methods adapt MoE ideas from large vision/vision–language models and combine them with parameter-efficient tuning or uncertainty modeling. Overall, the word cloud reveals a research landscape in which MoE is mainly deployed as expert-based feature modules for multi-modal, multi-scale classification and segmentation, with relatively fewer works addressing temporal modeling, low-level restoration or large unified foundation models. The MoE paradigm can be adapted to diverse remote sensing problems, but existing works are scattered across tasks and design choices, and a consolidated view is still lacking.

The goal of this survey is to provide a systematic overview of MoE methods for remote sensing, connecting general MoE developments with domain-specific requirements. We first review the fundamentals of MoE architectures, including expert design, gating strategies and training methods, with an emphasis on concepts that are most relevant for geospatial data [21, 22]. We then organize existing remote sensing MoE work by task type (e.g., classification, segmentation, detection, change detection, time series modeling, and vision–language understanding) [1, 2, 6, 41–48]. Finally, we discuss open challenges and future directions, including unified multi-modal and multi-task MoE foundations for Earth observation, expert interpretability and analysis, training strategies and efficient deployment on resource-constrained platforms. By linking

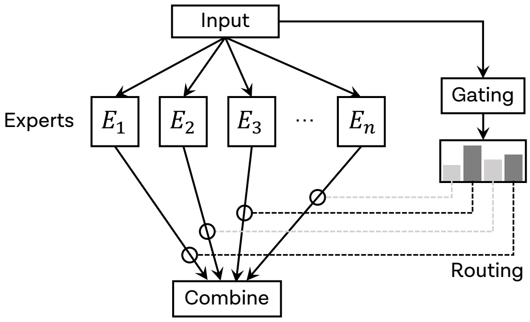

Figure 2. Basic architecture of Mixture-of-Experts (MoE).

advances in MoE modeling with the specific characteristics of remote sensing data, we aim to clarify both the current status and the potential of MoE as a general framework for remote sensing analysis.

2 Fundamentals of Mixture-of-Experts

A Mixture-of-Experts (MoE) model consists of a set of expert networks and a gating (routing) network (Figure 2). The gating network takes an input x (e.g., an image, a token embedding, or a feature vector) and produces nonnegative routing weights $g_1(x), \ldots, g_n(x)$ over n experts, typically obtained by normalizing routing scores (e.g., via a softmax). In the dense (soft) case, all experts may be active and the weights satisfy $\sum_{i=1}^{n} g_i(x) = 1$. In the sparse case, only a small subset of experts is selected (e.g., top-1 or top-k), and the remaining weights are set to zero, i.e., $g_i(x) = 0$ for $i \notin \mathcal{S}_k(x)$, where $\mathcal{S}_k(x)$ denotes the selected expert set for input x. Formally, if we denote the experts as $E_1, E_2, \ldots, E_n$ and the gating function as $G(x)$ producing weights $g_1(x), \ldots, g_n(x)$, the MoE output is a typically sparse weighted mixture of expert outputs:

$$\begin{aligned} y &= \sum_{i=1}^{n} g_i(x)\, E_i(x) \\ &= \sum_{i \in \mathcal{S}_k(x)} g_i(x)\, E_i(x), \end{aligned} \tag{1}$$

where the second equality emphasizes the common sparse-routing implementation. Here $g_i(x)$ is the routing weight assigned to expert i for input x. In practice, many MoE models use sparse gating to keep computation efficient even when the total number of experts n is large, e.g., selecting the top-1 or top-k experts per input [27]. Early MoE formulations used softmax gating trained via Expectation-Maximization (EM) [18], whereas modern implementations often use hard or top-k routing trained end-to-end with backpropagation, and may add noise or regularization to encourage load balancing across experts [25, 27]. A commonly used auxiliary objective for this purpose is a load-balancing loss $\mathcal{L}_{\text{balance}}$, which penalizes uneven expert utilization aggregated over a minibatch or the data distribution. Let

$$u_i = \mathbb{E}_{x \sim \mathcal{B}}[g_i(x)], \tag{2}$$

denote the expected routing importance of expert i over a minibatch $\mathcal{B}$, and let $\mathbf{u} = (u_1, \ldots, u_n)$. One simple imbalance measure is the Coefficient of Variation (CV) computed across experts:

$$\mathcal{L}_{\text{balance}} = \mathrm{CV}(\mathbf{u}) = \frac{\sqrt{\mathrm{Var}_i(\mathbf{u})}}{\mathbb{E}_i[\mathbf{u}]}, \tag{3}$$

where $\mathrm{Var}_i(\cdot)$ and $\mathbb{E}_i[\cdot]$ denote the variance and the expectation taken over the expert index i, respectively. When utilization differs strongly across experts, $\mathrm{CV}(\mathbf{u})$ becomes large, yielding a larger balancing penalty in the overall training objective and thus discouraging the gate from collapsing to a small subset of experts. For training stability, one often uses the squared coefficient of variation.

2.1 Taxonomy of MoE

Classical MoE models were introduced as modular regressors or classifiers, in which a set of expert models is combined by a gating network that assigns input-dependent weights to each expert. Early work such as adaptive mixtures of local experts and hierarchical mixtures of experts modeled each expert as a generalized linear model or shallow neural network, with the gate producing a softmax distribution over all experts and parameters estimated by variants of the EM algorithm or maximum likelihood [17, 18, 20]. These works mainly focused on approximation properties, statistical consistency, and small- to medium-scale tasks, and treated MoE as a probabilistic mixture with dense activation of the expert set [21, 22, 49]. Modern MoE systems extend this formulation along the aspect of model scale and sparsity. The transition from dense probabilistic MoE to sparse deep MoE can be described in terms of routing granularity, expert size, number of experts, and layer placement in Large Language Models (LLMs) [35–38]. Sparsely-gated MoE layers activate only a small subset of experts (typically top-k by the gating score) for each token, thereby decoupling total parameter count from per-token computation cost. Representative language and vision models include the sparsely-gated MoE layer [25], GShard [26], Switch Transformers [27], GLaM [28], V-MoE [54] and LIMoE [55], which demonstrate that conditional computation allows models with hundreds of billions to trillions of parameters to be trained at reasonable computational budgets.

Current MoE research also focuses on how experts are organized around tasks, domains and modalities. In multi-task and multi-objective settings, multi-gate MoE (MMoE) and progressive layered extraction (PLE) architectures use separate gating networks per task, together with shared and task-specific experts, to encourage both parameter sharing and task specialization in recommendation and advertising systems [50, 51]. Follow-up work shows that sparsely activated MoE with task-aware routing can improve transfer to low-resource tasks and robustness when many tasks are trained jointly [32, 52, 53]. In the multimodal setting, MoE backbones such as V-MoE [54], LIMoE [55] and Uni-Perceiver-MoE [56] combine image and text streams or more general modalities through shared Transformer [15] layers with MoE feed-forward blocks, where certain experts become specialized to modalities or sub-domains. For LLMs, architectures like DeepSeekMoE [57], Mixtral [58] and OLMoE [59] position MoE layers inside decoder-only Transformers [15] to obtain large effective capacity with relatively small active parameter counts per token.

System-level and implementation choices are also crucial for making large MoE models trainable and deployable. Works such as FastMoE [60], DeepSpeed-MoE [34], Tutel [61], HetuMoE [62], FasterMoE [63] and MegaBlocks [64] focus on efficient distributed training and inference, including all-to-all communication scheduling, CUDA kernel optimizations, expert placement, and parallelism strategies. Pre-gated MoE and related work co-design the algorithm and system by simplifying routing at runtime to reduce memory movement and improve latency, while preserving the advantages of conditional computation [65]. Other works such as BASE layers [66], hash layers [67] and sparse-upcycling [68] show that MoE-like conditional computation and expert modularity can be realized either by routing dense blocks or by converting pre-trained dense models into sparse expert collections. Current MoE research can be viewed as varying along model sparsity and scale, task and modality organization, and system realization. The following subsections focus on expert specialization, gating and training within this taxonomy.

2.2 Expert Specialization

In classical MoE models, expert specialization is usually interpreted as a partition of the input or covariate space into regions, each handled by a different local model. Adaptive mixtures of local experts and hierarchical mixtures of experts show that the gating network tends to assign nearby inputs to the same expert or leaf in a tree, and theoretical results establish approximation and consistency properties for such hierarchical mixtures-of-experts under suitable assumptions [17, 18, 20, 21, 49]. These works mostly use low-dimensional inputs and relatively small networks, but already make clear that a successful MoE should distribute data among experts in a way that yields simpler local mappings and avoids redundant experts.

In deep sparse MoE models, specialization emerges at the level of high-dimensional representations. The sparsely-gated MoE layer [25], GShard [26], Switch Transformers [27] and GLaM [28] all use token-level routing in Transformer [15] feed-forward sublayers, and empirical analyses show that experts tend to specialize to language, syntactic patterns or subsets of tokens, although the degree of specialization can vary with routing hyperparameters and auxiliary losses. Vision and multimodal MoE models such as V-MoE [54] and LIMoE [55] report that some experts focus on particular visual categories, image resolutions or modalities, while others act as more general experts, and that sparse expert utilization is crucial for scaling to large models without severe redundancy. Analytical studies confirm that, under certain conditions, MoE architectures can learn to partition data into clusters or domains and achieve better robustness and generalization than dense counterparts [69–71].

Expert specialization is particularly important in multi-task and multi-domain applications. MMoE [50] and PLE [51] explicitly factor experts into shared and task-specific groups, with multiple gating networks combining them to capture both common and idiosyncratic structure in recommendation tasks. Sparsely activated multi-task MoEs [52] and modular MoE [32] designs such as Mod-Squad [53] extend this idea by allowing tasks to share only subsets of experts and by routing different tasks or labels to different expert combinations, which can improve resistance to negative transfer and catastrophic forgetting when new tasks are added. Time-MoE pushes this idea to billion-scale time-series models, where experts specialize to temporal dynamics and domains while still being trained within a unified foundation model [74]. In these systems, specialization is typically encouraged through task-aware gating, auxiliary load-balancing losses and regularization, rather than by explicit hard constraints on which tasks an expert may serve. For LLMs, recent work has proposed architectural mechanisms that more directly control specialization. DeepSeekMoE separates shared experts, which are always active and capture common skills such as general syntax, from routed experts, which are selectively activated and encouraged to learn rarer capabilities or domain-specific knowledge [57]. Mixtral arranges experts in the feed-forward sublayers of a decoder-only Transformer [15] and shows that different experts specialize to language families, code versus natural language, or input length regimes, while using only a small number of active experts per token [58]. OLMoE conducts a detailed routing analysis and reports strong specialization across layers and experts, with some experts focusing on particular capability clusters or subsets of the training data distribution [59]. Recent works on MoE in LLMs emphasize that this kind of specialization is a key reason why sparse MoE LLMs can match or surpass dense models with substantially fewer active parameters at inference time [35, 36, 38]. The success of LLMs in NLP has inspired similar MoE architectures for vision and multimodal tasks, demonstrating that the LLM paradigm can be effectively adapted to other domains.

Parameter-efficient adaptation introduces another form of expert specialization, in which the experts are not full feed-forward blocks but low-rank (LoRA) [75] adapter modules. LoRA represents task-specific updates as low-rank matrices added to existing weights, and several works extend this idea to mixtures of LoRA experts. TT-LoRAMoE [76], MoLE [77] and HMoRA [78] treat each adapter or group of adapters as an expert and learn gating functions that select among them based on the input or task, which allows multiple domains or instruction styles to be captured within a single base LLM while keeping parameter overhead small. These methods demonstrate that expert specialization can be realized at the level of adapters as well as full layers, and can be learned during fine-tuning without retraining the base model from scratch.

Theoretical work further clarifies when and how MoE architectures specialize. Studies on convergence rates for Gaussian mixtures-of-experts and on the statistical behavior of softmax and sparse top-k gating provide conditions under which experts consistently approximate different parts of the data-generating function and under which gating parameters can be reliably estimated [70–73]. Together with empirical analyses in large-scale systems [28, 57, 59, 69], these results support the view that expert specialization is a central mechanism by which MoE models convert increased parameter count into improved accuracy and robustness.

2.3 Gating Strategies

Gating strategies determine how inputs are assigned to experts and are therefore central to the behavior of MoE models. Classical MoE work generally used softmax gating functions that output a full probability distribution over experts, combined with EM or gradient-based optimization. In these models, the output is usually a convex mixture of all experts, and the gate is trained to allocate responsibility for each data point across experts; theoretical analyses show how such softmax gates affect approximation rates and identifiability [17, 18, 21, 22, 49, 70]. More recent work on the statistical properties of softmax gating and its variants provides more rigorous guarantees for parameter estimation and convergence in Gaussian and generalized linear MoE models [71–73].

In deep sparse MoE architectures, the dominant strategy is sparse top-k token-level routing. The sparsely-gated MoE layer uses a linear gating network that outputs logits for each expert, applies a softmax, and then routes each token to the top-k experts (often $k = 1$ or 2), with additional noise and auxiliary load-balancing losses to prevent collapse onto a few experts [25]. GShard [26], Switch Transformers [27] and GLaM [28] adopt similar routing, differing in the exact loss design, capacity constraints and implementation details, but all select a small set of experts per token and rely on gradient-based training of the gate. Later work such as ST-MoE studies the stability of such routers, proposes auxiliary objectives such as the z-loss, and analyzes how routing schemes and noise influence expert utilization and transfer performance [69, 79]. Routing variations have been proposed to improve load balancing, robustness and flexibility. MoE with expert choice routing reverses the usual perspective and lets experts choose a fixed number of tokens rather than tokens choosing experts; this achieves better load balancing and allows the number of experts per token to vary [80]. Dynamic routing strategies adapt the number of experts or routing pattern to input difficulty, for example by activating more experts for harder examples [81]. Multi-gate designs such as MMoE [50] and PLE [51] use separate gating networks per task or tower, which allows different tasks or label spaces to see different mixtures of shared and task-specific experts while using the same underlying expert pool. Hash-based and static routing methods such as hash layers map tokens to experts using deterministic hash functions to eliminate learned gates, and BASE layers demonstrate that a fixed, random assignment can still achieve strong performance when combined with appropriate regularization and training [66, 67].

Recent work also aims to make sparse routing differentiable or more amenable to optimization. DSelect-k [82] replaces the hard top-k operator with a continuous relaxation based on a differentiable selection mechanism, enabling end-to-end gradient-based learning while still producing sparse expert assignments. Lory [83] and related fully differentiable MoE architectures merge experts in parameter space and route at the level of segments rather than individual tokens, removing the discrete selection step and simplifying router training in autoregressive language models. Other approaches, such as stochastic experts and locality-sensitive hashing-based routing, treat routing as a stochastic process, using randomness to regularize the model and to reduce communication overhead in large distributed systems [62, 84]. System-level designs such as Tutel and Pre-gated MoE show that simplifying routing computations, for example by precomputing routing decisions or using hierarchical gating, can substantially improve throughput without changing the overall model structure [61, 65].

Gating strategies and training objectives are closely linked. ST-MoE and subsequent work report that small changes to gating losses, capacity factors and noise can strongly affect training stability and final performance [69, 79, 80]. Theoretical analyses of softmax and sparse gating provide conditions under which gradient-based training converges to meaningful partitions, while empirical studies emphasize that practical MoE routers must balance model quality, communication cost and implementation simplicity [35–38].

2.4 Training Methods

Training methods for MoE models can be organized along three main dimensions: basic training strategies that establish fundamental load balancing and routing optimization, advanced optimization schemes that introduce sophisticated regularization and balancing mechanisms, and cross-scenario adaptation methods that enable flexible deployment and conversion from dense models.

Fundamental MoE training relies on standard load balancing and routing optimization approaches. The sparsely-gated MoE layer [25] introduced auxiliary load-balancing losses to prevent expert collapse, using capacity constraints and noise injection to encourage balanced expert utilization. Switch Transformers [27] refined these basic strategies with simplified top-k routing and improved load balancing through auxiliary losses that penalize uneven expert usage. GShard [26] further developed capacity constraints and load balancing mechanisms, ensuring that each expert receives a roughly equal number of tokens per batch. These basic strategies form the foundation for MoE training, focusing on preventing expert collapse and maintaining balanced utilization through simple auxiliary objectives and capacity constraints.

Recent work on training MoE models increasingly focuses on how routing decisions are made and how expert loads are controlled during optimization. Building on basic training strategies, recent work has developed more sophisticated optimization techniques. Wang et al. [85] propose an auxiliary-loss-free load balancing strategy that maintains expert-wise bias terms and updates them based on recent routing statistics, so that expert loads remain balanced without injecting additional gradients into the main objective. Thaman [86] formulates router balancing as a constrained optimization problem and derives a dual-ascent update with sparsemax gating that enforces target usage ratios while leaving the task loss unchanged. Omi et al. [87] replace uniform balancing losses by a similarity-preserving term that encourages similar tokens to be routed to similar expert sets, which improves convergence speed and reduces redundancy in expert usage. Complementary to these balancing-oriented approaches, SimSMoE [88] explicitly addresses representation collapse by minimizing similarity between expert representations, and shows that such regularization leads to more diverse experts under a fixed FLOPs budget during pretraining and fine-tuning. In parallel, several works use the MoE structure itself as a training regularizer: SMoE-Dropout [89] trains a dense Transformer [31] together with a sparse MoE layer driven by a randomly initialized and frozen router, gradually increasing the number of activated experts so that the final model becomes self-slimmable at inference while mitigating expert collapse during training, and MoEC [90] introduces expert clusters and variance-based constraints on the routing distribution, together with cluster-level dropout, so that experts within the same cluster specialize to complementary sub-regions of the data and maintain useful diversity even when data per expert is limited. In all of these cases, routing patterns, balancing mechanisms and expert activation schedules are treated as explicit components of the training procedure, rather than as purely passive structures optimized only through the main task loss.

Beyond optimizing routing within a fixed MoE architecture, recent work has explored methods for converting dense models to MoE and adapting MoE training to flexible inference-time behavior. On top of changing how routing is optimized, some works study how to convert or upcycle existing dense checkpoints into MoE models and how to adapt MoE training to flexible inference-time behavior and downstream optimization. MoEBERT [91] starts from a pre-trained BERT [99], partitions each feed-forward block into multiple experts, and trains the router with layer-wise distillation so that the converted MoE model matches the dense teacher while enabling sparse inference. D2DMoE [92] regularizes activation sparsity in a dense model and then converts groups of neurons into experts with a dynamic-k routing rule, learning routers that predict expert contributions and activating a variable number of experts per token to obtain large inference speedups with limited additional training. ToMoE [93] converts dense LLMs into MoE models through dynamic structural pruning that exposes experts implicitly present in the dense network and then fine-tunes the router and pruned structure to recover or improve the original accuracy. For text embedding models, Nussbaum and Duderstadt train Nomic Embed v2 [94] as a sparse MoE encoder using contrastive and distillation objectives, together with routing constraints that ensure balanced expert usage, showing that MoE-style training is also effective for retrieval-style encoders and not only for causal language models. Training procedures have also been adapted to support flexible inference-time configurations and reinforcement-learning fine-tuning: Elastic MoE [95] explicitly randomizes both the number and composition of active experts during pretraining so that the same checkpoint can later be evaluated with different expert budgets at inference, encouraging experts to collaborate under many activation patterns and enlarging the range over which activating more experts improves accuracy, while Ma et al. [96] show that in reinforcement learning for MoE LLMs the discrepancy between routing during training and routing during inference can destabilize policy optimization and address this by recording routing distributions from the inference engine and replaying them during training, which aligns the two phases and prevents collapse during reinforcement learning fine-tuning. Beyond single-model training, Gururangan et al. [97] demonstrate that one can train separate expert language models on clusters of documents and combine them as a sparse ensemble at inference time, effectively realizing a data-driven mixture-of-experts without tightly coupling expert training in a single network, and AutoMoE [98] integrates training with neural architecture search by learning where to place experts, how many experts to allocate and how much computation each token should receive under explicit compute constraints, showing that heterogeneous MoE layouts can be discovered automatically in neural machine translation.

3 Mixture-of-Experts in Remote Sensing

Mixture-of-Experts (MoE) models have been applied across a wide range of remote sensing tasks (refer to Figure 3). In this section, we survey their use in key application domains, including image classification, object detection and segmentation, multi-modal data fusion, change detection and temporal analysis, image restoration, and vision-language tasks. We organize the discussion by task type, highlighting how MoE architectures address specific challenges in each domain and summarizing representative studies. Notably, the idea of combining multiple expert networks in remote sensing is not entirely new, earlier works in the 1990s and 2000s explored multi-network and multi-model approaches for sensor fusion and image analysis [136, 144], but recent advances in Deep Learning (DL) and sparse gating have greatly expanded MoE's capabilities. Throughout these applications, a common theme is conditional model capacity: MoEs enable the model to activate different subsets of parameters (experts) depending on the input, allowing specialization for diverse data characteristics while keeping overall computation efficient.

3.1 Image Classification

Land-cover and scene classification is a foundational task in remote sensing, where each image or each pixel in an image is labeled as a certain category (such as water, urban, forest, agriculture, etc.).

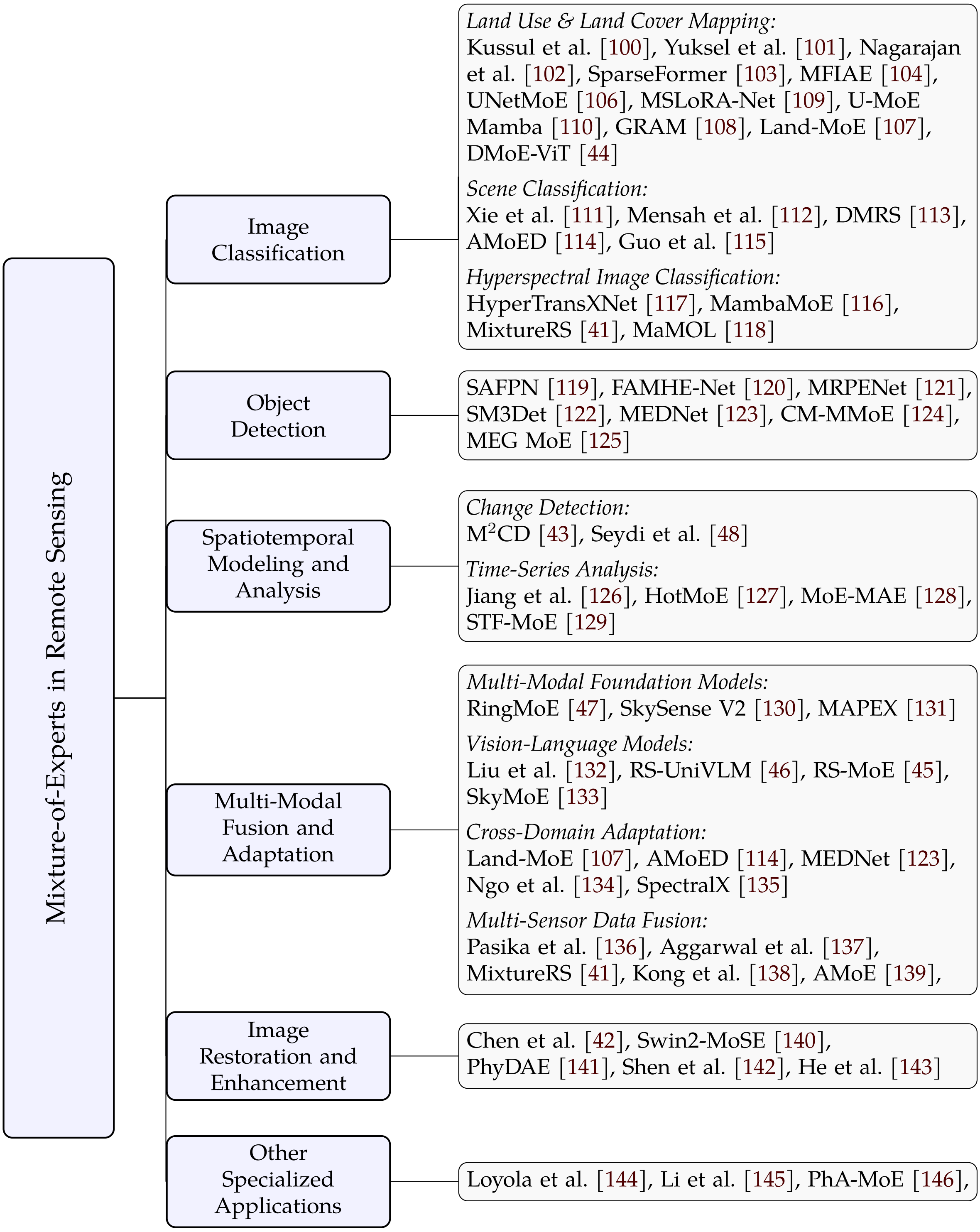
Mixture-of-Experts in Remote Sensing
Image Classification
Land Use & Land Cover Mapping:
Kussul et al. [100], Yuksel et al. [101], Nagarajan et al. [102], SparseFormer [103], MFIAE [104], UNetMoE [106], MSLoRA-Net [109], U-MoE Mamba [110], GRAM [108], Land-MoE [107], DMoE-ViT [44]
Scene Classification:
Xie et al. [111], Mensah et al. [112], DMRS [113], AMoED [114], Guo et al. [115]
Hyperspectral Image Classification:
HyperTransXNet [117], MambaMoE [116], MixtureRS [41], MaMOL [118]
Object Detection
SAFPN [119], FAMHE-Net [120], MRPENet [121], SM3Det [122], MEDNet [123], CM-MMoE [124], MEG MoE [125]
Spatiotemporal Modeling and Analysis
Change Detection:
M^{2}CD [43], Seydi et al. [48]
Time-Series Analysis:
Jiang et al. [126], HotMoE [127], MoE-MAE [128], STF-MoE [129]
Multi-Modal Fusion and Adaptation
Multi-Modal Foundation Models:
RingMoE [47], SkySense V2 [130], MAPEX [131]
Vision-Language Models:
Liu et al. [132], RS-UniVLM [46], RS-MoE [45], SkyMoE [133]
Cross-Domain Adaptation:
Land-MoE [107], AMoED [114], MEDNet [123], Ngo et al. [134], SpectralX [135]
Multi-Sensor Data Fusion:
Pasika et al. [136], Aggarwal et al. [137], MixtureRS [41], Kong et al. [138], AMoE [139],
Image Restoration and Enhancement
Chen et al. [42], Swin2-MoSE [140], PhyDAE [141], Shen et al. [142], He et al. [143]
Other Specialized Applications
Loyola et al. [144], Li et al. [145], PhA-MoE [146],

Figure 3. Overview of Mixture-of-Experts applications in remote sensing.

This task is challenging because of high intra-class variability, i.e., the same class can appear very different in different conditions, and inter-class similarity, i.e., different land-cover types can have confusingly similar appearances or spectral signatures. Traditional classification approaches struggled with these issues, especially when relying on fixed spectral thresholds or shallow classifiers. Early neural network methods and sensor fusion attempts provided some improvements but were limited by training data and model capacity. The advent of DL such as Convolutional Neural Networks (CNNs) [11] and Transformers [31] significantly boosted classification accuracy by extracting hierarchical features and learning complex spectral-spatial patterns. However, even advanced DL networks can hit performance plateaus on complex scenes or when faced with very heterogeneous data. While CNNs excel at capturing local spatial patterns, they may struggle with long-range dependencies in large-scale remote sensing imagery. This is where MoE offers a further boost: by partitioning the feature space among specialized expert networks, MoE classifiers can handle a wider variety of inputs more effectively.

3.1.1 Land Use and Land Cover Mapping

In Earth observation, the MoE paradigm has been used mainly for pixel-level labeling tasks, including semantic segmentation, land-cover and land-use (LULC) mapping, and other forms of pixel-wise classification. The central idea is to let multiple experts specialize in different features (e.g., sensor types, spatial scales, or scene conditions) and then combine their outputs through a gating mechanism to obtain more accurate and robust pixel predictions.

Early studies already explored MoE-like designs for geospatial pixel labeling. Kussul et al. [100] employed an ensemble of neural-network experts to produce large-scale land-cover maps. This model first clusters and denoises multi-sensor time series using self-organizing maps, and then feeds each partition into an ensemble of multilayer perceptrons for classification and optical-radar fusion, generating 30-m LULC maps for Ukraine from 1990 to 2010 and 2015. Yüksel et al. [101] considered landmine detection in ground-penetrating radar, modeling each radar trace as a sequence to be classified. They introduced a MoE framework using Hidden Markov Model experts, where each expert specializes in a particular landmine signature or soil condition, and a gating scheme combines their probabilistic outputs to distinguish landmine responses from clutter. Nagarajan et al. [102] applied an MoE approach to multiscale segmentation of elevation data, e.g. Light Detection and Ranging (LiDAR)-derived Digital Elevation Models (DEMs). In their framework, experts operate at different spatial scales to segment terrain features, and a gating function merges these expert segmentations into a final elevation segmentation map. Taken together, these works show that assigning pixel- or sample-level decisions to multiple specialized experts can improve robustness across sensors, acquisition conditions, and spatial scales.

More recent DL–based MoE architectures adopt the same principle for high-resolution remote sensing imagery. Chen et al. [103] proposed SparseFormer, a dual-CNN-expert-guided transformer for semantic segmentation. The model comprises three branches, where two CNN branches employ different attention mechanisms: convolutional block attention module and coordinate attention, to encourage diverse outputs and extract complementary features, including fine-grained spatial details and global context. A credible assessment mechanism combines the CNN outputs to produce high-quality pseudolabels that supervise a CNN-Transformer hybrid branch, which integrates global representations with local features for precise segmentation. On the Zurich Summer dataset, SparseFormer [103] reports mF1 and mIoU scores of 75.07% and 64.85%, respectively, with overall accuracy exceeding other weak supervision approaches, confirming the effectiveness of the dual-CNN-expert design for sparse annotation scenarios. Sun et al. [104] extended the idea to panoptic segmentation, jointly modeling object instances and semantic regions in remote sensing images. Their multi-scale feature interaction and adaptive experts framework introduces an adaptive disturbance sparse MoE module based on Transformer, where adaptive noise is introduced to disturb expert selection, enhancing randomness and exploration to improve generalization and robustness while reducing computational overhead. He et al. [105] further presented a resource-efficient MoE-based semantic segmentation network for remote sensing images, illustrating that expert specialization can also be leveraged under tight computational constraints. Ren et al. [106] proposed UNetMoE, a MoE-based semantic segmentation model for remote sensing images. The model employs category-specific expert submodels, where each expert is trained to perform binary classification for a single land-cover category, and a gating system based on ResNet with self-attention mechanisms dynamically allocates weights to combine expert outputs. This category-specific expert design enables targeted feature extraction for each land-cover type, improving per-class recognition precision while the gating system synthesizes outputs from all experts to optimize overall segmentation accuracy. Complementing these designs, Chen et al. [107] introduce Land-MoE, a frequency-aware mixture of low-rank [75] token experts used as parameter-efficient adapters on top of vision foundation models to improve multispectral land-cover classification under cross-sensor and cross-geospatial domain shifts, showing that token-level MoE adapters and shared frequency-aware filters can substantially enhance the generalization of LULC maps across sensors and regions. Building on this idea of region-aware specialization, Lee et al. [108] proposed GRAM, a region-aware MoE framework for slum segmentation that learns city-specific expert adapters on a large multi-city satellite dataset while a shared backbone captures universal informal-settlement morphology, and uses a region classifier plus cross-expert prediction consistency at test time to select reliable pseudo-labels for self-training, thereby improving the generalization of slum maps to previously unseen cities.

Beyond classification, MoE has also been used for specific pixel-wise segmentation applications. For building footprint extraction, Xu et al. [109] proposed MSLoRA-Net, which integrates a LoRA-based [75] MoE into a segmentation model. Low-rank adaptation experts are injected into the layers of a Vision Transformer [31], and a routing network decides which LoRA [75] expert weights to apply for each input patch. This effectively fine-tunes a large pre-trained model to specialize in building structures, enabling dynamic selection of adaptation parameters for different building shapes and sizes. In agricultural mapping, Lu et al. [44] developed an uncertainty-aware difficulty-based MoE framework (DMoE-ViT) for long-tail crop type mapping, where an autoencoder first estimates sample difficulty via reconstruction loss, crops are stratified into easy, moderate and hard subsets, and three Vision Transformer [31] experts are trained on different difficulty levels and fused by a gating network with evidence-based uncertainty weighting; experiments on heterogeneous agricultural regions show that this design improves classification accuracy and robustness for rare and hard crop samples compared with single-backbone CNN [11], LSTM [158], UNet [14] and ViT [31] baselines. In agricultural mapping, Li et al. [110] introduced U-MoE Mamba [14, 159, 160], a hybrid expert segmentation model for segmenting crop fields (specifically cabbage) from high-resolution drone imagery. The model integrates three expert paradigms, i.e., multi-scale convolution, attention mechanisms, and Mamba pathways through a lightweight gating network that dynamically combines expert outputs for adaptive feature aggregation. By leveraging expert diversity, the model better accommodates seasonal and appearance variations in crops than single-backbone approaches, leading to improved delineation of field boundaries across different growth stages.

3.1.2 Scene Classification

Beyond pixel-level labeling, MoE has also been widely explored for scene-level classification and object recognition in remote sensing. In this setting, multiple experts are typically designed to handle different levels of difficulty, class distributions, sensor modalities, or deployment environments, and a gating mechanism aggregates their predictions into a final scene label.

Xie et al. [111] introduced a stacked MoE network for fast aerial scene classification. In their design, multiple expert subnetworks are organized in a stacked architecture, where later experts refine the representations and predictions produced by earlier ones. By reusing shared features across stages and concentrating expert capacity on more discriminative processing in deeper layers, the model achieves a favorable trade-off between classification accuracy and computational cost for large-scale aerial scene classification. A related idea appears in wildlife monitoring: Mensah et al. [112] proposed an MoE-based mobile vision transformer for fine-grained bird species classification on edge devices. Their model modifies MobileViTV2 [161] to include patch-level MoE layers, where a router clusters patch embeddings and routes each image patch to a small subset of transformer experts, enabling conditional computation within a single network. Together, these approaches illustrate how MoE can trade off accuracy and efficiency by routing different parts of the input to specialized experts under resource and deployment constraints.

MoE-style expert frameworks have also been explored to alleviate class imbalance and hard categories in remote sensing recognition. Wang et al. [113] proposed DMRS, a diversity-oriented expert framework for long-tailed remote sensing scene recognition, where some categories appear much

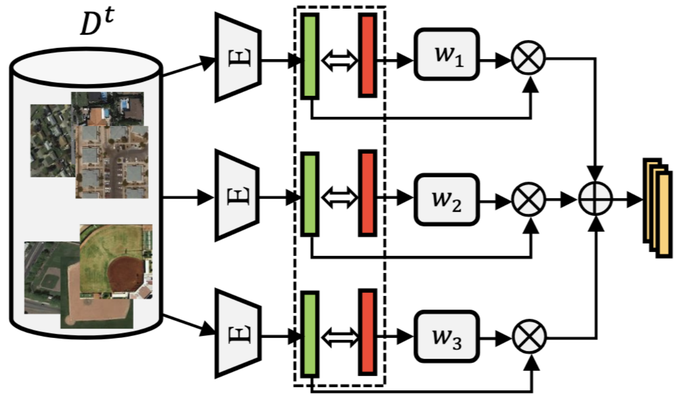

Figure 4. MoE in adaptive mixture-of-experts distillation (AMoED) [114] for cross-satellite generalizable incremental scene classification.

less frequently than others. DMRS builds multiple diversity experts together with a semantic-aware mixing strategy that exploits category-level semantic information to combine their outputs. By encouraging different experts to capture complementary semantic patterns across head and tail classes, the method reduces the bias toward frequent categories and markedly improves the recognition performance of rare classes compared with single-backbone baselines. On the NWPU-RESISC45 [162] dataset, DMRS [113] reports 88.3% overall accuracy, exceeding the next best method (MDCS [156] at 81.6%) by 6.7 percentage points, with tail class performance reaching 84.2%, substantially higher than traditional methods that typically achieve below 65%, highlighting the effectiveness of diversity experts for long-tailed recognition.

To address domain shifts and multi-source data variability, MoE has also been applied in cross-domain adaptation scenarios. Fu et al. [114] propose AMoED, an adaptive MoE distillation framework for cross-satellite generalizable incremental scene classification, where multiple domain-specific experts trained on different source domains provide coordinated guidance through a high-level semantic learning pipeline. The expert predictions are adaptively integrated based on domain-agnostic confidence measures to form universal class concepts, enabling stable knowledge acquisition without direct exposure to raw data streams. As shown in Figure 4, the framework trains a set of domain-specific experts $E = \{E_m\}_{m=1}^{M}$ independently on data from M distinct source domains, each specializing in recognizing newly emerged classes within its respective domain. The cross-domain classification confidence of each expert is then evaluated in a domain-agnostic manner, and the expert knowledge is adaptively mixed based on these confidence measures. An equi-partite subset is constructed by combining exemplars of previously learned classes with equally sampled instances of new classes from each source domain. Based on this subset, generalizable knowledge is acquired under the coordinated guidance of the mixed expert predictions through knowledge distillation, while knowledge consolidation is performed via class label supervision. Throughout the training process, a shallow style-mixing operation is applied to reduce geospatial and sensor-induced deviations, effectively mitigating domain shift and catastrophic forgetting across satellites. Guo et al. [115] addressed radar target recognition by fusing multiple radar modalities through a confidence fusion framework with representation distribution modeling and single-modal mixture of experts (sMoE). The sMoE structure employs a router to dynamically select appropriate feedforward network experts for processing different input tokens, while the representation distribution module extends point embeddings to distributional embeddings that capture uncertainty. The confidence fusion module then weights single-modal predictions based on relative confidence levels derived from evidential learning, yielding improved target classification accuracy over single-model baselines. These approaches show that MoE can handle the spectral, modal, and distributional variability inherent in multi-source remote sensing data by assigning domains or modalities to specialized experts.

3.1.3 *Hyperspectral Image Classification*

Hyperspectral image (HSI) classification by distinguishing materials or land-cover types from high-dimensional spectral data has greatly benefited from MoE strategies as well. Hyperspectral data is characterized by hundreds of spectral bands, and different subsets of bands often carry complementary information. Several works therefore use experts to focus on different spectral or spatial attributes. HyperTransXNet [117] is a representative HSI classification model that employs a dual-branch Transformer [15] with dynamic token mixers, effectively functioning as experts specializing in global vs. local spectral pattern. By learning separate expert transformers for broad spectral trends and fine-grained local variations, and then adaptively merging their outputs, this approach captures both global and local spectral dynamics to improve accuracy in HSI classification. In another study, Xu et

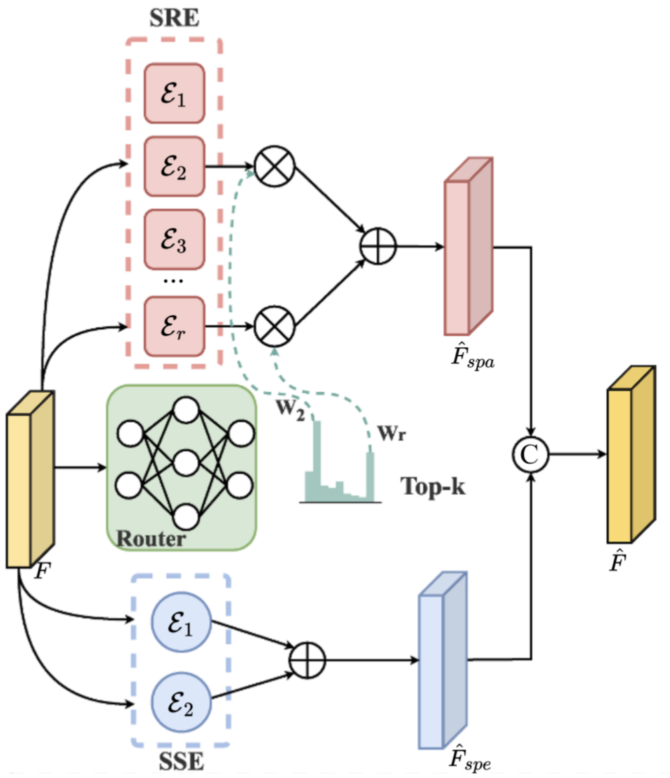

Figure 5. MoE in mixture-of-spectral-spatial-experts state space model (MambaMoE) [116] for hyperspectral image classification.

al. [116] developed MambaMoE, a MoE state-space model [159, 160] for HSI classification. As illustrated in Figure 5, MambaMoE employs a mixture of Mamba expert block (MoMEB) that splits input features F into spatial and spectral views. The spatial routed expert module contains multiple Mamba experts configured with distinct spatial scanning directions, where a router network dynamically selects the top-k experts and assigns routing weights to adaptively capture directional spatial context, producing spatially refined features $\hat{F}_{\text{spa}}$. The spectral shared expert module employs two spectral-directional Mamba branches (forward and backward) to extract shared spectral representations, generating $\hat{F}_{\text{spe}}$. The spatial and spectral features are then concatenated and integrated through a 1×1 convolution to form the final representation $\hat{F}$. This dual-branch design enables adaptive extraction of spectral-spatial joint features tailored to diverse land-cover characteristics. Both HyperTransXNet [117] and MambaMoE [116] showed that dividing the spectral feature space among specialized experts can significantly boost classification performance on HSI benchmarks compared to conventional single-expert models. On the Pavia University dataset, MambaMoE [116] yields a 3.67% improvement in baseline overall accuracy (OA), achieving 95.20% OA, demonstrating the effectiveness of this spectral-spatial expert decomposition approach. MixtureRS by Liu et al. [41] integrates LiDAR and hyperspectral imagery for land-cover classification using a multimodal MoE design. In MixtureRS [41], heterogeneous convolutional networks extract spectral-spatial features from hyperspectral data and elevation attributes from LiDAR data, which are then tokenized and processed through a cross-modality transformer encoder. The encoder incorporates multi-head cross-attention for inter-modal interaction and a sparse MoE feed-forward layer that selectively activates the most relevant experts for each token via top-k routing. This mixture-of-modalities approach achieved superior land classification accuracy by leveraging complementary spectral and 3D structural information from the two sensors, especially for complex landscapes. Notably, MixtureRS [41] highlighted that naive fusion can struggle when data are not perfectly co-registered, which can be mitigated by expert specializations for each modality. On a 15-class urban benchmark, MixtureRS [41] reports an overall accuracy of 88.64%, an average accuracy of 90.23%, and a Cohen's Kappa of 0.8767, exceeding the best homogeneous transformer by over 12 percentage points, demonstrating the advantage of modality-specific expert specialization. Beyond these architectures, Gao et al. [118] proposed a Missing-aware Mixture-of-Loras (MaMOL) framework, which treats multimodal, hyperspectral-based land-cover classification with incomplete modalities as an expert-selection problem rather than explicit modality reconstruction. MaMOL inserts lightweight LoRA-based [75] dynamic and static experts into a frozen Transformer [31] backbone and uses a dual routing strategy (task-aware dynamic pattern experts plus shared and modality-specific static experts) to adapt to arbitrary hyperspectral, Synthetic Aperture Radar (SAR), and Light Detection and Ranging (LiDAR) missing-patterns, achieving strong robustness and accuracy under high missing rates on benchmark datasets.

3.2 Object Detection

Object detection in remote sensing involves identifying and locating objects of interest (such as vehicles, buildings, ships, or airplanes) in aerial or satellite images. The task is challenging because remote sensing images are often high-resolution with small objects, complex backgrounds, and varied object orientations and scales. MoE models have started to be explored as a way to enhance detection by providing adaptable feature processing pipelines that adjust to different object types or image regions.

In object detection, one common strategy is to use experts at different feature pyramid levels or for different object orientations. Chai et al. [119] developed a scalable MoE attention feature pyramid network (SAFPN) to detect objects in remote sensing images. Each level of the feature pyramid has a group of expert subnetworks that attend to features of a specifical resolution. A gating mechanism then fuses information across levels, ensuring that small and large objects are all well-represented. SAFPN's [119] multi-level expert design results in mean average precision (mAP) improvements for detection and instance segmentation from 71.3% and 62.4% to 82.7% and 71.1%, respectively, on the Airbus Ship dataset, demonstrating the benefit of resolution-specific expert specialization. The substantial mAP gains highlight the effectiveness of multi-level expert architectures for object detection tasks. Similarly, Chen et al. [120] proposed FAMHE-Net for detecting oriented objects (such as rotated cars or ships) using a mixture of heterogeneous experts. FAMHE-Net [120] includes multiple expert detectors, each specializing in a particular object orientation or aspect ratio, and a gating network that adaptively combines their outputs. By augmenting multi-scale feature maps and letting experts handle different rotation angles, this model significantly improved detection accuracy for oriented bounding boxes compared to single-expert detectors. In object detection pipelines, MoE has also been integrated into the region proposal and classification stages. Lin et al. [121] introduced a multiple region proposal experts network for detecting small objects over large areas. They trained a committee of Region Proposal Networks (RPNs) as experts, where each expert focuses on proposals in a specific image region or for specific object sizes. A gating module then merges the proposals, and subsequent classification is done by an ensemble of expert classifiers as well. The *multiple experts at multiple stages* approach helped detect objects over wide scenes (e.g., detecting all cars in a city-scale image) by dividing the task among specialized RPN experts and classification experts, resulting in more reliable detection in wide-area images with experts covering different sub-scenes and object scales.

For cross-domain and multimodal object detection, MoE has also been leveraged to improve performance. Li et al. [122] presented SM3Det, a unified model for multi-modal remote sensing object detection that uses a grid-level sparse MoE backbone. Unlike traditional approaches that route entire images to modality-specific experts, SM3Det [122] operates at the feature grid level, enabling experts to process local spatial features adaptively. As shown in Figure 6, multi-modal feature maps are first divided into grids and dispatched by a dispatcher D to different experts based on local feature characteristics, and a collector C then reassembles the processed grids so that each spatial region benefits from the expert best suited to its local patterns and appearance. This grid-level routing mechanism allows SM3Det [122] to capture both shared knowledge and modality-specific representations simultaneously, as experts can specialize in different local patterns across modalities while still learning common spatial structures. The model integrates a dynamic learning rate adjustment strategy to handle varying learning difficulties across different modalities and tasks. By leveraging grid-level MoE, SM3Det [122] achieved strong results on detecting objects across SAR, optical, and infrared modalities, highlighting MoE's utility in unified multi-modal object detection. In a related vein, Lin et al. [123] developed MEDNet, a multi-expert detection network that addresses the challenge of leveraging diverse distinctive information for remote sensing object detection. MEDNet [123] employs multiple feature pyramids (MFPs) and multiple detection experts (MDEs), where a loss distance-based k-experts clustering (LD-kEC) strategy dynamically assigns training samples to different detection experts in an unsupervised manner based on loss distances. This strategy allows each expert to specialize in detecting objects with similar characteristics (e.g., appearance, internal texture, or external context) without requiring manual expert labels. Such specialized experts significantly improved detection performance compared to a single detection pipeline approach.

Beyond traditional objects, MoE-based detection extends to forgery detection in remote sensing imagery and anomalous target localization in marine monitoring. Zhang et al. [124] tackled the problem of copy-move forgery understanding in satellite images using a multimodal gated MoE model (CM-MMoE) for the remote sensing copy-move question answering task. In such forgeries, a region of an image is duplicated elsewhere to conceal information, e.g., cloning part of a satellite image to cover something. CM-MMoE employs

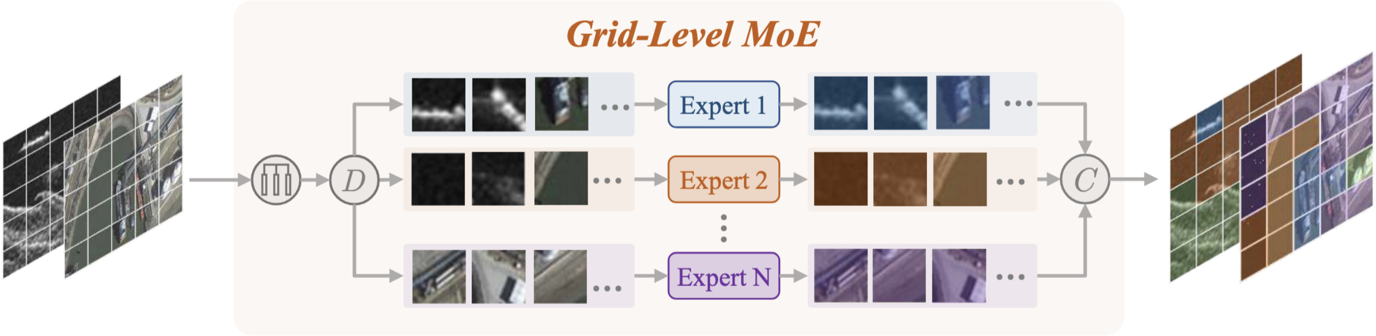

Figure 6. Grid-level MoE backbone used in single model for multi-modal datasets and multi-task object detection (SM3Det) [122] for multi-modal remote sensing object detection.

hierarchical visual representations that include the source region, tampered region, background, and original image features, which are integrated with textual question features. A multimodal auxiliary gating network, guided by cross-attention between visual and textual features, dynamically routes inputs to multiple forgery expert networks. Each expert processes the hierarchical visual features to achieve multi-level understanding of image semantics, enabling accurate answers to diverse questions about tampering scenarios. This multi-modal MoE approach outperformed single-expert methods in identifying subtle forgeries, underlining MoE's potential in remote sensing image forensics. Qian et al. [125] employed a multi-task, multi-expert and multi-gate (MEG) framework for underwater target detection and localization using acoustic signals. Their approach uses multiple expert networks with independent parameter spaces to specialize in different aspects of underwater acoustic signal processing. Unlike traditional multi-task learning with shared parameters, MEG employs multiple gating layers, where each task (recognition and localization) has its own gating network that dynamically learns task-specific weights to linearly combine expert outputs. This multi-gate design allows each task to obtain task-specific representations by adaptively selecting and weighting experts, while the top-k gating mechanism improves efficiency by activating only the most relevant experts. This multi-task MoE strategy improved underwater object detection performance by enabling specialized feature learning for classification and localization tasks while maintaining computational efficiency.

3.3 Spatiotemporal Modeling and Analysis

Many remote sensing applications involve analyzing how the Earth changes over time. Temporal analysis of remote sensing data may involve modeling time series of images to detect trends, seasonal variations, or anomalies. These tasks present unique challenges: one must account for differences in imaging conditions, align multi-temporal data, and distinguish meaningful changes (like deforestation or urban growth) from irrelevant ones (like seasonal vegetation cycles or shadows moving). MoE models can assist by allocating different experts to different temporal contexts or change types, and by handling multi-modal temporal input.

3.3.1 Change Detection

Change detection typically uses imagery from two or more time points to identify what has changed (for example, urban expansion, deforestation, or disaster impact). MoE models can contribute here by providing experts that specialize in particular change patterns or sensor modalities, which is particularly useful when dealing with multi-temporal and multi-source data.

Liu et al. [43] proposed M^{2}CD, a unified multimodal framework for optical-SAR change detection that leverages MoE modules integrated into the backbone network. Optical (e.g., visible spectrum) and SAR imagery have very different characteristics, and changes (like urban development or flooding) might appear differently in each modality. In M^{2}CD [43], MoE layers are inserted after each backbone block, where multiple experts adaptively handle images from different temporal phases through a gating function that selects the top-k experts based on similarity scores between input features and expert embeddings. The framework also introduces an optical-to-SAR path (O2SP) that generates simulated SAR images from pre-event optical images using the fully developed speckle assumption, serving as an intermediate representation to bridge the two modalities. During training, self-distillation is applied to minimize the feature space discrepancy between the optical path

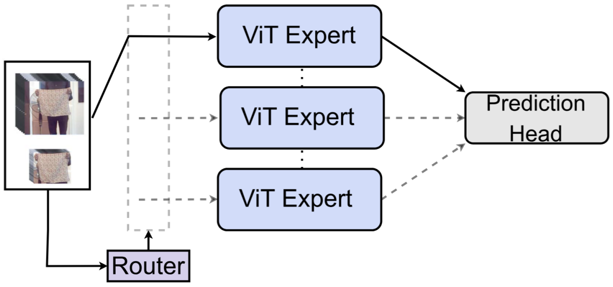

Figure 7. Sparse mixture-of-experts for hyperspectral object tracking (HotMoE) [127] framework for hyperspectral object tracking.

(OP), O2SP, and SAR path (SP), while O2SP is omitted during inference to avoid additional computational overhead. This mixture-of-experts approach addresses the problem of heterogeneous change detection, where one modality might be affected by season or weather (optical) while the other might suffer from speckle or different imaging geometry (SAR). By creating a sparser feature space that enables distinct representation learning for optical and SAR modalities, M^2CD [43] achieved more accurate change detection on optical and SAR datasets than methods that simply concatenate or transform one modality into the other, showing gains of 0.15% in OA, 0.36% in mF1, and 0.57% in mIoU compared to TTP, validating the effectiveness of modality-adaptive expert routing for cross-modal change detection. Another recent work by Seydi et al. [48] focused on detecting environmental changes due to disasters, specifically burned area mapping after wildfires, using a Siamese-based mixture of experts (SMoE) framework. They developed a deep siamese network that takes pre- and post-fire multispectral Sentinel-2 images as input through two deep feature extractor channels. The framework employs MoE layers where each expert is implemented as a convolution layer, and a gating network dynamically routes inputs to the most relevant experts based on local efficiency. The extracted features are further processed through position and channel attention modules before being fed into a dense-MoE layer for final classification. This approach yielded highly accurate burned area maps, as the mixture of experts could adaptively process different aspects of the bi-temporal data while the attention mechanisms enhanced feature representation. The MoE-based siamese model outperformed conventional change detectors by reducing false alarms and missed detections, especially in heterogeneous landscapes with complex backgrounds.

3.3.2 *Time-Series Analysis*

Remote sensing often involves time series analysis, such as predicting environmental variables over time or tracking moving objects across image sequences. While many MoE applications in remote sensing have focused on static imagery, a few have tackled challenges in the temporal dimension by exploiting MoE's ability to assign experts to different temporal patterns or tasks. A notable example is in climate and weather-related predictions. Jiang et al. [126] proposed a knowledge-guided adaptive MoE model for precipitation prediction. This model processes multi-source time-series data including meteorological satellite observations, reanalysis data, and ground sensor readings. The framework employs 16 independent MLP experts, with features organized into six physically meaningful categories based on domain knowledge: Momentum, Temperature, Moisture, Mass, Cloud, and Radiation. During training, a selective training strategy is applied where only two randomly selected experts are updated per epoch while others remain frozen, encouraging diversity through a custom loss function that penalizes weight similarity between active experts. A dynamic router learns to assign weights to expert outputs, computing final predictions as weighted sums of individual expert predictions. This knowledge-guided approach enables each expert to specialize in coherent subsets of climate features, leading to improved predictive accuracy and interpretability compared to baseline models. This demonstrates MoE's promise for complex time-series forecasting tasks in Earth science, where different dynamics govern different times or places. Another domain of temporal analysis is object tracking in aerial videos or sequential images. Sun et al. [127] introduced HotMoE, which explores a sparse MoE architecture for hyperspectral object tracking. As illustrated in Figure 7, HotMoE processes hyperspectral image inputs (template and search region) through a router module that dynamically selects the most suitable ViT expert from a pool of multiple experts. The router analyzes the input characteristics and routes all tokens to a single selected expert using a hard routing strategy, where only one expert is activated per inference. Each ViT expert consists of multiple transformer encoder layers with multi-head attention mechanisms, enabling specialized processing for different tracking scenarios. The outputs from the selected expert are then fed into a prediction head to generate the final tracking results. By sparsely activating only one expert per inference instead of computing all experts, HotMoE [127] achieved efficient and robust tracking performance on hyperspectral video data, achieving 43.7 FPS with an AUC of 0.704 on the HOT2022 dataset. This approach illustrates how MoE can effectively handle high-dimensional hyperspectral data while maintaining computational efficiency through conditional expert activation.

Some recent foundation models hint at incorporating temporal expertise. Albughdadi et al. [128], for example, in their MoE-MAE pretraining incorporated geo-temporal conditioning by encoding latitude, longitude, week-of-year, and hour-of-day as sinusoidal pairs to preserve cyclic structure. These metadata tokens are concatenated with patch embeddings and processed through MoE layers with NoisyTop-k routing, enabling the model to exploit spatio-temporal regularities inherent in Earth observation data. While that work is primarily about representation learning, it suggests a future direction where MoE experts could specialize not just on modalities, but on temporal segments, e.g., an expert for detecting changes in summer vs winter imagery. Time series analysis in remote sensing can also include phenological trend analysis, anomaly detection over time, and data assimilation. While not many works explicitly use MoE for these yet, the potential is clear. For example, an MoE could be devised for crop monitoring where one expert handles normal seasonal growth curves and another detects deviations (drought stress signals), with gating based on the current growth stage. In data assimilation for climate models, one could imagine experts specialized in different regimes (e.g., monsoon vs arid climate assimilation) being mixed. Li et al. [129] proposed STF-MoE, a spatio–temporal fusion MoE that couples an LSTM-Transformer [15, 158] hybrid framework with a heterogeneous mixture-of-experts mechanism. The model employs two parallel branches: a Transformer branch with multi-head self-attention to capture long-range temporal dependencies, and a bidirectional LSTM branch to extract local contextual features. The concatenated outputs from both branches are dynamically routed through an adaptive gating network to five structurally heterogeneous expert networks, with Top-2 sparse activation for computational efficiency. The model integrates multi-source remote sensing features (e.g., near-infrared reflectance vegetation index NIRv, fraction of photosynthetically active radiation absorption Fpar) and environmental variables (e.g., relative humidity, digital elevation model) for county-level wheat yield estimation across six major Chinese provinces, achieving $R^2 = 0.827$ and RMSE = 547.7 kg/ha in the most recent estimation year. This confirms that expert-based fusion can benefit quantitative agro-ecosystem prediction tasks built on multi-source remote sensing data.

In time-series and spatio-temporal remote sensing tasks, MoE applications show considerable promise. By employing experts that specialize in different temporal patterns or tracking subtasks, these models can address the high variability over time in Earth observation data. The precipitation prediction MoE [126] achieved improved accuracy and interpretability through knowledge-guided feature grouping that organizes climate variables into physically meaningful categories (e.g., Momentum, Temperature, Moisture, Mass, Cloud, Radiation), enabling each expert to specialize in coherent feature subsets. The hyperspectral tracker [127] enhanced robustness by dynamically routing inputs to specialized experts that handle distinct data distributions and address different tracking challenges, allowing the model to adapt seamlessly to various scenarios. These results indicate that MoEs can play a valuable role in temporal remote sensing analysis where different time-dependent processes need to be modeled concurrently.

3.4 Multi-Modal Fusion and Adaptation

Remote sensing often involves multi-modal data fusion by combining information from different sensors (e.g., optical, radar, LiDAR, multispectral), as well as adapting models across different data domains (e.g., different satellites or geographic regions). MoE models are naturally well-suited for these challenges, as they can assign dedicated experts to each modality or domain and learn how to best integrate them. A surge of recent work has applied MoE to multi-modal and cross-domain problems in remote sensing.

3.4.1 Multi-Modal Foundation Models

A clear example is in the construction of large remote sensing foundation models that incorporate multiple data modalities. Bi et al. [47] introduced RingMoE [47], a MoE-based modality experts model with an enormous 14.7 billion parameters, designed as a multi-modal foundation model for universal remote sensing image interpretation. RingMoE [47] employs a hierarchical MoE architecture called RMoE (Ring MoE) that incorporates three specialized expert types: modal-specialized experts that capture fine-grained intra-modal representations for each modality

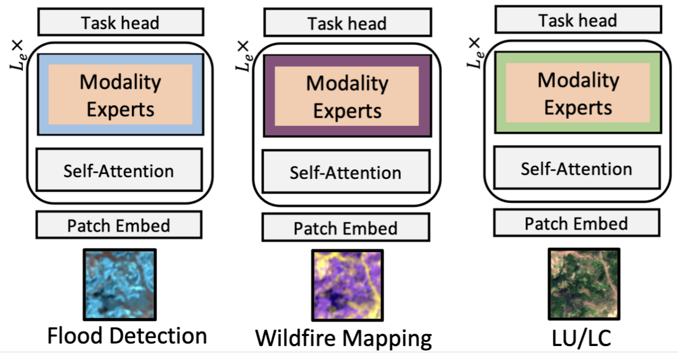

Figure 8. Modality-aware pruning of experts (MAPEX) [131] for multi-modal remote sensing foundation models.

(optical, SAR, multispectral, elevation), collaborative experts that model inter-modal correlations across related modalities, and a shared expert that distills common knowledge across all modalities. During pretraining on massive datasets comprising 400 million multi-modal images from nine satellites, modality-specific routing networks in RingMoE [47] learn to route tokens from each modality to its corresponding modal-specialized experts, while collaborative experts enable knowledge transfer between related modalities. This enables the model to capture a rich representation that spans visible, infrared, and radar signatures of Earth surfaces. RingMoE [47] achieved state-of-the-art results on a variety of downstream tasks (classification, segmentation, etc.) by virtue of its ability to flexibly combine modalities via expert specializations. Its success demonstrates how MoE is a powerful paradigm for scaling up multi-modal integration in remote sensing, ensuring that each data source is handled by networks optimized for its characteristics while still contributing to a unified representation. Similarly, Zhang et al. [130] developed SkySense V2, a unified foundation model for multi-modal remote sensing that employs a single transformer backbone to handle multiple modalities. SkySense V2 [130] integrates MoE modules into the last L transformer blocks, replacing the original feed-forward network layers with M experts and a learnable gating network that performs top-k expert selection. This design enables the model to scale up capacity efficiently while maintaining computational efficiency through sparse activation. This reflects a trend of expert pruning and selection in multi-modal MoE models: Hanna et al.'s [131] MAPEX specifically addresses this by performing modality-aware pruning of experts. As illustrated in Figure 8, MAPEX processes input images through patch embedding, self-attention layers with modality experts (where MoE replaces the feed-forward network layers), and task-specific heads for each downstream task (e.g., flood detection, wildfire mapping, land-use/land-cover mapping). MAPEX [131] employs a modality-conditioned token routing mechanism that uses learnable modality embeddings to route tokens from each modality to the same subset of experts, ensuring consistent expert-modality relationships. After pre-training, experts corresponding to unavailable modalities are pruned, and only the top-k experts for the downstream task modalities are retained, creating specialized models that are significantly smaller and easier to fine-tune. This kind of flexibility maintains many experts during training for generality, but selects a subset for a specific use, illustrating how MoE models can be both extensible and adaptable in multi-modal settings.

3.4.2 *Vision-Language Models*

Text is a special modality in remote sensing vision-language systems: unlike images or sensor signals, it typically appears as human-readable descriptions, questions, or ancillary metadata (e.g., scene tags, land-use labels, or reports), and encodes high-level semantics, reasoning cues, and task instructions rather than raw physical measurements. When combined with imagery, textual inputs enable more interactive forms of analysis such as querying, explanation, and semantic retrieval, but also introduce strong modality heterogeneity and require models that can align visual and linguistic information in a structured way. Cross-modal MoE designs are also prevalent in Visual Question Answering (VQA) and vision-language tasks for remote sensing. Liu et al. [132] proposed a unified transformer with cross-modal MoE for remote sensing VQA. In their model, cross-modal MoE experts (CMMEs) incorporate visual and textual experts that replace conventional feed-forward networks, with shared self-attention and cross-modal attention layers to capture intricate interactions between visual and language features. The modality experts generate fused features through cross-modal interaction, and their outputs are concatenated for answer prediction. This design improved performance on remote sensing visual question answering benchmarks by effectively modeling cross-modal attention and capturing complex semantic relationships between questions and images, rather than attempting to cram all modalities into one latent space. The effectiveness of these systems demonstrates the potential of MoE for interactive remote sensing analysis. Similarly, Lin et al. [45] and Liu & Lian [46] developed large-scale vision-language models for remote sensing (RS-MoE and RS-UniVLM, respectively) that use MoE layers to handle diverse inputs like captions, scene descriptions, and images. RS-MoE [45] employs an instruction router that dynamically generates task-specific prompts and multiple lightweight Large Language Models (LLMs) as expert models, where each expert focuses on distinct aspects of captioning: theme comprehension, object recognition, and relationship inference. On the RSIEval [45] dataset, RS-MoE-7B [45] leads other models across all evaluation metrics, while the lightweight RS-MoE-1B [45] variant surpasses BLIP2-13B [157] in most of the criterion scores, demonstrating remarkable efficiency of the instruction routing mechanism. RSUniVLM [46] introduces a granularity-oriented MoE (G-MoE) architecture with three experts specialized for different visual granularity levels: image-level expert for holistic understanding, region-level expert for localized patterns, and pixel-level expert for fine-grained semantic information. Building on the same granularity-aware specialization method, SkyMoE [133] further argues that a key bottleneck of remote sensing vision-language models is the persistent tension between global context reliance and local detail discrimination, and addresses it with an adaptive router that generates task- and granularity-aware routing instructions to activate specialized LLM experts. The development of these models represents a significant advancement in making remote sensing data more accessible through natural language interfaces. By leveraging LLM capabilities for natural language understanding, these models enable more intuitive interaction with remote sensing imagery. To explicitly encourage expert decoupling across local and global semantics, SkyMoE additionally introduces a context-disentangled augmentation strategy that forms contrastive local/global training pairs, and benchmarks generalization under multi-task, multi-granularity settings via MGRS-Bench. These vision-language MoEs are enabling more semantic-level interpretation of remote sensing data, moving beyond pure pixel classification to answering complex queries about images.

3.4.3 *Cross-Domain Adaptation*

Another area where MoE shines is cross-domain adaptation by ensuring models work well on data from different sources or distributions. We discussed how Land-MoE improved spectral domain generalization in classification [107] and how Fu et al. [114] and Lin et al. [123] approached domain adaptation via experts. Ngo et al. [134] also explored a related concept: using multiple experts for knowledge adaptation across multiple sources. In their approach, a shared feature extractor is combined with multiple domain-specific classifiers, where each classifier is trained on a particular source domain to specialize in that domain's characteristics. These domain-specific experts provide different views on the target domain, and collaborative learning is employed to connect these views by leveraging consistency regularization, enabling the experts to teach each other and recover missing label information. This collaboration of multiple experts was shown to yield better adaptation than single-source transfer, as each expert captured unique features of its source domain which, when combined through collaborative learning, provided a richer representation for the target. Zhang et al. [135] addressed domain generalization in the spectral domain with SpectralX, introducing parameter-efficient MoE fine-tunings for spectral shifts. SpectralX employs an attribute-oriented mixture of adapter that consists of an attribute-specific router bank and an attribute-shared adapter bank, where routing schemes dynamically allocate spatial and spectral attribute knowledge to different adapters. Rather than assigning experts to specific sensors, the model routes tokens based on their spatial and spectral attributes, enabling effective adaptation to diverse spectral imagery from different regions or seasons without retraining the whole foundation model. This demonstrates MoE's strength in capturing domain-specific nuances through attribute-oriented routing and applying them selectively.

3.4.4 *Multi-Sensor Data Fusion*

Historical uses of MoE in multi-sensor fusion provide important context. One of the earliest examples, Pasika et al. [136], applied multiple neural network methods (RBF, SVM, NDEKF-MLP, and backpropagation) to fuse diverse sensor measurements (including multispectral satellite data, temperature, and humidity) for cloud base height prediction. They demonstrated that combining information from multiple sensors could improve prediction accuracy, laying the foundation for later MoE-based fusion approaches. Aggarwal et al. [137] proposed a multiscale data fusion method regulated by a MoE network. Their approach employed multiple multiscale Kalman filters as experts, each with different parameter vectors, to estimate topography from InSAR and ALSM data. A gating network selected the most appropriate expert for each input, effectively adapting to non-stationary terrain variations. These pioneering efforts showed that adaptive expert selection could improve fusion performance across diverse data characteristics.

For multi-sensor data fusion in classification and mapping tasks, MoE approaches have also been key. Liu et al. [41] proposed MixtureRS, which integrates hyperspectral and LiDAR data through a cross-modality transformer enhanced with sparse MoE layers. The MoE layers replace conventional dense feed-forward blocks in the transformer, using Top-k routing to selectively activate the most relevant experts for each token, thereby improving model capacity without significant computational overhead. Kong et al. [138] tackled joint classification of hyperspectral and LiDAR data using a dual MoE framework. They designed a mixture of multimodal fusion experts mechanism, where multiple fusion experts correspond to different fusion strategies, and a gating network selects and mixes these experts to achieve diverse feature fusion. The gating mechanism learns to adaptively weight different fusion strategies based on the input characteristics, enabling effective complementary learning between hyperspectral and LiDAR modalities. He et al. [139] extended this work with an adaptive expert learning framework for HSI and MSI fusion. They introduced a modality-guided complementary module to establish bidirectional cross-attention pathways between HSI and MSI features, followed by an attribute-aware mixture of fusion experts module. In AMoFE, the fused features are decomposed into spectral, spatial, and edge attribute subspaces, each modeled by a dedicated expert network; a soft routing mechanism then dynamically adjusts each expert's contribution based on contextual cues (e.g., regional texture complexity). This design significantly improved fusion quality by ensuring that spectral details, spatial textures, and object boundaries were each given due attention by the respective experts, rather than forcing a single model to learn all at once.

3.5 Image Restoration and Enhancement

Beyond classification and detection, MoE models can also be applied into image restoration and enhancement tasks in remote sensing. Such tasks include super-resolution (enhancing image spatial resolution), dehazing or denoising (improving image clarity), pansharpening (fusion of panchromatic and multispectral images), and general image quality improvement under various degradations. The rationale for using MoE is that different experts can be trained to handle different types or levels of degradations, and a gating network can then apply the appropriate experts to any given image or region.

Super-resolution (SR) is a critical task for remote sensing, aiming to increase the resolution of satellite imagery. Chen et al. [42] proposed a heterogeneous MoE framework for remote sensing image super-resolution. Their model organizes experts into multiple expert groups, where experts within each group share identical structures while maintaining heterogeneity across groups. They introduced a multi-level feature aggregation strategy that aggregates multi-level features from the backbone network to estimate expert activation probabilities, and a dual-routing mechanism that first selects the most suitable expert group, then determines the optimal expert within the selected group. Experts across different groups utilize convolution kernels of varying sizes, enabling inter-group heterogeneous experts to provide different reconstruction scales while intra-group homogeneous experts offer multiple reconstruction patterns within the same scale. By doing so, the model was able to produce high-resolution images with both crisp details and low noise, outperforming single-expert SR networks especially on diverse terrain types. Rossi et al. [140] took a related approach with Swin2-MoSE, a single-image super-resolution model for remote sensing. They integrated MoE-SM, an enhanced sparsely-gated mixture-of-experts layer, to replace the feed-forward networks inside all Transformer blocks. The MoE-SM is designed with a smart merger layer to merge the output of individual experts, and employs a per-example strategy instead of the commonly used per-token one, where all tokens of an example are processed by the same experts. Swin2-MoSE's [140] per-example routing strategy yields gains of up to 0.377–0.958 dB in Peak Signal-to-Noise Ratio (PSNR) and 0.0006–0.0031 in Structural Similarity Index Measure (SSIM) over any Swin-derived models on tasks of 2×, 3×, and 4× resolution upscaling, demonstrating superior performance especially for more complex tasks. The improvements in both PSNR and SSIM metrics indicate enhanced perceptual quality in the super-resolved images. Image restoration under multiple degradations is another scenario where

MoE shines. Dong et al. [141] introduced PhyDAE, a physics-guided degradation-adaptive expert model for all-in-one remote sensing image restoration. Instead of training separate models for denoising, deblurring, super-resolving, etc., PhyDAE [141] uses a set of physics-aware expert modules, each explicitly designed for a specific degradation type: dehazing, denoising, deblurring, and low-light enhancement. Each expert is rigorously designed based on corresponding physical models (e.g., atmospheric scattering model for dehazing, Retinex theory for low-light enhancement). During training, images with various simulated degradations (noise, haze, blur, low-resolution) are fed, and the gating network learns to route each degraded image to the expert best suited to restore it. Importantly, PhyDAE [141] incorporates physics-based constraints (e.g., sensor noise models, point spread functions) into the experts, guiding them to adhere to known degradation processes. The result is a single MoE model that can restore remote sensing images suffering from different problems. It can clean up a hazy image or sharpen a blurry one or super-resolve a coarse image, by internally selecting the appropriate expert pathway. This approach offers a flexible and efficient alternative to having separate models for each task, and the inclusion of physics knowledge in experts improved the realism and reliability of restorations.

In the realm of atmospheric correction and dehazing, MoE techniques have also been explored. Shen et al. [142] proposed a spatial-frequency adaptive network for remote sensing image dehazing that effectively employs an MoE principle. Their method incorporates a mixture of modulation experts in the spatial domain and a decoupled frequency learning block (DFLB) in the frequency domain. The DFLB employs a dual-branch structure to facilitate independent learning of low-frequency and high-frequency features, where low-frequency features are processed for global haze removal while high-frequency features are enhanced for detail reconstruction. A mixture of fusion experts then adaptively combines the outputs of these frequency-domain branches for the final haze-free image. By doing so, the model can address both the global effects of haze (which require low-frequency corrections) and the local effects (which benefit from high-frequency detail enhancement). This frequency-domain decoupled learning approach achieved clearer and more information-rich results on hazy satellite images than single-technique methods, highlighting how expert specialization in frequency domains is advantageous. Another important enhancement task is pan-sharpening, where a high-resolution panchromatic image is fused with lower-resolution multispectral images to produce a high-res multispectral image. He et al. [143] presented a frequency-adaptive pan-sharpening method using MoE. In their approach, a frequency mask predictor generates adaptive frequency masks that separate the image into high-frequency and low-frequency parts. The frequency experts module employs two MoE components: low-frequency MoE and high-frequency MoE, which exclusively process low-frequency and high-frequency information, respectively. An experts mixture module then dynamically fuses the high-frequency and low-frequency features, as well as PAN and MS features, using multiple fusion experts with adaptive gating. A gating network, guided by frequency analysis of the input, decides how to weight the fusion experts. This ensures that texture-rich areas where high-frequency detail is crucial get more contribution from the high-frequency expert, whereas homogeneous areas rely more on the low-frequency expert. This MoE-based pan-sharpening yielded images with both sharp details and accurate spectral colors, reducing typical artifacts like spectral distortion or ringing, and surpassed traditional pan-sharpening algorithms in evaluations.

In image restoration and enhancement, MoE approaches enable adaptive processing that is well-suited to the varied and complex degradations encountered in remote sensing imagery. By having experts that each excel at certain conditions, e.g., a particular noise level, frequency band, or degradation type, and gating between them, a single MoE model becomes highly versatile. This versatility is crucial for remote sensing, where images can be affected by different atmosphere, sensor, and resolution issues. The success of MoE in super-resolution, dehazing, pan-sharpening, and multi-problem restoration shows that expert specialization can yield higher fidelity outputs than one-fits-all networks, providing users with clearer, more detailed imagery.

3.6 Other Specialized Applications

In addition to the major categories above, MoE models show potential in several other specialized remote sensing applications. These often involve physical modeling, geoscientific data analysis, and novel tasks that benefit from expert decomposition.

One such area is geophysical parameter retrieval using remote sensing data to estimate environmental or geophysical variables via inverse modeling. Loyola et al. [144] provided an early demonstration of MoE in this context, applying neural network MoE to the processing of satellite data for atmospheric parameter retrieval. In their work, the complex inversion problem of predicting parameters like total column ozone from satellite radiances was broken into sub-problems handled by different networks. The input space was divided into three independent regions: aerosol types (maritime and rural), total ozone levels across different latitude bands (low, mid, and high), and satellite viewing angles (normal-view and polar-view). A total of 12 neural networks were combined via a gating network that weighted each expert's contribution based on its distance to the center of the overlapping region. This modular approach yielded both accurate and fast retrievals, as each expert could be simpler and more tuned to a subset of the problem, compared to a monolithic inversion model. It essentially proved that MoE can seamlessly merge data-driven models with the divide-and-conquer strategy often employed in physical sciences.

Seismic inversion, another form of geophysical remote sensing using seismic data to infer subsurface properties, has also adopted MoE. Li et al. [145] proposed a pertinent multi-gate MoE for prestack seismic inversion, aiming to estimate multiple subsurface parameters from seismic signals. In their model, the expert network consists of one shared expert and three special experts, each corresponding to a specific parameter (P-wave velocity, S-wave velocity, and density). Each task has its own gating network that assigns weights to the shared expert and its corresponding special expert. The shared expert receives seismic data and all three-parameter initial models, while each special expert receives seismic data and only its corresponding initial model. This design ensures that each task can learn task-specific knowledge from its special expert while benefiting from shared information through the shared expert. This approach improves inversion accuracy for each parameter and makes the overall model more robust to variations in seismic data, compared with a single network that attempts to predict all parameters simultaneously. It illustrates how MoE can effectively handle multi-output regression problems in which different outputs follow different patterns in the data.

In the domain of oceanography, Wang et al. [146] introduced PhA-MoE, an MoE model to enhance the retrieval of phytoplankton absorption coefficients from hyperspectral imagery. Retrieving ocean biochemical properties from spectral data is challenging due to variable water conditions, data scarcity, and heterogeneity. PhA-MoE addresses this by using a noisy top-k gating network that dynamically selects the most relevant experts for each input, where each expert network handles a specific subset of the data distributions. As depicted in Figure 9, each pre-processed reflectance sample $\tilde{\mathbf{R}}$ is first embedded by an MoE-based embedding module into a latent vector $\mathbf{Y}$ via top-k routing over multiple expert networks. The gating network determines the relevance of each expert based on the input reflectance, and the top-k selected experts form a weighted sum to produce the latent embedding. The latent vector $\mathbf{Y}$ is then fed into an MDN-based predictor, which uses MLP layers to generate mixture parameters $(\alpha_i, \boldsymbol{\mu}_i, \mathbf{Cov}_i)$ that model the conditional probability distribution of absorption coefficients. A combination function then produces the final prediction $\hat{\mathbf{a}}_\phi$ by selecting the mean of the Gaussian component with the largest mixing coefficient. PhA-MoE's [146] MoE-based embedding approach leads to superior performance compared to other state-of-the-art models, with Normalized Root Mean Square Error (NRMSE) of 1.11 and a slope of 0.98, indicating no overestimation or underestimation, and demonstrates the effectiveness of the MoE structure in handling data heterogeneity across different evaluation metrics. The low NRMSE value confirms the model's accuracy in retrieving phytoplankton absorption coefficients. This yielded more accurate and generalizable estimates of phytoplankton absorption across diverse water types, illustrating MoE's ability to handle heterogeneous data distributions in ocean color remote sensing.

The MoE paradigm extends to a variety of specialized remote sensing applications. It provides a framework for tackling complex inverse problems by splitting them into simpler ones handled by experts, and for injecting domain knowledge into learning via dedicated experts for known conditions. It also resonates with the ensemble nature of many remote sensing analyses, offering a trainable mechanism to combine different models or data sources. The versatility observed, from atmospheric and ocean parameter retrieval to seismic inversion and beyond, indicates that wherever a remote sensing task can be divided into sub-tasks or conditioned on context, MoE could be a beneficial approach. The next section discusses the challenges that remain in using MoE for remote sensing and potential future research

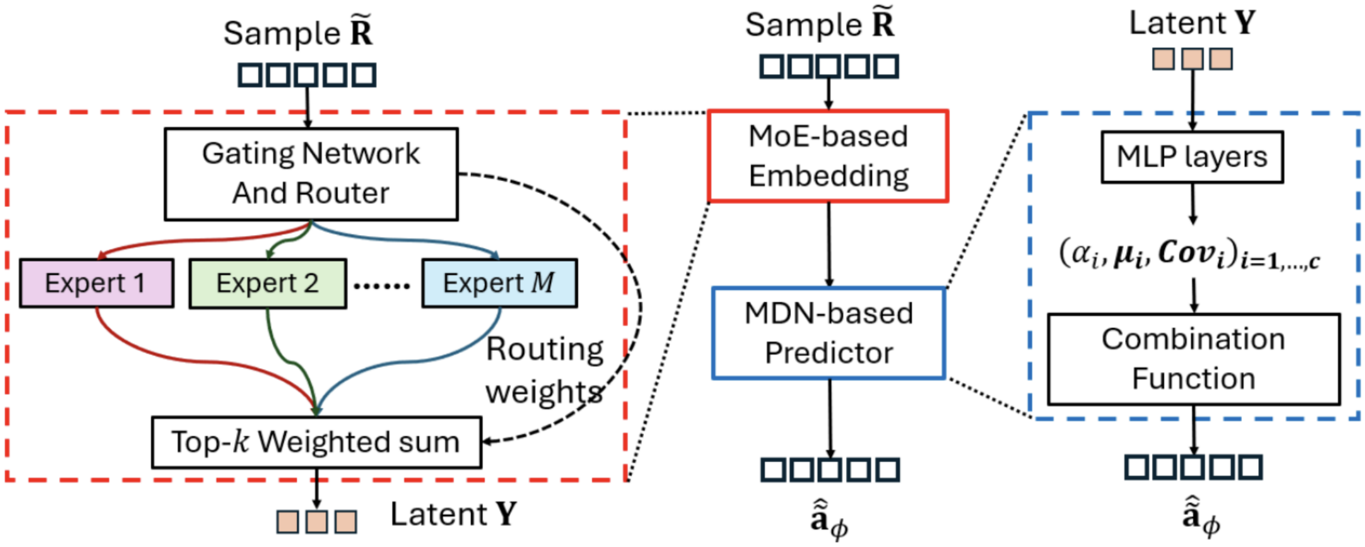

Figure 9. Phytoplankton absorption mixture-of-experts (PhA-MoE) [146] architecture for hyperspectral retrieval of phytoplankton absorption coefficients.

directions.

4 Discussion

The evidence surveyed here indicates that the performance impact of Mixture-of-Experts (MoE) in Remote Sensing (RS) varies across datasets, modality pairings, and task formulations. RS data are heterogeneous due to sensor physics, acquisition geometry, spatial scale variation, seasonality, and regional domain shift, which provides a plausible basis for conditional specialization. These same factors also increase the risk that routing decisions reflect correlations unrelated to the target task rather than task-relevant structure, and they complicate reproducible evaluation at scale [1, 6].

4.1 Assessment of MoE models

A rigorous assessment of MoE requires comparisons to simpler alternatives under matched capacity and matched compute. In practice, a substantial portion of reported gains can also be obtained by scaling a dense backbone, adding explicit multi-branch or multi-scale pathways, or using ensembles at inference time. For object-centric RS tasks, multi-expert detection and proposal mechanisms often resemble structured multi-branch designs that already encourage specialization across scales [121, 123]. For multi-source settings, carefully designed fusion pipelines can remain competitive when modality heterogeneity is moderate or supervision is limited [4]. MoE is most appropriate when the regimes that benefit from specialization are difficult to predefine, including wide geographic coverage, mixed sensors, and multi-granularity vision-language supervision, as reflected in recent RS foundation models and vision-language models that introduce modality-oriented or granularity-oriented experts [45–47, 130, 133].

Sparse activation is often motivated as a way to increase parameter capacity without proportionally increasing floating-point operations [25–27]. For RS workloads, the practical trade-offs are shaped by large spatial tensors, multi-temporal stacks, and multi-modal inputs, which amplify memory pressure and inter-device communication costs. Routing introduces dispatch and combination overhead, and its system-level impact depends on the distributed strategy and implementation. Efficient MoE training and inference therefore relies on optimized kernels and communication-aware designs [34, 61, 64, 65]. For this reason, accuracy comparisons are more interpretable when accompanied by throughput, peak memory, and wall-clock training cost under consistent hardware and parallelism settings.

Assessment of effect sizes in MoE. Effect sizes in MoE are tightly coupled to expert count and sparsity settings. Increasing the number of experts expands representational capacity, but it also increases the risk of under-trained experts, routing collapse, and severe imbalance, especially when training data are long-tailed across regions, land-cover types, and acquisition conditions [25, 27, 54, 79]. In RS, this trade-off can be unfavorable when additional experts fragment supervision without sufficient per-expert diversity. This is consistent with RS studies that focus on usage control, regularization, or pruning rather than increasing expert counts without constraint [42, 131, 138].

To make expert-count effects interpretable, reporting should include scaling curves over expert count and top-k routing, together with usage summaries such as load balance, assignment entropy, and per-expert patch counts. Router design and balancing strategy often dominate these curves. Expert-choice routing, assignment-based balancing, auxiliary-loss-free balancing, and router rebalancing explicitly target utilization issues that otherwise worsen as expert count grows. These considerations are particularly relevant for RS foundation models trained on heterogeneous global corpora, where naive routing can correlate with factors such as sensor identity, orbit, or latitude rather than task-relevant semantics [47, 133].

Statistical significance and robust evaluation. Small reported gains are difficult to interpret without statistical evidence. MoE models can exhibit higher run-to-run variance than comparable dense models because routing stochasticity and uneven expert training introduce additional sources of instability. Robust claims therefore benefit from repeated runs with confidence intervals and, when comparing across multiple datasets or tasks, appropriate multiple-comparison procedures [149, 150]. Evaluation design is equally important. Random splits in geospatial data can be overly optimistic due to spatial autocorrelation, obscuring whether improvements transfer to new regions and acquisition regimes. Spatially or environmentally separated validation is often more aligned with deployment settings and reduces leakage between training and evaluation folds [147, 148]. For time series products, temporally separated evaluation can help distinguish robust phenological learning from short-term correlations [39]. Taken together, significance analysis is most informative when paired with split strategies that match the intended transfer setting, including geographic transfer, seasonal transfer, and cross-sensor transfer.

Interpretability and diagnosing expert specialization. MoE provides a direct handle for model inspection via routing decisions. Expert assignment maps over space, time, or modality, combined with usage statistics, can reveal whether specialization aligns with meaningful regimes [33, 69]. However, routing-based interpretability is not guaranteed. Routers may respond to correlations that are not related to the target task, and routing patterns can change across retraining, which complicates monitoring and maintenance [79, 85]. Complementary explanation tools can help verify that expert selection is supported by semantically plausible evidence, rather than artifacts, including gradient-based localization and attribution methods [151–153]. For operational RS systems, routing diagnostics are also more actionable when combined with calibrated uncertainty estimates and structured failure auditing [154, 155].

4.2 Challenges of MoE in remote sensing

The same heterogeneity that motivates MoE in RS also makes training and deployment demanding, because it affects supervision quality, routing stability, and the operational reliability of the resulting system.

Data heterogeneity, preprocessing, and supervision constraints. RS data differ from natural images in radiometric calibration, atmospheric effects, viewing geometry, and sensor-specific noise, all of which introduce structured variability that can dominate learning signals [1]. Multi-modal settings add co-registration and resolution harmonization requirements; imperfect alignment can cause routers to specialize to misregistration patterns rather than geophysical content [4, 139]. Supervision is often sparse, noisy, or updated infrequently, and domain shift across regions can be substantial. Under these conditions, specialization is beneficial only when each expert receives sufficiently consistent supervision within its regime; otherwise experts can diverge into under-constrained solutions, and performance can become sensitive to sampling choices.

Failure cases under domain, scale, and modality shift. RS systems are frequently evaluated under distribution shift, including new geography, different seasonal dynamics, and atypical atmospheric conditions. In MoE models, such shifts can lead to routing failures when out-of-distribution samples are assigned to experts that were not trained for the relevant regimes, leading to confident but incorrect predictions. This risk increases in global and multi-sensor pipelines, where missing or degraded modalities, including cloud-obscured optical imagery and gaps in Synthetic Aperture Radar (SAR) acquisitions, can cause routing to rely on spurious signals [47, 118]. These observations motivate explicit shift evaluation and conservative decision rules in operational workflows, particularly when downstream actions are sensitive to false positives and false negatives.

Training stability and system-level constraints. MoE introduces stability issues that are amplified at RS scales, including expert imbalance and representational collapse, and these issues can intensify as expert count increases or as the training corpus becomes more heterogeneous. Large imagery and spatiotemporal stacks further increase memory and communication demands, making efficient dispatch and optimized kernels important for feasibility [34, 61, 64]. Advances in balancing and router optimization show that utilization can be improved without extensive tuning of auxiliary losses, but systematic validation in RS-specific pipelines remains limited [66, 85, 86].

Operational reliability, monitoring, and maintenance. Deployment introduces requirements beyond average benchmark performance. Sparse routing can yield non-uniform behavior across regions, which increases the need for geographically stratified validation, drift detection, and periodic recalibration [154, 155]. Accountability and auditing also matter in RS applications, where users may need to diagnose why outputs differ across regions or sensors, and routing summaries alone may be insufficient without additional evidence. Maintenance is also more complex than in dense models because routing patterns can change under modest data or code updates, altering expert specialization and downstream product consistency. These considerations support deployment practices that couple model-centric diagnostics with geospatially stratified evaluation and explicit update protocols [6, 147, 148].

4.3 Benchmarking protocols and reproducibility

In Computer Vision (CV) and Natural Language Processing (NLP), MoE benchmarking has largely converged on reporting practices that disentangle model capacity from realized computation, because sparse routing makes parameter count and inference cost non-equivalent. Canonical MoE studies therefore evaluate against dense baselines under matched training budgets, and report both total parameters and the activated subset per token, together with routing-related diagnostics such as expert utilization and load balance [25–27, 54, 80]. RS MoE research has started to adopt similar protocol choices, while also introducing domain-specific benchmarks that reflect geospatial heterogeneity and multimodality. SkyMoE explicitly proposes MGRS-Bench as a geospatial vision-language benchmark intended to stress multi-granularity interpretation and cross-domain generalization under an MoE foundation-model setting [133]. For multi-modal detection, SM3Det establishes a unified benchmark dataset to enable single-model evaluation across modalities and detection tasks, and uses a sparse MoE backbone to support joint training without collapsing modality-specific representations [122]. For integrity and forensics, Zhang et al. [124] release a challenging dataset for RS copy-move forgery understanding and pair it with a multimodal gated MoE model to benchmark robustness under realistic manipulations. Beyond explicitly new benchmarks, several RS MoE works emphasize breadth-of-evaluation as a benchmarking signal: RingMoE reports universal interpretation performance across many existing RS benchmarks under a mixture-of-modality-experts design [47], and SkySense V2 reports results on a wide collection of multimodal datasets under a unified foundation-model interface [130]; similarly, RS-MoE and RSUniVLM benchmark vision-language MoE designs on established RS captioning and visual question answering datasets to measure generalization across tasks and prompts [45, 46, 132].

Reproducibility for MoE models depends on faithfully specifying the routing algorithm and its training-time constraints, because small changes in capacity factors, top-k routing, auxiliary balancing objectives, and token batching can shift both convergence and expert specialization. In CV and NLP, reproducibility has been strengthened by releasing optimized training stacks that make sparse dispatch deterministic and scalable, including DeepSpeed-MoE, Tutel, FastMoE, and MegaBlocks [34, 60, 61, 64]. The lessons learned from NLP and CV communities provide valuable guidance for establishing reproducibility standards in RS MoE research. RS amplifies these sensitivities through additional degrees of freedom that sit outside the network: geospatial tiling, coordinate reference systems, cross-sensor co-registration, temporal alignment, and modality-dependent normalization can all interact with routing and produce different expert usage patterns even when the model code is unchanged [47, 131, 133]. As a result, several RS MoE papers foreground artifact release as part of their experimental methodology. MAPEX provides an MoE foundation model with modality-conditioned routing and modality-aware pruning, and the authors release code to support faithful replication of the full pre-training and adaptation pipeline [131]. Likewise, SM3Det releases code alongside its benchmark construction and sparse MoE training recipe, which is particularly important because the protocol couples multi-dataset sampling with sparse expert activation [122]. These practices do not eliminate all sources of variance, but they materially reduce ambiguity around the two dominant reproducibility bottlenecks for RS MoE studies: the routing configuration that determines conditional computation, and the geospatial preprocessing pipeline that determines what tokens the router actually sees [27, 47, 80, 133].

5 Future Directions

Given the established progress and diverse applications of Mixture-of-Experts (MoE) in remote sensing, several promising avenues and critical challenges emerge for future research. This section outlines key directions for advancing the capabilities, efficiency, and applicability of MoE models within the unique context of Earth observation.

5.1 Unified Multi-Modal and Multi-Task MoE

Remote sensing workflows are increasingly multi-modal (optical, Synthetic Aperture Radar (SAR), hyperspectral, Light Detection and Ranging (LiDAR), Digital Elevation Model (DEM), text/metadata) and multi-task (classification, segmentation, detection, change detection, retrieval, captioning, Visual Question Answering (VQA)), but most existing MoE models are still designed for relatively narrow settings (e.g., a single downstream task and a small number of modalities). Integrating DEM data with optical and SAR imagery can provide complementary topographic information that enhances land-cover classification and terrain analysis. For instance, MixtureRS focuses on fusing hyperspectral imagery and LiDAR for land-use classification via modality-specific experts and cross-attention fusion, while the model proposed by Liu et al. [132] adopts cross-modal mixture experts for remote-sensing visual question answering. These works clearly show that MoE is well suited to handle heterogeneous inputs, but they do not yet deliver a single, unified *all-in-one* architecture that can serve as a foundation model across modalities and tasks.

A key future direction is the design of unified multi-modal, multi-task MoE frameworks tailored to Earth observation. Such a framework could include a backbone with separate expert pools for different modalities (e.g., optical, SAR, hyperspectral, LiDAR, text) and another set of experts for different task families (e.g., dense prediction, detection, sequence modeling). A hierarchical gating mechanism would first route tokens based on modality and acquisition metadata (sensor, orbit, season, incidence angle) and then further route within task-specific experts depending on the query (classification, segmentation, captioning, VQA, etc.). Joint pre-training of such a model with mixed objectives (masked reconstruction, contrastive alignment, language modeling, detection and segmentation heads) could produce a single MoE backbone reusable for a wide range of remote sensing tasks, including those with missing modalities or weak supervision.

In addition, future multi-modal MoE models could explicitly exploit metadata-aware routing (e.g., using acquisition time, location or orbit as router inputs) to activate experts specialized for specific regions, seasons or imaging conditions. This would be particularly valuable for large-scale monitoring, where the same architecture must handle very different landscapes and sensor configurations while sharing as much capacity as possible. Lessons from large-scale MoE in Natural Language Processing (NLP) and vision, such as sparsely-gated layers and switch-style routing [27], provide a strong blueprint for building such unified remote sensing MoE backbones at scale.

5.2 Expert Specialization for Low-Level Vision

Most MoE applications in remote sensing have so far emphasized mid- and high-level tasks such as classification, semantic segmentation, or VQA. Low-level vision problems, such as super-resolution, denoising, dehazing, pansharpening, atmospheric correction, and multi-degradation restoration—are equally important in operational pipelines, yet only a few works exploit MoE explicitly. A recent example is the multi-level feature guided heterogeneous MoE [42] for remote sensing image super-resolution, which introduces heterogeneous experts in the upsampling stage and a dual-routing mechanism guided by multi-level features to adapt reconstruction to different ground-object characteristics. This example indicates that MoE can be very effective in handling the diverse textures, frequencies and structures that appear within a single remote sensing scene.

Future research can push this idea much further by designing expert sets that explicitly specialize in: (i) different degradation types (sensor noise, blur, haze, downsampling kernels, compression artefacts), (ii) different spectral types (visible, NIR, SWIR, LiDAR, etc.), and (iii) different spatial structures (water, vegetation, urban, mountainous terrain). In such models, the gating network could operate at patch- or even pixel-level to assign each region to the most appropriate combination of restoration experts. Physics-guided experts that embed sensor models, radiative transfer approximations, or priors on atmospheric scattering could coexist with purely data-driven experts, with the router learning where each is most reliable (e.g., physics-based experts for clear-sky radiometry, data-driven experts for heavy haze or mixed pixels).

Another promising line is to couple low-level MoE with downstream tasks. For example, one could jointly train a restoration MoE and a segmentation or detection head, allowing some experts to specialize in task-aware enhancement (e.g., sharpening building edges or small vessels) rather than purely perceptual quality. End-to-end training would encourage experts to learn restorations that preserve or amplify discriminative cues important for land-cover mapping, object detection, or change detection, rather than optimizing generic image quality metrics alone.

5.3 Efficient and Robust MoE Architectures

A well-known attraction of MoE is its ability to increase model capacity without proportionally increasing computation, by activating only a small subset of experts per input. Sparsely-gated MoE layers and Switch Transformers have shown that very large conditional-capacity models can be trained efficiently with simple routing and load-balancing strategies [27]. However, remote sensing deployments impose additional constraints: models may need to run on satellites, aircraft, or edge devices with tight power and memory budgets, and they must remain robust under strong domain shifts (different sensors, regions, seasons, or acquisition geometries).

This motivates future work on resource-aware and robust MoE architectures specifically designed for Earth observation. On the efficiency side, interesting directions include: token- or patch-wise routing that prunes uninformative regions (e.g., uniform ocean), expert pruning or distillation for specific missions, and dynamic compute allocation where the router chooses not only which experts to activate but also how much computation to spend per patch. On the robustness side, domain-aware routing, conditioning on sensor ID, incidence angle, or geolocation, could help activate experts specialized for particular acquisition regimes, while auxiliary losses could regularize the router to avoid collapse (e.g., overusing one expert) and to remain stable under distribution shift. Models like [41] already hint that modality-aware expert design improves robustness to imperfect coregistration and heterogeneous landscapes; similar ideas could be extended to multi-sensor change detection or cross-satellite adaptation.

Another open issue is fault tolerance and safety in operational settings. Future MoE architectures for remote sensing should include mechanisms to detect unreliable routing decisions (e.g., when the input is far from any expert's training distribution) and fall back to simpler, more conservative experts or ensembles. This could be combined with uncertainty estimation at both expert and gating levels to support risk-aware decision-making in high-stakes applications such as disaster response or maritime surveillance.

5.4 Effectiveness of Experts and Training Strategies

Despite the growing number of MoE-based models, we still lack a clear understanding of what remote sensing experts actually learn, how many experts are needed, and how best to train and regularize them. General MoE work has emphasized issues such as expert specialization, load balancing, routing stability and training dynamics, proposing auxiliary losses and simplified routers (e.g., top-k gating, single-expert routing) to keep training stable while encouraging diversity. Remote sensing MoEs often report ablations on the number of experts or routing variants, but systematic analyses across tasks and modalities are still rare.

A valuable research direction is to explicitly probe and visualize expert specialization in Earth observation models: for example, analyzing whether experts align with land-cover types, geographic regions, seasons, sensor families, or task types. Works such as MFG-HMoE [42] and MixtureRS [41] already incorporate non-trivial routing designs (dual routing, cross-attention fusion, cross-modal mixture experts) and report improvements, but they do not fully characterize how experts partition the data space. Future studies could combine MoE with probing tasks, clustering of expert activations, and metadata conditioning to obtain more interpretable "maps of expertise" over the Earth.

Training strategies themselves are another open frontier. Beyond standard load-balancing losses, remote sensing MoE could benefit from curriculum routing (gradually increasing the number of active experts or the complexity of inputs), expert dropout (forcing robustness to missing experts), or teacher–student setups where a dense foundation model teaches a sparse MoE to allocate capacity where it matters most. In multi-task settings, it may be advantageous to share some experts across tasks while dedicating others to task-specific nuances (e.g., instance-level, pixel-level reasoning), or to learn separate routers for tasks that compete for the same features. Finally, because high-quality labels are expensive in remote sensing, semi-supervised and self-supervised pretraining for MoE using large archives of unlabeled imagery and metadata remains largely unexplored and could significantly improve expert quality and sample efficiency across downstream tasks.

6 Conclusion

This survey has provided a comprehensive overview of the Mixture-of-Experts (MoE) paradigm and its growing applications in Remote Sensing (RS). The core advantages of MoE in RS lie in its ability to handle data heterogeneity through conditional computation and expert specialization. MoE architectures can allocate different subsets of parameters to different regions of the input space, classes, modalities, or tasks, making them natural candidates for heterogeneous RS data. By activating only a small subset of experts per input, MoE models can scale up capacity efficiently while maintaining computational efficiency, which is crucial for processing diverse Earth observation data across multiple sensors, resolutions, and temporal scales.

Despite these advantages, several unresolved key issues remain. Most existing works focus on specific tasks or limited modalities, lacking unified frameworks capable of generalizing across diverse Earth observation scenarios. The field still needs efficient and robust MoE architectures designed specifically for resource-constrained environments common in RS deployments. Understanding and visualizing expert specialization remains challenging, with limited systematic analyses of what RS experts actually learn and how they partition the data space. Training strategies that ensure stability and generalization with limited labeled data are still underdeveloped, particularly for multi-modal and multi-temporal scenarios. Additionally, benchmarking protocols and reproducibility standards need to be established to enable fair comparisons and reliable reproduction of results across different studies.

Addressing these challenges requires focused research efforts on unified multi-modal and multi-task MoE foundations, extending expert specialization to low-level vision tasks, creating resource-aware deployment strategies, and deepening the interpretability and efficiency of expert routing mechanisms. Progress in these areas will be essential for fully harnessing MoE's potential in scalable, accurate, and efficient remote sensing analysis, ultimately enabling more robust and generalizable Earth observation systems.

Data Availability Statement

Not applicable.

Funding

This work was supported in part by the National Natural Science Foundation of China under Grant U2243222 and Grant 61731022, in part by the National Key Research and Development Program of China under Grant 2024YFF1307204, and in part by the Project of Comprehensive Site Selection System under Grant KY24004.

Conflicts of Interest

The authors declare no conflicts of interest.

AI Use Statement

The authors declare that no generative AI was used in the preparation of this manuscript.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Zhu, X. X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., & Fraundorfer, F. (2017). Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources. *IEEE Geoscience and Remote Sensing Magazine, 5*(4), 8-36. [CrossRef]

[2] Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., & Johnson, B. A. (2019). Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS Journal of Photogrammetry and Remote Sensing, 152*, 166-177. [CrossRef]

[3] Zhao, S., Tu, K., Ye, S., Tang, H., Hu, Y., & Xie, C. (2023). Land Use and Land Cover Classification Meets Deep Learning: A Review. *Sensors, 23*(21), 8966. [CrossRef]

[4] Gao, J., Li, P., Chen, Z., & Zhang, J. (2020). A Survey on Deep Learning for Multimodal Data Fusion. *Neural Computation, 32*(5), 829-864. [CrossRef]

[5] Osco, L. P., Junior, J. M., Ramos, A. P. M., de Castro Jorge, L. A., Fatholahi, S. N., de Andrade Silva, J., ... & Li, J. (2021). A review on deep learning in UAV remote sensing. *International Journal of Applied Earth Observation and Geoinformation, 102*, 102456. [CrossRef]

[6] Lu, S., Guo, J., Zimmer-Dauphinee, J. R., Nieusma, J. M., Wang, X., VanValkenburgh, P., ... & Huo, Y. (2025). Vision foundation models in remote sensing: A survey. *IEEE Geoscience and Remote Sensing Magazine, 13*(3), 190-215. [CrossRef]

[7] Jiang, H., Peng, M., Zhong, Y., Xie, H., Hao, Z., Lin, J., Ma, X., & Hu, X. (2022). A Survey on Deep Learning-Based Change Detection from High-Resolution Remote Sensing Images. *Remote Sensing, 14*(7), 1552. [CrossRef]

[8] Shafique, A., Cao, G., Khan, Z., Asad, M., & Aslam, M. (2022). Deep Learning-Based Change Detection in Remote Sensing Images: A Review. *Remote Sensing, 14*(4), 871. [CrossRef]

[9] Ding, L., Hong, D., Zhao, M., Chen, H., Li, C., Deng, J., Yokoya, N., Bruzzone, L., & Chanussot, J. (2025). A Survey of Sample-Efficient Deep Learning for Change Detection in Remote Sensing: Tasks, strategies, and challenges. *IEEE Geoscience and Remote Sensing Magazine, 13*(3), 164-189. [CrossRef]

[10] Peng, D., Liu, X., Zhang, Y., Guan, H., Li, Y., & Bruzzone, L. (2025). Deep learning change detection techniques for optical remote sensing imagery: Status, perspectives and challenges. *International Journal of Applied Earth Observation and Geoinformation, 136*, 104282. [CrossRef]

[11] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems, 25*.

[12] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770-778. [CrossRef]

[13] Shelhamer, E., Long, J., & Darrell, T. (2016). Fully Convolutional Networks for Semantic Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*(4), 640-651. [CrossRef]

[14] Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* (pp. 234-241). Cham: Springer international publishing. [CrossRef]

[15] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6000-6010.

[16] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research, 21*(140), 1-67.

[17] Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive Mixtures of Local Experts. *Neural Computation, 3*(1), 79-87. [CrossRef]

[18] Jordan, M. I., & Jacobs, R. A. (1994). Hierarchical Mixtures of Experts and the EM Algorithm. *Neural Computation, 6*(2), 181-214. [CrossRef]

[19] Miller, D. J., & Uyar, H. S. (1997). A mixture of experts classifier with learning based on both labelled and unlabelled data. In *10th Annual Conference on Neural Information Processing Systems, NIPS 1996* (pp. 571-577). Neural information processing systems foundation.

[20] Jiang, W., & Tanner, M. A. (1999, January). Hierarchical mixtures-of-experts for generalized linear models: some results on denseness and consistency. In *Seventh International Workshop on Artificial Intelligence and Statistics*. PMLR.

[21] Yuksel, S. E., Wilson, J. N., & Gader, P. D. (2012). Twenty Years of Mixture of Experts. *IEEE Transactions on Neural Networks and Learning Systems, 23*(8), 1177-1193. [CrossRef]

[22] Nguyen, H. D., & Chamroukhi, F. (2018). Practical and theoretical aspects of mixture-of-experts modeling: An overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8*(4), e1246. [CrossRef]

[23] Chamroukhi, F. (2017). Skew t mixture of experts. *Neurocomputing, 266*, 390-408. [CrossRef]

[24] Fung, T. C., & Tseung, S. C. (2025). Mixture of experts models for multilevel data: Modeling framework and approximation theory. *Neurocomputing, 626*, 129357. [CrossRef]

[25] Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., & Dean, J. (2017). Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

[26] Lepikhin, D., Lee, H., Xu, Y., Chen, D., Firat, O., Huang, Y., ... & Chen, Z. (2020). Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*.

[27] Fedus, W., Zoph, B., & Shazeer, N. (2022). Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *Journal of Machine Learning Research, 23*(120), 1-39. [CrossRef]

[28] Du, N., Huang, Y., Dai, A. M., Tong, S., Lepikhin, D., Xu, Y., ... & Cui, C. (2022, June). Glam: Efficient scaling of language models with mixture-of-experts.

In *International conference on machine learning* (pp. 5547-5569). PMLR.

[29] Wang, X., Yu, F., Dunlap, L., Ma, Y. A., Wang, R., Mirhoseini, A., ... & Gonzalez, J. E. (2020, August). Deep mixture of experts via shallow embedding. In *Uncertainty in artificial intelligence* (pp. 552-562). PMLR.

[30] Fan, Z., Sarkar, R., Jiang, Z., Chen, T., Zou, K., Cheng, Y., ... & Wang, Z. (2022). M^{3}vit: Mixture-of-experts vision transformer for efficient multi-task learning with model-accelerator co-design. *Advances in Neural Information Processing Systems, 35*, 28441-28457.

[31] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

[32] Chen, Z., Shen, Y., Ding, M., Chen, Z., Zhao, H., Learned-Miller, E., & Gan, C. (2023, June). Mod-Squad: Designing Mixtures of Experts As Modular Multi-Task Learners. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 11828-11837). IEEE Computer Society. [CrossRef]

[33] Pavlitskaya, S., Hubschneider, C., Weber, M., Moritz, R., Hüger, F., Schlicht, P., & Zöllner, J. M. (2020, June). Using Mixture of Expert Models to Gain Insights into Semantic Segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (pp. 1399-1406). IEEE. [CrossRef]

[34] Rajbhandari, S., Li, C., Yao, Z., Zhang, M., Aminabadi, R. Y., Awan, A. A., ... & He, Y. (2022, June). Deepspeed-moe: Advancing mixture-of-experts inference and training to power next-generation ai scale. In *International conference on machine learning* (pp. 18332-18346). PMLR.

[35] Cai, W., Jiang, J., Wang, F., Tang, J., Kim, S., & Huang, J. (2025). A Survey on Mixture of Experts in Large Language Models. *IEEE Transactions on Knowledge & Data Engineering, 37*(07), 3896-3915. [CrossRef]

[36] Cai, W., Jiang, J., Wang, F., Tang, J., Kim, S., & Huang, J. (2024). A Survey on Mixture of Experts. *arXiv preprint arXiv:2407.06204*.

[37] Gan, W., Ning, Z., Qi, Z., & Yu, P. S. (2025). Mixture of experts (moe): A big data perspective. *Information Fusion*, 103664. [CrossRef]

[38] Dimitri, V., Regina, B., & Alfonz, M. (2025). A Survey on Mixture of Experts: Advancements, Challenges, and Future Directions. *Authorea Preprints*. [CrossRef]

[39] Dou, P., Shen, H., Li, Z., & Guan, X. (2021). Time series remote sensing image classification framework using combination of deep learning and multiple classifiers system. *International Journal of Applied Earth Observation and Geoinformation, 103*, 102477. [CrossRef]

[40] Jia, Y., Ge, Y., Ling, F., Guo, X., Wang, J., Wang, L., Chen, Y., & Li, X. (2018). Urban Land Use Mapping by Combining Remote Sensing Imagery and Mobile Phone Positioning Data. *Remote Sensing, 10*(3), 446. [CrossRef]

[41] Liu, Y., Wu, C., Guan, M., & Wang, J. (2025). MixtureRS: A Mixture of Expert Network Based Remote Sensing Land Classification. *Remote Sensing, 17*(14), 2494. [CrossRef]

[42] Chen, B., Chen, K., Yang, M., Zou, Z., & Shi, Z. (2025). Heterogeneous Mixture of Experts for Remote Sensing Image Super-Resolution. *IEEE Geoscience and Remote Sensing Letters, 22*, LGRS-2025. [CrossRef]

[43] Liu, Z., Zhang, J., Wang, W., & Gu, Y. (2025). M 2 CD: A Unified MultiModal Framework for Optical-SAR Change Detection With Mixture of Experts and Self-Distillation. *IEEE Geoscience and Remote Sensing Letters, 22*, LGRS-2025. [CrossRef]

[44] Lu, Q., Zhao, W., Chen, J., Chen, X., & Zhang, L. (2025). Uncertainty Mixture of Experts Model for Long Tail Crop Type Mapping. *Remote Sensing, 17*(22), 3752. [CrossRef]

[45] Lin, H., Hong, D., Ge, S., Luo, C., Jiang, K., Jin, H., & Wen, C. (2025). 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, 63*, 1-18. [CrossRef]

[46] Liu, X., & Lian, Z. (2024). Rsunivlm: A unified vision language model for remote sensing via granularity-oriented mixture of experts. *arXiv preprint arXiv:2412.05679*.

[47] Bi, H., Feng, Y., Tong, B., Wang, M., Yu, H., Mao, Y., ... & Sun, X. (2025). RingMoE: Mixture-of-modality-experts multi-modal foundation models for universal remote sensing image interpretation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. [CrossRef]

[48] Seydi, S. T., Hasanlou, M., & Chanussot, J. (2024). A novel deep Siamese framework for burned area mapping Leveraging mixture of experts. *Engineering Applications of Artificial Intelligence, 133*, 108280. [CrossRef]

[49] Masoudnia, S., & Ebrahimpour, R. (2014). Mixture of experts: a literature survey. *Artificial Intelligence Review, 42*(2), 275-293. [CrossRef]

[50] Ma, J., Zhao, Z., Yi, X., Chen, J., Hong, L., & Chi, E. H. (2018). Modeling Task Relationships in Multi-task Learning with Multi-gate Mixture-of-Experts. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1930-1939. [CrossRef]

[51] Tang, H., Liu, J., Zhao, M., & Gong, X. (2020, September). Progressive layered extraction (ple): A novel multi-task learning (mtl) model for personalized recommendations. In *Proceedings of the*

14th ACM conference on recommender systems (pp. 269-278). [CrossRef]

[52] Gupta, S., Mukherjee, S., Subudhi, K., Gonzalez, E., Jose, D., Awadallah, A. H., & Gao, J. (2022). Sparsely activated mixture-of-experts are robust multi-task learners. *arXiv preprint arXiv:2204.07689*.

[53] Kudugunta, S., Huang, Y., Bapna, A., Krikun, M., Lepikhin, D., Luong, M. T., & Firat, O. (2021, November). Beyond distillation: Task-level mixture-of-experts for efficient inference. In *Findings of the association for computational linguistics: EMNLP 2021* (pp. 3577-3599). [CrossRef]

[54] Riquelme, C., Puigcerver, J., Mustafa, B., Neumann, M., Jenatton, R., Susano Pinto, A., ... & Houlsby, N. (2021). Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems, 34*, 8583-8595.

[55] Mustafa, B., Riquelme, C., Puigcerver, J., Jenatton, R., & Houlsby, N. (2022). Multimodal contrastive learning with limoe: the language-image mixture of experts. *Advances in Neural Information Processing Systems, 35*, 9564-9576.

[56] Zhu, J., Zhu, X., Wang, W., Wang, X., Li, H., Wang, X., & Dai, J. (2022). Uni-perceiver-moe: Learning sparse generalist models with conditional moes. *Advances in Neural Information Processing Systems, 35*, 2664-2678.

[57] Dai, D., Deng, C., Zhao, C., Xu, R. X., Gao, H., Chen, D., ... & Liang, W. (2024, August). Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1280-1297). [CrossRef]

[58] Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., ... & Sayed, W. E. (2024). Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

[59] Muennighoff, N., Soldaini, L., Groeneveld, D., Lo, K., Morrison, J., Min, S., ... & Hajishirzi, H. (2024). Olmoe: Open mixture-of-experts language models. *arXiv preprint arXiv:2409.02060*.

[60] He, J., Qiu, J., Zeng, A., Yang, Z., Zhai, J., & Tang, J. (2021). Fastmoe: A fast mixture-of-expert training system. *arXiv preprint arXiv:2103.13262*.

[61] Hwang, C., Cui, W., Xiong, Y., Yang, Z., Liu, Z., Hu, H., ... & Xiong, Y. (2023). Tutel: Adaptive mixture-of-experts at scale. *Proceedings of Machine Learning and Systems, 5*, 269-287.

[62] Nie, X., Zhao, P., Miao, X., Zhao, T., & Cui, B. (2022). HetuMoE: An efficient trillion-scale mixture-of-expert distributed training system. *arXiv preprint arXiv:2203.14685*.

[63] He, J., Zhai, J., Antunes, T., Wang, H., Luo, F., Shi, S., & Li, Q. (2022). FasterMoE: modeling and optimizing training of large-scale dynamic pre-trained models. *Proceedings of the 27th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*, 120-134. [CrossRef]

[64] Gale, T., Narayanan, D., Young, C., & Zaharia, M. (2023). Megablocks: Efficient sparse training with mixture-of-experts. *Proceedings of Machine Learning and Systems, 5*, 288-304.

[65] Hwang, R., Wei, J., Cao, S., Hwang, C., Tang, X., Cao, T., & Yang, M. (2024, June). Pre-gated moe: An algorithm-system co-design for fast and scalable mixture-of-expert inference. In *2024 ACM/IEEE 51st Annual International Symposium on Computer Architecture (ISCA)* (pp. 1018-1031). IEEE. [CrossRef]

[66] Lewis, M., Bhosale, S., Dettmers, T., Goyal, N., & Zettlemoyer, L. (2021, July). Base layers: Simplifying training of large, sparse models. In *International Conference on Machine Learning* (pp. 6265-6274). PMLR.

[67] Roller, S., Sukhbaatar, S., Szlam, A., & Weston, J. (2021, December). Hash layers for large sparse models. In *Proceedings of the 35th International Conference on Neural Information Processing Systems* (pp. 17555-17566).

[68] Komatsuzaki, A., Puigcerver, J., Lee-Thorp, J., Ruiz, C. R., Mustafa, B., Ainslie, J., ... & Houlsby, N. (2022). Sparse upcycling: Training mixture-of-experts from dense checkpoints. *arXiv preprint arXiv:2212.05055*.

[69] Chen, Z., Deng, Y., Wu, Y., Gu, Q., & Li, Y. (2022). Towards Understanding the Mixture-of-Experts Layer in Deep Learning. *Advances in Neural Information Processing Systems, 35*, 23049-23062.

[70] Ho, N., Yang, C.-Y., & Jordan, M. I. (2022). Convergence Rates for Gaussian Mixtures of Experts. *Journal of Machine Learning Research, 23*(323), 1-81.

[71] Nguyen, H., Nguyen, T., & Ho, N. (2023). Demystifying Softmax Gating Function in Gaussian Mixture of Experts. *Advances in Neural Information Processing Systems, 36*, 4624-4652.

[72] Nguyen, H., Akbarian, P., Yan, F., & Ho, N. (2023). Statistical perspective of top-k sparse softmax gating mixture of experts. *arXiv preprint arXiv:2309.13850*.

[73] Nguyen, H., Akbarian, P., Nguyen, T., & Ho, N. (2023). A general theory for softmax gating multinomial logistic mixture of experts. *arXiv preprint arXiv:2310.14188*.

[74] Shi, X., Wang, S., Nie, Y., Li, D., Ye, Z., Wen, Q., & Jin, M. (2024). Time-moe: Billion-scale time series foundation models with mixture of experts. *arXiv preprint arXiv:2409.16040*.

[75] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2022). Lora: Low-rank adaptation of large language models. *Iclr, 1*(2), 3.

[76] Kunwar, P., Vu, M. N., Gupta, M., Abdelsalam, M., & Bhattarai, M. (2025, November). TT-LoRA MoE: Using Parameter-Efficient Fine-Tuning and Sparse Mixture-Of-Experts. In *Proceedings of the International Conference for High Performance Computing, Networking,*

Storage and Analysis (pp. 1332-1350). [CrossRef]

[77] Wu, X., Huang, S., & Wei, F. (2024). Mixture of lora experts. *arXiv preprint arXiv:2404.13628*.

[78] Liao, M., Chen, W., Shen, J., Guo, S., & Wan, H. (2025, April). Hmora: Making llms more effective with hierarchical mixture of lora experts. In *The Thirteenth International Conference on Learning Representations*.

[79] Zoph, B., Bello, I., Kumar, S., Du, N., Huang, Y., Dean, J., ... & Fedus, W. (2022). St-moe: Designing stable and transferable sparse expert models. *arXiv preprint arXiv:2202.08906*.

[80] Zhou, Y., Lei, T., Liu, H., Du, N., Huang, Y., Zhao, V., ... & Laudon, J. (2022). Mixture-of-experts with expert choice routing. *Advances in Neural Information Processing Systems, 35*, 7103-7114.

[81] Huang, Q., An, Z., Zhuang, N., Tao, M., Zhang, C., Jin, Y., ... & Feng, Y. (2024, August). Harder task needs more experts: Dynamic routing in MoE models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (*Volume 1: Long Papers*) (pp. 12883-12895). [CrossRef]

[82] Hazimeh, H., Zhao, Z., Chowdhery, A., Sathiamoorthy, M., Chen, Y., Mazumder, R., Hong, L., & Chi, E. (2021). DSelect-k: Differentiable Selection in the Mixture of Experts with Applications to Multi-Task Learning. *Advances in Neural Information Processing Systems, 34*, 29335-29347.

[83] Zhong, Z., Xia, M., Chen, D., & Lewis, M. (2024). Lory: Fully differentiable mixture-of-experts for autoregressive language model pre-training. *arXiv preprint arXiv:2405.03133*.

[84] Zuo, S., Liu, X., Jiao, J., Kim, Y. J., Hassan, H., Zhang, R., ... & Gao, J. (2021). Taming sparsely activated transformer with stochastic experts. *arXiv preprint arXiv:2110.04260*.

[85] Wang, L., Gao, H., Zhao, C., Sun, X., & Dai, D. (2024). Auxiliary-loss-free load balancing strategy for mixture-of-experts. *arXiv preprint arXiv:2408.15664*.

[86] Thaman, K. (2025). One Must Imagine Experts Happy: Rebalancing Neural Routers via Constrained Optimization. *Sparsity in LLMs* (*SLLM*)*: Deep Dive into Mixture of Experts, Quantization, Hardware, and Inference*.

[87] Omi, N., Sen, S., & Farhadi, A. (2025). Load balancing mixture of experts with similarity preserving routers. *arXiv preprint arXiv:2506.14038*.

[88] Do, G., Le, H., & Tran, T. (2025). SimSMoE: Toward Efficient Training Mixture of Experts via Solving Representational Collapse. *Findings of the Association for Computational Linguistics: NAACL 2025*, 2012-2025. [CrossRef]

[89] Chen, T., Zhang, Z., Jaiswal, A., Liu, S., & Wang, Z. (2023). Sparse moe as the new dropout: Scaling dense and self-slimmable transformers. *arXiv preprint arXiv:2303.01610*.

[90] Xie, Y., Huang, S., Chen, T., & Wei, F. (2023, June). Moec: Mixture of expert clusters. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 37, No. 11, pp. 13807-13815). [CrossRef]

[91] Zuo, S., Zhang, Q., Liang, C., He, P., Zhao, T., & Chen, W. (2022). MoEBERT: from BERT to Mixture-of-Experts via Importance-Guided Adaptation. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1610-1623. [CrossRef]

[92] Szatkowski, F., Wójcik, B., Piórczyński, M., & Scardapane, S. (2024). Exploiting activation sparsity with dense to dynamic-k mixture-of-experts conversion. *Advances in Neural Information Processing Systems, 37*, 43245-43273.

[93] Gao, S., Hua, T., Shirkavand, R., Lin, C. H., Tang, Z., Li, Z., ... & Hsu, Y. C. (2025). ToMoE: Converting Dense Large Language Models to Mixture-of-Experts through Dynamic Structural Pruning. *arXiv preprint arXiv:2501.15316*.

[94] Nussbaum, Z., & Duderstadt, B. (2025). Training sparse mixture of experts text embedding models. *arXiv preprint arXiv:2502.07972*.

[95] Gu, N., Zhang, Z., Feng, Y., Chen, Y., Fu, P., Lin, Z., ... & Wang, H. (2025). Elastic MoE: Unlocking the Inference-Time Scalability of Mixture-of-Experts. *arXiv preprint arXiv:2509.21892*.

[96] Ma, W., Zhang, H., Zhao, L., Song, Y., Wang, Y., Sui, Z., & Luo, F. (2025). Stabilizing moe reinforcement learning by aligning training and inference routers. *arXiv preprint arXiv:2510.11370*.

[97] Gururangan, S., Li, M., Lewis, M., Shi, W., Althoff, T., Smith, N. A., & Zettlemoyer, L. (2023). Scaling expert language models with unsupervised domain discovery. *arXiv preprint arXiv:2303.14177*.

[98] Jawahar, G., Mukherjee, S., Liu, X., Kim, Y. J., Abdul-Mageed, M., Lakshmanan, L., ... & Gao, J. (2023, July). Automoe: Heterogeneous mixture-of-experts with adaptive computation for efficient neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023* (pp. 9116-9132). [CrossRef]

[99] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1* (*long and short papers*) (pp. 4171-4186). [CrossRef]

[100] Kussul, N., Shelestov, A., Lavreniuk, M., Butko, I., & Skakun, S. (2016, July). Deep learning approach for large scale land cover mapping based on remote sensing data fusion. In *2016 IEEE international geoscience and remote sensing symposium* (*IGARSS*) (pp. 198-201). IEEE. [CrossRef]

[101] Yuksel, S. E., & Gader, P. D. (2012, July). Mixture of hmm experts with applications to landmine detection. In *2012 IEEE International Geoscience and Remote Sensing Symposium* (pp. 6852-6855). IEEE. [CrossRef]

[102] Nagarajan, K., & Slatton, K. C. (2009). Multiscale Segmentation of Elevation Images Using a Mixture-of-Experts Framework. *IEEE Geoscience and Remote Sensing Letters*, *6*(4), 865-869. [CrossRef]

[103] Chen, Y., Cui, H., Zhang, G., Li, X., Xie, Z., Li, H., & Li, D. (2024). SparseFormer: A credible Dual-CNN Expert-Guided transformer for remote sensing image segmentation with sparse point annotation. *IEEE Transactions on Geoscience and Remote Sensing*, *63*, 1-16. [CrossRef]

[104] Sun, Z., Liu, J., Zhang, W., Liu, F., Yang, J., & Xiao, L. (2025, April). Multi-scale Feature Interaction and Adaptive Experts for Panoptic Segmentation in Remote Sensing Images. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1-5). IEEE. [CrossRef]

[105] He, S., Cheng, Q., Huai, Y., Zhu, Z., & Ding, J. (2024, July). Mixture-of-experts for semantic segmentation of remoting sensing image. In *International Conference on Image Processing and Artificial Intelligence (ICIPAl 2024)* (Vol. 13213, pp. 478-483). SPIE. [CrossRef]

[106] Ren, J., Zai, K., Li, H., Wang, H., Du, J., Mu, W., & Qin, F. (2025). A Mixture of Experts Model for Image Classification Based on High-Resolution Remote Sensing Image. *2025 32nd International Conference on Geoinformatics*, 1-7. [CrossRef]

[107] Chen, X., Yan, S., Zhu, J., Chen, C., Liu, Y., & Zhang, M. (2025). Generalizable multispectral land cover classification via frequency-aware mixture of low-rank token experts. *arXiv preprint arXiv:2505.14088*.

[108] Lee, S., Park, S., Yang, J., Kim, J., & Cha, M. (2025). Generalizable Slum Detection from Satellite Imagery with Mixture-of-Experts. *arXiv preprint arXiv:2511.10300*.

[109] Xu, H., Xue, B., Liu, R., Zhang, Q., & Lu, W. (2025, June). Multi-Scale Mixture-of-Experts With Lora for Building Extraction from Optical Remote-Sensing Images. In *2025 32nd International Conference on Geoinformatics* (pp. 1-9). IEEE. [CrossRef]

[110] Li, R., Ding, X., Peng, S., & Cai, F. (2025). U-MoEMamba: A Hybrid Expert Segmentation Model for Cabbage Heads in Complex UAV Low-Altitude Remote Sensing Scenarios. *Agriculture*, *15*(16), 1723. [CrossRef]

[111] Xie, J., Yu, F., & Wang, H. (2022, April). Stacked Mixture-of-Expert Networks for Fast Aerial Scene Classification. In *2022 3rd International Conference on Geology, Mapping and Remote Sensing (ICGMRS)* (pp. 121-126). IEEE. [CrossRef]

[112] Mensah, E. A., Lee, A., Zhang, H., Shan, Y., & Heimerl, K. (2024). Towards vision mixture of experts for wildlife monitoring on the edge. *arXiv preprint arXiv:2411.07834*.

[113] Wang, Y., Zhang, F., Zhao, Q., Hu, W., & Ma, F. (2025). DMRS: Long-tailed remote sensing recognition via semantic-aware mixing and diversity experts. *International Journal of Applied Earth Observation and Geoinformation*, *141*, 104623. [CrossRef]

[114] Fu, Y., Yang, R., Liu, Z., & Ng, M. K. (2025). Adaptive Mixture-of-Experts Distillation for Cross-Satellite Generalizable Incremental Remote Sensing Scene Classification. *IEEE Transactions on Circuits and Systems for Video Technology, 36*(1), 233-247. [CrossRef]

[115] Guo, S., Chen, T., Wang, P., Yan, J., & Liu, H. (2025). Confidence Fusion With Representation Distribution and Mixture of Experts for Multimodal Radar Target Recognition. *IEEE Transactions on Aerospace and Electronic Systems*, *61*(5), 13251-13268. [CrossRef]

[116] Xu, Y., Wang, D., Jiao, H., Zhang, L., & Zhang, L. (2025). MambaMoE: Mixture-of-spectral-spatial-experts state space model for hyperspectral image classification. *Information Fusion*, 103811. [CrossRef]

[117] Dai, X., Li, Z., Li, L., Xue, S., Huang, X., & Yang, X. (2025). HyperTransXNet: learning both global and local dynamics with a dual dynamic token mixer for hyperspectral image classification. *Remote Sensing*, *17*(14), 2361. [CrossRef]

[118] Gao, Q., Qu, J., Li, Y., & Dong, W. (2025). Rethinking Efficient Mixture-of-Experts for Remote Sensing Modality-Missing Classification. *arXiv preprint arXiv:2511.11460*.

[119] Chai, B., Zhou, Q., Nie, X., Qiao, Q., Wu, W., Shi, Y., & Li, X. (2025). Scalable Mixture-of-Experts Attention Feature Pyramid Network for Detection and Segmentation. [CrossRef]

[120] Chen, Y., Jiang, W., & Wang, Y. (2025). FAMHE-Net: Multi-Scale Feature Augmentation and Mixture of Heterogeneous Experts for Oriented Object Detection. *Remote Sensing*, *17*(2), 205. [CrossRef]

[121] Lin, Q., Huang, H., Zhu, D., Chen, N., Fu, G., & Yu, Y. (2025). Multiple Region Proposal Experts Network for Wide-Scale Remote Sensing Object Detection. *IEEE Transactions on Geoscience and Remote Sensing*, *63*, 1-16. [CrossRef]

[122] Li, Y., Li, X., Li, Y., Zhang, Y., Dai, Y., Hou, Q., ... & Yang, J. (2026, March). Sm3det: A unified model for multi-modal remote sensing object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 40, No. 8, pp. 6717-6725). [CrossRef]

[123] Lin, Q., Zhao, J., Du, B., Fu, G., & Yuan, Z. (2021). MEDNet: Multiexpert detection network with unsupervised clustering of training samples. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1-14. [CrossRef]

[124] Zhang, Z., Zhao, E., Jiang, Y., Jie, N., & Liang, X. (2025, June). Challenging Dataset and Multi-Modal Gated Mixture of Experts Model for Remote Sensing Copy-Move Forgery Understanding. In *2025 IEEE International Conference on Multimedia and Expo (ICME)* (pp. 1-6). IEEE. [CrossRef]

[125] Qian, P., Wang, J., Liu, Y., Chen, Y., Wang, P., Deng, Y., Xiao, P., & Li, Z. (2025). Multi-Task Mixture-of-Experts Model for Underwater Target Localization and Recognition. *Remote Sensing, 17*(17), 2961. [CrossRef]

[126] Jiang, C., Osei, K., Yeddula, S. D., Feng, D., & Ku, W.-S. (2025). Knowledge-Guided Adaptive Mixture of Experts for Precipitation Prediction. *arXiv preprint arXiv:2509.11459*.

[127] Sun, W., Tan, Y., Li, J., Hou, S., Li, X., Shao, Y., Wang, Z., & Song, B. (2025). HotMoE: Exploring Sparse Mixture-of-Experts for Hyperspectral Object Tracking. *IEEE Transactions on Multimedia, 27*, 4072-4083. [CrossRef]

[128] Albughdadi, M. (2025). Lightweight Metadata-Aware Mixture-of-Experts Masked Autoencoder for Earth Observation. *arXiv preprint arXiv:2509.10919*.

[129] Li, J., Kang, J., Lu, J., Fu, H., Li, Z., Liu, B., ... & Liu, Z. (2025). Dynamic gating-enhanced deep learning model with multi-source remote sensing synergy for optimizing wheat yield estimation. *Frontiers in Plant Science, 16*, 1640806. [CrossRef]

[130] Zhang, Y., Ru, L., Wu, K., Yu, L., Liang, L., Li, Y., & Chen, J. (2025). SkySense V2: A unified foundation model for multi-modal remote sensing. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9136-9146.

[131] Hanna, J., Scheibenreif, L., & Borth, D. (2026). MAPEX: Modality-aware pruning of experts for remote sensing foundation models. *IEEE Transactions on Geoscience and Remote Sensing, 64*, 1-11. [CrossRef]

[132] Liu, G., He, J., Li, P., Zhong, S., Li, H., & He, G. (2023). Unified transformer with cross-modal mixture experts for remote-sensing visual question answering. *Remote Sensing, 15*(19), 4682. [CrossRef]

[133] Liu, J., Fu, R., Sun, L., Liu, H., Yang, X., Zhang, W., ... & Yang, B. (2026, March). Skymoe: A vision-language foundation model for enhancing geospatial interpretation with mixture of experts. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 40, No. 9, pp. 7168-7178). [CrossRef]

[134] Ngo, B. H., Kim, J. H., Park, S. J., & Cho, S. I. (2022). Collaboration between multiple experts for knowledge adaptation on multiple remote sensing sources. *IEEE Transactions on Geoscience and Remote Sensing, 60*, 1-15. [CrossRef]

[135] Zhang, Y., Li, W., Zhang, M., Han, J., Tao, R., & Liang, S. (2025). SpectralX: Parameter-efficient Domain Generalization for Spectral Remote Sensing Foundation Models. *arXiv preprint arXiv:2508.01731*.

[136] Pasika, H., Haykin, S., Clothiaux, E., & Stewart, R. (1999). Neural networks for sensor fusion in remote sensing. *IJCNN'99. International Joint Conference on Neural Networks. Proceedings, 4*, 2772-2776. [CrossRef]

[137] Aggarwal, V., Nagarajan, K., & Slatton, K. C. (2004). Multiple-model multiscale data fusion regulated by a mixture-of-experts network. *IGARSS 2004. 2004 IEEE International Geoscience and Remote Sensing Symposium, 1*. [CrossRef]

[138] Kong, Y., Yu, S., Cheng, Y., Philip Chen, C. L., & Wang, X. (2025). Joint Classification of Hyperspectral Images and LiDAR Data Based on Candidate Pseudo Labels Pruning and Dual Mixture of Experts. *IEEE Transactions on Geoscience and Remote Sensing, 63*, 1-12. [CrossRef]

[139] He, W., Cai, Y., Ren, Q., Ruze, A., & Jia, S. (2025). Adaptive Expert Learning for Hyperspectral and Multispectral Image Fusion. *IEEE Transactions on Geoscience and Remote Sensing, 63*, 1-15. [CrossRef]

[140] Rossi, L., Bernuzzi, V., Fontanini, T., Bertozzi, M., & Prati, A. (2025). Swin2-MoSE: A new single image supersolution model for remote sensing. *IET Image Processing, 19*(1), e13303. [CrossRef]

[141] Dong, Z., Zhang, Z., Sun, Y., Jiang, H., Liu, T., & Gu, Y. (2026). PhyDAE: Physics-Guided Degradation-Adaptive Experts for All-in-One Remote Sensing Image Restoration. *IEEE Transactions on Geoscience and Remote Sensing*. [CrossRef]

[142] Shen, H., Ding, H., Zhang, Y., Cong, X., Zhao, Z.-Q., & Jiang, X. (2024). Spatial-Frequency Adaptive Remote Sensing Image Dehazing With Mixture of Experts. *IEEE Transactions on Geoscience and Remote Sensing, 62*, 1-14. [CrossRef]

[143] He, X., Yan, K., Li, R., Xie, C., Zhang, J., & Zhou, M. (2024, March). Frequency-adaptive pan-sharpening with mixture of experts. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 38, No. 3, pp. 2121-2129). [CrossRef]

[144] Loyola R., D. G. (2006). Applications of neural network methods to the processing of earth observation satellite data. *Neural Networks, 19*(2), 168-177. [CrossRef]

[145] Li, Z., Chen, X., Li, J., & Zhang, J. (2022). Pertinent Multigate Mixture-of-Experts-Based Prestack Three-Parameter Seismic Inversion. *IEEE Transactions on Geoscience and Remote Sensing, 60*, 1-15. [CrossRef]

[146] Wang, W., Liu, B., Gao, S., Li, J., Zhou, Y., Zhang, S., & Ding, Z. (2025). PhA-MOE: Enhancing Hyperspectral Retrievals for Phytoplankton Absorption Using Mixture-of-Experts. *Remote Sensing, 17*(12), 2103. [CrossRef]

[147] Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., ... & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic

structure. *Ecography, 40*(8), 913-929. [CrossRef]

[148] Valavi, R., Elith, J., Lahoz-Monfort, J. J., & Guillera-Arroita, G. (2018). blockCV: An r package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. *Biorxiv*, 357798. [CrossRef]

[149] Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research, 7*(Jan), 1-30.

[150] Dror, R., Baumer, G., Shlomov, S., & Reichart, R. (2018). The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1383-1392. [CrossRef]

[151] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *2017 IEEE International Conference on Computer Vision (ICCV)*, 618-626. [CrossRef]

[152] Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, 3319-3328.

[153] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why Should I Trust You?: Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135-1144. [CrossRef]

[154] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, 1321-1330.

[155] Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., ... & Snoek, J. (2019). Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems, 32*.

[156] Zhao, Q., Jiang, C., Hu, W., Zhang, F., & Liu, J. (2023). MDCS: More Diverse Experts with Consistency Self-distillation for Long-tailed Recognition. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 11563-11574. [CrossRef]

[157] Li, J., Li, D., Savarese, S., & Hoi, S. (2023, July). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning* (pp. 19730-19742). PMLR.

[158] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735-1780. [CrossRef]

[159] Gu, A., & Dao, T. (2024, May). Mamba: Linear-time sequence modeling with selective state spaces. In *First conference on language modeling*.

[160] Dao, T., & Gu, A. (2024, July). Transformers are SSMs: generalized models and efficient algorithms through structured state space duality. In *Proceedings of the 41st International Conference on Machine Learning* (pp. 10041-10071).

[161] Mehta, S., & Rastegari, M. (2022). Separable self-attention for mobile vision transformers. *arXiv preprint arXiv:2206.02680*.

[162] Cheng, G., Han, J., & Lu, X. (2017). Remote Sensing Image Scene Classification: Benchmark and State of the Art. *Proceedings of the IEEE, 105*(10), 1865-1883. [CrossRef]

Yongchuan Cui received the B.E. degree in data science and big data technology from China University of Geosciences, Wuhan, China, in 2023. He is currently pursuing the master's degree in signal and information processing from Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. His current research interests include image processing, deep learning, multimodal, unified foundation models for earth vision, and geospatial AI. (Email: yongchuancui@gmail.com)

Peng Liu received the M.S. and Ph.D. degrees in signal processing from the Chinese Academic of Science, Beijing, China, in 2004 and 2009, respectively. He is currently an Associate Professor at the Aerospace Information Research Institute, Chinese Academy of Sciences. From May 2012 to May 2013, he was with the Department of Electrical and Computer Engineering, George Washington University, Washington, DC, USA, as a Visiting Scholar. His research is focused on big data, sparse representation, compressive sensing, deep learning and their applications to remote sensing data processing. (Email: liupeng202303@aircas.ac.cn)

Lajiao Chen is an associate professor at Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. Her research interests are focused on geographic information systems and remote sensing techniques and their application to water resource management. (Email: chenlj@radi.ac.cn)