DISSERTATION

TOWARDS SCIENCE-GUIDED KNOWLEDGE DISTILLATION FOR MULTIMODAL LEARNING UNDER CROSS-DOMAIN DISTRIBUTION SHIFT

Submitted by

Abdul Matin

Department of Computer Science

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Summer 2026

Doctoral Committee:

Advisor: Sangmi Lee Pallickara
Co-Advisor: Shrideep Pallickara

Sudipto Ghosh
Jeffrey D. Niemann

ABSTRACT

TOWARDS SCIENCE-GUIDED KNOWLEDGE DISTILLATION FOR MULTIMODAL LEARNING UNDER CROSS-DOMAIN DISTRIBUTION SHIFT

Modern multimodal artificial intelligence has demonstrated unprecedented generalization across broad domains. However, deploying these massive architectures in specialized scientific and operational settings, such as earth observation and precision agriculture, presents prohibitive computational and memory bottlenecks. Furthermore, these models often experience substantial performance degradation when confronted with extreme label sparsity and persistent distribution shifts arising from evolving sampling devices, diverse data modalities, and dynamic operational characteristics. This dissertation explores these challenges and presents a comprehensive methodology to bridge the adaptability gap by advancing the paradigm of Knowledge Distillation (KD). KD serves as a critical mechanism to extract broad intelligence from massive pre-trained foundation models, distilling it into lightweight, domain-specific student networks that respect strict computational limits. Building upon this concept, this work introduces three complementary frameworks designed to handle severe domain mismatches.

First, we develop strategies for heterogeneous domain-shift mitigation that ensure representational consistency across varied spatiotemporal datasets. The proposed cross-resolution distillation framework distills coarse satellite observations into lightweight, locally precise models anchored by sparse in-situ measurements, demonstrating a 40% reduction in soil moisture prediction error. Second, we propose inverse knowledge distillation, which reverses the standard distillation paradigm by transferring architectural and representational knowledge from models trained on abundant, low-spectral-resolution observations to specialized models operating on high-spectral-resolution hyperspectral data. This approach achieves a 26% gain in reconstruction fidelity for 218-band imagery. Third, we establish science-guided domain adaptation through proposed physics-

guided distillation frameworks, which embed domain-theoretic physical laws, specifically the Linear Spectral Mixing Model, directly into the neural architecture as differentiable constraints, improving structural fidelity by 23.6% and generalizing across geographically distinct regions.

Collectively, these contributions enable the training of efficient neural networks that adapt to distribution shifts while remaining grounded in physical theory. Evaluations across diverse Earth observation datasets demonstrate substantial improvements in model robustness, computational efficiency, and representational fidelity over purely data-driven baselines. This research provides a principled path for deploying reliable, lightweight, and physically grounded AI systems in dynamic scientific domains.

ACKNOWLEDGEMENTS

I am deeply indebted to my advisors, Dr. Sangmi Lee Pallickara and Dr. Shrideep Pallickara, for their steadfast guidance, scientific rigor, and unwavering support throughout my doctoral studies. Their mentorship shaped not only this dissertation but my entire approach to research.

I thank my committee members, Dr. Sudipto Ghosh and Dr. Jeffrey D. Niemann, for their thoughtful feedback and constructive criticism at every stage of this work. Their expertise in software engineering and hydrology, respectively, helped ground the technical contributions of this dissertation in broader scientific context.

I am grateful to my labmates and collaborators in the Distributed Systems and Data Science (DS2) group at Colorado State University, particularly Rupasree Dey, Tanjim Bin Faruk, Andrei Bachinin, and Mu Hong, whose intellectual energy and collaboration enriched every project. The discussions we shared, formal and informal alike, were among the most productive of my graduate experience.

I acknowledge the computational resources provided by the Colorado State University High Performance Computing (HPC) cluster and the National Science Foundation ACCESS program, which made the large-scale experiments in this dissertation possible.

Finally, I am grateful to my family for their patience and encouragement across the many years of this undertaking. This work is dedicated to them.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 The Era of Big Scientific Data and Foundational AI

Advancements in global positioning systems (GPS), remote sensing, and computational simulation have led to a significant increase in the collection of spatiotemporal data from diverse fields, including Earth sciences, agriculture, smart cities, and public safety [1]. A prominent example of this data deluge is NASA's Earth Observing System Data and Information System (EOSDIS), which alone stores over 90 petabytes of Earth science data, a repository that is continuously expanding on a daily basis [2]. These newly acquired big spatiotemporal data possess distinct characteristics in terms of volume, velocity, and variety, surpassing the capabilities of conventional computational platforms [3].

The analysis of this geospatial and spatiotemporal data has captured researchers' attention for decades, aiming to uncover significant patterns or make precise predictions that were previously obscured by the sheer scale of the data. Recent progress in deep learning technologies has opened up distinctive opportunities to achieve this. The abundance of large training datasets, coupled with efficient computational architectures like GPUs and distributed learning frameworks such as TensorFlow [4] and PyTorch [5], has led to significant successes in computer vision and natural language processing [6].

This evolution has been fundamentally accelerated by the emergence of large-scale Foundational Models (FMs), such as Vision Transformers (ViT) and Masked Autoencoders (MAE). These models, trained on massive, diverse datasets, exhibit remarkable capabilities for generalized representation learning [7]. Deep learning has already shown promising advancements in spatial and spatiotemporal domains, contributing to areas like Earth system process modeling [8], drug discovery [9], and transportation planning [10, 11].

However, scientific deployment exposes a sharper question: not whether deep models can learn powerful representations, but whether those representations remain valid when the data source, scale, sensor, or physical regime changes. Yet these models are built from statistical patterns alone. This distinction matters because Earth observation is not merely an image-recognition problem at planetary scale; it is a measurement problem governed by physical processes. Earth system processes operate under physical laws, from radiative transfer equations that determine how light reflects off surface materials to hydrological equations that constrain soil moisture dynamics. A model that ignores these constraints may produce outputs that satisfy empirical metrics under in-distribution conditions but fail in entirely predictable ways when sensor configurations change or geographic coverage expands. Grounding representation learning in domain-theoretic physical knowledge is a prerequisite for trustworthy scientific AI, not an optional refinement.

The central challenge is to combine the flexibility of foundation models with the discipline imposed by scientific structure.

1.2 The Adaptability Gap in Scientific Applications

Despite broad gains made by modern multimodal AI models in general domains, applying them to specialized scientific and operational settings reveals a persistent *Adaptability Gap*. The characteristics of scientific data introduce challenges that violate the independent and identically distributed assumptions underlying most conventional machine learning paradigms. In such settings, generalization is not a single property of a trained model. It is a negotiation among sensor physics, spatial scale, temporal cadence, label availability, and computational feasibility.

1.2.1 Data Heterogeneity and Sparsity

While the volume of Earth observation data is massive, high-quality, physically grounded labels are paradoxically scarce. This creates a particularly difficult asymmetry: the easiest data to collect are not always the data most needed for scientific inference. Label sparsity, low spatial or temporal resolution, and a lack of ancillary data create scenarios where the direct application

of standard deep learning techniques often fails [8, 12]. For example, accurate measurements of soil moisture (SM) are necessary for precision agriculture, yet are far too sparse to serve as direct training samples for deep neural networks. Capturing SM at the resolution required to represent true soil variability demands in-situ sensor installations every 1–100 meters, which is impractical at any meaningful scale due to prohibitive cost and maintenance constraints [13]. The result is a familiar scientific compromise: dense but indirect measurements from satellites, and precise but sparse measurements from the ground.

While satellite-based monitoring provides global coverage, it is inherently limited by coarse spatial resolution (often on the scale of kilometers) and attenuation by atmospheric phenomena, dense vegetation, and cloud cover [14]. Deep learning methods for reconstructing cloud-occluded satellite observations can partially alleviate atmospheric data loss [15], yet they cannot resolve the fundamental spatial resolution and spectral coverage constraints inherent to multispectral sensor design. This creates a Cross-Modal Asymmetry, where we possess abundant low-resolution data but lack the dense, high-resolution ground truth required to fine-tune large foundation models effectively. The learning problem is not merely missing data imputation. It is the transfer of useful structure across observational regimes that disagree in resolution, coverage, and physical fidelity.

1.2.2 Computational Constraints and Deployment

Concurrently, the architectural complexity of state-of-the-art models presents a deployment barrier. With the improvement of prediction performance in advanced networks, trainable parameters have crossed into the billions. Deploying such well-performing models with huge computational requirements is often not suitable for real-world scientific applications, particularly in edge devices deployed in remote observational stations. Even when large models are scientifically promising, their size can make them operationally impractical. Scientific AI must therefore be accurate enough to be trusted, small enough to be deployed, and constrained enough to remain physically meaningful.

1.3 Bridging the Gap

To address these multifaceted challenges, this dissertation proposes a principled methodological shift centered on **Domain-Adaptive Knowledge Distillation (DA-KD)**. The unifying premise is that adaptation should not be left to fine-tuning alone. Instead, useful knowledge must be deliberately selected, transferred, and constrained as it moves from source domains to target domains.

Knowledge Distillation (KD) [16] is traditionally viewed as a model compression technique where a high-capacity model (the teacher) transfers its learned representations to a smaller, more efficient model (the student). However, in the context of scientific AI, we position KD not merely as a compression tool, but as a sophisticated mechanism for cross-modal feature alignment, representational stability, and domain adaptation. In this dissertation, the teacher is not merely a larger network, and the student is not merely a compressed copy. The teacher provides a source of structure; the student learns how to express that structure under the constraints of the target scientific setting.

Knowledge transfer is an effective technique for bridging the gap between abundant, low-fidelity data and sparse, high-fidelity ground truth. A deep learning model trained on ancillary data that is widely available can be adapted to a target dataset even without missing annotation [17]. For instance, recent geospatial AI foundation models pretrained on massive satellite data (such as the HLS foundation model developed by IBM and NASA [18]) can be effective for knowledge distillation. However, reusing the massive training data for every downstream task is impractical. Distillation offers a practical alternative: it allows the representational benefits of large-scale pretraining to be reused without requiring the full pretraining corpus, architecture, or deployment cost.

1.3.1 Science-Guided Knowledge Transfer

A second limitation of ordinary transfer is that statistical alignment alone does not guarantee scientific validity. Two representations may appear close in feature space while still violating the physical process that generated the measurements. Beyond model compression and domain

alignment lies a less-explored challenge: ensuring that learned representations remain physically plausible [19]. The solution space in Earth observation is not arbitrary. Optical reflectance spectra are governed by the material composition of the Earth's surface, and any valid reconstruction must respect compositional constraints such as the non-negativity of fractional abundances and the requirement that material contributions sum to unity across each pixel [20].

Science-Guided Knowledge Distillation (SGKD) addresses this by embedding domain-theoretic physical laws within the neural architecture as differentiable constraints, rather than treating them as auxiliary regularization terms that can be overridden by statistical pressure [21]. In the spectral domain, the Linear Spectral Mixing Model (LSMM) offers a principled mechanism: any observed reflectance spectrum is expressible as a linear combination of endmember signatures weighted by fractional abundances that satisfy non-negativity and sum-to-one constraints [22]. Implemented as a differentiable decoder layer, this constraint prevents the student model from producing physically implausible outputs regardless of what training statistics reward. This shift from soft empirical fitting to physically constrained representation learning is central to the dissertation's notion of science-guided distillation.

Combining distillation-based knowledge transfer with architectural physical constraints advances this work beyond conventional transfer learning. The resulting frameworks must simultaneously satisfy statistical objectives and physical invariants, a requirement that forms the central scientific motivation behind the third research thrust of this dissertation. In other words, the goal is not only to make the student imitate the teacher, but to make the student inherit what is scientifically reusable from the teacher while rejecting what is sensor-specific, scale-specific, or physically implausible.

1.4 Research Thrusts

This work is structured around three complementary research thrusts designed to enable the transfer of intelligence across disparate data characteristics, moving the field from foundational model adaptation toward the integration of domain-theoretic knowledge for robust generalization.

The three thrusts differ in their immediate technical targets, but they share a common purpose: each uses distillation to stabilize learning under a different form of scientific mismatch.

1. **Heterogeneous Domain-Shift Mitigation:** We propose strategies to ensure representational consistency across heterogeneous spatiotemporal datasets. This enables the transfer of knowledge from models trained on high-fidelity sampling processes to real-world settings characterized by noisy and incomplete observations. This thrust asks whether coarse but spatially continuous knowledge can be made useful for localized prediction where ground truth is accurate but sparse.

2. **Inverse Knowledge Distillation:** We explore methods to transfer architectural and representational knowledge from models trained on low–spectral-resolution observations (abundant data) to models operating on high–spectral-resolution data (sparse data). This addresses the dimensional scaling challenge inherent in moving from multispectral to hyperspectral analysis. This thrust asks whether a model trained on spectrally limited but abundant observations can serve as a scaffold for learning richer hyperspectral representations.

3. **Science-Guided Domain Adaptation:** We incorporate domain-theoretic scientific knowledge (such as physical mixing models) to introduce new data modalities into the modeling environment. This enhances a model's ability to generalize in complex scientific systems by grounding statistical learning in physical reality. This thrust asks whether physical laws can be made part of the transfer mechanism itself, rather than added afterward as interpretive justification.

1.4.1 Problem Formulation

To realize robust adaptation for scientific AI, three primary, interrelated challenges must be formally defined. These challenges map directly to the three research thrusts of this dissertation.

P1: Latent Representational Drift in Heterogeneous Domain-Shifts

(Aligns with Research Thrust 1)

Domain shifts in scientific data are not monolithic but multi-axial, involving co-varying changes in spectral characteristics, spatial context, and statistical distributions arising from evolving operational conditions. These problems are best understood as failures of transfer under constraint: the source domain contains useful information, but that information must be reshaped before it can support the target task. When models are adapted to new heterogeneous domains (e.g., transferring from high-fidelity lab settings to noisy real-world observations), their latent representations often drift. This leads to a loss of the intrinsic structural and spectral coherence of the target data, which is critical because slight changes in learned spectral signatures directly translate to severe errors in physical parameter estimation.

Mitigating this *Representational Drift* necessitates novel architectural strategies that ensure consistency. We propose developing principled techniques to ensure representational stability across heterogeneous spatiotemporal datasets. This involves introducing mechanisms such as Spectral Adaptation Units (SAU) to project heterogeneous modalities into a unified latent representation suitable for cross-domain comparison. Similarly, we aim to enforce structural coherence (via SSIM loss) and spectral fidelity (via Spectral Information Divergence loss) to stabilize representations, mitigating inconsistencies that arise when moving from high-fidelity sampling to incomplete observational settings.

P2: Cross-Modal Asymmetry and the Challenge of Inverse Knowledge Distillation

(Aligns with Research Thrust 2)

Scientific data often involves integrating information from sensors operating across heterogeneous modalities and widely varying scales, leading to profound asymmetries in dimensionality, resolution, spectral content, and noise profiles. Traditional Knowledge Distillation (KD) assumes a symmetric transfer, typically from a large, high-fidelity teacher to a smaller student on the same domain. However, scientific contexts often demand **Inverse Knowledge Distillation**. This critical variant involves leveraging generalized knowledge from a low-dimensional teacher (e.g., pretrained on 6-band multispectral data) to guide a high-dimensional student (e.g., a 218-band hyperspectral model).

This constitutes a substantial dimensional scaling challenge. The low-dimensional knowledge must serve as a stable scaffold for learning more complex, domain-specific features without restricting the student's capacity to recognize nuances in the expanded feature space. Our investigations directly address this by developing a framework for Inverse KD that guides a high-dimensional student using a low-dimensional teacher. We also explore applying KD to bridge the gap between coarse-resolution satellite predictions and high-resolution outputs for lightweight student models anchored by sparse in-situ measurements.

P3: Lack of Physical Grounding and Science-Guided Adaptation

(Aligns with Research Thrust 3)

The reliance on purely data-driven foundational models carries the intrinsic risk of statistical optimization based on brittle, context-dependent correlations. In high-stakes scientific decision systems, a lack of physical grounding undermines model trust, reliability, and interpretability. To enhance a model's ability to generalize in complex systems, we must incorporate domain-theoretic scientific knowledge directly into the modeling environment.

The critical problem is how to formalize scientific principles (e.g., spectral physics, flow conservation) as differentiable constraints. Our proposed **Science-Guided Domain Adaptation (SGDA)** framework tackles this by using physics-informed constraints to steer learning. This approach explores integrating the Linear Spectral Mixing Model (LSMM) and Spectral-Angle-Aware reconstruction (SAM) as constraints within self-supervised processes (such as ViT-MAE). This mechanism ensures mathematical stability and explainability, enabling the introduction of new data modalities grounded in scientific theory rather than mere statistical coincidence.

1.4.2 Core Research Questions

The research trajectory is guided by four core research questions that collectively span data exploitation, model efficiency, cross-modal knowledge transfer, and science-guided adaptation. The research questions therefore move from availability, to efficiency, to modality transfer, to

physical grounding. This ordering mirrors the dissertation's progression from practical adaptation toward principled scientific generalization.

RQ1: Exploiting Low-Resolution Earth Observation Data Under Sparse Labels

How can low-resolution Earth observation data be effectively leveraged to support accurate scientific modeling when high-quality ground-truth labels are sparse and the target domain distribution differs from the source?

Low-resolution satellite products such as SMAP soil moisture at 3 km resolution are globally available but spectrally and spatially coarse relative to measurements needed for local precision. The question is how to extract physically meaningful representations from these low-fidelity sources and use them to supervise learning under sparse ground-truth conditions, without overfitting to sensor-specific artifacts or spatial aggregation effects.

RQ2: Knowledge Distillation for Computationally Efficient Inference

How can Knowledge Distillation be applied to compress large scientific models into lightweight student models that retain predictive accuracy while meeting the inference constraints of operational Earth observation deployments?

Foundation models with millions to billions of parameters are impractical on edge sensors and resource-constrained observational platforms. This question targets distillation strategies that transfer the teacher's learned representations into a student at a fraction of the computational cost, with efficiency measured concretely in inference latency, memory footprint, and prediction accuracy on domain-shifted data.

RQ3: Transferring Knowledge from Multispectral to Hyperspectral Foundation Models

How can knowledge distillation mechanisms be engineered to transfer representations from foundation models trained on abundant multispectral observations to student models operating on high-dimensional hyperspectral imagery?

Multispectral foundation models such as Prithvi-100M [18] are pretrained on 6-band HLS data at a 2–3 day revisit frequency, while hyperspectral sensors such as EnMAP provide 218 usable spectral bands at a 27-day revisit cycle. The design challenge is a transfer pathway that uses the low-dimensional teacher's spatial and structural representations as a scaffold for the high-dimensional student, without restricting the student's ability to capture fine-grained spectral variance.

RQ4: Integrating Scientific Knowledge for Interpretability and Generalization

How can domain-theoretic scientific knowledge be formalized as differentiable architectural constraints within foundation model objectives to improve interpretability and geographic generalization under distribution shift?

This question addresses the physical grounding problem. Embedding the Linear Spectral Mixing Model (LSMM) as a differentiable decoder component constrains the model's hypothesis space to physically plausible material compositions, preventing statistically convenient but spectroscopically invalid reconstructions. The expected outcome is a model whose representations generalize across geographically distinct regions without additional fine-tuning, because material-level physical invariants govern its behavior rather than region-specific statistical correlations.

1.4.3 Fundamental Research Challenges

Successful execution of this research agenda requires overcoming several deep, non-trivial challenges that reflect the cutting edge of domain adaptation and knowledge transfer research.

Challenge 1: Fidelity Preservation of High-Frequency Information

Scientific phenomena often rely on localized, fine-grained variations, including high-frequency components such as sharp spectral absorption features. Research indicates that during KD, lightweight student models exhibit a tendency toward spectral imbalance, often overfitting high-frequency noise or underfitting critical transients [23, 24]. The challenge is to devise distillation losses that

explicitly decouple spectral features, enforcing frequency-aligned knowledge transfer to ensure the student retains critical local physical responses.

Challenge 2: The Dimensional Barrier and Stability in Inverse KD

While our studies have demonstrated the feasibility of Inverse KD for dimensional scaling, the inherent complexity remains. The core difficulty lies in ensuring that the distilled low-dimensional knowledge provides sufficient guidance without paradoxically restricting the high-dimensional student's capacity to learn the subtle, domain-specific features embedded in the expanded data. Achieving optimal knowledge transfer across a $36\times$ dimensional increase requires precisely calibrated distillation temperatures and specialized spatio-spectral loss functions that manage the asymmetric feature importance between the teacher and student models.

Challenge 3: Integrating Scientific Theory for Multimodal Disentanglement

Extending Science-Guided Domain Adaptation (SGDA) to sophisticated Multimodal Foundation Models (MFMs) presents a fundamental integration challenge. Direct adaptation often results in a "mixture of semantic and domain-specific information," leading to significant domain bias [25]. The fundamental challenge is formalizing physical constraints that act as a disentanglement mechanism, constraining the model's reasoning using precise visual and physical signals (e.g., LSMM constraints) to filter statistical noise and suppress domain-irrelevant semantic biases.

1.5 Dissertation Organization

Together, these chapters develop the dissertation's central argument: robust scientific AI requires not only larger models, but better mechanisms for transferring, constraining, and interpreting what those models learn. The remainder of this dissertation is organized as follows:

- **Chapter 2: Background and Related Work** establishes the technical and scientific context for this research. It reviews the characteristics of hyperspectral and multispectral remote sensing modalities, the landscape of Earth observation foundation models, and the foun-

dational machine learning paradigms of Knowledge Distillation (KD), Domain Adaptation (DA), and Science-Guided Domain Adaptation.

- **Chapter 3: Methodology** details the technical frameworks developed for this dissertation to address three critical barriers: (1) heterogeneous domain shifts inherent in multi-resolution sensing environments; (2) cross-modal asymmetry between abundant low-fidelity and scarce high-fidelity data; and (3) the lack of physical grounding in purely data-driven architectures.

- **Chapter 4: Experimental Setup and Results** reports experimental protocols, validates the proposed methodologies using multiple Earth observation datasets, and presents comprehensive results analysis across the three research thrusts.

- **Chapter 5: Conclusion** summarizes the dissertation's contributions, discusses broader impact and limitations, and outlines directions for future research.

Chapter 2

Background and Related Work

The methodological framework of this dissertation lies at the intersection of remote sensing, efficient deep learning, and domain adaptation. This chapter establishes the technical and scientific context relevant to this work. It begins with the characteristics of optical remote sensing modalities (Section 2.1) and the landscape of Earth observation foundation models (Section 2.2). The chapter then reviews Knowledge Distillation (Section 2.3), Domain Adaptation (Section 2.4), and Science-Guided Domain Adaptation (Section 2.5), analyzing their capabilities and limitations in the context of scientific data constraints.

2.1 Remote Sensing Modalities: Hyperspectral and Multispectral Imaging

Earth observation (EO) systems acquire reflected and emitted electromagnetic radiation from the Earth's surface using imaging sensors aboard satellites, aircraft, and uncrewed platforms. The two dominant families of optical sensors, multispectral and hyperspectral imagers, differ fundamentally in how they sample the electromagnetic spectrum and what physical properties they can retrieve from a scene [26].

2.1.1 Multispectral Imaging

Multispectral sensors capture reflected radiation in a small number of broad, non-contiguous spectral bands, typically between 4 and 36 [26]. Each band integrates reflected energy across a wavelength range spanning tens to hundreds of nanometers. Prominent examples include NASA's Landsat series and the European Space Agency's Sentinel-2, whose observations are combined into the Harmonized Landsat Sentinel-2 (HLS) dataset [27]. HLS provides 6 spectral bands covering the visible and short-wave infrared spectrum (Blue: 443 nm; Green: 560 nm; Red: 665 nm; NIR:

865 nm; SWIR-1: 1610 nm; SWIR-2: 2190 nm), at 30 m spatial resolution and a 2–3 day revisit frequency.

The broad spectral bands of multispectral sensors limit their ability to resolve narrow spectral features, such as absorption troughs associated with specific mineral compositions or vegetation stress states. However, high temporal revisit frequency and global spatial coverage make multispectral data well-suited as the pretraining basis for large geospatial foundation models [18, 28].

2.1.2 Hyperspectral Imaging

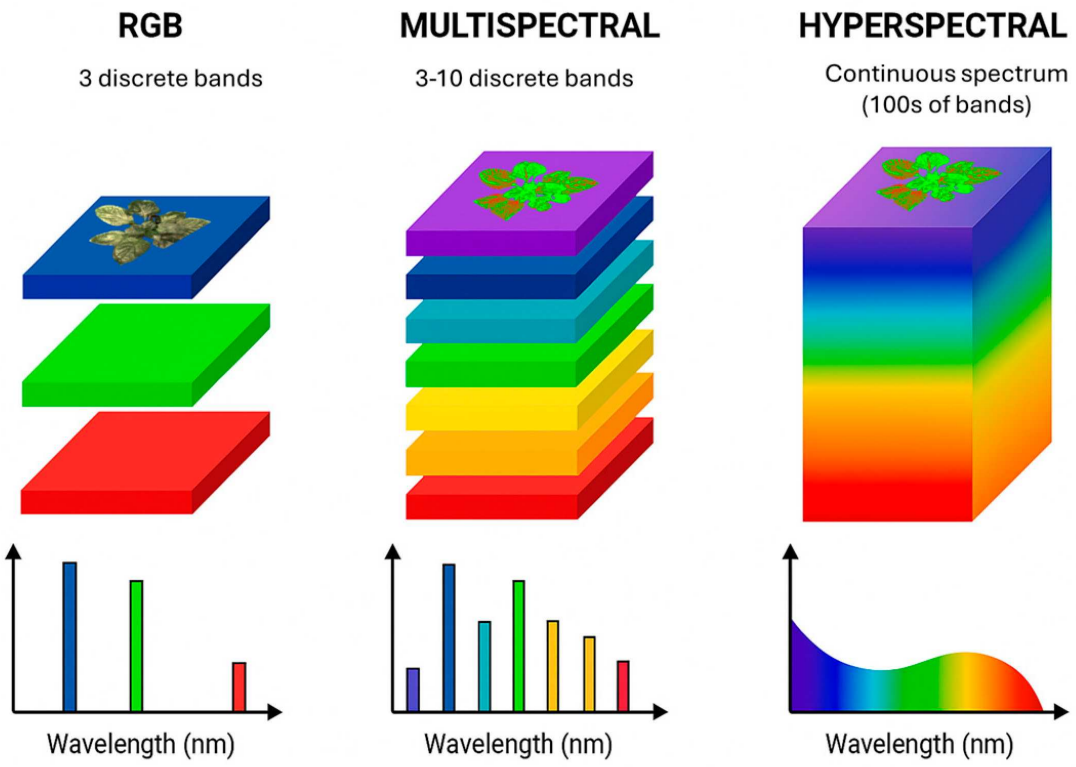

Figure 2.1: Spectral sampling comparison across three optical imaging modalities. Standard RGB sensors capture three broad visible bands. Multispectral sensors extend coverage to 4–36 wider bands spanning the visible and infrared spectrum. Hyperspectral sensors record hundreds of adjacent narrow bands ($<$10 nm each), assigning a complete reflectance spectrum to every pixel for per-pixel material discrimination. Image credit: [29].

Hyperspectral sensors collect measurements across a much larger number of adjacent, narrow spectral bands, typically 100 to 300 channels with individual bandwidths below 10 nm [26].

This dense spectral sampling assigns a continuous reflectance spectrum to every pixel in the scene (Figure 2.2), enabling per-pixel discrimination of surface materials based on their spectral fingerprints. Spectral libraries maintained by organizations such as the USGS provide reference spectra for thousands of mineral, soil, and vegetation types against which observed pixel spectra can be compared [30].

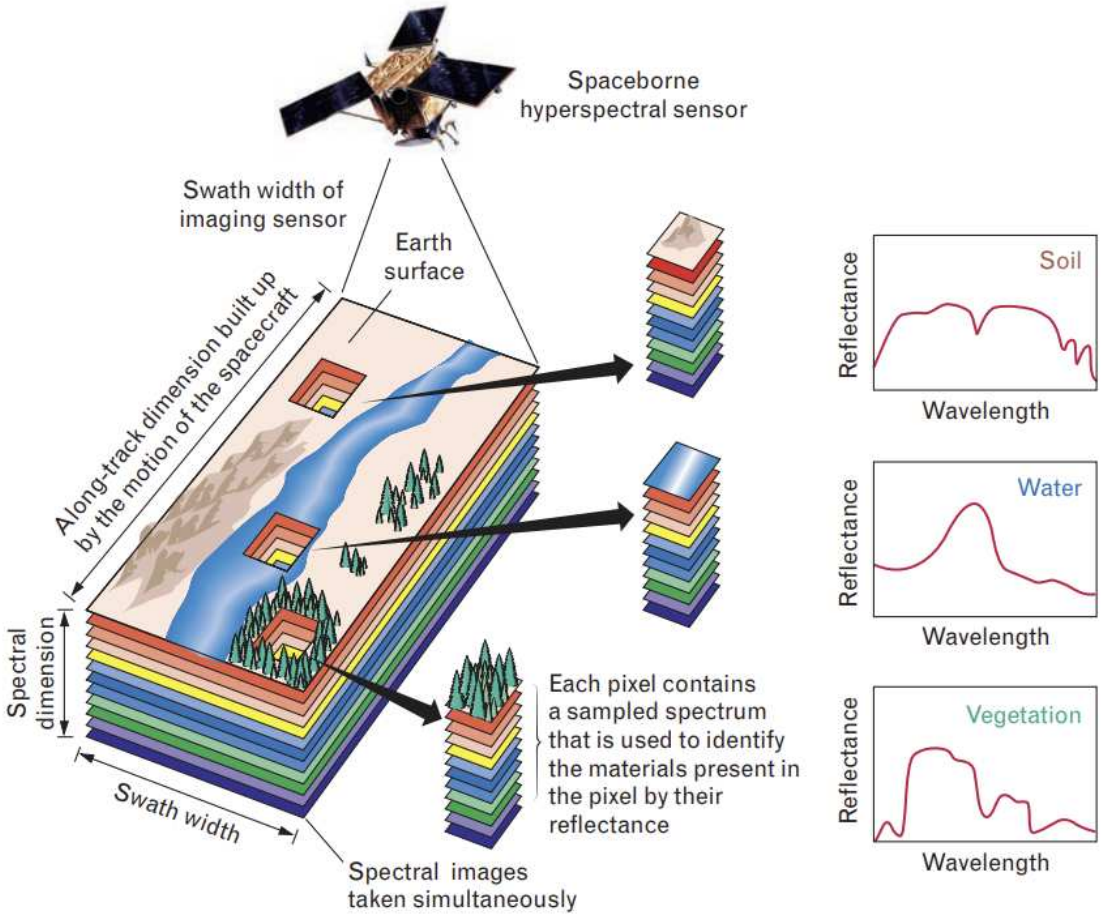

Figure 2.2: Principles of hyperspectral imaging. Unlike broadband multispectral sensors that capture a small number of wide spectral bands, hyperspectral imagers record a continuous narrow-band reflectance spectrum for each spatial pixel, enabling material identification from spectral fingerprints. Image credit: [26].

The Environmental Mapping and Analysis Program (EnMAP) satellite [31], which provides the primary hyperspectral data in this dissertation, acquires 224 spectral bands across the 420–2450 nm wavelength range at 30 m spatial resolution. After excluding six noisy bands in atmo-

spheric water-vapor absorption windows, 218 usable bands remain. EnMAP's revisit frequency is 27 days, compared to the 2–3 day cadence of HLS. This combination of high spectral resolution and low temporal availability defines the data scarcity challenge that motivates the inverse knowledge distillation approach in Thrust 2.

2.1.3 Spectral Data Cubes

Hyperspectral imagery is organized as three-dimensional spectral data cubes comprising two spatial dimensions and one spectral dimension (Figure 2.3). Successive cross-track scans at increasing wavelengths are stacked along the spectral axis to form the cube. Extracting the reflectance values at any spatial location along the spectral axis yields a continuous spectral signature that characterizes the material composition of that pixel [26]. This architecture is both the source of hyperspectral data's interpretive power and the root of its computational demands: a single EnMAP scene tiled at 224×224 pixels with 218 spectral channels contains approximately 10.9 million spectral measurements per tile.

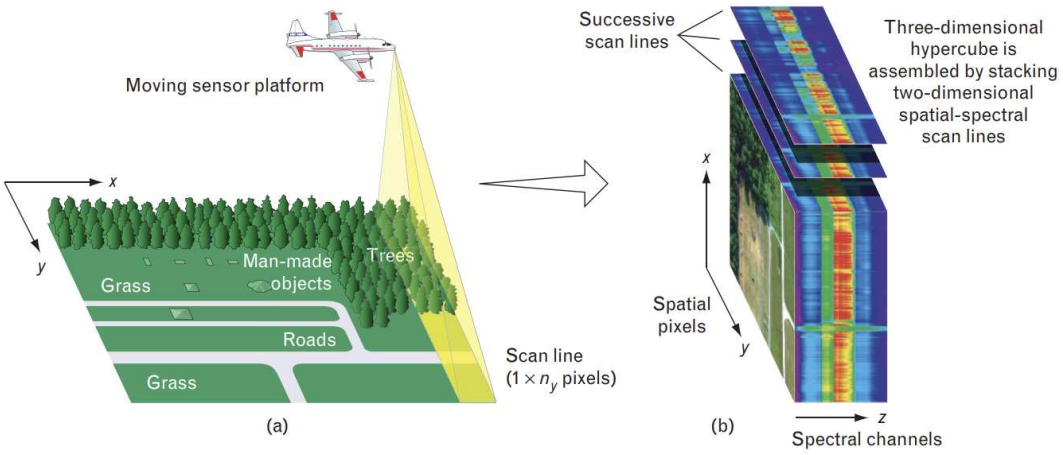

Figure 2.3: Three-dimensional hyperspectral data cube structure. Two spatial axes (across-track and along-track) define the scene footprint, while the spectral axis stacks 218 narrow-band reflectance images from 420 nm to 2450 nm. Measuring reflectance along the spectral axis at any pixel location yields a continuous spectral signature that encodes the material composition of that ground location. Image credit: [26].

Table 2.1: Comparison of multispectral and hyperspectral imaging characteristics for the HLS and EnMAP sensors used in this dissertation.

Property	Multispectral (HLS)	Hyperspectral (EnMAP)
Spectral bands	6	218 (usable)
Band width	20–200 nm	$<$10 nm
Wavelength range	443–2201 nm	420–2450 nm
Spatial resolution	30 m	30 m
Revisit frequency	2–3 days	27 days
Data availability	High	Low
Spectral fingerprinting	Limited	Full
Foundation model training	Feasible	Challenging

2.1.4 Applications of Hyperspectral Imaging

The narrow-band spectral fidelity of hyperspectral sensors supports Earth observation tasks that broadband multispectral sensors cannot address:

- **Vegetation and Crop Analysis:** Narrow spectral bands resolve chlorophyll absorption features, leaf water content dips, and lignin/cellulose absorption signatures, enabling crop stress detection, species-level plant identification, and biophysical variable retrieval [26].
- **Soil and Mineral Mapping:** Clay mineralogy, soil organic carbon, iron oxide content, and carbonate minerals produce diagnostic absorption features in the 2000–2500 nm range, accessible only to hyperspectral sensors.
- **Land Cover Classification:** Downstream classification benchmarks used in this dissertation, including the NLCD and USDA Cropland Data Layer (CDL), benefit from richer spectral inputs for separating spectrally similar land cover categories.
- **Coastal and Hydrological Monitoring:** Narrow-band measurements support shallow-water bathymetry, coral reef assessment, and inland water quality retrieval [26].

Table 2.1 summarizes the technical differences between the HLS multispectral and EnMAP hyperspectral sensors used in this dissertation.

2.1.5 The Spectral Domain Gap

The dimensional disparity between multispectral and hyperspectral observations creates a *spectral domain gap* that direct transfer learning cannot bridge. A foundation model pretrained on 6-band HLS inputs learns representations optimized for 6-channel spectral tokens; deploying it on 218-band EnMAP data requires a $36\times$ expansion of spectral channels with no direct architectural correspondence. Standard channel-wise feature alignment strategies fail because most of the EnMAP spectral range has no counterpart band in HLS. Closing this gap is the core technical problem addressed by the inverse knowledge distillation framework described in Chapter 3.

2.2 Foundation Models and Earth Observation

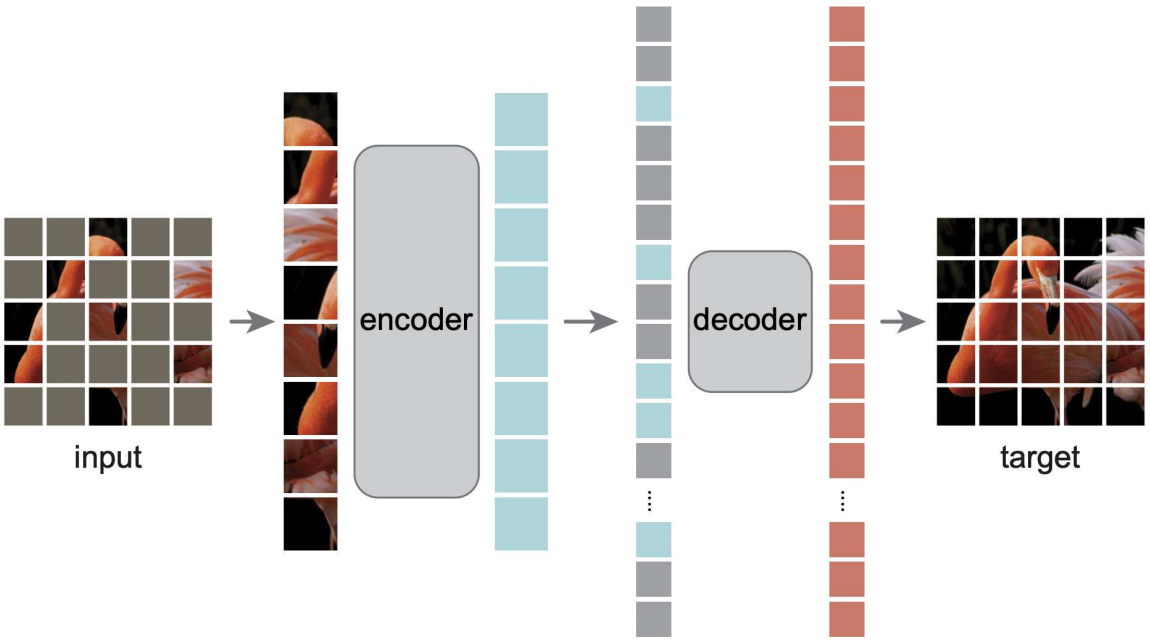

Figure 2.4: Masked Autoencoder (MAE) pretraining framework for self-supervised representation learning [32]. A large fraction (typically 75%) of input image patches are masked at random. The Vision Transformer (ViT) encoder processes only the visible patches, while a lightweight decoder reconstructs the pixel content of the masked regions from the latent tokens. This aggressive masking forces the encoder to learn high-level spatial and structural representations without any labeled data, producing features that transfer effectively to downstream Earth observation tasks.

Foundation models are large neural networks pretrained on broad, unlabeled datasets, with the expectation that the learned representations transfer effectively to diverse downstream tasks via fine-tuning or zero-shot inference (Figure 2.4) [7]. Within Earth observation (EO), this paradigm has motivated the development of geospatial foundation models like Prithvi [18] trained on continental- to global-scale satellite archives, fundamentally shifting how remote sensing tasks are approached (Figure 2.5).

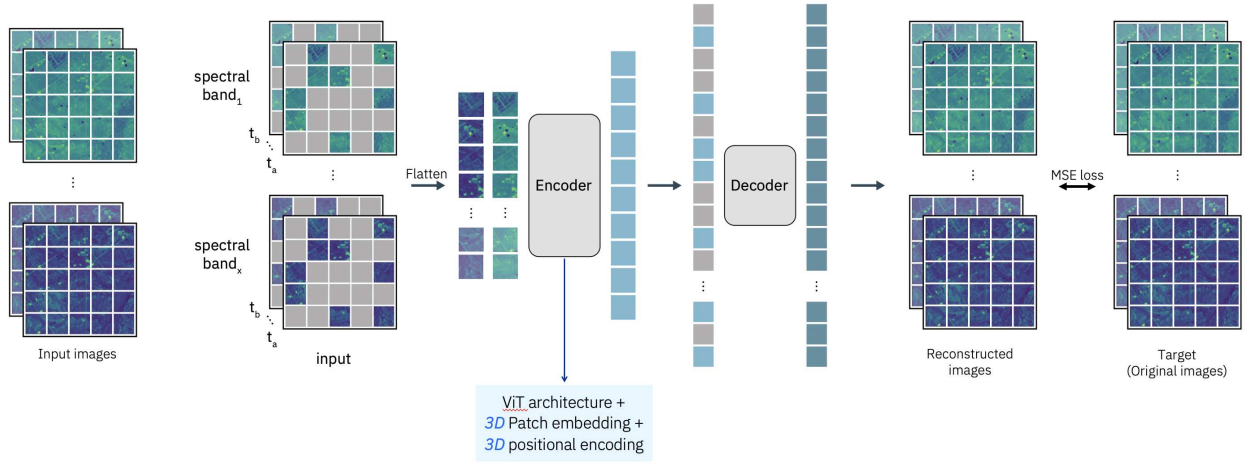

Figure 2.5: Architectural overview of the NASA-IBM Prithvi Geospatial Foundation Model framework. The multi-spectral, multi-temporal input remote sensing imagery is structured as a spatiotemporal tensor $\mathbf{X} \in \mathbb{R}^{B \times C \times T \times H \times W}$ [18]. Input patches are converted into tokenized representations using a 3D convolutional patch embedding layer. Self-supervised pre-training is driven by a Masked Autoencoder (MAE) strategy [32], where a high ratio of spatiotemporal tokens are masked out prior to encoding. Unmasked tokens are processed through a heavy Vision Transformer (ViT) encoder featuring joint spatial and temporal attention mechanisms. Metadata variables, including center latitude, longitude, and day-of-year (DoY), are embedded independently via 2D sine/cosine positional encodings and fused with the latent tokens. Finally, a lightweight transformer decoder reconstructs the missing pixels to compute the Mean Squared Error (MSE) loss. Image Credit: [18].

2.2.1 The Necessity and Mechanism of Foundation Models

Historically, deep learning applied to remote sensing relied heavily on supervised learning, which requires massive, meticulously annotated datasets. In Earth observation, acquiring high-quality labels (e.g., pixel-level land cover, crop types, or soil moisture) is prohibitively expensive,

time-consuming, and geographically biased, often requiring extensive field surveys or expert interpretation. Conversely, satellite constellations generate petabytes of raw, unlabeled imagery daily.

Foundation models resolve this data bottleneck through Self-Supervised Learning (SSL), effectively turning the raw data into its own supervisory signal. By learning to solve a "pretext task"—such as reconstructing missing parts of an image—the model is forced to internalize the underlying physical, spatial, and temporal dynamics of the Earth's surface. Once this generalized "foundation" of knowledge is established, the model requires substantially less labeled data to be fine-tuned for specific, localized downstream tasks.

2.2.2 Architecture and Pretraining

The dominant architecture for EO foundation models combines the Vision Transformer (ViT) [33] encoder with a Masked Autoencoder (MAE) pretraining objective [32].

During pretraining, the MAE algorithm masks a large fraction of input image patches—typically up to 75%. The encoder processes only the visible patches, while a lightweight decoder attempts to reconstruct the masked, missing content based solely on the latent representations. This aggressive masking strategy forces the network to look beyond trivial local interpolations; it must learn deep, spatially and spectrally coherent representations. In the context of satellite imagery, this means the model implicitly learns semantics such as how rivers flow, how vegetation indices correlate with specific spectral bands, and how landscapes change across temporal stacks.

2.2.3 Existing Geospatial Foundation Models

The rapid acceleration of this field has led to several prominent geospatial foundation models, each demonstrating strong performance on multispectral satellite data:

- **Prithvi-100M** [18]: A ViT-Large MAE developed jointly by IBM and NASA, pretrained on over one million Harmonized Landsat Sentinel-2 (HLS) tiles. Prithvi accepts 6-band multispectral inputs at three temporal timestamps (18 input channels total), capturing multi-temporal land surface dynamics at a continental scale.

- **Prithvi 2.0 (Prithvi-EO)** [34]: The evolution of the original Prithvi model, scaled up to 300M and 600M parameter variants. Prithvi 2.0 introduces expanded temporal context windows, improved spatial resolution handling, and broader geographic diversity in its pretraining data, significantly enhancing zero-shot and few-shot capabilities.

- **DOFA (Dynamic One-For-All)** [35]: A recent and highly flexible foundation model designed to natively handle varying spectral bands across different satellite sensors. DOFA utilizes a dynamic weight generation mechanism based on the wavelengths of the input bands, allowing it to ingest data from diverse sensors without requiring structural architectural changes.

- **Clay** [36]: An open-source, spatio-temporal foundation model that directly embeds latitude, longitude, and timestamp data into the tokenization process, allowing the MAE to explicitly condition its representations on geography and time.

- **SatMAE** [28]: A ViT-based MAE pretrained on temporally stacked multispectral imagery, introducing spectral-temporal pretext tasks that explicitly exploit the temporal dimension of satellite archives.

- **ScaleMAE** [37]: A resolution-aware MAE that conditions the decoder on the ground sampling distance of the input, enabling multi-resolution generalization during fine-tuning.

- **SpectralGPT** [38]: A generative pretraining approach with a 3D patch tokenization scheme explicitly designed to accommodate higher spectral band counts than standard RGB-based foundation models.

Despite these advancements, a common limitation remains: the vast majority of these models (with the partial exception of DOFA) are pretrained exclusively on multispectral data. Directly applying them to hyperspectral inputs requires resolving a massive spectral domain gap. Recent efforts have begun addressing this directly: spatially adaptive masking based on gradient-derived

saliency concentrates MAE reconstruction effort on high-variance hyperspectral patches, improving pretraining fidelity without modifying the encoder architecture [39], while the TerraMAE framework extends adaptive masking to learn joint spatial-spectral representations from EnMAP observations at scale [40].

2.2.4 Challenges for Hyperspectral Adaptation

Adapting established multispectral foundation models to hyperspectral data introduces three critical structural challenges [18]:

1. **Data Scarcity:** Hyperspectral missions (e.g., EnMAP, PRISMA) typically feature long revisit cycles (e.g., 27 days for EnMAP) and narrow geographic footprints. This yields pretraining archives that are orders of magnitude smaller than the globally available HLS data used to train models like Prithvi.

2. **Spectral Dimensionality:** Adapting a standard ViT-MAE to process 218 input channels increases the spectral parameter count by approximately $36\times$ relative to a standard 6-band model. This drastic increase raises computational costs and introduces a severe risk of overfitting, particularly given the small size of hyperspectral datasets.

3. **Architectural Incompatibility:** Existing geospatial foundation models do not natively support hundreds of input channels. Adaptation requires novel projection strategies or distillation pathways that can transfer knowledge across the spectral dimension without discarding the rich spatial and semantic representations learned during pretraining.

These challenges motivate the inverse knowledge distillation strategy explored in Thrust 2: rather than pretraining a hyperspectral foundation model from scratch—which is computationally and data-prohibitive—the proposed approach uses the representations of Prithvi as a structural scaffold. By transferring spatial and structural knowledge across the spectral domain gap, we can effectively initialize and guide high-fidelity hyperspectral representation learning.

2.3 Knowledge Transfer via Distillation

Knowledge Distillation (KD) is a widely used technique for compressing model capacity while retaining representational fidelity. Originally proposed by Hinton et al. [16], the paradigm involves training a compact "student" network to mimic the behavior of a high-capacity "teacher" network (Figure 2.6). The knowledge transferred can be categorized into response-based, feature-based, and relation-based distillation.

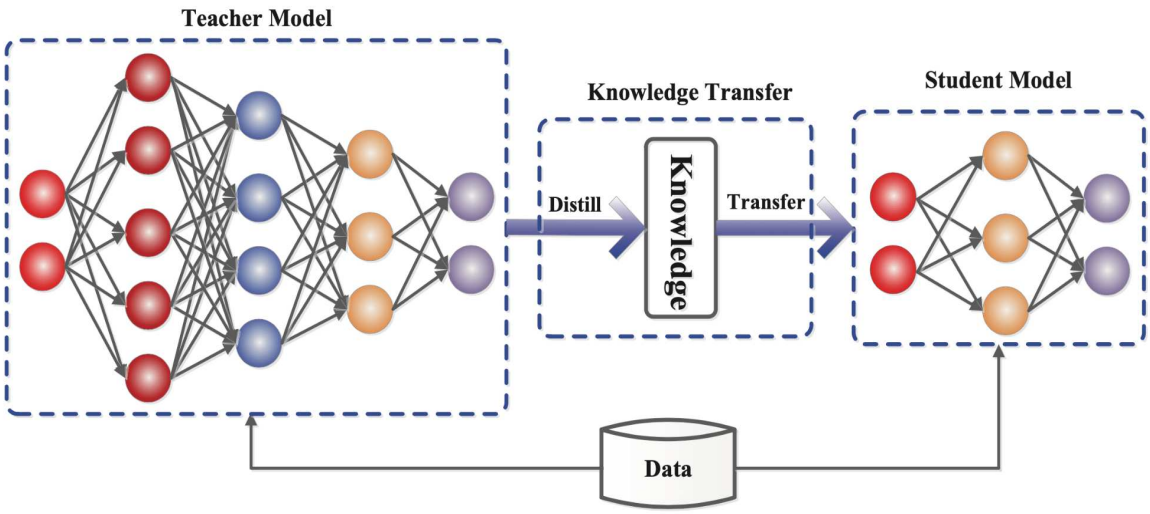

Figure 2.6: Standard Knowledge Distillation (KD) framework [16]. A high-capacity teacher network (left) supervises a compact student network (right) by transferring soft output probability distributions in addition to hard ground-truth labels. The student minimizes a weighted combination of cross-entropy loss against true labels and KL divergence against the teacher's temperature-scaled logits, which encode inter-class relationship structure invisible in one-hot label encodings.

2.3.1 Response-Based Knowledge Distillation

Response-based KD leverages the neural output of the teacher's final layer. It is the most straightforward form of distillation, where the student learns to mimic the soft probability distributions produced by the teacher (Figure 2.7). **Formulation:** The loss function combines standard cross-entropy with a Kullback-Leibler (KL) divergence term to align the student's logits (z_{student}) with the teacher's softened logits (z_{teacher}):

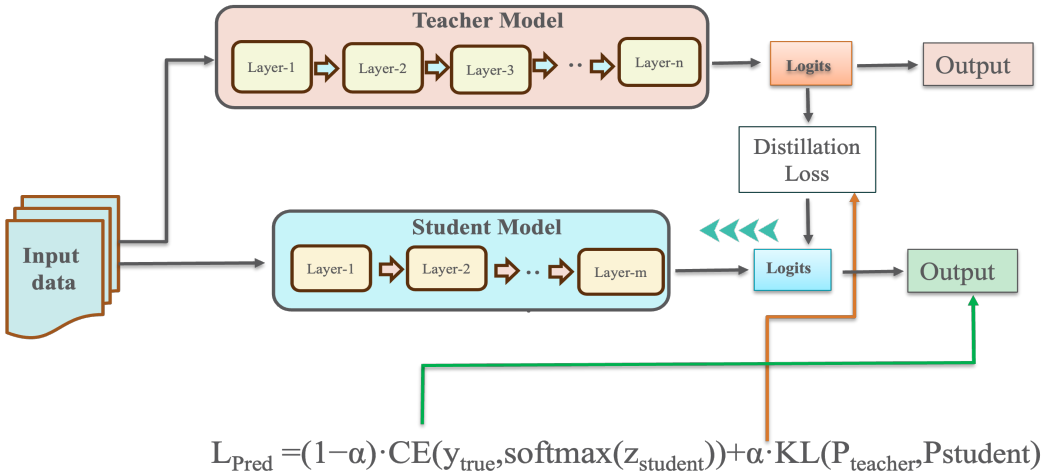

Figure 2.7: Response-based Knowledge Distillation. The student is jointly supervised by the ground-truth one-hot label (cross-entropy loss, CE) and the teacher's temperature-softened output distribution (KL divergence). The temperature parameter T amplifies small inter-class probability differences in the teacher's output, exposing "dark knowledge" about similarity relationships between categories that cannot be recovered from sparse label annotations alone.

$$L_{Pred} = (1-\alpha) \cdot \text{CE}(y_{\text{true}}, \text{softmax}(z_{\text{student}})) + \alpha \cdot \text{KL}(P_{\text{teacher}} \| P_{\text{student}}) \tag{2.1}$$

Here, α balances the terms. The distributions P_{teacher} and P_{student} are calculated using a temperature parameter T to soften the probability distribution, revealing the "dark knowledge" or inter-class relationships learned by the teacher:

$$P_{\text{teacher}}(i) = \frac{\exp(z_{\text{teacher},i}/T)}{\sum_j \exp(z_{\text{teacher},j}/T)} \tag{2.2}$$

$$P_{\text{student}}(i) = \frac{\exp(z_{\text{student},i})}{\sum_j \exp(z_{\text{student},j})} \tag{2.3}$$

Limitations: While efficient, response-based KD relies solely on the final output, ignoring intermediate representational structures. This can lead to the student inheriting teacher biases or failing to generalize when the teacher's predictions are noisy.

2.3.2 Feature-Based Knowledge Distillation

To address the superficiality of response-based methods, Feature-Based KD aligns the intermediate feature maps of the student with those of the teacher (Figure 2.8). First proposed in FitNets [41], this approach utilizes a "hint" loss to guide the student's hidden layers.

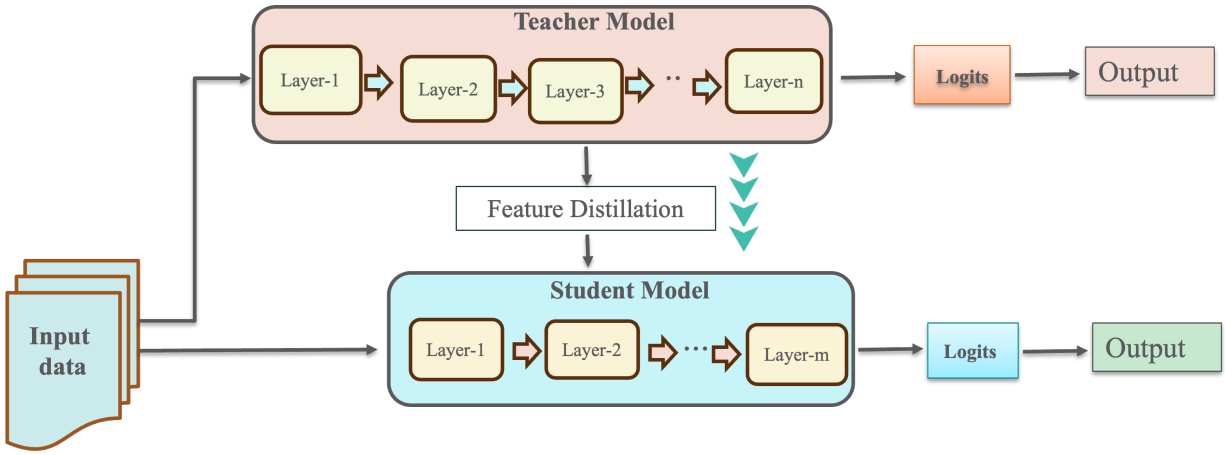

Figure 2.8: Feature-based Knowledge Distillation (FitNets [41]). Beyond matching final outputs, the student is trained to mimic the teacher's intermediate layer activations at designated hint layers. Projection functions $\Phi(\cdot)$ resolve dimensionality mismatches between teacher and student feature spaces, allowing the student to inherit structural representations from mid-network layers that are discarded by response-based methods.

The general loss function measures the dissimilarity between transformed feature maps:

$$L_{\text{Feat}}(ft(x), fs(x)) = L_F(\Phi_t(ft(x)), \Phi_s(fs(x))) \tag{2.4}$$

where $ft(x)$ and $fs(x)$ are feature maps, and $\Phi(\cdot)$ are transformation functions used to match dimensions (e.g., 1×1 convolutions).

Advanced Alignment (KD++): Wang et al. [42] introduced KD++, which aligns student features with the class means of teacher features (Figure 2.9). This method utilizes an ND loss to maximize feature norm and align directions via cosine distance, achieving state-of-the-art results on ImageNet and COCO.

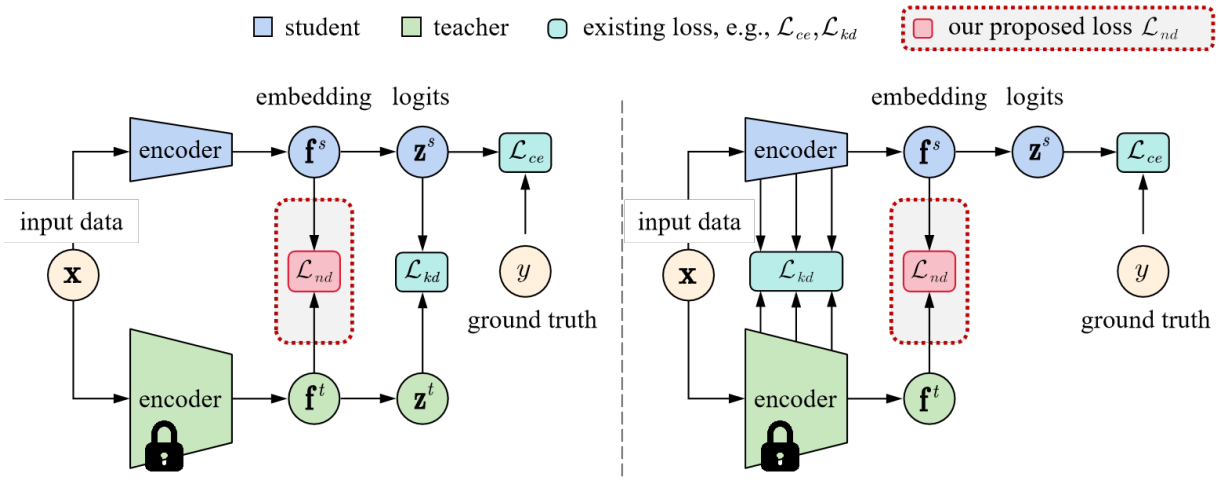

Figure 2.9: Normalized Direction (ND) loss in KD++ [42]. Student features are aligned with class-conditional mean feature vectors of the teacher by simultaneously maximizing feature norm and minimizing cosine distance to the class centroids. This dual objective prevents the student from collapsing to low-norm trivial representations while preserving the directional structure of the teacher's class-discriminative feature space.

2.3.3 Relation-Based Knowledge Distillation

Relation-based KD moves beyond individual samples to transfer the structure of the data manifold. It captures how the teacher models relationships between data examples (Figure 2.10) or between network layers [43].

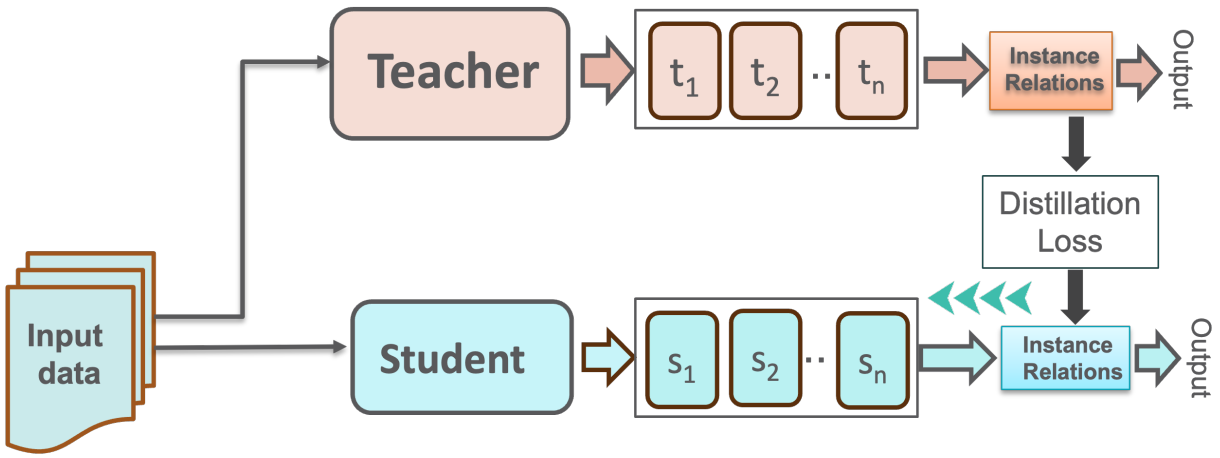

Figure 2.10: Relation-based Knowledge Distillation. Rather than aligning individual sample representations, the student learns to reproduce the geometric relationships among data samples as modeled by the teacher. The distillation loss measures pairwise distance and angular relations in the teacher's feature space and enforces structural consistency in the student's feature space, transferring the topology of the data manifold rather than point-wise feature coordinates.

The loss incorporates a relation term:

$$L_{\text{Rel}} = (1-\alpha) \cdot L_{\text{Base}}(y_{\text{true}}, \text{softmax}(z_{\text{student}})) + \alpha \cdot L_{\text{Relation}}(\mathcal{R}_{\text{teacher}}, \mathcal{R}_{\text{student}}) \tag{2.5}$$

Relational Knowledge Distillation (RKD): Park et al. [44] formalized this by transferring distance-wise and angle-wise relations between samples (Figure 2.11), ensuring the student learns the structural topology of the feature space rather than just point-wise features.

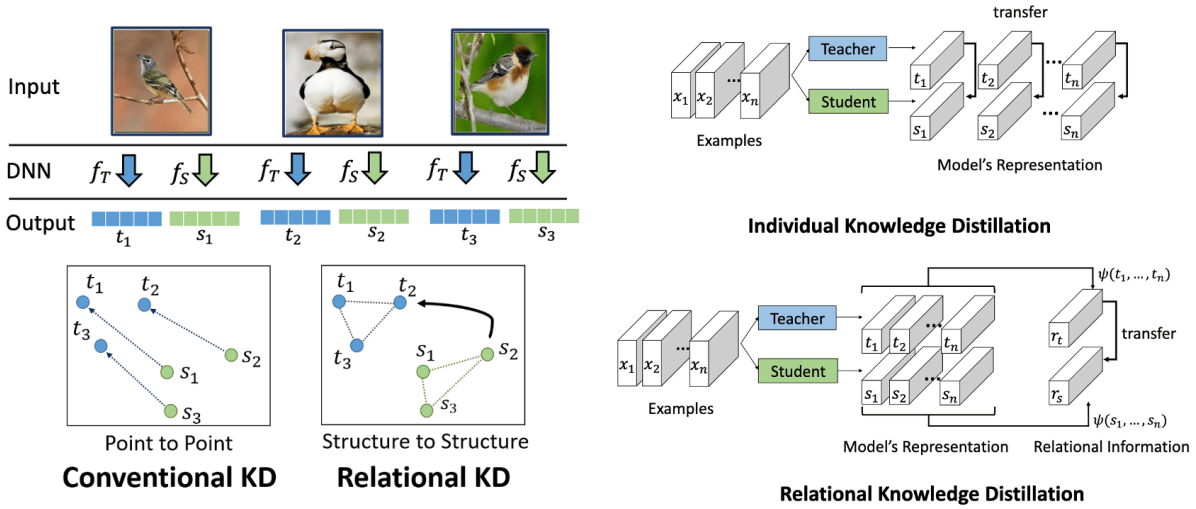

Figure 2.11: Conventional vs RKD. Relations of the outputs of teacher model f_T are transferred to a student model f_s in RKD instead of point-wise knowledge transfer.

2.3.4 Hybrid Knowledge Distillation

Hybrid approaches combine architectural and data-based distillation. A prominent example is the Hierarchical Multi-Attention Transfer (HMAT) framework [45] (Figure 2.12), which distills knowledge at three granularity levels:

1. **Position-based (PKT):** Captures spatial context from early layers.

2. **Activation-based (AKT):** Extracts local features from middle layers.

3. **Channel-based (CKT):** Captures semantic channel-wise correlations from high-level layers.

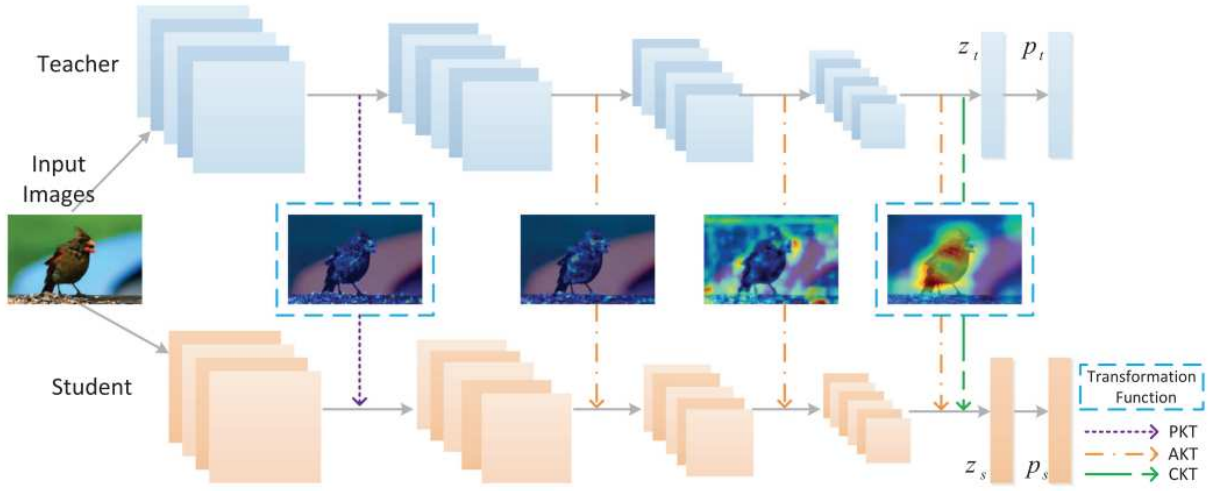

Figure 2.12: Hierarchical Multi-Attention Transfer (HMAT) framework [45]. Knowledge is distilled at three granularity levels: Position-based attention (PKT) from early layers captures spatial context; Activation-based attention (AKT) from middle layers captures local feature responses; Channel-based attention (CKT) from high-level layers captures semantic channel correlations. The composite loss $L_{MAT} = \alpha L_{AKT} + \beta L_{CKT} + \gamma L_{PKT}$ enables independent weighting of each level.

The composite loss function allows for fine-grained control over the distillation process:

$$L_{MAT} = \alpha L_{AKT} + \beta L_{CKT} + \gamma L_{PKT} \tag{2.6}$$

$$L_{total} = L_{MAT} + L_{task} + L_{KL} \tag{2.7}$$

2.3.5 Data-Free Knowledge Distillation (DFKD)

In scientific domains where original training data is restricted due to privacy or size (e.g., massive satellite archives), DFKD allows knowledge transfer using only the teacher's weights.

DFKD typically involves generating synthetic data that approximates the original training distribution via Noise Optimization or Generative Networks [46, 47]. A leading method is DeepIn-

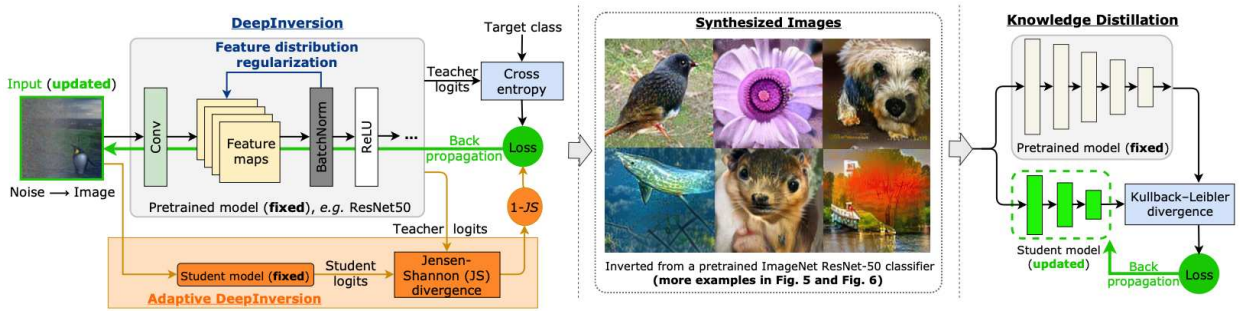

Figure 2.13: DeepInversion framework that generate class-specific input images from random noise using pretrained teacher network.

version [46] (Figure 2.13), which inverts a pre-trained network to synthesize class-specific images from random noise by regularizing feature map statistics (Batch Normalization).

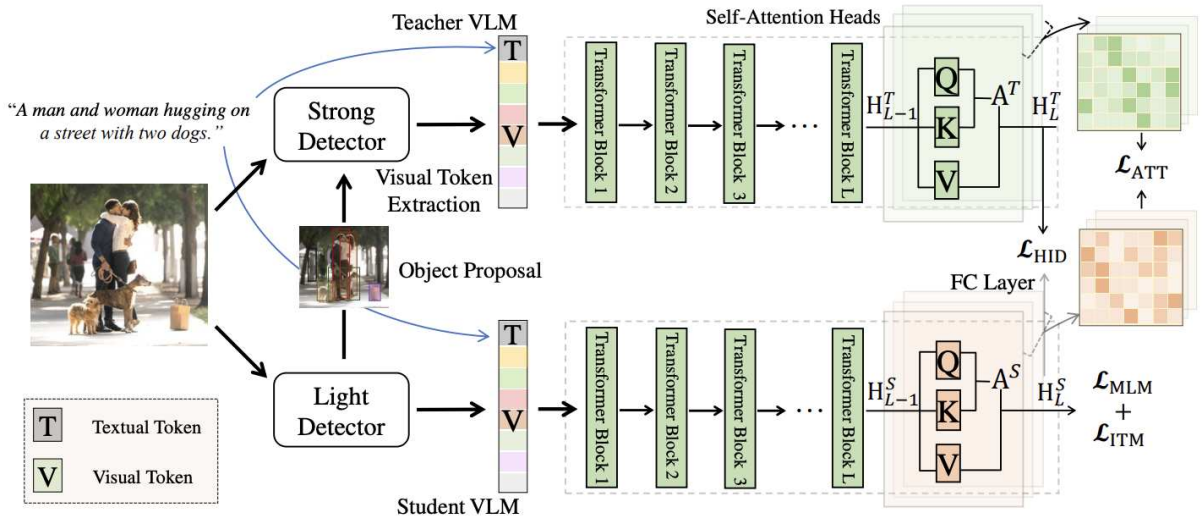

Figure 2.14: DistillVLM architecture for compressing Vision-Language Models [48]. A lightweight student detector generates region proposals that are injected into the frozen teacher detector, aligning region-level features between the two networks. This cross-detector feature matching transfers the teacher's cross-modal alignment between visual regions and textual semantics into the more efficient student architecture.

2.3.6 Distilling Large Foundation Models (LLMs and VLMs)

The rapid scaling of Large Language Models (LLMs) and Vision-Language Models (VLMs) has established a new paradigm in AI, characterized by unprecedented zero-shot generalization, complex reasoning, and multimodal understanding. Deploying these massive architectures (of-

ten comprising tens to hundreds of billions of parameters) presents prohibitive computational, memory, and latency bottlenecks. In specialized scientific domains, such as edge-deployed sensor networks or high-throughput earth observation systems, strict constraints on hardware and power consumption make the direct application of these foundational models infeasible.

Knowledge Distillation provides an effective mechanism for circumventing these deployment barriers, enabling the transfer of generalized knowledge from massive pre-trained teachers into lightweight, domain-specific student networks. In the context of large foundation models, KD has expanded well beyond traditional logit matching. For LLMs, advanced techniques such as "reasoning distillation" have emerged, where a smaller student model is trained to mimic not just the final output, but the step-by-step rationales and intermediate inferential chains generated by the massive teacher [49]. This allows the lightweight model to inherit complex reasoning capabilities while operating at a fraction of the computational cost.

Similarly, adapting VLMs for specialized domains requires transferring cross-modal alignments, such as the mapping between textual semantics and spatial-spectral visual features. Distillation frameworks for VLMs enforce consistency between the multimodal embedding spaces of the teacher and the student, compressing visual-linguistic knowledge into streamlined architectures [48], as illustrated in Figure 2.14.

By leveraging KD, researchers can utilize massive LLMs and VLMs as "knowledge engines" rather than deployment targets. The teacher model processes expansive, multi-domain data to create rich representations, which are then distilled into a compact model tightly optimized for the target domain's specific modalities and computational limits. This paradigm forms a foundational motivation for the methodologies proposed in this dissertation, demonstrating how KD can synthesize robust, lightweight systems from unwieldy foundational models.

2.4 Domain Adaptation in Deep Learning

Domain Adaptation (DA) addresses the Adaptability Gap, which describes the performance degradation that deep learning models experience when deployed on a target domain whose dis-

tribution differs from the source domain on which they were trained. This failure occurs because real-world deployment often violates the fundamental assumption that training and test data are independent and identically distributed. To bridge this gap, adaptation strategies generally fall into two broad categories: distribution alignment and self-supervised consistency.

2.4.1 Categories of Domain Adaptation

A primary approach to DA involves aligning the source and target distributions to learn domain-invariant representations. **Adversarial Alignment** techniques, exemplified by Domain-Adversarial Neural Networks (DANN), introduce a domain discriminator that competes with the feature extractor in a minimax game. By utilizing a gradient reversal layer, DANN forces the model to learn features that are discriminative for the task but indistinguishable regarding their domain origin [50]. Complementary to feature-level alignment, **Image Translation** methods operate at the pixel level. Frameworks such as CyCADA employ generative adversarial networks to translate source images into the visual style of the target domain while preserving semantic content [51]. This imposes cycle-consistency constraints, ensuring that the input space is aligned before the data reaches the classifier.

Alternative strategies leverage the target data itself to guide adaptation through semi-supervised or self-supervised signals. **Supervised Selection**, often referred to as self-training, iteratively generates pseudo-labels for the unlabeled target data. By filtering for high-confidence predictions and treating them as ground truth, methods like those proposed by Chen et al. refine the decision boundary to better fit the target distribution [52]. More recently, the field has shifted toward **Unsupervised Domain Adaptation (UDA)** via masked image modeling. State-of-the-art approaches such as Masked Image Consistency (MIC) utilize random masking as a data augmentation strategy [17]. By forcing the network to yield consistent predictions between masked and unmasked views of the same target image, MIC encourages the model to learn robust, context-aware features rather than relying on domain-specific spurious correlations.

2.4.2 Source-Free Domain Adaptation (SFDA)

SFDA applies when the source data is unavailable, relying solely on a pre-trained source model. Kim et al. [53] proposed a framework (Figure 2.15) using Adaptive Prototype Memory (APM) to generate pseudo-labels for the target domain. This allows the target model to self-adapt without direct access to the source dataset, a necessary capability for decentralized scientific AI applications.

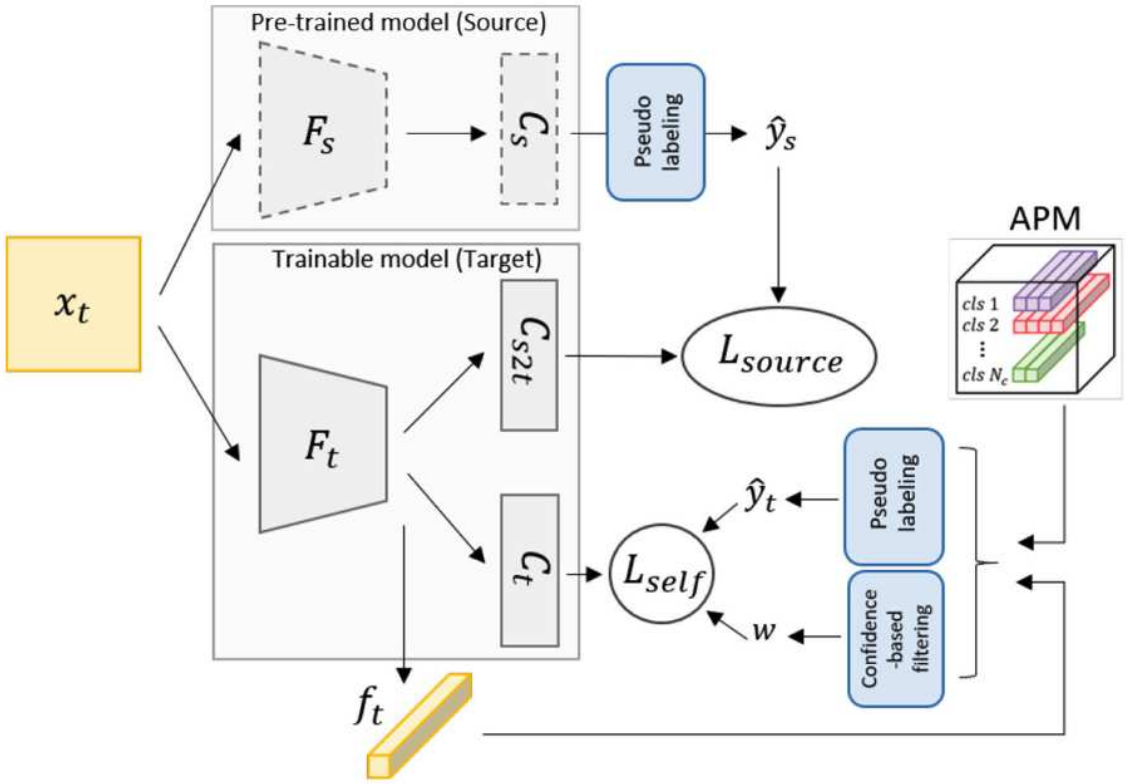

Figure 2.15: Overview of the SFDA framework (fixed model parameters are indicated by dashed line).

At the intersection of domain adaptation and knowledge distillation, teacher-student frameworks have been deployed for Earth observation soil property estimation, bridging the distributional gap between laboratory reference spectra and satellite-observed surface reflectance at continental scale [54, 55]. These works demonstrate that resolution and sensor mismatches endemic to scientific EO tasks are tractable through principled cross-modal transfer learning.

2.5 Science-Guided Domain Adaptation and Physics-Informed Learning

The intersection of Domain Adaptation (DA) and scientific machine learning represents a paradigm shift from purely statistical alignment to physically grounded representation learning. This section reviews the foundational technologies enabling this shift: self-supervised masked modeling, physics-informed learning strategies, and the specific application of spectral unmixing as a domain-invariant mechanism.

2.5.1 Self-Supervised Vision and Masked Modeling

The advent of Masked Autoencoders (MAE) has fundamentally altered the landscape of representation learning. By masking a high proportion of input patches and reconstructing missing pixels, MAEs force the model to learn rich, generalized semantic features rather than relying on superficial low-level statistics [32, 33].

These approaches have shown strong transferability to downstream tasks, prompting extensions into multi-spectral and temporal remote sensing. For instance, SatMAE [28] demonstrates that spectral-temporal pretext tasks can effectively leverage the temporal dynamics of satellite imagery. However, standard MAE approaches remain purely data-driven. While effective, they lack explicit constraints to ensure that the learned latent representations adhere to the physical laws governing the data generation process. The present study departs from purely distillation-based strategies to introduce a physics-guided mechanism, embedding domain-theoretic models directly within the architecture to enhance reconstruction fidelity and interpretability.

2.5.2 Physics-Informed and Knowledge-Guided Learning

Integrating scientific priors into machine learning models, formalized under Physics-Informed Machine Learning (PIML) and Knowledge-Guided Machine Learning (KGML), addresses the limitations of black-box models in scientific discovery. These approaches aim to improve physical realism, generalization to out-of-distribution samples, and scientific consistency [56, 57]. Sev-

eral efforts have explored knowledge-guided machine learning (KGML) methods for modeling spatiotemporally evolving phenomena. These methods often operate in concert with mechanistic or process-based domain-theoretic models, and frequently incorporate multipart loss functions that encode scientific knowledge directly into model training. Examples include physical constraints from soil hydrology, such as the van Genuchten water-retention equations and hydraulic-conductivity models based on Richards' Equation [58–60]; vegetation indices [61]; evapotranspiration dynamics [62]; preservation of graph-theoretic properties such as betweenness centrality [11]; cloud-occlusion masking in satellite imagery [15]; perceptual constraints for visualization [63]; soil salinity estimation [54]; hyperspectral imagery modeling [64]; correlations among soil spectroscopic properties [65], and time-series projections [66, 67].

Two primary strategies dominate this field:

1. **Soft Constraints (Loss-Based):** Physics-Informed Neural Networks (PINNs) [68] typically incorporate differential equations (PDEs) as regularization terms in the loss function. While effective for continuous fields, this "soft" constraint does not guarantee that the model's architecture inherently respects physical laws.

2. **Hard Constraints (Architecture-Based):** Recent advancements focus on embedding physical rules directly into the network structure. For example, ensuring mass conservation in flow networks or thermodynamic consistency in climate modeling [69].

2.5.3 Spectral Unmixing as a Domain Invariant

In the context of optical remote sensing and spectroscopy, the Linear Spectral Mixing Model (LSMM) provides a robust physical ground truth. LSMM posits that any observed pixel spectrum is a linear combination of pure material signatures (endmembers) weighted by their fractional abundance [70].

Unlike statistical features, which may drift significantly between sensors (domain shift), the fundamental material composition (abundance) of a scene remains invariant. Classical formulations and modern deep unmixing approaches, such as EndNet [71], demonstrate how material

signatures can be recovered for interpretability. However, these have traditionally been treated as standalone unmixing tasks rather than mechanisms for domain adaptation.

2.5.4 Addressing the Gap

Current literature highlights a significant disconnect between general-purpose domain adaptation strategies and domain-specific physical theory. While standard unsupervised domain adaptation effectively minimizes statistical discrepancies between source and target distributions, it often ignores the underlying physical data generation processes, leading to features that are statistically aligned but physically implausible.

Although recent surveys such as [72] discuss the broad potential of Knowledge-Guided Machine Learning (KGML), existing implementations often treat physics merely as an auxiliary regularization term or a "soft" constraint. There is limited exploration of using well-defined physical laws (specifically physical mixing models) as the primary engine for driving unsupervised adaptation within modern Transformer architectures.

Our research targets this critical intersection. By embedding the physics of spectral mixing directly into the learning loop, we move beyond simple feature matching. We utilize the Linear Spectral Mixing Model (LSMM) to disentangle sensor-specific artifacts from invariant physical properties (material abundances). This approach allows us to guide the adaptation of Foundation Models across heterogeneous scientific instruments, ensuring that the transferred knowledge respects the immutable laws of spectroscopy rather than just the statistical patterns of the dataset.

Chapter 3

Methodology

3.1 Overview of Proposed Methodology

This chapter details the technical frameworks developed to address the three core research questions outlined in the introduction: (1) How to mitigate domain shifts in heterogeneous environments (Shift Mitigation); (2) How to transfer knowledge across asymmetric modalities (Inverse KD); and (3) How to enforce physical consistency in deep learning models (Science-Guided Adaptation). While each thrust targets a specific barrier (dimensionality, resolution, or interpretability), they collectively operate under a unified principle: leveraging prior knowledge (whether from foundation models or physical laws) to regularize and guide the learning of student models in data-constrained scientific domains. The following sections describe the specific architectural innovations, loss functions, and training strategies designed for each thrust.

3.2 Thrust 1: Heterogeneous Domain-Shift Mitigation

3.2.1 Overview and Conceptual Foundation

The first thrust addresses the **Heterogeneous Domain Shift** arising from the inherent disparity in environmental sensing modalities. The framework developed for this thrust [55] was published at the 2023 IEEE International Conference on Big Data. In geospatial and environmental monitoring, a critical gap exists between satellite data and ground-truth measurements. Satellite observations (the source or teacher domain) provide spatially contiguous but coarse coverage (e.g., 3km grids). Conversely, in-situ sensor networks (the target or student domain) provide high-fidelity ground truth but are geographically sparse, point-based measurements.

Directly training a deep learning model solely on sparse point data leads to overfitting and poor spatial generalization across diverse terrains. Conversely, models trained exclusively on satellite data lack the localized precision required for micro-climate analysis. To mitigate this shift, we

propose a cross-resolution knowledge distillation framework. This approach utilizes a complex, coarse-resolution teacher model to learn the global, nonlinear physics of environmental interactions (e.g., the relationship between soil properties, topography, and weather). This global wisdom is subsequently distilled into a lightweight student model anchored by the sparse point data. The core methodology enables the transfer of representational consistency from a grid-based modality to a point-based modality, effectively bridging the spatial resolution gap.

3.2.2 Data Integration and Preprocessing

Aligning heterogeneous datasets is a prerequisite for resolving the domain shift. The input data comprises a diverse set of variables, including satellite imagery (gNATSGO at 10m resolution), terrain attributes (DEM at 30m resolution), and climate data (gridMET at 4km resolution). To synchronize these sources, we establish a common temporal range (2020–2021) and geospatial extent (covering diverse terrains in Colorado and California).

For the teacher model, the higher-resolution datasets (gNATSGO and DEM) are downscaled, and gridMET is merged with the DEM to align with the 3km spatial resolution of the SMAP satellite labels. Filtering for temporal validity yields a robust training set of over 13 million input-output samples. To facilitate knowledge transfer while preserving fine-grained localized accuracy, both the teacher and student models ingest the same input tensor shape of $32 \times 32 \times 10$ (representing the spatial window and 10 integrated feature bands), though the student model utilizes the highest possible resolution available at the specific ground station point.

3.2.3 Asymmetric Teacher-Student Architecture

To address the resolution disparity while optimizing computational efficiency, the proposed framework employs a highly asymmetric dual-model architecture (illustrated in Figure 3.1).

The Global Teacher Model (VGG13)

The teacher model is tasked with learning the complex, non-linear relationships between meteorological factors, topographical conditions, and environmental variables over a broad geospatial

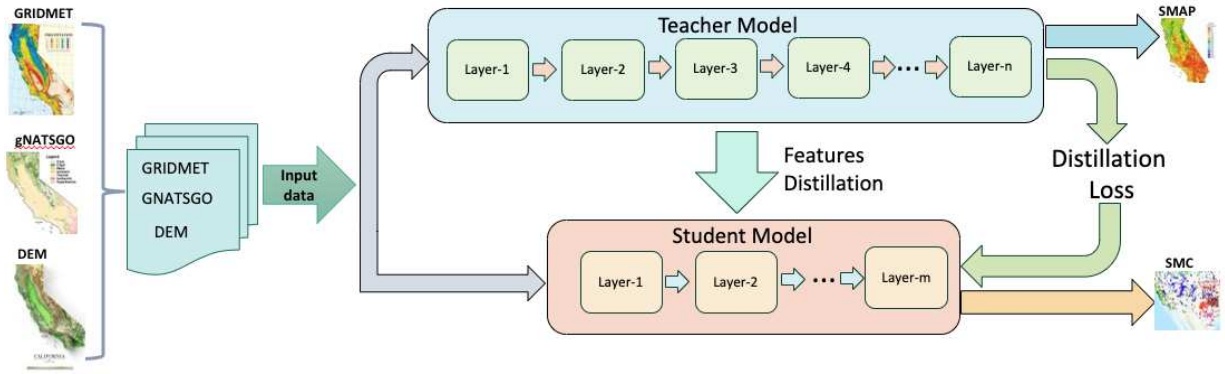

Figure 3.1: Cross-resolution Knowledge Distillation architecture for soil moisture estimation. A VGG13 teacher (9.41M parameters) trained on coarse SMAP satellite observations (3 km) transfers intermediate feature-map representations to a ResNet8 student (78.5K parameters) supervised by sparse in-situ ground station measurements. Feature-Map Distillation (FMD) aligns intermediate convolutional activations across the two architectures, enabling the student to inherit the teacher's geospatial modeling capacity at a fraction of the inference cost.

extent. Through empirical evaluation against deeper architectures like ResNet34, the VGG13 architecture [73] was selected for its optimal balance of predictive accuracy and memory footprint. Consisting of 13 convolutional layers followed by three fully connected layers, the VGG13 teacher contains approximately 9.41M learnable parameters. It optimizes an L1 loss against the coarse 3km SMAP targets, capturing generalized environmental patterns.

The Local Student Model (ResNet8)

The student model is engineered for high-precision local estimation and efficient edge deployment. We utilize a highly lightweight ResNet8 architecture [74], containing only 8 residual blocks and roughly 78.5K parameters, a fraction of the teacher's size. For comparative analysis in specific ablation scenarios, MobileNetV2 [75] (2.1M parameters) is also utilized as an intermediate complexity student. The student model is trained on the highly sparse in-situ station data. By ingesting the same $32 \times 32 \times 10$ input dimensional space, the student model is structurally primed to receive spatial representations from the teacher.

3.2.4 Knowledge Transfer via Feature-Map Distillation

To transfer the structural and spatial knowledge from the coarse-resolution teacher to the fine-resolution student, we employ Feature-Map Distillation (FMD). Feature maps represent the high-level, generalized environmental mappings formed by the intermediate layers of the teacher. Distilling this information enables the student model to replicate the teacher's representational capacity without requiring massive amounts of point-level data. The FMD loss is formulated as [43]:

$$L_{FMD} = \sum_{s_l, t_l} Dist(T^t(F^t_{t_l}), T^s(F^s_{s_l})), \tag{3.1}$$

where $F^t_{t_l}$ and $F^s_{s_l}$ denote the feature maps of the target teacher layer and corresponding student layer, respectively. The functions $T^t(\cdot)$ and $T^s(\cdot)$ are transformations that convert these feature maps into pairwise similarity matrices [76]. Based on these matrices, layers exhibiting similarity scores above a dynamic threshold are selected for distillation. Architectural dimension mismatches between the VGG13 teacher and ResNet8 student are resolved via dimension reshaping techniques, such as applying an extra convolution operation to project the required number of features from the teacher's latent space to the student's.

3.2.5 Unified Multi-Source Loss Function

Training the student model on extremely sparse ground truth requires rigorous regularization. We introduce a unified loss function, L_{Total}, which guides the student network using three distinct supervisory signals:

$$L_{Total} = LF_1 L_S + LF_2 L_{KD} + LF_3 L_{FMD}, \tag{3.2}$$

where LF_1, LF_2, and LF_3 act as dynamic weighting factors for each corresponding loss component. Because the task functions fundamentally as a regression problem, we rely on the standard L1 formulation for numerical disparities:

$$\ell(x, y) = \{l_1, ..., l_N\}^T, \quad l_n = |x_n - y_n|, \tag{3.3}$$

where x represents the model input, y represents the target, and N represents the batch size.

The three primary components of the unified loss are:

- L_S **(Student Loss):** The absolute L1 disparity between the student model's high-resolution prediction and the true in-situ ground station measurements.
- L_{KD} **(Logit Loss):** The L1 discrepancy between the student model's prediction and the *soft prediction* generated by the teacher model for the same input batch.
- L_{FMD} **(Feature Loss):** The structural alignment loss between intermediate layers, ensuring the student mimics the teacher's internal spatial reasoning.

3.2.6 Nested Training for Spatial Scalability

Training a single student model to capture continental-scale phenomena while maintaining micro-climate accuracy is computationally prohibitive. To ensure spatial scalability, the framework implements a **Nested Training** strategy anchored by the Köppen Climatic Classification system [77].

1. **Global Teacher:** A singular, complex teacher model is trained over a broad geographic extent (e.g., California and Colorado) utilizing the abundant, coarse SMAP data.
2. **Localized Student Ensembles:** An ensemble of independent student models is trained in parallel. Each student is specialized for a distinct Köppen climate zone (e.g., semi-arid Bsk, Mediterranean Csa/Csb, or continental Dfb/Dfc).

This nested hierarchy is especially useful for regions with extreme label scarcity. By leveraging the hierarchical structure of the Köppen system, the framework can dynamically assign a point location lacking sufficient training data to a student model optimized for the most structurally similar climatic condition, ensuring robust generalization and mitigating geographic overfitting.

3.3 Thrust 2: Inverse Knowledge Distillation for Cross-Modal Asymmetry

3.3.1 Overview and Conceptual Foundation

The second research thrust addresses the fundamental challenge of **Cross-Modal Knowledge Asymmetry** in scientific deep learning. In traditional Knowledge Distillation (KD), the flow of information is typically symmetric or compressive: a high-capacity "teacher" model transfers its learned representations to a smaller "student" model to improve efficiency. However, in domains like Earth Observation, we encounter the *Inverse Domain Shift* problem.

We possess powerful Foundational Models (FMs) pre-trained on abundant, low-dimensional data (e.g., multispectral satellite imagery with 6 bands). Conversely, we aim to train models for high-dimensional, scarce data (e.g., hyperspectral imagery with 218 bands). The standard KD paradigm is insufficient here because the student's feature space is significantly more complex than the teacher's.

Inverse Knowledge Distillation reverses this paradigm. It leverages the generalized spatial-spectral correlations captured by the low-dimensional teacher to scaffold the training of a high-dimensional student. The core hypothesis is that while the teacher cannot resolve fine-grained spectral detail, it carries a robust understanding of macro-scale geospatial structure. This thrust develops the mechanisms of *Spectral Domain Alignment* and *Feature-Guided Masking* to project the teacher's knowledge into the student's high-fidelity latent space without constraining the student's capacity to learn novel spectral signatures.

3.3.2 Proposed Architecture

The core architecture [78] builds upon the Masked Autoencoder (MAE) framework with a Vision Transformer (ViT) backbone. As illustrated in Figure 3.2, the system comprises two distinct asymmetric networks:

- **The Teacher (Prithvi-100M):** A foundational geospatial model pre-trained on 6-band HLS data with a temporal depth of 3 timestamps [18]. It utilizes a ViT backbone with 12 transformer blocks.

- **The Student:** A specialized ViT-MAE designed for 218-band EnMAP data [31]. To maintain architectural compatibility for feature alignment, the student utilizes a patch size of 16×16, an embedding dimension of 768, and 12 transformer layers with 12 attention heads.

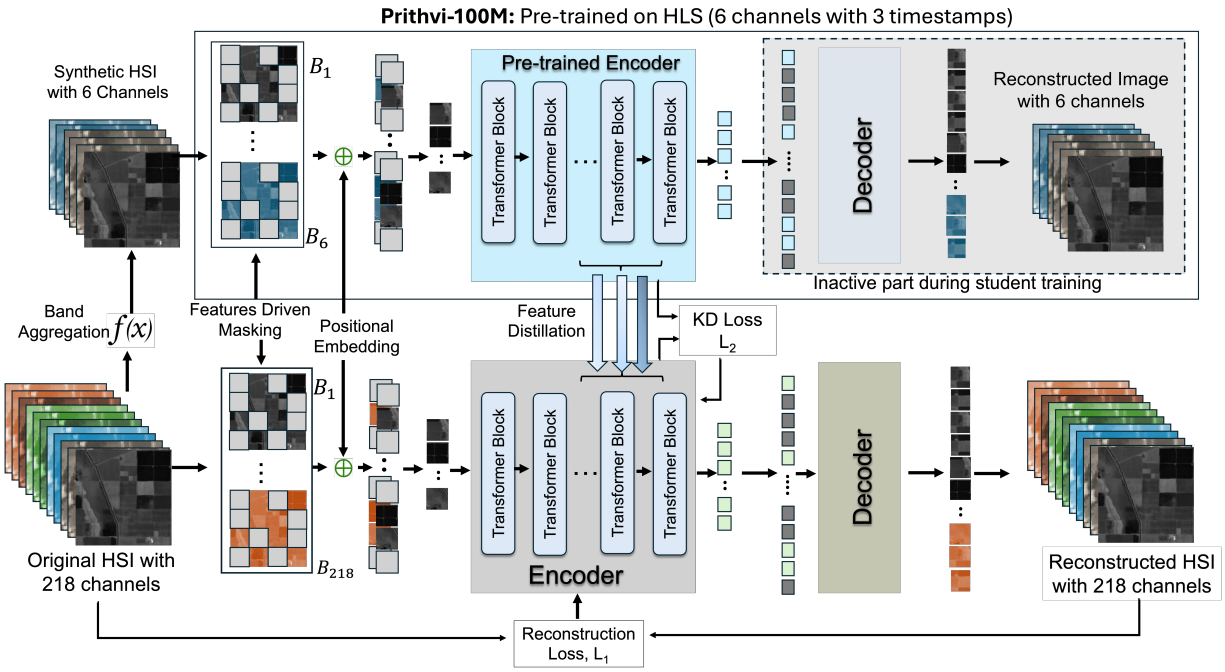

Figure 3.2: Proposed architecture for inverse knowledge distillation: a cross-domain framework for hyperspectral image reconstruction that transfers intermediate-layer knowledge from the Prithvi-100M multispectral foundation model to a high-dimensional student network via guided masking and a composite distillation loss.

3.3.3 Spectral Domain Alignment and Aggregation

A fundamental obstacle in this inverse distillation is the spectral misalignment between the HLS teacher (6 bands) [27] and the EnMAP student (218 bands). Without alignment, the teacher cannot effectively process the student's input domain to provide guidance. We employ a spectral matching strategy that aggregates EnMAP bands to synthesize the teacher's input space.

Formally, let $B_{\text{HLS}} = \{b_1, \ldots, b_6\}$ and $B_{\text{EnMAP}} = \{e_1, \ldots, e_{218}\}$ represent the spectral bands of the teacher and student, respectively. For each teacher band b_i, we identify the subset $E_i \subseteq B_{\text{EnMAP}}$ of overlapping EnMAP bands:

$$E_i = \{e_j \in B_{\text{EnMAP}} \mid \lambda_j^{\min} \geq \lambda_i^{\min} \text{ and } \lambda_j^{\max} \leq \lambda_i^{\max}\}. \tag{3.4}$$

We then compute a synthetic band $\hat{b}_i$ via aggregation:

$$\hat{b}_i = \frac{1}{|E_i|} \sum_{e_j \in E_i} e_j, \tag{3.5}$$

where $|E_i|$ denotes the cardinality of the subset. This process generates a synthetic input tensor $\hat{B}_{\text{EnMAP}}$ that is spectrally consistent with the teacher's pre-training while retaining the student's high-dimensional characteristics.

3.3.4 Spatial Feature-Guided Masking Strategy

Standard random masking in MAEs often discards high-frequency spatial details that are important for scientific modeling. To maximize the information gain from the teacher, we introduce a *Spatial Feature-Guided Masking* strategy. We utilize Gabor filters [79] and Wavelet transforms [80] to compute a significance score S_p for each patch, prioritizing regions with high structural complexity.

The Gabor filter response, capable of capturing edge and texture information [81], is defined as:

$$G(x, y; \lambda, \theta, \psi, \sigma, \gamma) = \exp\left(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right) \cos\left(2\pi \frac{x'}{\lambda} + \psi\right) \tag{3.6}$$

where x' and y' are rotated coordinates. Alternatively, we employ discrete wavelet transforms (DWT) to extract multi-scale frequency components:

$$W_\psi(a, b) = \int_{-\infty}^{\infty} X(t) \psi_{a,b}^*(t) dt \tag{3.7}$$

The patch significance score is computed as the norm of the filter response:

$$S_p = \|G_p\| \text{ or } S_p = \|W_p\| \tag{3.8}$$

We rank patches by S_p and mask the top 75% (ratio $r = 0.75$). This forces the student to reconstruct the most challenging, information-rich regions under the teacher's guidance.

3.3.5 Feature Distillation and Specialized Loss

To reconcile the dimensional discrepancy between the teacher and student latent spaces, we introduce a fully connected (FC) alignment layer. Through ablation studies, we determined that distilling from the upper-mid encoder layer (Layer 8 of 12) provides the optimal balance between low-level feature extraction and high-level semantic abstraction.

The optimization objective is a composite loss function designed to enforce both reconstruction fidelity and distributional alignment. The reconstruction loss $\mathcal{L}_{\text{recon}}$ combines Mean Squared Error (MSE) for spectral accuracy and the Structural Similarity Index (SSIM) [82] for spatial coherence:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N}(X_i - \hat{X}_i)^2 \tag{3.9}$$

$$\mathcal{L}_{\text{SSIM}} = 1 - \frac{(2\mu_X\mu_{\hat{X}} + C_1)(2\sigma_{X\hat{X}} + C_2)}{(\mu_X^2 + \mu_{\hat{X}}^2 + C_1)(\sigma_X^2 + \sigma_{\hat{X}}^2 + C_2)} \tag{3.10}$$

For the distillation term $\mathcal{L}_{\text{KD}}$, we employ the Kullback–Leibler Divergence (KLD) [83] to minimize the information loss between the teacher's probability distribution P_T and the student's distribution P_S. The total weighted loss function is:

$$\mathcal{L}_{\text{total}} = \alpha(\lambda_1\mathcal{L}_{\text{MSE}} + \lambda_2\mathcal{L}_{\text{SSIM}}) + \beta\mathcal{L}_{\text{KD}} \tag{3.11}$$

This composite loss allows the proposed inverse distillation framework to significantly outperform baselines in recovering fine spectral details.

3.4 Thrust 3: Science-Guided Domain Adaptation

3.4.1 Overview and Conceptual Foundation

The final thrust focuses on **Science-Guided Domain Adaptation**, addressing the lack of physical interpretability in standard deep learning models. In scientific applications, maximizing statistical accuracy (e.g., minimizing MSE) is insufficient if the model violates fundamental physical laws. This creates a risk of generalization failure when the model encounters out-of-distribution data.

Instead of relying solely on data-driven optimization, this approach acts as a bridge between neural approximation and physical theory. We formally embed the *Linear Spectral Mixing Model (LSMM)* [70], a foundational principle in spectroscopy, directly into the neural architecture as a differentiable constraint layer. This forces the model to learn latent representations that correspond to physical material abundances (e.g., percentage of water, soil, vegetation) rather than arbitrary feature vectors, ensuring both interpretability and physical consistency.

3.4.2 Physics-Informed Architecture

The proposed approach modifies the standard ViT-MAE architecture [32] by introducing a physics-guided branch within the decoder.

Integration of Linear Spectral Mixing Model (LSMM)

In hyperspectral analysis, an observed pixel $\mathbf{r} \in \mathbb{R}^C$ is modeled as a linear combination of pure material signatures (endmembers) [84]. The LSMM equation is:

$$\mathbf{r} = \mathbf{A}\mathbf{x} + \mathbf{e}, \tag{3.12}$$

where $\mathbf{A} \in \mathbb{R}^{C \times M}$ is the endmember matrix (signatures of M materials), $\mathbf{x} \in \mathbb{R}^M$ is the abundance vector, and $\mathbf{e}$ is noise.

To enforce physical validity, we implement a specialized *abundance head* $f_\theta(\cdot)$ in the decoder. This is a compact MLP ($D \to D/2 \to M$) that predicts the abundance vector for each patch.

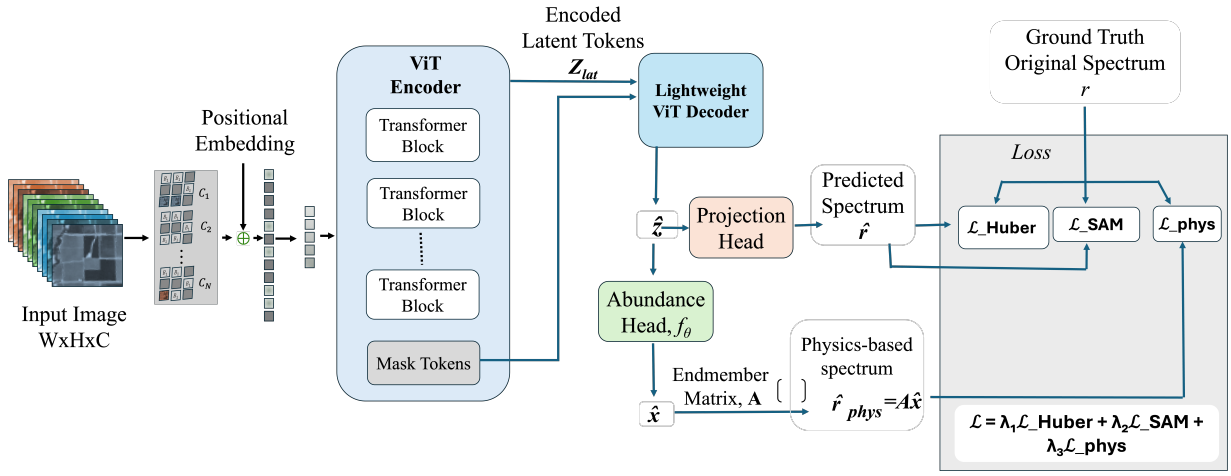

Figure 3.3: Overview of the Science Guided Foundation Model Framework. A knowledge-guided ViT-MAE that integrates the Linear Spectral Mixing Model (LSMM) within the decoder. The model jointly optimizes Huber loss, SAM, and physics-consistency losses to enhance spectral fidelity and interpretability.

Crucially, we apply a Softmax activation to enforce the non-negativity ($\mathbf{x} \geq 0$) and sum-to-one ($\mathbf{1}^\top \mathbf{x} = 1$) constraints inherent to physical mixtures:

$$\hat{\mathbf{x}} = \text{softmax}(f_\theta(\mathbf{z})), \tag{3.13}$$

where $\mathbf{z}$ is the decoder patch token. The physics-guided reconstruction is computed as $\hat{\mathbf{r}}_{\text{phys}} = \mathbf{A}\hat{\mathbf{x}}$.

3.4.3 Geometric Alignment via Spectral Angle Mapper (SAM)

To prioritize spectral shape preservation over simple intensity matching (which can be affected by illumination conditions), we incorporate the Spectral Angle Mapper (SAM) loss [85, 86]. SAM measures the angular distance between the reconstructed vector $\hat{\mathbf{r}}$ and the ground truth $\mathbf{r}$:

$$\mathcal{L}_{\text{SAM}} = \frac{1}{N}\sum_{i=1}^{N} \arccos\left(\frac{\langle \hat{\mathbf{r}}_i, \mathbf{r}_i \rangle}{\|\hat{\mathbf{r}}_i\|_2 \, \|\mathbf{r}_i\|_2 + \epsilon}\right). \tag{3.14}$$

This aligns the vectors in the high-dimensional spectral space, ensuring the model learns discriminative features for materials with similar spectral profiles.

3.4.4 Hybrid Physics-Consistent Objective

The optimization is governed by a hybrid objective that balances numerical accuracy, geometric alignment, and physical consistency. The unified loss function is:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{Huber}} + \lambda_2 \mathcal{L}_{\text{SAM}} + \lambda_3 \mathcal{L}_{\text{phys}}, \tag{3.15}$$

The components include:

1. **Huber Loss $\mathcal{L}_{\text{Huber}}$:** Chosen for robust regression to handle spectral outliers.

$$\mathcal{L}_{\text{Huber}} = \frac{1}{N} \sum_{i=1}^{N} \begin{cases} \frac{1}{2} \|\hat{\mathbf{r}}_i - \mathbf{r}_i\|_2^2, & \text{if } \|\hat{\mathbf{r}}_i - \mathbf{r}_i\|_2 \leq \delta, \\ \delta \|\hat{\mathbf{r}}_i - \mathbf{r}_i\|_2 - \frac{1}{2}\delta^2, & \text{otherwise}, \end{cases} \tag{3.16}$$

2. **Physics Consistency Loss $\mathcal{L}_{\text{phys}}$:** This term explicitly minimizes the residual between the observed signal and the LSMM-reconstructed signal, forcing the latent representation to adhere to the mixing model:

$$\mathcal{L}_{\text{phys}} = \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{r}_i - \mathbf{A}\hat{\mathbf{x}}_i\|_2^2. \tag{3.17}$$

By jointly optimizing these terms, the methodology ensures that the model learns representations that are not only statistically predictive but also grounded in the physical principles of spectroscopy.

3.4.5 Science-Guided Knowledge Distillation

To address the performance gap when deploying hyperspectral models across heterogeneous sensor domains (e.g., transitioning from high-resolution EnMAP to coarser HLS data), we propose a science-guided knowledge distillation framework (see Figure 3.4). Unlike standard distillation which focuses on matching logit outputs, the proposed approach leverages a high-capacity, frozen

teacher model to guide a compact, trainable student model through a combination of latent feature alignment and physical consistency.

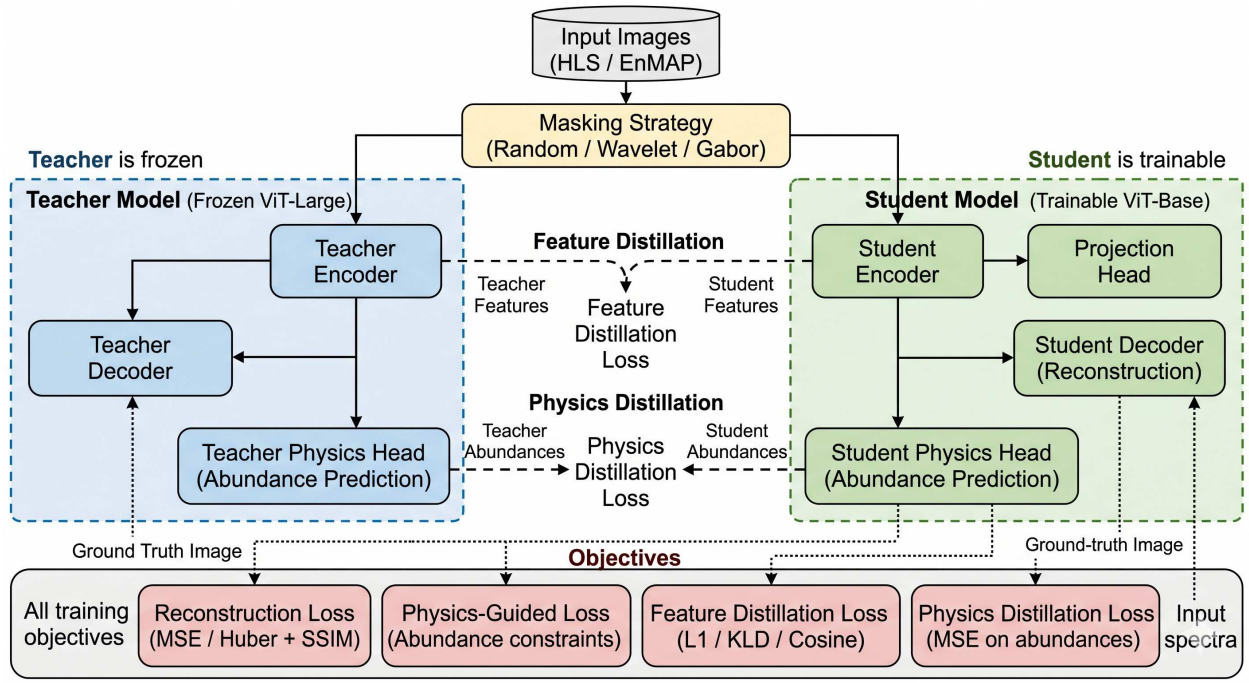

Figure 3.4: The proposed science-guided distillation framework: a dual-branch architecture utilizing a frozen ViT-Large teacher to guide a trainable ViT-Base student through feature-level and physics-based distillation paths.

Distillation Mechanisms

The core of the proposed approach lies in the simultaneous transfer of abstract neural representations and concrete physical parameters.

- **Latent Feature Distillation:** To ensure the student model inherits the robust spatial-spectral feature extraction capabilities of the teacher, we minimize the distance between their latent representations. Given the dimensionality mismatch between the ViT-Large teacher and ViT-Base student backbones, a projection head $P(\cdot)$ is applied to the student features:

$$\mathcal{L}_{\text{feat}} = \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{z}_i^{\text{teacher}} - P(\mathbf{z}_i^{\text{student}})\|_1, \tag{3.18}$$

where $\mathbf{z}$ represents the encoder output tokens.

- **Physical Abundance Distillation:** The framework enforces domain-invariant physical knowledge by distilling the abundance maps $\hat{\mathbf{x}}$ predicted by the Teacher Physics Head. Since the Linear Spectral Mixing Model (LSMM) is inherently sensor-independent, this ensures the student learns universal material signatures rather than sensor-specific artifacts:

$$\mathcal{L}_{\text{phys-distill}} = \frac{1}{N} \sum_{i=1}^{N} \|\hat{\mathbf{x}}_i^{\text{teacher}} - \hat{\mathbf{x}}_i^{\text{student}}\|_2^2. \tag{3.19}$$

Multi-Objective Optimization for Domain Transfer

The student model is optimized using a composite loss function that balances reconstruction accuracy in the target domain with knowledge retention from the source domain:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{feat}}\mathcal{L}_{\text{feat}} + \lambda_{\text{phys}}\mathcal{L}_{\text{phys-distill}} + \lambda_{\text{cons}}\mathcal{L}_{\text{physics}}. \tag{3.20}$$

By incorporating the physics-consistency loss $\mathcal{L}_{\text{physics}}$, the model remains constrained by the physical properties of the target domain's spectra. Simultaneously, the distillation terms $\mathcal{L}_{\text{feat}}$ and $\mathcal{L}_{\text{phys-distill}}$ act as regularizers to prevent overfitting to sensor-specific noise, ensuring the student model remains interpretable and physically consistent across out-of-distribution satellite imagery.

3.5 Downstream Tasks and Representational Transferability

Following the unsupervised pretraining and physics-guided knowledge distillation frameworks developed in Thrusts 2 and 3 (Phase I), it is imperative to evaluate the utility, robustness, and generalizability of the learned hyperspectral representations. To rigorously assess the quality of these encoded features, we employ a two-stage downstream evaluation protocol, often referred to as linear or convolutional probing.

3.5.1 Phase II: The Frozen Encoder Protocol

As illustrated in Figure 3.5, the transition from Phase I (reconstruction and distillation) to Phase II (downstream application) relies on isolating the pretraining gains from the downstream optimization. To achieve this, the weights of the pretrained encoder are strictly frozen. The encoder thus functions as a universal, static feature extractor that projects high-dimensional hyperspectral inputs into a compressed, semantically rich latent space.

Freezing the encoder is a critical methodological choice for two reasons. First, it prevents the catastrophic forgetting of the physics-informed and structural features acquired during Phase I. Second, it ensures that any observed performance improvements on downstream tasks are strictly attributable to the representational power of the pretraining phase, rather than the capacity of a heavily fine-tuned network memorizing the downstream training set.

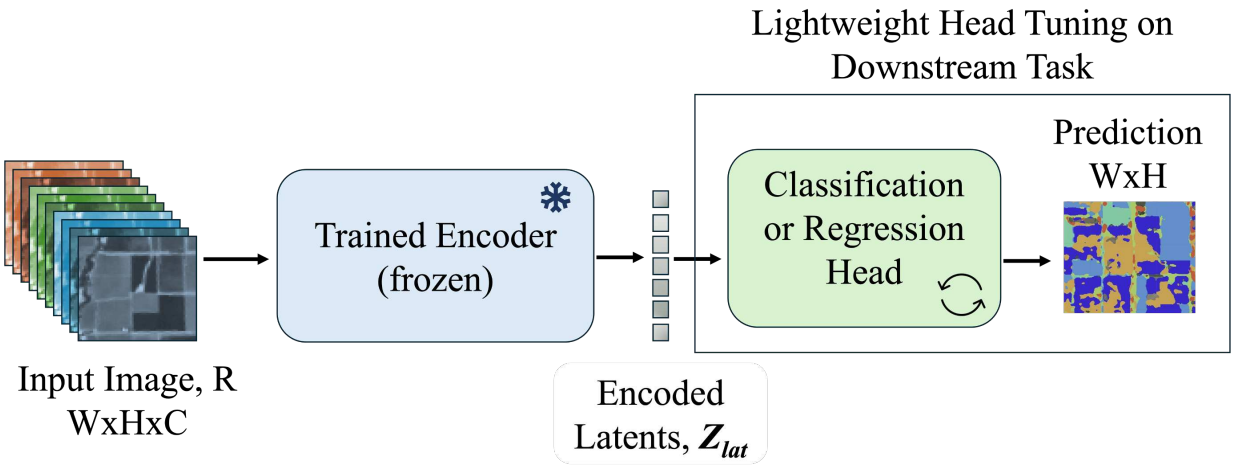

Figure 3.5: Downstream evaluation architecture for Thrusts 2 and 3 (Phase II). The pretrained encoder is frozen after Phase I reconstruction training and paired with a lightweight convolutional classification head trained on labeled targets. This two-stage protocol assesses representational transferability across classification and regression without updating encoder weights.

3.5.2 Task-Specific Convolutional Heads

To map the frozen latent representations to actionable Earth observation insights, the encoder is paired with task-specific, lightweight convolutional heads. Because the encoder weights remain

static, only the parameters within these shallow downstream heads are updated during Phase II training. This approach is highly computationally efficient and heavily penalizes poor pretraining, as the lightweight head lacks the parameter depth to compensate for weak latent features.

This architecture is seamlessly adapted to target a diverse array of downstream applications by modifying the terminal layers and loss functions of the convolutional head:

- **Classification Tasks:** For discrete semantic mapping applications, such as land cover classification, crop type identification, or anomaly detection, the convolutional head projects the latent embeddings into categorical logits. The head is optimized utilizing a standard Cross-Entropy loss. Superior performance in this regime indicates that the Phase I encoder successfully clustered spectrally similar materials into distinct, linearly separable boundaries within the latent space.

- **Regression Tasks:** For continuous biophysical and geophysical parameter retrieval, such as fractional abundance estimation, soil moisture mapping, or biomass prediction, the head is configured to output continuous scalar or vector values. The optimization relies on regression-centric objectives such as Mean Squared Error (MSE) or Huber loss. Success in these tasks demonstrates that the frozen representations preserved the fine-grained, physical gradients of the input spectra, verifying the efficacy of the science-guided constraints introduced in Thrust 3.

By evaluating the framework across both discrete classification and continuous regression paradigms, this methodology provides a comprehensive assessment of how effectively the proposed cross-modal distillation strategies generalize to real-world precision agriculture and environmental monitoring pipelines.

Chapter 4

Experiments and Results

4.1 Overview of Experimental Design

This dissertation proposes a unified framework of **Science-Guided Knowledge Distillation** to bridge the adaptability gap in scientific AI. The experimental validation is structured around three distinct but interconnected research thrusts, each addressing a specific barrier to deploying deep learning in Earth observation:

1. **Thrust 1 (Heterogeneous Domain-Shift Mitigation):** Addresses the *Resolution Gap*. We evaluate whether global wisdom learned from coarse satellite grids (teacher) can be distilled into a point-based student model to predict soil moisture at specific ground station locations, thereby overcoming extreme label sparsity.

2. **Thrust 2 (Inverse Knowledge Distillation):** Addresses *Cross-Modal Asymmetry*. We test whether a teacher model trained on abundant, low-resolution multispectral data can successfully guide a student model operating on scarce, high-resolution hyperspectral data. This effectively reverses the standard distillation paradigm to solve the "data poverty" issue in hyperspectral imaging.

3. **Thrust 3 (Science-Guided Domain Adaptation):** Addresses the *Physical Consistency Gap*. We experiment with embedding the Linear Spectral Mixing Model (LSMM) directly into the neural architecture. This thrust verifies if physics-based constraints improve the model's ability to disentangle complex material mixtures compared to purely statistical baselines.

The following sections detail the datasets, implementation, and results analysis for each of these thrusts.

4.2 Datasets and Study Areas

4.2.1 Soil Moisture and Ancillary Data

Soil moisture content is defined as the amount of water present in the soil on a volumetric basis (i.e., volume of water per unit volume of soil, typically expressed as m^3/m^3). Although SM is a critical parameter in various domains such as agriculture, hydrology, and climate science, using highly accurate in-situ sensors is not a suitable solution for obtaining SM due to the high spatial variability of SM measurements.

Traditionally, physics-based models have been widely adopted to estimate SM content [87, 88]. These models simulate the movement of water in the vadose zone by considering physical processes such as infiltration, evapotranspiration, and drainage. Environmental parameters, such as soil properties, weather conditions, vegetation characteristics, and topography, are critical for obtaining accurate values using physics-based models.

Remote sensing techniques, including microwave and thermal remote sensing, have been actively used to capture spatially distributed estimates of SM over large areas [89–91]. These techniques leverage the interaction between electromagnetic radiation and soil properties to infer SM content. For instance, microwave remote sensing exploits the sensitivity of SM to the dielectric properties of the soil, which affects the propagation and scattering characteristics of microwave radiation. Thermal remote sensing, on the other hand, utilizes the difference in thermal properties between wet and dry soil to estimate moisture content. However, the spatial and temporal heterogeneity of soil properties necessitates the need for SM measurements at higher spatial resolution. Soil Moisture Active Passive (SMAP), a widely known remote sensing observatory is designed to carry two instruments that map SM and determine the freeze or thaw state of the same area being mapped. SM content can be mapped via the radiometer data at a spatial resolution of 36 km every 2–3 days. Figure 4.1 shows a snapshot of the distributional heterogeneity of soil moisture between SMAP and in-situ sensors. Details about such satellite datasets are discussed in the following section.

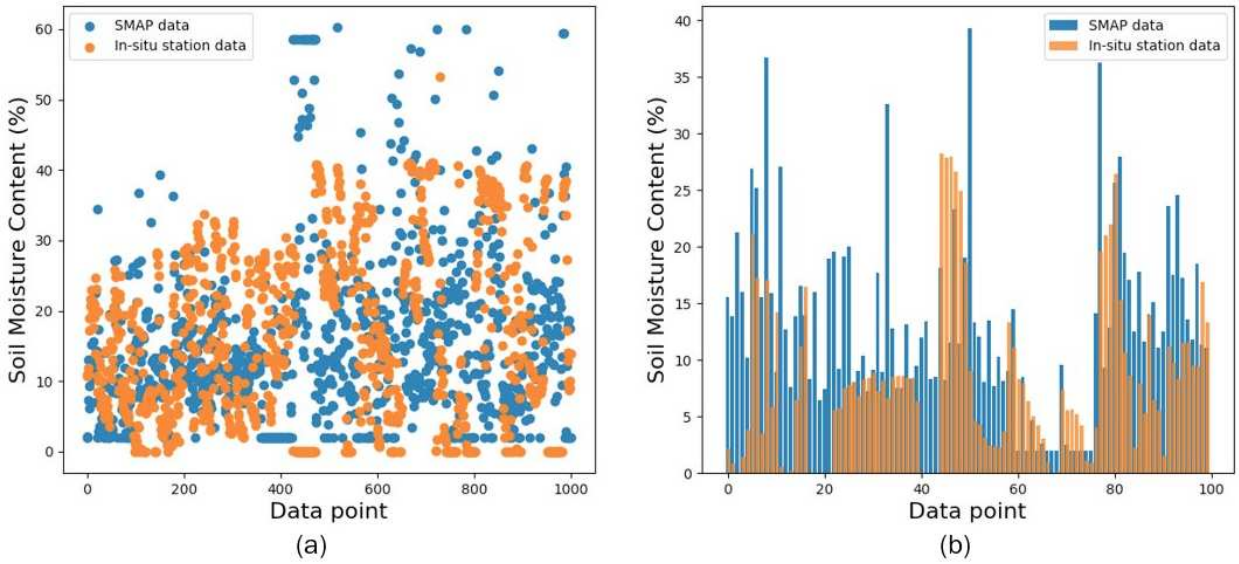

Figure 4.1: Comparison of SMAP satellite soil moisture estimates and in-situ ground station measurements for Colorado across the 2020–2021 period. Scatter plots show (a) 1,000 and (b) 100 randomly sampled data points. The systematic offset between the coarse-resolution SMAP product (3 km) and precise point-level in-situ readings quantifies the heterogeneous domain shift that motivates the proposed cross-resolution distillation framework.

In this study, one of the primary challenges is the heterogeneous spatial and temporal resolution of different datasets that we integrate. As depicted in Table 4.1, we used the Gridded National Soil Survey Geographic (gNATSGO) Database [92] which is a 10m resolution with multiple bands (9 bands were used for this study) representing different properties of soil. The Gridded Surface Meteorological (gridMET) dataset [93] offers daily surface meteorological data such as temperature, wind, humidity at the spatial resolution of 4km, 1/24 degree. We used 10 different bands of gridMET data.

We also integrate the USGS Shuttle Radar Topography Mission (SRTM) Digital Elevation model (DEM) [94], which is a depiction of topographic characteristics that specifically represents the natural terrain, excluding trees, structures, and other surface features covering entire US. We use elevations at 30m resolution. We use SMAP [89] which directly measures the amount of water in the top 5cm of surface soil everywhere on Earth. SMAP was first launched in January 2015 and data operation started from April 2015. It has several products starting from half orbit 36km. In our study we use L3_SM_A product which is a 3km daily product. The last dataset that we use is in-situ

Table 4.1: Multi-source datasets integrated for the cross-resolution knowledge distillation framework. The five heterogeneous inputs span soil composition (gNATSGO, 10 m), surface meteorology (gridMET, 4 km daily), terrain elevation (DEM, 30 m), satellite soil moisture (SMAP, 3 km daily), and sparse in-situ station observations. Spatial resolutions are downscaled to the SMAP grid (3 km) for teacher model training; the in-situ measurements serve as ground truth for student supervision.

Dataset	Source	Spatial Resolution	Temporal Resolution
gNATSGO	NRCS USDA	10m	one time
gridMET	NRCS USDA	4km	daily
DEM	USGS	30m	one time
SMAP	NASA	3km	daily
In-situ station	NOAA	single point	daily

sensor based weather station data that we primarily collect from National Oceanic and Atmospheric Administration (NOAA) data repository for 146 stations across California and Colorado. The temporal range for our study is 2020 to 2021 [95]. According to the experimental scenario, these states are divided into multiple sub-regions based on climatic characteristics. We organize the datasets into two scenarios. The first parts of experiments were performed on scenario-1 which includes data from Colorado that has an area of 269,837 km^2. For scenario-2, we utilized the Köppen Climatic Classification system [77] to segment our study area. The Köppen Classification is a widely employed climate classification system that offers hierarchical climatic regions. we chose five different climate classes (Bsk, Csa, Csb, Dfb, Dfc) and incorporate dataset for those classes from California and Colorado area.

To rigorously evaluate our proposed frameworks, we utilize a diverse set of multimodal Earth observation datasets spanning different spectral resolutions, spatial scales, and sensing modalities.

4.2.2 Hyperspectral and Multispectral Data

Experiments in Thrusts 2 and 3 use paired hyperspectral and multispectral imagery to evaluate cross-modal knowledge transfer. The two data sources are spectrally misaligned by design, representing the inverse domain shift that the proposed frameworks address.

EnMAP Hyperspectral Imagery

The **Environmental Mapping and Analysis Program (EnMAP)** satellite [31] provides the high-dimensional student domain. EnMAP offers contiguous spectral coverage from 420 nm to 2450 nm across 224 spectral bands at a spatial resolution of 30 meters. Six bands located within atmospheric water-vapor absorption windows (approximately 1350–1460 nm and 1800–2000 nm) exhibit systematic noise and are excluded prior to training, yielding **218 usable bands**. The satellite revisits each location approximately every 27 days, necessitating a temporal alignment strategy when pairing with higher-frequency multispectral acquisitions.

Preprocessing pipeline. Raw EnMAP Level-2A surface reflectance tiles are divided into non-overlapping 224×224 pixel patches to match the ViT-MAE input dimensions. Each patch covers a 6.72×6.72 km^2 ground footprint at 30 m resolution. Tiles with residual cloud contamination are excluded through a per-pixel quality-flag filter. Band-wise mean–standard deviation normalization is applied across the training set to standardize reflectance values over the full 218-band spectrum. The normalized patches are stored in Zarr format for efficient on-demand retrieval during GPU training.

Data splits and study regions. The primary training set covers California, comprising 5,000 tiles drawn from diverse land-cover types including urban areas, agricultural fields, dense forests, and shrubland. A disjoint validation set (500 tiles) and a primary test set (Test Dataset 1, 2,000 tiles) are held out with no spatial overlap with the training region. Geographic generalization is assessed with a second test set (Test Dataset 2) spanning distinct California sub-regions unseen during training, and a third test set (Test Dataset 3) that combines data from California, Colorado, and Kansas.

HLS Multispectral Imagery (Teacher Domain)

The **Harmonized Landsat Sentinel-2 (HLS)** dataset [27] provides the low-dimensional teacher domain. HLS harmonizes surface reflectance observations from the Landsat Operational Land Imager (OLI) and the Sentinel-2 Multi-Spectral Instrument (MSI) into a six-band product at approximately 30 m spatial resolution, with a combined revisit frequency of 2–3 days. The six HLS bands

used by the Prithvi-100M foundation model [18] correspond to the Blue (≈480 nm), Green (≈560 nm), Red (≈660 nm), Near-Infrared (≈860 nm), SWIR-1 (≈1610 nm), and SWIR-2 (≈2200 nm) spectral channels.

Temporal alignment. For each EnMAP tile, the closest cloud-free HLS observation within a ± 14-day window is selected to minimize phenological mismatch. Because the Prithvi model was pretrained on stacks of three consecutive HLS timestamps (18 input channels), the student training setting uses only the single matched HLS timestamp (6 channels) to maximize spatial correspondence with the EnMAP tile and accommodate the 27-day EnMAP revisit cadence.

Spectral domain alignment. HLS and EnMAP employ different spectral sampling schemes. A spectral aggregation step synthesizes HLS-equivalent bands from EnMAP data: for each HLS band b_i covering wavelength range $[\lambda_i^{\min}, \lambda_i^{\max}]$, the subset of overlapping EnMAP bands is averaged to form a synthetic counterpart $\hat{b}_i$ (see Section 3.3.3). This six-band synthetic representation enables the frozen Prithvi teacher to compute latent features from EnMAP-derived input, providing spectrally consistent supervisory signals during student training.

Downstream task datasets. Downstream evaluation uses two publicly available reference products. The **Cropland Data Layer (CDL)** [96] provides annual, crop-specific land-cover labels at 30 m resolution for the contiguous United States. The **National Land Cover Database (NLCD)** [97] offers broader ecological land-cover classifications at 30 m resolution. Both datasets are spatially aligned with the EnMAP test tiles for classification head evaluation.

4.3 Evaluation Protocol

The three research thrusts address qualitatively distinct learning problems and require different evaluation strategies. Thrust 1 evaluates a point-wise regression task; Thrusts 2 and 3 evaluate dense spectral reconstruction and downstream semantic classification. This section formalizes the evaluation structure and provides complete definitions of all performance metrics used throughout Chapter 4.

4.3.1 Thrust 1: Regression Evaluation for Soil Moisture Estimation

For the cross-resolution distillation framework, evaluation is structured as a direct regression comparison between models trained with and without knowledge distillation, using held-out in-situ station measurements as ground truth.

Experimental scenarios. Two geographic configurations are evaluated. Scenario 1 covers the state of Colorado (269,837 km^2). Scenario 2 extends to both Colorado and California, partitioned by the Köppen climate classification system [77] into five climate zones (Bsk, Csa, Csb, Dfb, Dfc). In both scenarios, training and test spatial extents are strictly non-overlapping to prevent information leakage.

Baselines. The teacher model (VGG13) is evaluated against SMAP-resolution labels. The student models (ResNet8 and MobileNetV2) are evaluated against sparse in-situ station measurements both with and without the distillation objective.

4.3.2 Thrusts 2 and 3: Two-Phase Evaluation for Spectral Reconstruction

Thrusts 2 and 3 share a two-phase evaluation structure.

Phase I: Reconstruction quality. The pretrained encoder-decoder pair is evaluated on its ability to reconstruct masked EnMAP patches, with both average and per-channel band-wise metrics reported.

Phase II: Downstream transfer. The pretrained encoder is frozen and paired with a lightweight convolutional classification head (Figure 3.5). Two downstream tasks assess representational quality: (1) **Crop Type Identification** using the Cropland Data Layer (CDL) [96], a multi-class classification problem at the species or crop level; and (2) **Land Cover Classification** using the National Land Cover Database (NLCD) [97], which spans broader ecological categories. For the physics-guided framework (Thrust 3), geographic generalization is assessed by training on California data and evaluating on Colorado and Kansas.

4.3.3 Performance Metrics

All metrics are computed on held-out test sets. For regression metrics, lower values indicate better performance; for reconstruction and classification metrics, higher values are preferable unless stated otherwise.

Mean Absolute Error (MAE)

Mean Absolute Error (MAE) is the primary metric for Thrust 1. It measures the average absolute discrepancy between the predicted soil moisture value $\hat{y}_i$ and the measured in-situ value y_i, both expressed as volumetric percentage:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|. \tag{4.1}$$

Mean Absolute Error (MAE) is preferred over mean squared error for this task because it penalizes all errors uniformly without inflating large outliers, and its units (percent soil moisture) are directly interpretable against field instrument precision. The dielectric probe method, the most reliable in-situ SM technology, reports a measurement uncertainty of approximately 4% [98]; model accuracy is benchmarked against this threshold. To quantify the contribution of knowledge distillation (KD), we report the percentage accuracy improvement:

$$I_{\text{acc}} = \frac{\text{MAE}_{\text{w/o KD}} - \text{MAE}_{\text{w/ KD}}}{\text{MAE}_{\text{w/o KD}}} \times 100. \tag{4.2}$$

Peak Signal-to-Noise Ratio (PSNR)

PSNR measures pixel-level reconstruction fidelity in decibels (dB) relative to the maximum possible signal. For reconstructed image $\hat{X}$ and reference X over N pixels with maximum value L:

$$\text{PSNR}(X, \hat{X}) = 10 \cdot \log_{10}\left(\frac{L^2}{\frac{1}{N}\sum_{i=1}^{N}(X_i - \hat{X}_i)^2}\right). \tag{4.3}$$

For hyperspectral data, PSNR is computed both as a macro-average across all 218 bands (Avg. PSNR) and per-channel, with the maximum per-channel value reported separately to highlight spectral band performance extremes. Higher PSNR indicates more faithful spectral recovery [99].

Structural Similarity Index Measure (SSIM)

SSIM assesses perceptual reconstruction quality by comparing local luminance, contrast, and structural statistics between the reconstructed and reference images [99]. For image patches X and $\hat{X}$ with means μ_X, $\mu_{\hat{X}}$, variances σ_X^2, $\sigma_{\hat{X}}^2$, cross-covariance $\sigma_{X\hat{X}}$, and small stability constants C_1, C_2:

$$\text{SSIM}(X, \hat{X}) = \frac{(2\mu_X\mu_{\hat{X}} + C_1)(2\sigma_{X\hat{X}} + C_2)}{(\mu_X^2 + \mu_{\hat{X}}^2 + C_1)(\sigma_X^2 + \sigma_{\hat{X}}^2 + C_2)}. \tag{4.4}$$

SSIM is bounded in $[-1, 1]$, with a value of 1 indicating perfect structural identity. Unlike PSNR, which treats all pixel-level deviations as equally significant, SSIM accounts for the joint statistics of luminance, contrast, and spatial correlation that are perceptually relevant and ecologically meaningful when evaluating spectral reconstructions used in downstream land-cover classification. Both average SSIM across all bands and per-channel maximum SSIM are reported.

Spectral Angle Mapper (SAM)

SAM quantifies the angular distance between the reconstructed spectrum $\hat{\mathbf{r}}$ and the reference spectrum $\mathbf{r}$ in the high-dimensional spectral space [85, 86]:

$$\text{SAM}(\mathbf{r}, \hat{\mathbf{r}}) = \arccos\left(\frac{\langle \mathbf{r}, \hat{\mathbf{r}} \rangle}{\|\mathbf{r}\|_2 \cdot \|\hat{\mathbf{r}}\|_2 + \epsilon}\right). \tag{4.5}$$

SAM is expressed in radians (lower is better). Because it computes the angle between spectral vectors rather than their magnitudes, SAM is invariant to multiplicative illumination effects which is a property that makes it particularly suitable for material identification tasks in which spectral shape, not absolute brightness, encodes land-cover class. SAM serves dual roles in this disser-

tation: as a component of the physics-guided training loss (Section 3.4) and as a supplementary reconstruction evaluation metric.

Top-1 Accuracy

For downstream classification, Top-1 Accuracy measures the proportion of test samples for which the highest-confidence class prediction matches the ground-truth label:

$$\text{Top-1 Acc} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[\hat{c}_i = c_i], \tag{4.6}$$

where $\hat{c}_i$ is the predicted class and c_i is the true class. Per-class accuracy is reported alongside the macro-average to expose performance variation across spectrally ambiguous land-cover categories such as Shrubland and Cultivated Crops, which are frequently confused by purely data-driven models.

Mean Intersection over Union (mIoU)

mIoU, also termed the Jaccard Index, measures the spatial overlap between predicted and ground-truth segmentation regions across all C land-cover classes [100]:

$$\text{mIoU} = \frac{1}{C}\sum_{c=1}^{C} \frac{|P_c \cap G_c|}{|P_c \cup G_c|}, \tag{4.7}$$

where P_c and G_c are the predicted and ground-truth pixel sets for class c. mIoU is stricter than Top-1 Accuracy: a model that classifies labels correctly at the pixel level but produces spatially diffuse boundaries will incur a lower mIoU score. This makes mIoU particularly informative for geospatial monitoring applications, where precise land-cover boundary delineation affects downstream agronomic and ecological analysis.

4.4 Thrust 1: Heterogeneous Domain-Shift Mitigation

4.4.1 Experimental Setup

Model Configuration

The teacher model is a **VGG13** convolutional network [73] containing 13 convolutional layers and three fully connected layers, totaling approximately 9.41 million trainable parameters. It accepts input tensors of shape $32 \times 32 \times 10$ and is trained to predict coarse-resolution SMAP soil moisture values as a regression target. For comparative ablation, a **ResNet34** backbone was also evaluated but yielded no material advantage over VGG13 at considerably higher memory cost.

The student model is a **ResNet8** network [74] with 8 residual blocks and approximately 78.5K parameters, roughly $120\times$ fewer than the teacher. For selected ablation experiments, a **MobileNetV2** model (2.1M parameters) is used as an intermediate-capacity student. Both student architectures accept the same $32 \times 32 \times 10$ input shape as the teacher to facilitate feature-map distillation without dimensional mismatch.

Training Details

The teacher is trained first on SMAP-labeled data until validation loss converges. The student is then trained in a second phase with the teacher frozen, using the combined distillation loss L_{Total} from Equation (3.2). The student loss backpropagates exclusively through student parameters. All models are trained using an L1 regression objective against the respective ground-truth labels (SMAP for teacher, in-situ station measurements for student). Hyperparameter weights LF_1, LF_2, LF_3 are selected empirically through grid search over the CA+CO validation split. A 70–30% train-test split is used for the teacher, and an 80–20% split for the student, with strict spatial non-overlap enforced between partitions.

Hardware

All experiments are conducted on a single NVIDIA GPU with standard deep-learning configurations. The lightweight student model (ResNet8, 78.5K parameters) requires substantially less

memory than the teacher (VGG13, 9.41M parameters), achieving a training time of 0.4 ms per sample at inference compared to 2.2 ms per sample for the teacher, a 5.5× reduction in inference latency.

Data Preprocessing

Our data preprocessing involves several steps, including aligning the spatial resolutions of the datasets through downscaling (e.g., gNATSGO and DEM), as well as cropping aligned images to focus on the target regions. For the model accuracy evaluation (Section 4.4.2), we generated approximately 13 million training samples to train the teacher model using observations from the state of Colorado. The spatial resolutions were adjusted to match the resolution of the target dataset, SMAP satellite images (3km). We downscaled gNATSGO (10m) and DEM (30m) using aggregation to align with the 3km resolution of SMAP. The SM maps from SMAP were cropped to a pixel dimension of 6×4. Training student models with the SM measurements from the ground stations involves additional steps of data preprocessing. Due to the limited number of stations, only corresponding small geospatial bounds of the input and SMAP satellite data are available at a time in our study area (each bound covering an approximate area of 18×12 km^2). As student models estimate SM values at specific locations, we selected values from each dataset based on their geospatial proximity to the target location. We used a 70-30% train-test split for the teacher model and 80%-20% train-test split for the student model to conduct the experiments. The model sensitivity evaluation (Section 4.4.3) and the performance analysis of the nested KD model training (Section 4.4.4) involve extensive comparisons across seasons and climactic areas. To support these evaluations, we generated approximately 500,000 samples for the teacher model and 73,000 samples for the student model from datasets covering the states of Colorado and California. This broader geographic scope led us to use a total of 123 unique small geospatial bounds (each containing at least one in-situ weather station) for generating training samples for the student model. In selecting SMAP image samples for inclusion, we considered only those small geospatial bounds containing soil surfaces and less than 20% null values. Areas with more than 20% invalid values in the gNATSGO dataset and GRIDMet were excluded. For the teacher model, we chose a pixel

from the SMAP image based on its proximity to the weather station with an in-situ SM sensor. We reduced the spatial extent for the teacher model to account for possible variability in soil properties. We ensured that the test dataset had no overlapping spatial extent with the dataset used for model training to avoid information leakage. We conducted a 70-30% train-test split for teacher model and consider samples from the unique areas for testing the teacher model to accomplish the experiments.

Table 4.2: Soil moisture prediction accuracy with and without Knowledge Distillation for Colorado. The VGG13 teacher is evaluated against coarse SMAP satellite labels; the ResNet8 and MobileNetV2 student models are evaluated against sparse in-situ station measurements. MAE (%) is reported for a held-in validation region and a spatially unseen test region, with the resulting accuracy improvement from applying KD.

Model	**Ground Truth**	**Validation MAE(%)**		**Test MAE(%) for unseen region**		**Improvement(%)**	
		without KD	With KD	without KD	with KD	Validation	Test
VGG13	SMAP	5.1		6.1			
ResNet8	Station SMC	6.7	4.6	20.6	17.7	31	14
MobileNetV2	Station SMC	9.2	5.5	20.7	16.2	40	22

Model Training Details

As illustrated in Figure 3.1, training proceeds in two sequential phases. The teacher model is trained first until it reaches stable performance. The student model is then trained by passing inputs through both networks in parallel: the distillation loss is computed from intermediate feature maps and the output logits of the frozen teacher, while the student supervision loss uses the same ground-truth labels. The combined loss from Equation (3.2) is backpropagated exclusively through the student model.

Evaluation Metrics

Thrust 1 uses Mean Absolute Error (MAE) as the primary quantitative metric; see Section 4.3 for the formal definition and justification. Briefly, MAE computes the average absolute difference between predicted and measured soil moisture values (in %). We also compute the percentage

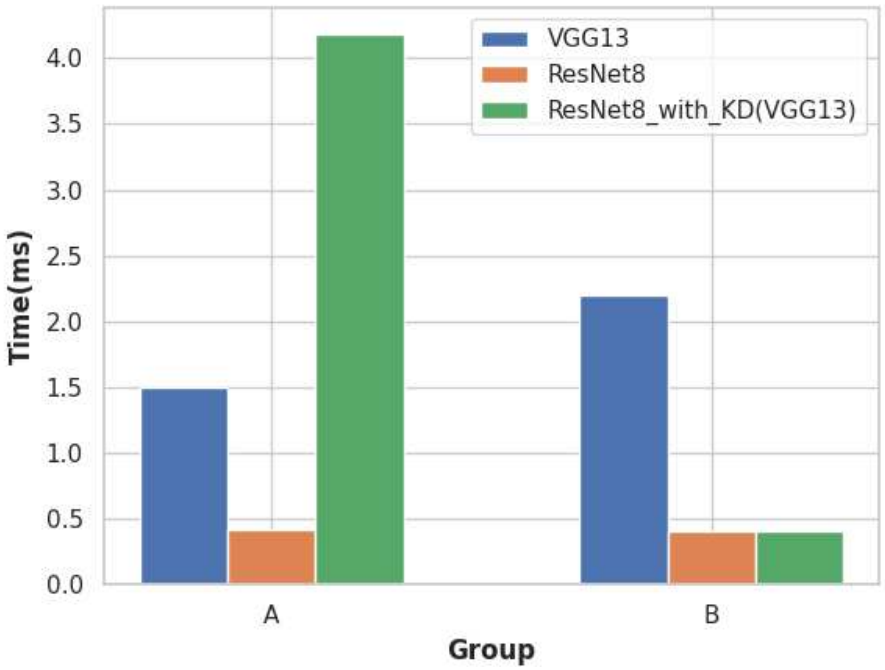

Figure 4.2: Per-sample average training and inference latency (ms) for the VGG13 teacher and ResNet8 student models, with and without Knowledge Distillation (California and Colorado). Although KD increases student training time due to the additional teacher forward pass, inference latency is identical to the standalone student (0.4 ms), yielding a 5.5× reduction over the VGG13 teacher (2.2 ms) at deployment.

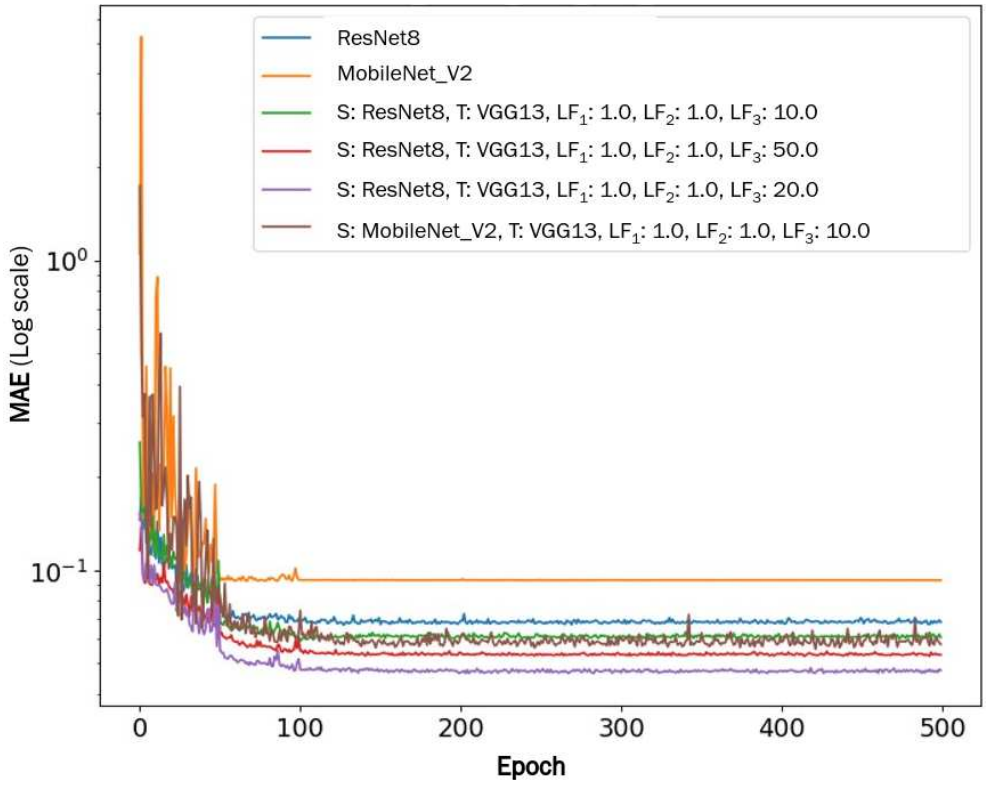

Figure 4.3: Accuracy of the student model and the proposed distillation framework under different combinations of loss weighting factors LF1, LF2, and LF3 (Area: Colorado).

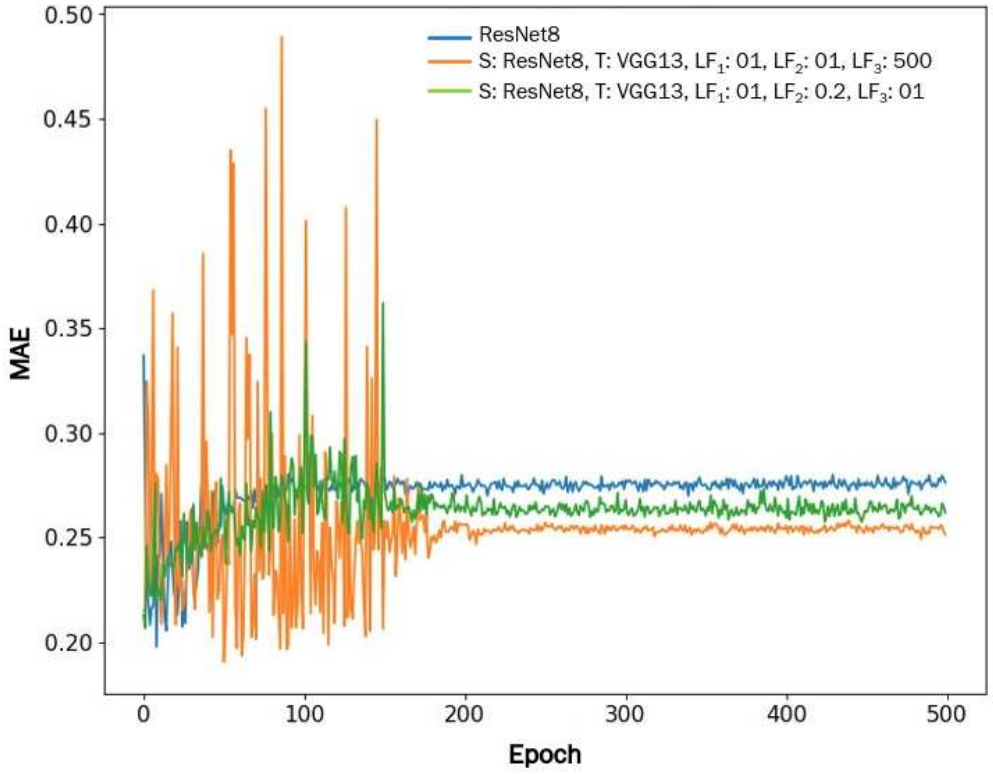

Figure 4.4: Accuracy of the student model and the proposed distillation framework under different loss weighting combinations for spatially unseen regions (Area: Colorado).

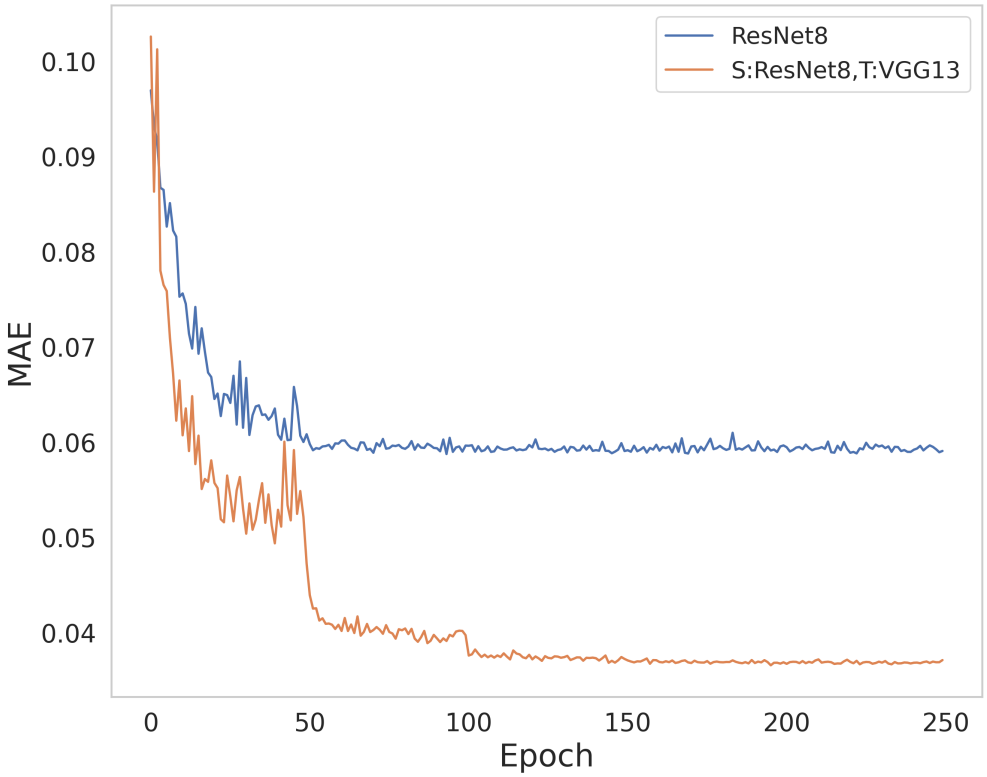

Figure 4.5: Comparative model accuracy of ResNet8 (student without distillation) and the proposed cross-resolution distillation framework (Area: part of California and Colorado).

accuracy improvement I_{acc} to isolate the distillation contribution:

$$I_{acc} = \frac{MAE_{Model_{w/oKD}} - MAE_{Model_{withKD}}}{MAE_{Model_{w/oKD}}} \times 100. \tag{4.8}$$

Inference latency (ms per sample) and trainable parameter count are additionally reported as efficiency metrics relevant to edge-deployment requirements.

4.4.2 Model Performance Analysis

Model Accuracy

In this evaluation scenario, we employed VGG13 as the teacher model and ResNet8 as the student model using 70% of all the available data covering the entire area of Colorado. As shown in Table 4.2, the proposed cross-resolution distillation framework demonstrates a validation accuracy of 95.4%. Generally, SM ranges from 10% to 45%, except during or after watering. Considering that the most reliable in-situ SM sensor technology, the dielectric probe method, estimates surface SM with an error level of 4% [98], our model's performance is highly reliable for various applications. We also compared the model accuracies with and without the KD approach. The validation accuracy improved by up to 40% when we employed the KD approach for model training.

To assess the effectiveness of our approach in regions that were not part of the model training, we evaluated our model's accuracy in areas remote from the ones used for training. To create the testing data samples for this experiment, we divided the entire Colorado region into unique 18×12km sub-regions and trained and tested our model with data samples from groups of these sub-regions without any overlap. As shown in Table 4.2, the model's accuracy improved by 14-20% when we applied KD technology to the unseen area.

Model Training Time, Inference Latency, and Memory Requirement

In our methodology, the teacher model is a sophisticated deep-learning network, while the student model adopts a space-efficient architecture. In Figure 4.2, we compared the training time

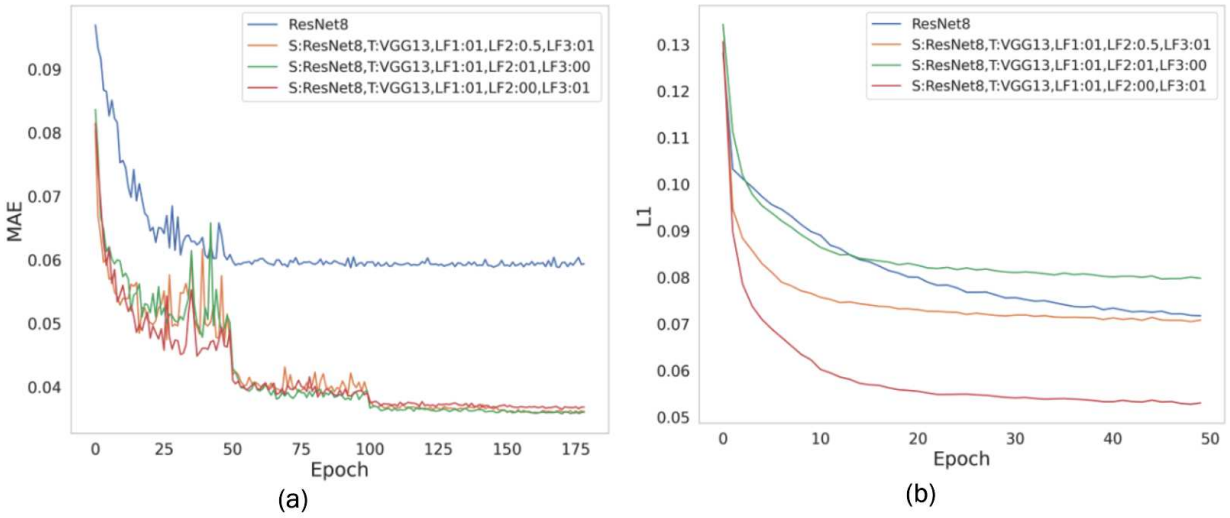

Figure 4.6: Effect of different learning factors (LF1, LF2, LF3) on (a): Model accuracy and (b): Model Convergence (Area: part of California and Colorado)

(for a single sample) and inference latency for VGG13 (teacher model) and ResNet8 (student model) with and without the KD approach.

Because the teacher models have a higher number of learnable parameters (VGG13 has approximately 9.41M parameters in our evaluation), they require more time for training and consume more memory. In contrast, student models are relatively lightweight (ResNet8 has approximately 78.5K parameters in our evaluation) and require less time for training, as well as frugal memory consumption [101]. Due to the involvement of VGG13 while training a student model, the training time for the student model was longer compared to the case without the KD approach. However, our model achieves equivalent inference latency (0.4 milliseconds) compared to ResNet8 without KD. Overall, our approach provides highly accurate predictions with significantly lower latency (5.5$\times$ lower than VGG13) compared to the complex teacher model (VGG13 at 2.2 milliseconds).

Effectiveness of the KD components

To understand how each component of KD in the distillation loss contributes to model training, we conducted a microbenchmark by applying different sets of weights within our loss function while using ResNet8 as the student model. In Figure 4.3, we compare the effectiveness of the KD approaches collectively employed by the proposed framework. We have repeated the same

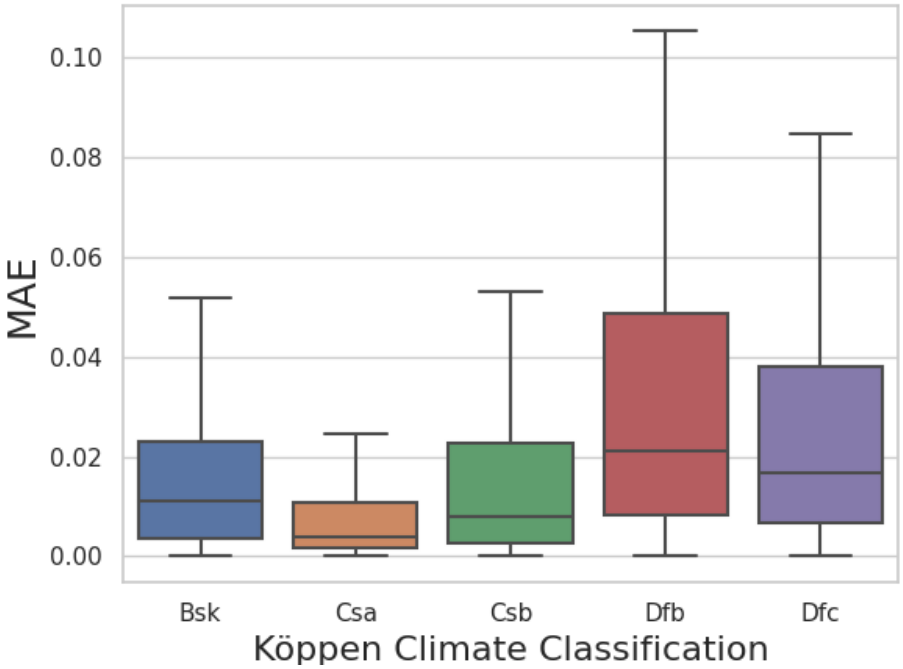

Figure 4.7: Distribution of model accuracy across different Köppen climate zones. The boxplots illustrate the performance variability and spatial robustness of the framework within specific climatic regions across the California and Colorado study areas.

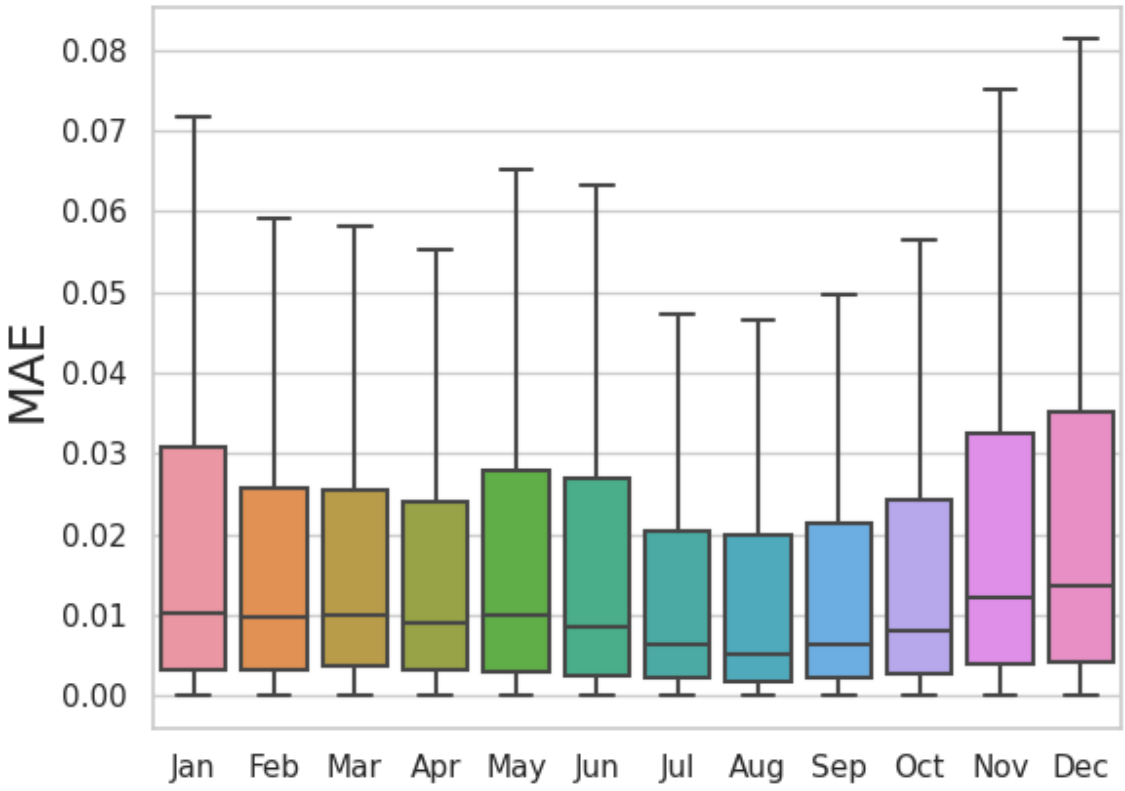

Figure 4.8: Temporal analysis of model accuracy. The month-wise boxplots demonstrate the framework's performance consistency and robustness to seasonal variations over the studied timeframes in California and Colorado.

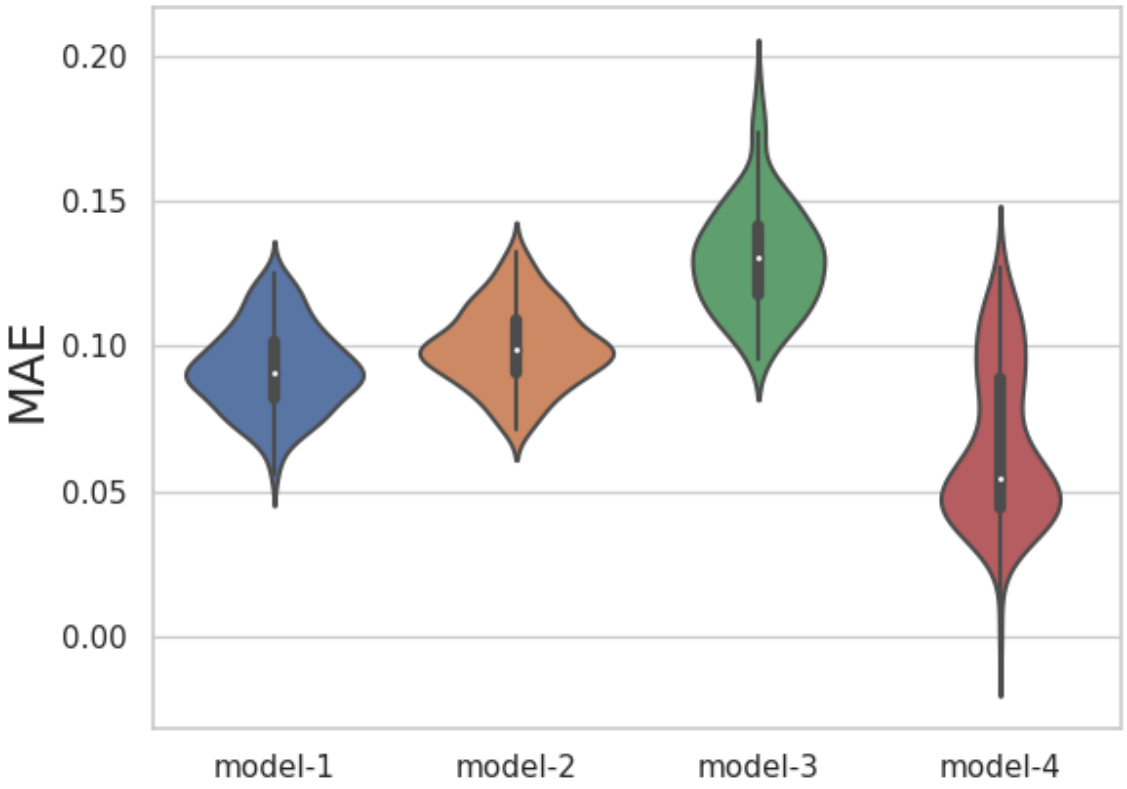

Figure 4.9: Scalable transfer learning evaluation for the proposed framework across unseen climatic areas. The violin plots depict the spatial generalization capabilities of the model when trained on specific climate zones and tested on disparate, unseen regions. The transfer learning scenarios evaluated are: (1) Trained on CA (Bsk, Csa, Csb) and tested on CO (Bsk, Dfb, Dfc); (2) Trained on CO (Dfc) and tested on CO (Bsk, Dfb); (3) Trained on CA (Csb) and tested on CA (Bsk, Csa); and (4) Trained on CO (Bsk) and tested on CA (Bsk).

evaluation with the dataset scenario-2 listed in Section 4.3 (Figure 4.6). In the figures, LF1, LF2, LF3 are the three multiplicative factors which decide how much loss is considered from each of student model loss, teacher model output loss and feature distillation loss respectively (equation. 3.2). The results in Figure 4.3 to Figure 4.6 show that KD not only improve the model accuracy but also guide the student model to converge earlier. Besides, the combination of the KD and feature distillation losses with different weights contribute to the student model training and performance differently. Empirically we can find the appropriate combination that gives the best improvement.

4.4.3 Model Sensitivity Analysis

To assess model performance under various conditions, we conducted sensitivity analysis across seasons and climatic zones. To enhance diversity and have sufficient training samples, we expanded our target area to include both Colorado and California in this evaluation. For this, we utilized VGG13 as the teacher model and ResNet8 as the student model. With the extended spatial extent, our student model achieved a similar overall accuracy, with a mean absolute error (MAE) of 5.9% (Figure 4.5). Furthermore, the use of KD improved the student model's accuracy by 37% compared to ResNet8 without KD.

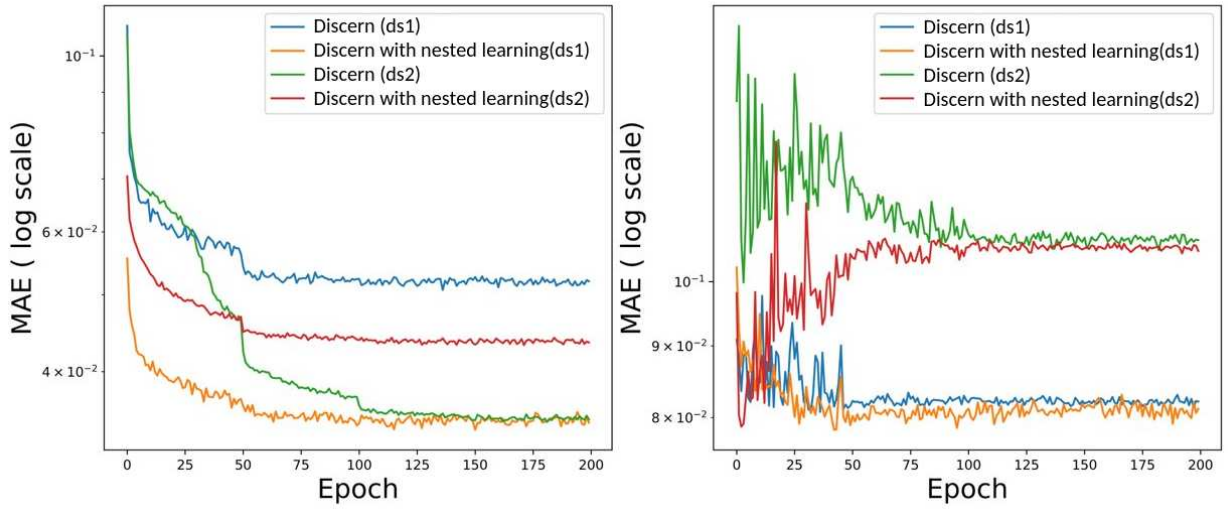

Figure 4.10: (a): Training accuracy and (b): Test accuracy of the proposed distillation framework with and without nested training for unseen regions. Scenario 1: trained on CO(Bsk), tested on CO(Dfc); Scenario 2: trained on CO(Dfc), tested on CO(Bsk, Dfb).

We evaluated the model performance over different climatic zones specified in the Koppen climate classification system. We compared the model performance across five climate zones that are most popular in Colorado and California. Our model performs the best in the Cs (Csa and Csb, dry summer climate) and Bsk (Semi-arid climate). These climate zones are the majority of Colorado and California. The individual average MAE of each climate class was lower than 2% throughout these regions (Figure 4.7). The model accuracy in the Dfb (warm summer humid continental climate) region was relatively lower than other climate zones due to the relatively low number of samples from the area.

Figure 4.8 presents a monthly sensitivity analysis, which assesses the model's ability to predict SM effectively across different months of the year. The analysis highlights the model's strongest performance during the months of July, August, and September when the weather patterns in California and Colorado exhibit notable consistency. In contrast, predictions during the winter months exhibit a comparatively wider range of mean absolute error (MAE) variation when compared to other seasons. These variations can be attributed to unique outliers primarily caused by snow effect. During this time, usually the soil surface is covered with snow most of the time and the external factors like temperature, wind have little effect on the SM [102].

4.4.4 Performance Evaluation of Nested Training using KD

The proposed distillation framework harnesses insights captured by the sophisticated teacher model using coarse-resolution observations. In this section, we assess the effectiveness of the proposed framework in training diverse student models, which is applicable for model training over a large spatial extent. As described in Section 3.2.6, our nested training involves a single teacher model and shared by multiple student models based on the climate zones. In this experiment, the teacher model was trained over the Colorado and California regions. Using this teacher model, we subsequently trained student models over four distinct scenarios; (1) California (with the climate zones of Bsk, Csa, and Csb , (2) California (with the climate zone of Csb only), (3) Colorado (with

the climate area of Dfc only), and (4) Colorado (with the climate area of Bsk only). All these student models were trained with the same teacher model.

To evaluate our model's capability to handle areas with extremely sparse labels, we assessed model accuracy using teacher models trained in different areas or under similar climatic zones. Figure 4.9 and 4.10 demonstrates the consistently high accuracy observed across these diverse experiments. Notably, we observed that models trained with data from climates similar to the target region tend to exhibit superior accuracy. In the climate zone Bsk in Colorado, where labels are extremely scarce, instead of training a student model from the limited available labels, we employed the California-trained student model, which yielded promising accuracy.

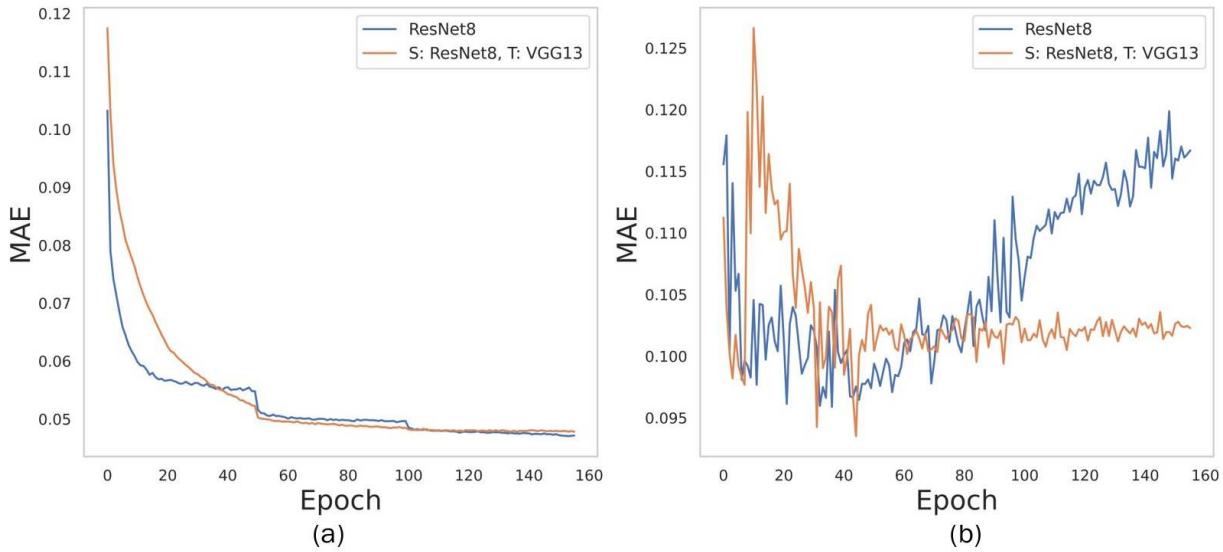

Figure 4.11: (a) Training accuracy and (b) test accuracy of the student model with and without KD when trained on California climate zones (Bsk, Csa, Csb) and evaluated on geographically unseen Colorado zones (Bsk, Dfb, Dfc). KD maintains prediction accuracy in regions entirely outside the training distribution.

4.4.5 Model Performance with Calibration

Post-hoc model calibration directly impacts prediction accuracy across the combined California and Colorado datasets. As shown in Figure 4.12(a), the calibration process reduces both the average and maximum Mean Absolute Error (MAE) of the predictions. Furthermore, the frequency

distribution of these errors shifts following calibration. Figure 4.12(b) demonstrates an increased proportion of low-error predictions across the MAE bins. This distribution shift indicates that the post-hoc calibration consistently corrects larger deviations and improves the overall reliability of the model outputs.

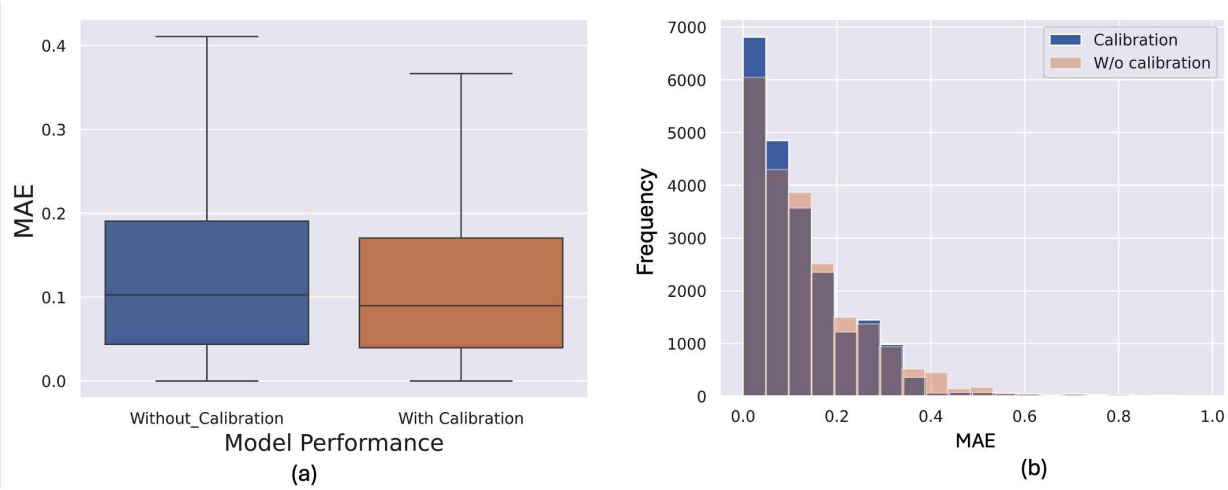

Figure 4.12: Effect of post-hoc model calibration on prediction error (combined California and Colorado). (a) Average and maximum MAE reduction achieved after calibration. (b) Frequency distribution shift showing increased proportion of low-error predictions following calibration across MAE bins.

4.5 Thrust 2: Inverse Knowledge Distillation

4.5.1 Experimental Setup

Model Configurations

The teacher network for this framework is the **Prithvi-100M** foundational model. Prithvi is constructed on a Masked Autoencoder (MAE) framework using a Vision Transformer (ViT) backbone, consisting of 12 transformer blocks optimized for 18-channel inputs (6 spectral bands across 3 temporal steps). While the teacher and student share the same fundamental ViT architecture, the student network is designed to process significantly higher-dimensional data, accepting 218 hyperspectral channels.

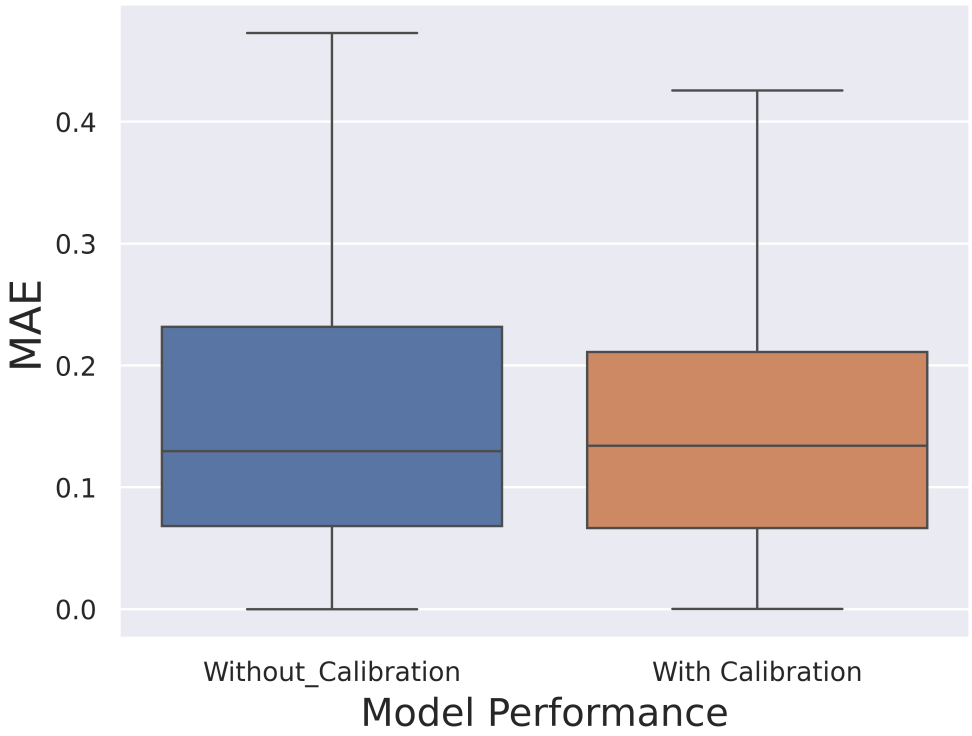

Figure 4.13: Distribution of prediction accuracy (MAE) for the cross-climate transfer scenario trained on California (Csb) and evaluated on California (Bsk, Csa). Boxplots summarize MAE variability across test locations within each target climate zone.

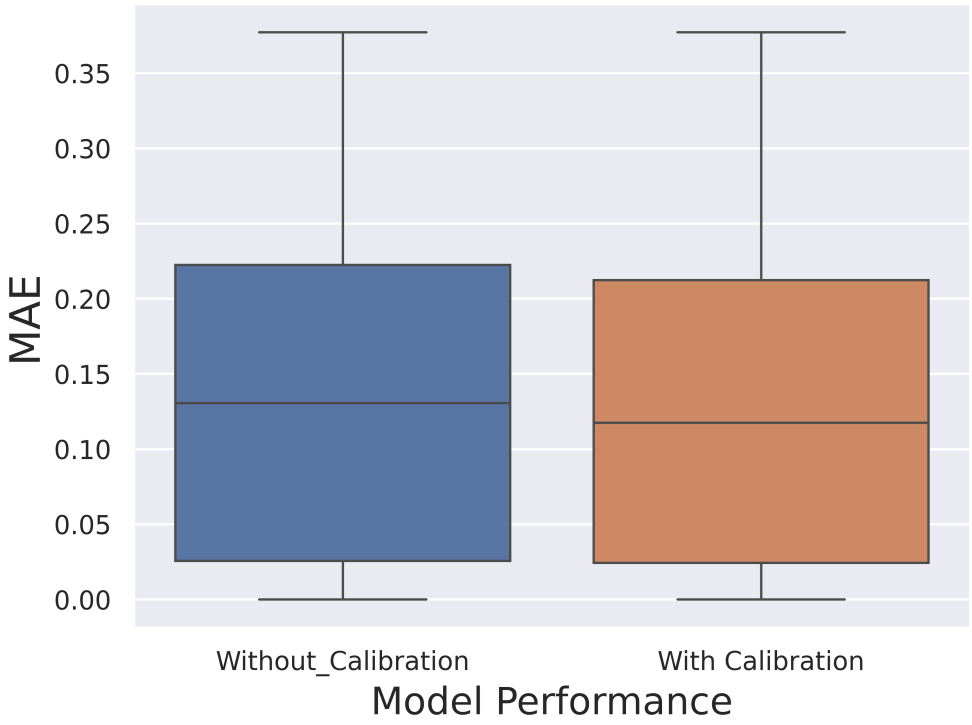

Figure 4.14: Scalable transfer learning evaluation of the proposed framework across unseen climatic areas. The boxplots illustrate generalization capabilities across distinct spatial and climatic boundaries (1: CO(Bsk, Dfb, Dfc), 2: CO(Bsk, Dfb), 3: CA(Bsk)).

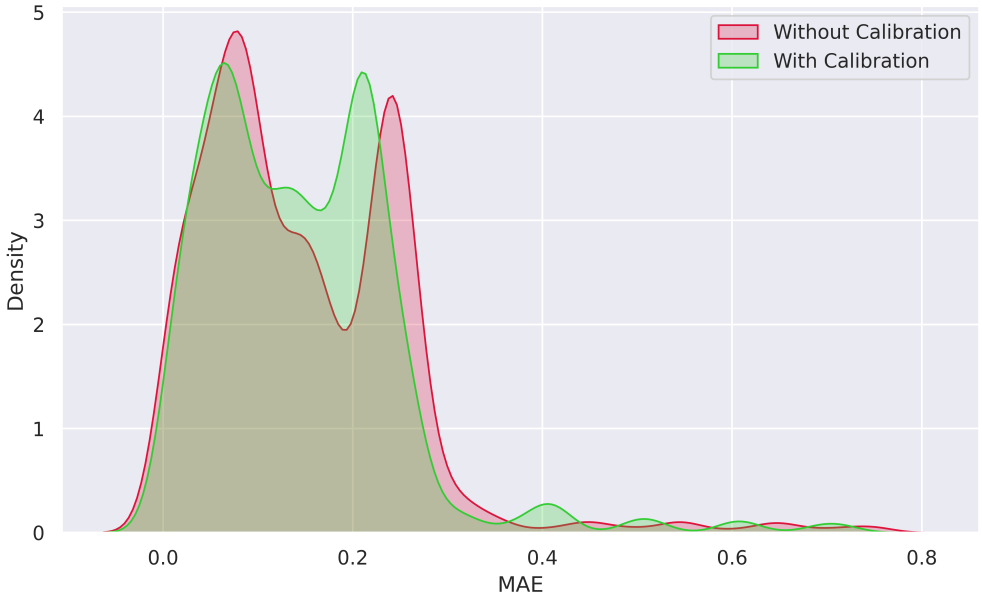

Figure 4.15: Kernel density estimate of prediction MAE for the transfer scenario trained on California (Csb) and tested on California (Bsk, Csa). The distribution illustrates the concentration of low-error predictions when the student is deployed in climatically similar but spatially unseen regions.

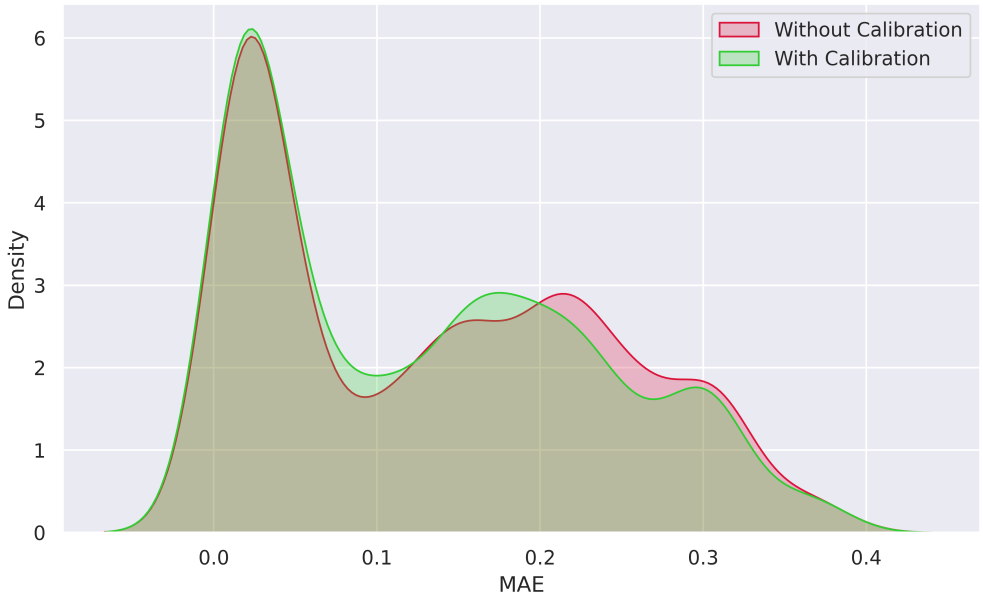

Figure 4.16: Kernel density estimate of prediction MAE for the cross-climate transfer scenario trained on Colorado (Dfc) and tested on Colorado (Bsk, Dfb). The distribution shows generalization from a subarctic continental training climate to arid and humid continental test zones.

To ensure a rigorous comparison and maintain complexity parity with the teacher, the baseline model (**BaseKD**) utilizes a small MAE configuration. The encoder employs a 16×16 patch size with an embedding dimensionality of 768, 12 transformer layers, 12 attention heads, and a Multi-Layer Perceptron (MLP) hidden dimensionality of 3072. The corresponding decoder uses an embedding dimensionality of 512 across 12 transformer layers and 16 attention heads, with an MLP hidden dimensionality of 2048. This specific backbone is identical across the student model and the proposed framework, regardless of the increased input dimensionality.

Implementation Details

All models are trained using the Adam optimizer with a weight decay of 0.01. We utilize a custom learning rate scheduler that initiates at 1×10^{-4} and decays to a minimum of 1×10^{-6}, supported by a 10-epoch warm-up phase. Input imagery, sized at $224 \times 224 \times 218$ (height, width, and spectral bands), is partitioned into 16×16 patches, and a 75% masking ratio is applied via a Gabor filter-based strategy. The student model processes unabridged EnMAP data, whereas the

frozen teacher model operates on $224 \times 224 \times 6$ HLS bands; both models utilize mean–standard deviation normalization.

Training is conducted on a single NVIDIA A100 (40GB) GPU for 300 epochs. To achieve an effective batch size of 32, we employ a hardware batch size of 4 with gradient accumulation over 8 iterations. Optimization for the student involves a composite reconstruction loss of Mean Squared Error (MSE) and Structural Similarity Index (SSIM), supplemented by embedding-level distillation from the teacher using weighted Kullback–Leibler Divergence (KLD). For efficient I/O, data is stored in Zarr format and fetched on-demand to minimize memory overhead.

Evaluation Metrics

Thrust 2 adopts the two-phase evaluation protocol described in Section 4.3. Reconstruction quality is assessed using **PSNR** (Eq. 4.3) and **SSIM** (Eq. 4.4), reported as both overall averages and per-channel values across the 218 EnMAP bands. Downstream classification performance is measured by Top-1 Accuracy and mIoU on the CDL and NLCD datasets. Complete metric definitions are provided in Section 4.3.

4.5.2 Ablation Experiments

Because high-dimensional hyperspectral training incurs substantial computational cost, the ablation studies were conducted on a fixed subset of 1,000 training, 100 validation, and 300 test samples. Training was limited to 40 epochs for the KD loss function comparisons and 150 epochs for the KD layer selection study to allow rapid iteration while keeping the experimental conditions controlled.

All ablation trials used a 75% masking ratio and Mean Squared Error (MSE) as the primary reconstruction loss. The first ten epochs of each trial served as a warm-up phase with a reduced learning rate to stabilize early convergence.

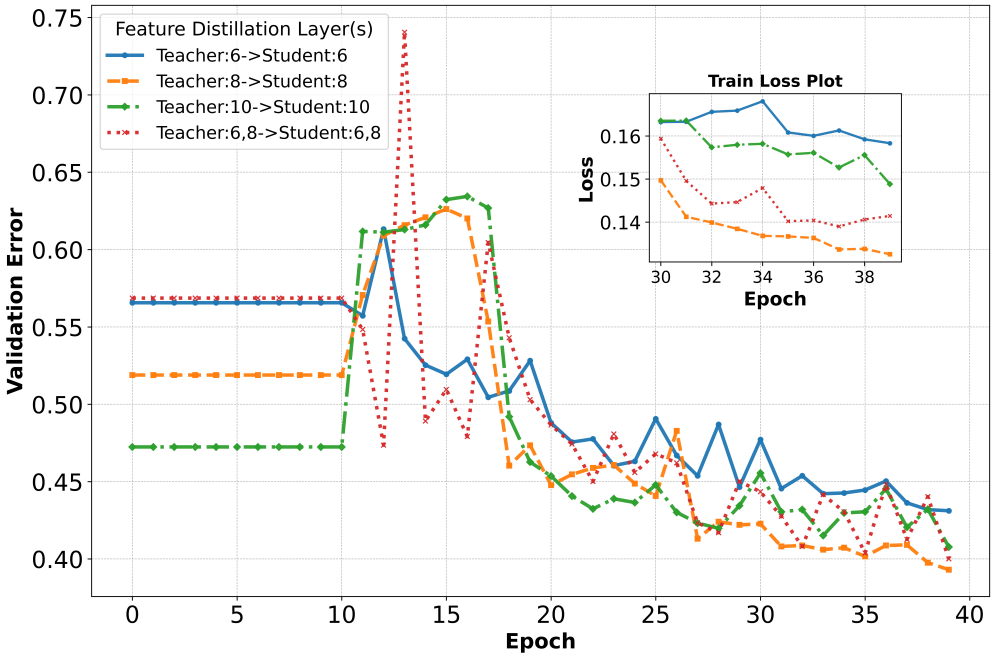

Figure 4.17: Impact of Teacher-Student Layer Pairings on Feature Distillation Convergence. The plot compares convergence trends across various layer combinations, with stability enhanced by an initial 10-epoch warm-up phase.

Selecting the Masking Ratio

The masking ratio (MR) determines the difficulty of the pretext reconstruction task and directly affects both the quality of learned representations and training throughput. We evaluated several MR configurations to identify an appropriate balance between reconstruction fidelity, computational efficiency, and generalization to unseen spectral patterns, as detailed in Table 4.3.

Table 4.3: Ablation Study: Effect of Train/Test Mask Ratios on Reconstruction and Error Metrics.

Train MR	Test MR	PSNR	SSIM	MAP
50%	50%	**32.76**	**0.824**	**2.32**
50%	75%	30.39	0.756	3.21
85%	85%	29.42	0.718	3.73
85%	75%	30.30	0.742	3.30
75%	75%	30.37	0.748	3.23

Empirical results indicate that while a lower MR (50%) produces the highest raw performance scores (PSNR: 32.76, SSIM: 0.824, MAP: 2.32), it is suboptimal from a training efficiency per-

spective. Low masking ratios provide the encoder with a denser visible token set, which reduces the information bottleneck and the computational pressure to learn robust, high-level features. In contrast, higher MRs (75%–85%) enforce a more aggressive information compression, which not only accelerates the training process but also encourages the model to learn complex spatial-spectral dependencies. For these trials, we utilized random masking coupled with a wavelet transform to identify patch priority. Consequently, we adopted a 75% MR for all subsequent experiments as the best compromise between reconstruction accuracy and feature scalability.

Selecting the Optimal KD Layer

The teacher and student architectures both utilize a ViT-MAE backbone composed of 12 transformer blocks. Identifying the specific layer from which to extract distillation features matters considerably, as lower layers generally encode generic low-level features while higher layers capture task-specific abstractions. We conducted an ablation to identify the most informative representations within the mid-to-top hierarchy of the transformer.

As evidenced by the training and validation convergence curves in Figure 4.17, feature distillation performed at Layer 8 resulted in superior student model convergence and stability compared to either earlier or later blocks. Layer 8 appears to represent a "sweet spot" where features are sufficiently abstract to be informative yet general enough to remain transferable across different spectral dimensions. Thus, Layer 8 was finalized as the KD anchor layer for all following experiments.

Choosing the Best KD Function

The mathematical formulation of the KD objective significantly influences the alignment of the student's feature space with the teacher's. We evaluated three common divergence metrics: L_1 loss, Kullback-Leibler Divergence (KLD), and Jensen-Shannon (JS) Divergence.

The findings summarized in Table 4.4 reveal that KLD provides the most effective supervision, achieving the highest PSNR (27.39) and SSIM (0.65). This suggests that matching the probability distributions of the latent features allows the student to better capture the structural nuances of the

Table 4.4: Ablation Study: Impact of Different KD Loss Functions on Reconstruction Metrics.

KD Loss Function	Avg. PSNR	Max Channel PSNR	Avg. SSIM	Max Channel SSIM
L_1	26.86	32.78	0.63	0.84
KLD	**27.39**	**33.48**	**0.65**	**0.85**
JS	27.08	33.36	0.64	0.85

teacher's output. While L_1 provided moderate performance, JS Divergence proved less effective in our specific high-dimensional spectral context. Therefore, KLD was selected as the primary distillation objective.

Table 4.5: Image reconstruction comparison: the proposed inverse distillation framework, with specialized loss functions and masking strategies, outperforms the student trained without KD and the Baseline KD model on two datasets.

ID	Model	Recon. Loss	KD Loss	Masking Strategy	Salient Patches	Start Rand. Masking	Test Dataset 1				Test Dataset 2 (Unseen Area)	
							Avg. PSNR	Per Channel PSNR (Max)	Avg. SSIM	Per Channel SSIM (Max)	Avg. PSNR	Avg. SSIM
1	Student	HUBER	-	-	-	-	24.61	32.45	0.55	0.82	25.48	0.52
2	Base KD	HUBER	L1	Random	-	epoch 1	27.10	33.49	0.65	0.85	27.56	0.62
3	Ours†	MSE + SSIM	KLD	Wavelet	visible	-	27.70	34.32	0.70	0.87	28.05	0.68
4	Ours‡	MSE + SSIM	KLD	Wavelet	masked	epoch 100	30.80	37.39	0.76	0.91	30.64	0.73
5	**Ours**	MSE + SSIM	KLD	Gabor Filter	masked	epoch 100	**31.02**	**37.56**	**0.77**	**0.91**	**30.62**	**0.73**

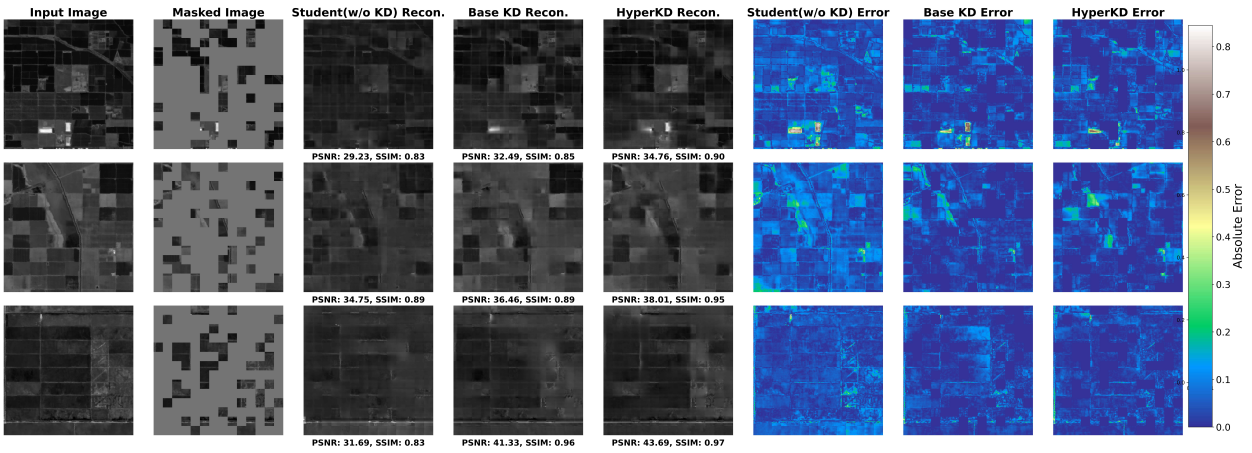

Figure 4.18: Image reconstruction quality comparison: the proposed inverse distillation framework versus Baseline KD and the student model trained without distillation. The proposed method achieves the highest PSNR and SSIM scores across all test samples in 218-band spectral analysis.

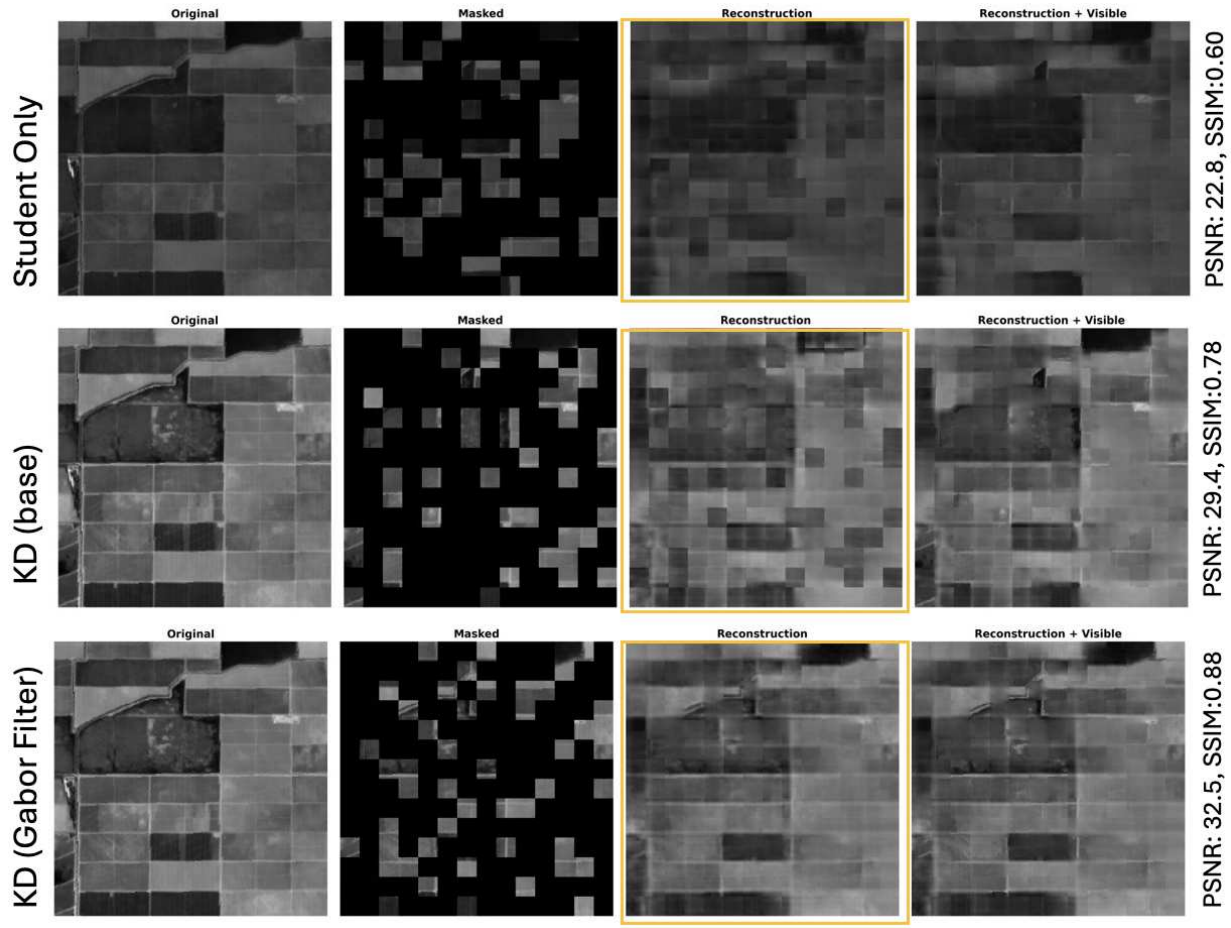

Figure 4.19: Qualitative spectral reconstruction comparison on held-out EnMAP test patches. False-color composites (representative spectral bands mapped to RGB) compare the original hyperspectral imagery against reconstructions by the student without KD, the Baseline KD model, and the proposed inverse distillation framework. The proposed method recovers finer spectral contrast and sharper spatial boundaries, particularly in spectrally heterogeneous terrain.

Table 4.6: Reconstruction performance of the proposed inverse distillation framework across extensive geospatial regions (CA, CO, KS), demonstrating superior generalization compared to baseline methods.

ID	Model	Recon. Loss	Masking Strategy	**Test Dataset 3** (CA+CO+KS)	
				Avg. PSNR	Avg. SSIM
1	Student	HUBER	-	25.57	0.41
2	Base KD	HUBER	Random	31.73	0.80
3	**Ours**	MSE + SSIM	Gabor Filter	**33.62**	**0.85**

4.5.3 Effectiveness of the Specialized Loss Function

The composite objective function in Equation 3.11 addresses the dual demands of spectral accuracy and spatial coherence in HSI reconstruction. Adding the Structural Similarity Index Measure (SSIM) as a loss component had the most substantial effect: as shown in Table 4.5, average SSIM rose from 0.65 to 0.77 while PSNR improved from 27.10 to 31.02, reflecting the student model's ability to recover fine-grained spatial structure rather than minimizing pixel-wise error alone. The KLD distillation loss further aligns the student's latent representations with the teacher's, acting as a regularizer that prevents the student from collapsing to trivial spectral reconstructions.

4.5.4 Impact of Spatial Feature-Driven Masking

The transition from uniform random masking to spatially-aware masking strategies produced consistent gains in reconstruction quality. By employing Gabor filters and wavelet transforms to prioritize high-variability patches, the model concentrates reconstruction effort on spectral regions where information density is highest. As shown in Table 4.5, wavelet-guided masking with the composite loss reached a PSNR of 30.80 and an SSIM of 0.76, compared to 27.10 PSNR and 0.65 SSIM under random masking. The highest performance (PSNR: 31.02) was obtained through a two-stage schedule: an initial phase of Gabor-based masking followed by random masking from epoch 100 onward. This schedule allows the model to capture both high-frequency structural

patterns and broader spectral trends, as confirmed by its consistency on the geographically unseen test set (Table 4.5, Test Set 2).

4.5.5 Inverse-Shift Spectral Domain Adaptation

The proposed inverse distillation framework is specifically designed to address the "inverse domain shift," where a high-dimensional student model (218 EnMAP channels) learns from a low-dimensional teacher (6 HLS synthetic channels). Despite the massive dimensionality gap, the distillation methodology enables the student to reconstruct the full EnMAP spectrum with exceptional fidelity (PSNR: 31.02, SSIM: 0.77). This confirms that the proposed approach successfully bridges disparate spectral domains, allowing the rich spatial-spectral priors of the teacher model to guide the student in high-dimensional HSI reconstruction. The consistency of these results across California, Colorado, and Kansas (Table 4.6) proves the robustness of this transfer learning strategy.

4.5.6 Spectral Channel Level Reconstruction Performance

As illustrated in Figure 4.20, we conducted a band-by-band analysis of reconstruction accuracy to understand how KD influences specific spectral ranges. By aligning the spectral windows of HLS and EnMAP (subsection 3.3.3), we ensured that the teacher provided direct supervision where the sensors overlap.

The experimental results show that the proposed inverse distillation framework significantly outperforms a student model without KD across the entire EnMAP spectrum. The performance gains are naturally most prominent in the overlapping spectral bands, where the teacher's pre-trained HLS knowledge is directly applicable. However, we also observe meaningful improvements in non-overlapping bands, which indicates that the teacher's guidance helps the student learn a more generalized latent representation of the Earth's surface, effectively addressing the inverse domain gap and enhancing the cross-domain transferability of the framework.

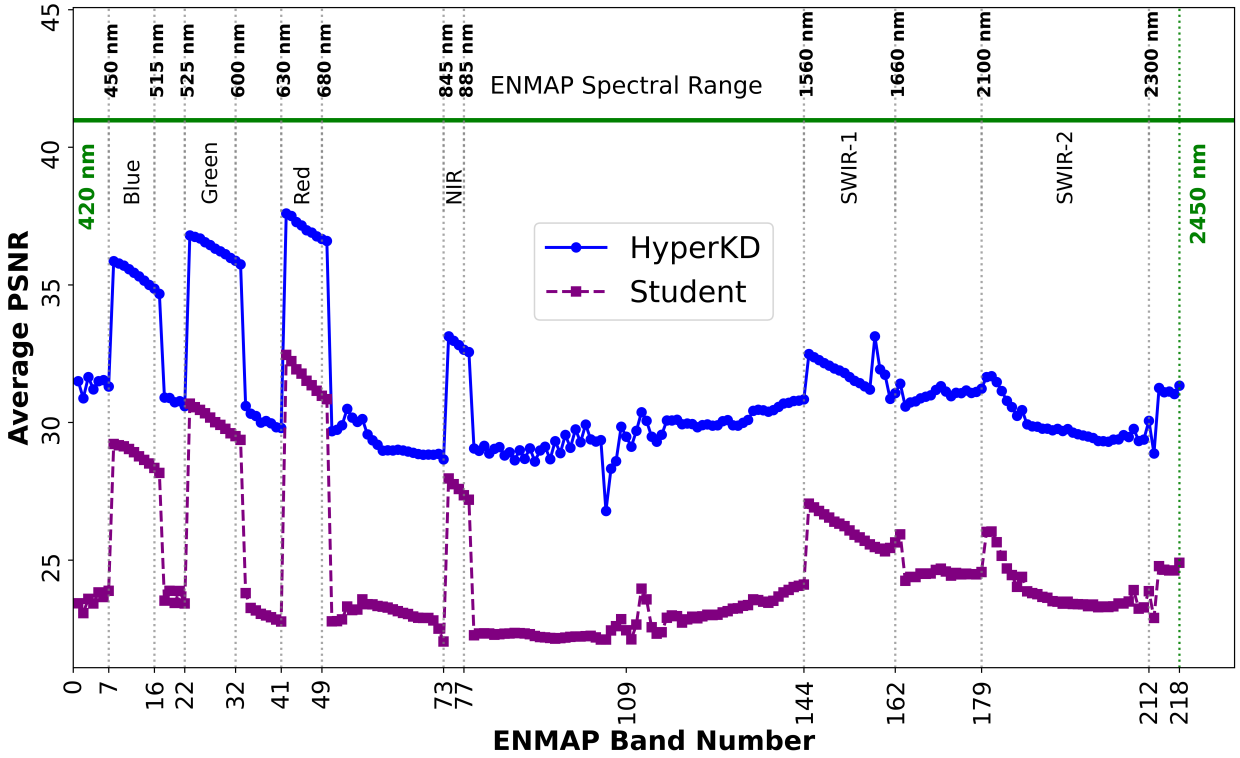

Figure 4.20: Band-wise PSNR across all 218 EnMAP spectral channels for the proposed inverse distillation framework versus the student trained without KD. Gains are largest in the spectral windows that directly overlap with HLS teacher bands, where the teacher provides explicit supervision. PSNR improvements also extend into non-overlapping bands, indicating that the teacher's spatial priors generalize beyond the HLS spectral range to improve reconstruction of channels the teacher never directly observed.

4.5.7 Computational and Hardware Requirements

GPU and Memory Utilization. To evaluate the accessibility of the proposed framework, we tested it on both enterprise-grade (NVIDIA A100, 40GB) and consumer-grade (NVIDIA RTX 4000, 20GB) hardware. While a standalone student model requires approximately 6.3 GB of VRAM, the proposed framework increases this requirement to 9.7 GB (a $\sim$1.5$\times$ increase). This overhead is primarily attributed to the storage of activations from the frozen teacher during the distillation pass. Notably, because the teacher model's parameters are not being updated, the increase in trainable parameters is negligible (182.2M vs. 182.8M). By utilizing gradient accumulation (8 iterations), we maintained a stable batch size of 32 even on hardware with 10 GB of memory.

Training and Inference Efficiency. The training runtime per step for the proposed framework is approximately 2.6$\times$ slower than the student model due to the extra forward pass through the teacher shown in Table 4.7. However, the resulting convergence quality and performance gains justify this overhead. Crucially, during inference, the teacher model is discarded. This means that the student model provides high-accuracy, teacher-guided performance at a deployment cost identical to a standard model, making it ideal for real-time edge applications.

Table 4.7: Hardware and runtime comparison illustrating the computational trade-offs of our approach. The framework requires approximately 1.5$\times$ more GPU memory and 2.6$\times$ more per-step runtime due to the frozen teacher network, but introduces negligible overhead to the trainable parameter count (Batch Size = 4, Gradient Accumulation = 8).

Metric	Student Only	Ours	Relative Overhead
Total Parameters	182.4 M	183.0 M (S) + 112.9 M (T)	$\sim$1.6$\times$ larger
Trainable Parameters	182.2 M	182.8 M	$\approx$ Negligible
Max GPU Memory	6.3 GB	9.7 GB	$\sim$1.5$\times$ higher
Per-step Runtime	0.55 Sec	1.45 Sec	$\sim$2.6$\times$ slower

4.5.8 Downstream Task Evaluation

The pre-trained encoder serves as a robust backbone for diverse hyperspectral applications. By integrating the encoder with task-specific heads, we can address both classification (Land Cover,

Crop Identification) and regression (Soil Organic Carbon prediction) problems. In this work, we specifically evaluate the frozen encoder's ability to provide discriminative spectral-spatial features for environmental monitoring.

Land Cover Classification (NLCD)

Table 4.8: Land cover classification accuracy (Top-1, %) on the NLCD dataset across California, Colorado, and Kansas (Test Dataset 3). The proposed inverse distillation framework (Ours) is compared against the student encoder without KD and the standard KD baseline (BaseKD) on three spectrally ambiguous land cover classes.

Class Name	Top-1 Accuracy (%)		
	Student	**BaseKD**	**Ours**
Shrub/Scrub	63.18	62.00	**63.56**
Herbaceous	35.43	38.70	**46.25**
Cultivated Crops	56.70	77.18	**85.50**

We utilized the 30m NLCD dataset [97] for broad land cover classification, focusing on five primary classes in California. As shown in Table 4.8, the proposed method consistently outperforms the student-only baseline, particularly in spectrally complex categories such as "Cultivated Crops" and "Other." Even when using a single EnMAP timestamp, our model achieves results comparable to the Prithvi model, which requires multi-temporal inputs. This demonstrates that our KD approach effectively captures the structural textures necessary for accurate land cover mapping.

Crop Type Identification (CDL)

The CDL dataset [96] provides a more granular challenge by focusing on specific crop types. Results in Table 4.9 show that the proposed framework surpasses the baseline across major crop categories. The improvements are most evident in structurally intricate classes where subtle inter-class boundaries are defined by fine-grained spectral signatures. Despite the known label noise and

Table 4.9: Crop type identification accuracy on the USDA Cropland Data Layer (CDL) for the top four land-cover classes in California. Top-1 Accuracy (%) and mean Intersection over Union (mIoU, %) are reported for the student encoder without KD and the proposed inverse distillation framework.

Class Name	Student + CNN Head		Ours + CNN Head	
	Top-1 Acc (%)	mIoU (%)	Top-1 Acc (%)	mIoU (%)
Shrubland	40.31	34.74	**55.93**	**46.00**
Grassland/Pasture	22.74	16.85	**25.95**	**19.50**
Evergreen Forest	81.72	53.04	**93.62**	**56.44**
Other	90.87	62.02	87.68	72.44
Mean	67.41	41.66	**73.27**	**48.60**

class imbalance in CDL, the proposed framework remains robust, confirming its adaptability for specialized agricultural monitoring.

Soil Organic Carbon (SOC) Prediction

Table 4.10: Soil organic carbon (SOC) regression accuracy on the gNATSGO dataset, reporting Mean Absolute Error (MAE) for top-30 cm soil layer predictions. The proposed framework (Gabor Filter masking) achieves the lowest MAE (0.038) compared to the student without KD (0.045) and the standard KD baseline (0.046), demonstrating that the teacher's spectral-spatial priors transfer effectively to continuous-valued regression tasks.

Model	SOC Prediction (Mean Absolute Error)
Student w/o KD	0.045
Baseline KD	0.046
Ours w/ Gabor Filter	**0.038**

To demonstrate the framework's versatility in regression, we predicted top 30-cm SOC content using the gNATSGO dataset [92]. The results in Table 4.10 show that the proposed method achieves a Mean Absolute Error (MAE) of 0.038, significantly outperforming both the student-only model (0.045) and standard KD (0.046). This indicates that the knowledge transferred from

the teacher successfully bridges the spectral gap, creating features that are highly effective for continuous value regression in remote sensing.

4.6 Thrust 3: Science-Guided Domain Adaptation

4.6.1 Experimental Setup

Model Configuration

The physics-guided masked autoencoder is built on a ViT-MAE backbone with a ViT-Large configuration for the teacher and a ViT-Base configuration for the student. The teacher encoder is pretrained and frozen throughout student training. Both backbones use a patch size of 16×16. The student decoder incorporates a dedicated abundance head: a compact multi-layer perceptron of shape $D \to D/2 \to M$, where D is the decoder embedding dimension and M is the number of endmember components. A Softmax activation enforces the non-negativity and sum-to-one constraints required by the Linear Spectral Mixing Model (LSMM). The endmember matrix $\mathbf{A} \in \mathbb{R}^{218 \times M}$ is initialized randomly and learned end-to-end alongside the decoder, functioning as a low-rank physics bottleneck.

The Science-Guided Knowledge Distillation (SGKD) extension applies the frozen physics-guided teacher to guide a compact student through two additional distillation paths: latent feature alignment via $\mathcal{L}_{\text{feat}}$ and physical abundance alignment via $\mathcal{L}_{\text{phys-distill}}$, as defined in Section 3.4.5.

Implementation Details

All models are trained using the AdamW optimizer with a weight decay of 0.01 and a cosine-decaying learning rate schedule starting at 1×10^{-4}. Training is conducted for 300 epochs on a single NVIDIA RTX 4000 GPU (20 GB) with a batch size of 32. A 75% masking ratio is applied to input patches during pretraining. The hybrid training objective combines Huber loss, SAM loss, and physics-consistency loss, with weights λ_1, λ_2, λ_3 selected empirically on the validation set. The endmember count M is determined through the ablation study described in Section 4.6.2.

Data and Evaluation

The primary pretraining dataset consists of 5,000 EnMAP tiles from California, with 500 tiles held for validation and 2,000 for testing. Downstream classification tasks (CDL and NLCD) are evaluated using the two-phase protocol described in Section 4.3. Geographic generalization is assessed by evaluating the California-pretrained encoder on held-out tiles from Colorado and Kansas without any fine-tuning.

4.6.2 Ablation Study

The abundance head, which predicts the fractional material composition of each patch, is parameterized by the number of endmember components M. This hyperparameter governs the model's capacity to represent mixed-material pixels: too few components risks underfitting complex spectral mixtures, while an excess can cause over-parameterization and impede convergence. We conducted an ablation over several values of M to identify the configuration that best balances physical expressiveness and training stability.

4.6.3 Physics-Informed Reconstruction Analysis

Reconstruction performance is evaluated using PSNR (Eq. 4.3), SSIM (Eq. 4.4), and SAM (Eq. 4.5), as defined in Section 4.3. The efficacy of the proposed science-guided distillation framework is primarily evidenced by its ability to reconstruct complex hyperspectral signals while adhering to physical constraints. As demonstrated in Table 4.11, the integration of the **Linear Spectral Mixing Model (LSMM)** and **Spectral Angle Mapper (SAM)** loss provides significant improvements over purely data-driven baselines. The proposed method achieves an average PSNR of 27.38 dB, representing a relative gain of 11.3% over the ViT-MAE baseline (24.61 dB).

The most compelling evidence for science-guided modeling is observed in the **Structural Similarity Index Measure (SSIM)**. The proposed framework improves the SSIM from 0.55 to 0.68, a 23.6% relative increase. These higher structural fidelity scores suggest that by distilling knowledge through a physics-informed teacher, the student model preserves meaningful spectral-spatial

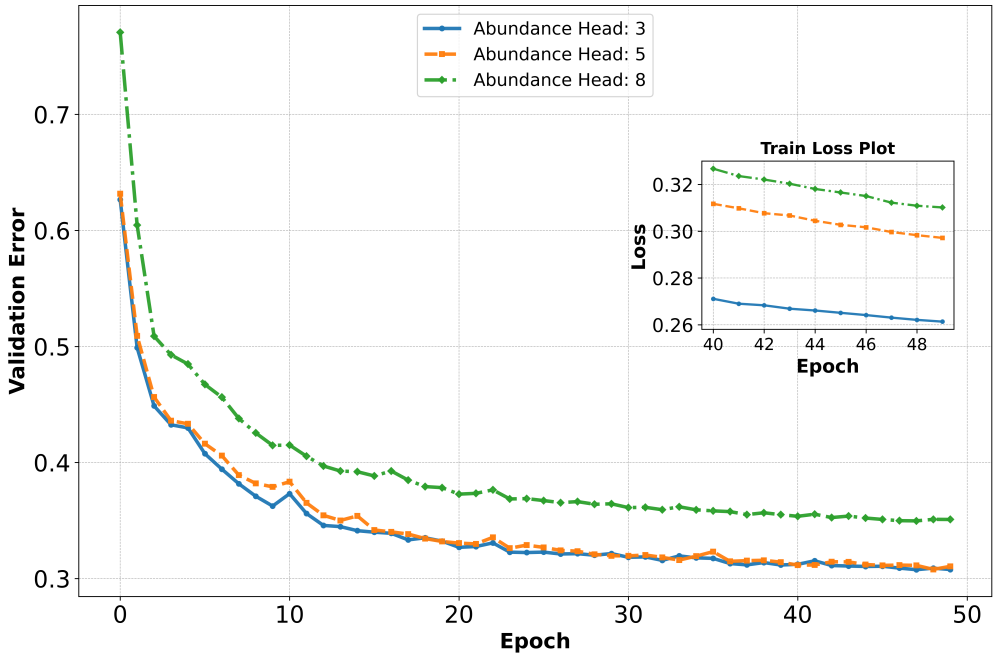

Figure 4.21: Ablation study on the number of endmember abundance head components M. Performance improves as M increases from small values, enabling richer material decomposition, but plateaus and declines at large values due to redundancy and optimization difficulty. The configuration $M = 10$ achieves the best balance between reconstruction fidelity and training efficiency and was adopted for all main experiments.

patterns that purely statistical methods overlook. By explicitly minimizing the residual between observed signals and LSMM-reconstructed signals via $\mathcal{L}_{\text{phys}}$, the latent representations are forced to align with physical theory.

Table 4.11: Reconstruction accuracy comparison between the proposed physics-guided framework and ViT-MAE on hyperspectral data. Physics-informed learning objectives consistently yield superior reconstruction fidelity over the data-driven baseline.

ID	Model	Reconstruction Loss	Physical Constraints	Avg. PSNR (↑)	Per-Channel PSNR (Max) (↑)	Avg. SSIM (↑)	Per-Channel SSIM (Max) (↑)
1	VitMAE	HUBER	-	24.61	32.45	0.55	0.82
2	**Ours**	HUBER	SAM + LSMM	27.38	35.10	0.68	0.88
Improvement (% over ViTMAE)				**11.26**	**8.17**	**23.64**	**7.32**

Table 4.12: Crop Type Identification on CDL data (top three classes in California). The proposed physics-guided framework with a lightweight CNN significantly outperforms the ViT-MAE baseline across all spectrally complex land-cover classes.

Class Name	**VitMAE + Head**		**Ours + Head**	
	Top-1 Acc(%) (↑)	mIoU (%) (↑)	Top-1 Acc(%) (↑)	mIoU (%) (↑)
Shrubland	40.31	34.74	71.18	51.35
Grassland/Pasture	22.74	16.85	44.22	31.47
Evergreen Forest	81.72	53.04	85.03	56.30
Average	48.26	34.88	**66.81**	**46.37**

Table 4.13: Land Cover Classification on NLCD data (top three classes across CA, CO, and KS). The proposed physics-guided encoder, trained solely on California data, generalizes effectively to Colorado and Kansas, demonstrating robust cross-regional transfer.

Class Name	**Top-1 Accuracy (%)** (↑)	
	ViTMAE + Head	**Ours** + Head
Shrub/Scrub	63.18	70.50
Herbaceous	35.43	46.59
Cultivated Crops	56.70	91.59
Average	51.77	**69.56**

4.6.4 Physics-Based Interpretability and Generalization

A core contribution of the proposed science-guided distillation approach is the transformation of arbitrary feature vectors into physically interpretable abundance maps. This grounding in physical laws (e.g., material percentages for soil, water, and vegetation) significantly enhances the transferability of learned features. As shown in Figure 4.22, the physics-guided student converges more stably and reaches a lower validation loss than the data-driven baseline, indicating that the physics-informed teacher provides a coherent supervisory signal (see the comparison in Table 4.14) that prevents the overfitting common when fine-tuning foundation models on limited labeled data.

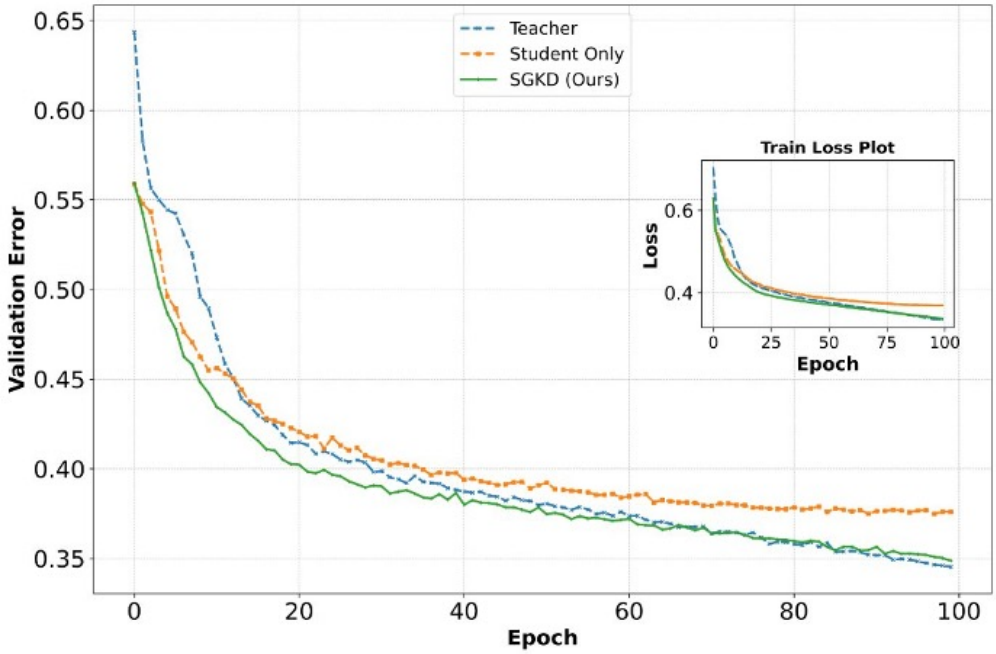

Figure 4.22: Convergence comparison of the proposed science-guided distillation framework versus the student-only baseline. The physics-informed supervisory signal from the teacher accelerates convergence and reduces final validation loss.

Cross-Domain Downstream Performance (CDL)

The benefits of distillation are most visible in downstream **Crop Type Identification**. As shown in Table 4.12, the distilled student, augmented with a lightweight CNN head, outperforms

Table 4.14: Reconstruction and estimation accuracy of the proposed Science-Guided Knowledge Distillation (SGKD) framework versus the student-only baseline. PSNR, SSIM, MAE, and Mean Average Precision (MAP) are reported; ↓ denotes metrics where lower is better. The physics-informed supervisory signal delivers a 28.83% PSNR gain and a 46.19% SSIM improvement over the student baseline.

Model	Mean PSNR	Mean SSIM	Mean MAE	Mean MAP
Student Only	23.13	0.533	0.478	5.05
Ours	**29.80**	**0.780**	**0.296**	**3.07**
% Improvement	+28.83%	+46.19%	-38.06%*	-39.11%*

the baseline across all spectrally complex classes. On average, Top-1 Accuracy rises from 48.26% to 66.81%, while the mean IoU increases from 34.88% to 46.37%. These gains confirm that **Physics-Based Distillation** produces discriminative features that are more resilient to the noisy labels common in satellite datasets.

Geographic Generalization (NLCD)

To evaluate the robustness of the science-guided inductive biases, the model was tested on geographically distinct regions (Colorado and Kansas) after training solely on California data. In **Land Cover Classification** (Table 4.13), the proposed framework achieves a notable 61.5% relative increase in accuracy for "Cultivated Crops," rising to 91.59%. This confirms that distilling the teacher's physics-based abundance predictions allows the student to capture universal spectral signatures that remain stable across diverse terrains.

4.6.5 Computational Efficiency of Distillation

While the proposed science-guided distillation introduces additional complexity during the training phase, the use of a teacher-student architecture provides a favorable trade-off for deployment. As detailed in Table 4.15, the total overhead for the full physics-guided model is approximately 31.7%, with an average training time of 9.47 ms per sample.

Crucially, this overhead is **sublinear** relative to the substantial performance improvements in reconstruction and downstream accuracy. Because the high-capacity **Teacher Model (ViT-Large)** is frozen, the computational cost is concentrated on the training of the efficient **Student Model**

(ViT-Base), enabling high-performance HSI modeling on resource-constrained hardware without sacrificing physical consistency.

Table 4.15: Computational trade-offs of science-guided components per sample. The full framework adds 2.28 ms over the baseline, with the physical mixing model (LSMM) contributing minimal overhead (0.41 ms) beyond the SAM loss calculation.

Model Configuration	Avg. Time per sample (ms)
Baseline ViT-MAE	7.19
ViT-MAE + SAM Loss	9.06
Ours (SAM+LSMM)	**9.47**

4.7 Summary of the Findings

The experimental results across all three research thrusts provide strong empirical support for the proposed research agenda, validating the core hypothesis that foundation models can serve as effective knowledge priors for specialized scientific tasks even under conditions of extreme data asymmetry.

- **Thrust 1 (Heterogeneous Domain-Shift Mitigation):** The proposed cross-resolution distillation framework [55] validates that knowledge transfer can bridge the gap between coarse, contiguous satellite coverage (SMAP, 3 km) and sparse in-situ ground truth. In experiments covering diverse climatic zones in Colorado, the student model achieved a **31% to 40% reduction in Mean Absolute Error (MAE)** for soil moisture estimation compared to models trained solely on sparse station data.

  The framework also addresses the operational deployment constraint. The lightweight ResNet8 student (78.5K parameters) achieves a **5.5× speedup in inference latency** (0.4 ms vs. 2.2 ms per sample for VGG13) while maintaining accuracy within the measurement error of the best available in-situ sensors. This confirms that cross-resolution distillation can produce efficient, edge-deployable soil moisture models without sacrificing local precision.

- **Thrust 2 (Inverse Knowledge Distillation):** The proposed inverse distillation framework [78] demonstrates that this approach reliably bridges a 36× spectral dimensionality gap. By distilling spatial semantics from the Prithvi-100M teacher (6 HLS bands) into the 218-band EnMAP student, the proposed method achieves a **26% improvement in PSNR** and a **40% improvement in SSIM** compared to the student trained from scratch.

  These gains extend to downstream classification tasks. The learned feature representations improved Land Cover Classification accuracy on the NLCD dataset by over **28% for spectrally ambiguous classes** such as Cultivated Crops (77.18% vs. 56.70%), confirming that the spatial scaffolding from the teacher stabilizes student learning in high-dimensional feature spaces.

- **Thrust 3 (Science-Guided Domain Adaptation):** The proposed science-guided distillation framework [64] demonstrates that embedding physical constraints into the distillation process bridges heterogeneous sensor domains more effectively than purely statistical alignment. Across diverse geographic regions, the physics-guided student achieved a **30.3% to 61.5% relative improvement in Top-1 Accuracy** for land cover classification over data-driven baselines.

  By distilling abundance maps derived from the LSMM, the lightweight student (ViT-Base) achieved a **23.6% relative gain in Structural Similarity (SSIM)** over the vanilla ViT-MAE. The training overhead is only 31.7% above the baseline, making the framework practical for deployment on research computing clusters without dedicated GPU farms.

Chapter 5

Conclusion

5.1 Summary of Contributions

Large-scale Foundation Models have produced strong generalization across broad domains, yet applying them directly to specialized scientific settings, such as Earth observation and precision agriculture, remains a persistent challenge. The core difficulty is the *Adaptability Gap*: a compound of mismatched data modalities, resolution asymmetries, extreme label sparsity, and the absence of physically grounded inductive biases. This dissertation identifies three distinct barriers that constitute this gap and proposes a corresponding set of methodologies under the unified paradigm of **Science-Guided Knowledge Distillation**.

5.1.1 Thrust 1: Heterogeneous Domain-Shift Mitigation

The first thrust addresses the resolution and modality mismatch between coarse satellite observations and sparse in-situ measurements. The proposed cross-resolution distillation framework [55] employs an asymmetric teacher-student architecture in which a VGG13 model trained on abundant SMAP satellite data (3 km resolution) transfers global environmental knowledge to a lightweight ResNet8 student anchored by sparse station measurements. A nested training strategy based on the Köppen climate classification system allows the framework to handle regions with very few ground-truth labels by routing predictions through student models optimized for climatically similar zones.

Empirical results covering Colorado and California demonstrate a 31–40% reduction in Mean Absolute Error for soil moisture estimation relative to the student trained on station data alone. The distilled student model (ResNet8, 78.5K parameters) achieves 5.5$\times$ lower inference latency than the teacher, with an equivalent deployment memory footprint. This confirms that cross-resolution

knowledge distillation is a practical strategy for producing high-fidelity local estimates from globally trained models in data-sparse scientific environments.

5.1.2 Thrust 2: Inverse Knowledge Distillation for Cross-Modal Asymmetry

The second thrust addresses a form of domain shift that does not appear in standard distillation settings: the teacher operates in a lower-dimensional space than the student. We term this the *inverse shift* problem. The proposed inverse distillation framework [78] transfers spatial-semantic knowledge from a 6-band multispectral teacher (Prithvi-100M, pretrained on HLS data) into a 218-band hyperspectral student (EnMAP). Spectral domain alignment maps overlapping wavelength ranges to synthetic teacher bands, enabling direct supervision across the spectral gap. Spatial feature-guided masking using Gabor filters or wavelet transforms concentrates reconstruction effort on high-variability patches, and a composite loss combining MSE, SSIM, and KLD governs both reconstruction fidelity and latent alignment.

The proposed method achieves a PSNR of 31.02 and SSIM of 0.77 on the primary test set, representing a 26% and 40% improvement over a student trained without distillation, respectively. These gains carry forward into downstream tasks: Land Cover Classification accuracy on the NLCD dataset improved by up to 28% for spectrally ambiguous classes such as Cultivated Crops. Results are consistent across three geographically distinct regions (California, Colorado, Kansas), demonstrating that the spatial scaffolding from the multispectral teacher generalizes beyond the training domain.

5.1.3 Thrust 3: Science-Guided Domain Adaptation

The third thrust addresses the absence of physical consistency in data-driven models. The proposed physics-guided masking framework [64] embeds the Linear Spectral Mixing Model (LSMM) as a differentiable constraint within the decoder of a ViT-MAE architecture. An abundance head predicts the fractional material composition of each patch, subject to non-negativity and sum-to-one constraints enforced via softmax. A physics-consistency loss minimizes the residual between observed and LSMM-reconstructed spectra. The Spectral Angle Mapper (SAM) loss ad-

ditionally enforces preservation of spectral shape over absolute intensity, making the model robust to illumination variation.

The proposed method achieves a PSNR of 27.38 dB and SSIM of 0.68, representing relative gains of 11.3% and 23.6% over the vanilla ViT-MAE baseline. Downstream crop type identification (CDL dataset) improves by an average of 18.6 percentage points in Top-1 Accuracy, with the largest gains in spectrally ambiguous categories such as Shrubland and Grassland/Pasture. Trained on California data alone, the proposed method generalizes effectively to Colorado and Kansas in the NLCD land cover classification task, achieving 91.59% accuracy on Cultivated Crops. The science-guided distillation extension of this framework applies the physics-guided teacher to guide a compact student in cross-sensor adaptation, achieving a 28.8% improvement in PSNR and a 46.2% improvement in SSIM over a student-only baseline with a training overhead of only 31.7%.

5.2 Broader Impact and Limitations

The methodologies developed in this dissertation address bottlenecks that are common to scientific AI more broadly: sparse ground truth, heterogeneous sensing modalities, and the need for physically interpretable predictions. The cross-resolution distillation strategy developed in the first thrust generalizes to any multi-resolution Earth observation problem where coarse gridded data is abundant and fine-scale labeled data is scarce. The inverse distillation concept developed in the second thrust applies to any setting where a foundation model trained on a lower-dimensional modality can supervise learning in a higher-dimensional one. The physics-constraint approach developed in the third thrust is extensible to other physical mixing models beyond LSMM, including those governing SAR backscatter or thermal emission.

5.3 Future Directions

Several directions emerge naturally from this work.

Unified Multimodal Foundation Models. The three thrusts in this dissertation operate on spectral imagery and point measurements independently. A natural extension is to integrate SAR,

LiDAR, and textual metadata into a unified pretraining regime where each modality informs the others through cross-modal distillation.

Bidirectional Knowledge Transfer. The frameworks in this dissertation treat knowledge transfer as one-directional. In federated or distributed monitoring scenarios, edge-deployed student models accumulate domain-specific experience that could refine the central teacher. Establishing bidirectional distillation, where the student contributes back to the teacher, would allow the system to continuously adapt to sensor drift, seasonal change, and land-cover transitions without requiring centralized data aggregation.

Causal and Mechanistic Interpretability. The abundance maps produced by the physics-guided framework correspond to interpretable physical quantities (material fractions). This opens a path toward causal analysis: if the model's predictions can be traced back to specific material abundances, then changes in those abundances over time can be attributed to physical drivers such as drought, urbanization, or agricultural practice. Linking the latent physical representation to downstream causal inference is a natural next step.

Trustworthy Deployment and Calibration. Deploying the proposed models in operational Earth observation pipelines requires well-calibrated uncertainty estimates, not just point predictions. Conformal prediction methods and evidential deep learning approaches could be applied to the existing student architectures to produce prediction intervals that are provably valid under distribution shift, making the systems suitable for high-stakes decision contexts such as wildfire monitoring or flood mapping.

Bibliography

[1] E. Eftelioglu, Z. Jiang, X. Tang, and S. Shekhar, "The nexus of food, energy, and water resources: Visions and challenges in spatial computing," in *Advances in Geocomputation: Geocomputation 2015–The 13th International Conference*, Springer, 2017, pp. 5–20.

[2] Z. Jiang, "Deep learning for spatiotemporal big data: A vision on opportunities and challenges," *arXiv preprint arXiv:2310.19957*, 2023.

[3] Z. Jiang and S. Shekhar, "Spatial big data science," *Schweiz: Springer International Publishing AG*, 2017.

[4] M. Abadi, P. Barham, J. Chen, Z. Chen, et al., "{Tensorflow}: A system for {large-scale} machine learning," in *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, 2016, pp. 265–283.

[5] A. Paszke, S. Gross, F. Massa, et al., "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[6] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[7] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. S. Chatterji, A. S. Chen, K. Creel, J. Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, et al., "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

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

[9] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al., "Highly accurate protein structure prediction with alphafold," *Nature*, vol. 596, no. 7873, pp. 583–589, 2021.

[10] X. Yin, G. Wu, J. Wei, Y. Shen, H. Qi, and B. Yin, "Deep learning on traffic prediction: Methods, analysis, and future directions," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 6, pp. 4927–4943, 2021.

[11] A. Matin, S. Armstrong, S. Mitra, S. Pallickara, and S. L. Pallickara, "Rapid betweenness centrality estimates for transportation networks using capsule networks," in *2022 Fourth International Conference on Transdisciplinary AI (TransAI)*, IEEE, 2022, pp. 89–96.

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

[13] T. E. Ochsner, M. H. Cosh, R. H. Cuenca, W. A. Dorigo, C. S. Draper, Y. Hagimoto, Y. H. Kerr, K. M. Larson, E. G. Njoku, R. Rollenbeck, and M. Zreda, "State of the art in large-scale soil moisture monitoring," *Soil Science Society of America Journal*, vol. 77, no. 6, pp. 1888–1919, 2013.

[14] J. Peng, A. Loew, O. Merlin, and N. E. Verhoest, "A review of spatial downscaling of satellite remotely sensed soil moisture," *Reviews of Geophysics*, vol. 55, no. 2, pp. 341–366, 2017.

[15] P. Khandelwal, S. Armstrong, A. Matin, S. Pallickara, and S. L. Pallickara, "CloudNet: A deep learning approach for mitigating occlusions in Landsat-8 imagery using data coalescence," in *Proceedings of the 18th IEEE International Conference on eScience*, IEEE, 2022.

[16] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[17] L. Hoyer, D. Dai, H. Wang, and L. Van Gool, "Mic: Masked image consistency for context-aware domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 11 721–11 732.

[18] J. Jakubik, S. Roy, C. E. Phillips, P. Fraccaro, D. Godwin, B. Zadrozny, D. Szwarcman, C. Gomes, G. Nyirjesy, B. Edwards, D. Bernstein, N. Kimura, N. Simumba, L. Chu, et al., "Foundation models for generalist geospatial artificial intelligence," *arXiv preprint arXiv:2310.18660*, 2023.

[19] G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang, "Physics-informed machine learning," *Nature Reviews Physics*, vol. 3, no. 6, pp. 422–440, 2021.

[20] N. Keshava and J. F. Mustard, "Spectral unmixing," *IEEE signal processing magazine*, vol. 19, no. 1, pp. 44–57, 2002.

[21] A. Karpatne, G. Atluri, J. H. Faghmous, M. Steinbach, A. Banerjee, A. Ganguly, S. Shekhar, N. Samatova, and V. Kumar, "Theory-guided data science: A new paradigm for scientific discovery from data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 10, pp. 2318–2331, 2017.

[22] J. M. Bioucas-Dias, A. Plaza, N. Dobigeon, M. Parente, Q. Du, P. Gader, and J. Chanussot, "Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 5, no. 2, pp. 354–379, 2012.

[23] X. Li, W. Chen, H. Zhang, F. Wang, and J. Liu, "Frequency-aligned knowledge distillation for lightweight spatiotemporal forecasting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

[24] Y. Zhao, R. Liu, J. Chen, L. Wang, and T. Zhang, "Spectral-decoupled knowledge distillation for spatiotemporal forecasting," *arXiv preprint arXiv:2507.02939*, 2025.

[25] Y. Li, W. Zhang, X. Chen, J. Liu, and M. Wang, "Counterfactual knowledge maintenance for unsupervised domain adaptation with large visual-language models," in *Proceedings of the 34th International Joint Conference on Artificial Intelligence (IJCAI)*, 2025.

[26] ESA EOPortal Directory, *Hyperspectral imaging*, https://www.eoportal.org/other-space-activities/hyperspectral-imaging, Accessed: May 2026, 2024.

[27] M. Claverie, J. Ju, J. G. Masek, J. L. Dungan, E. F. Vermote, J.-C. Roger, S. V. Skakun, and C. Justice, "The harmonized landsat and sentinel-2 surface reflectance data set," *Remote sensing of environment*, vol. 219, pp. 145–161, 2018.

[28] Y. Cong, S. Khanna, C. Meng, P. Liu, B. Roziere, M. Maire, J. Jiao, and S. Ermon, "Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 197–211.

[29] C. L. Nkwocha and A. K. Chandel, "Towards an end-to-end digital framework for precision crop disease diagnosis and management based on emerging sensing and computing technologies: State over past decade and prospects," *Computers*, vol. 14, no. 10, 2025, ISSN: 2073-431X. DOI: 10.3390/computers14100443. [Online]. Available: https://www.mdpi.com/2073-431X/14/10/443.

[30] R. F. Kokaly, R. N. Clark, G. A. Swayze, K. E. Livo, T. M. Hoefen, N. C. Pearson, R. A. Wise, W. M. Benzel, H. A. Lowers, R. L. Driscoll, and A. J. Klein, "USGS Spectral Library Version 7," U.S. Geological Survey, Tech. Rep. Data Series 1035, 2017. DOI: 10.3133/ds1035.

[31] T. Storch, H.-P. Honold, S. Chabrillat, M. Habermeyer, P. Tucker, M. Brell, A. Ohndorf, K. Wirth, M. Betz, M. Kuchler, et al., "The EnMAP imaging spectroscopy mission towards operations," *Remote Sensing of Environment*, vol. 294, p. 113 632, 2023. DOI: 10.1016/j.rse.2023.113632. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0034425723001839.

[32] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 000–16 009.

[33] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[34] D. Szwarcman, S. Roy, P. Fraccaro, Þ. E. Gíslason, B. Blumenstiel, R. Ghosal, P. H. de Oliveira, J. L. de Sousa Almeida, R. Sedona, Y. Kang, et al., "Prithvi EO 2.0: A versatile multi-temporal foundation model for earth observation applications," *arXiv preprint arXiv:2412.02732*, 2024.

[35] Z. Xiong, Y. Wang, F. Zhang, A. J. Stewart, J. Hanna, D. Borth, I. Papoutsis, B. Le Saux, G. Camps-Valls, and X. X. Zhu, "DOFA: Dynamic one-for-all pre-training for remote sensing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[36] Clay Foundation Team, *Clay foundation model for earth observation*, Open-source release, https://github.com/Clay-foundation/model, 2024.

[37] C. J. Reed, R. Gupta, S. Li, S. Brockman, C. Funk, B. Clipp, K. Keutzer, S. Candido, M. Uyttendaele, and T. Darrell, *Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning*, 2023. arXiv: 2212.14532 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2212.14532.

[38] D. Hong, B. Zhang, X. Li, Y. Li, C. Li, J. Yao, N. Yokoya, H. Li, P. Ghamisi, X. Jia, A. Plaza, P. Gamba, J. A. Benediktsson, and J. Chanussot, "Spectralgpt: Spectral remote sensing foundation model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[39] T. B. Faruk, A. Matin, S. Pallickara, and S. L. Pallickara, "Accounting for spatial variability with the histogram of oriented gradients based masking improves performance of masked autoencoder over hyperspectral satellite imagery (student abstract)," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 28, pp. 29 365–29 367, Apr. 2025. DOI: 10.1609/aaai.v39i28.35253. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/35253.

[40] T. B. Faruk, A. Matin, S. Pallickara, and S. L. Pallickara, "TerraMAE: Learning spatial-spectral representations from hyperspectral earth observation data via adaptive masked autoencoders," in *Proceedings of the 33rd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM SIGSPATIAL)*, 2025.

[41] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," *arXiv preprint arXiv:1412.6550*, 2014.

[42] Y. Wang, L. Cheng, M. Duan, Y. Wang, Z. Feng, and S. Kong, "Improving knowledge distillation via regularizing feature norm and direction," *arXiv preprint arXiv:2305.17007*, 2023.

[43] J. Yim, D. Joo, J. Bae, and J. Kim, "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4133–4141.

[44] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3967–3976.

[45] J. Gou, L. Sun, B. Yu, S. Wan, and D. Tao, "Hierarchical multi-attention transfer for knowledge distillation," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 20, no. 2, pp. 1–20, 2023.

[46] H. Yin, P. Molchanov, J. M. Alvarez, Z. Li, A. Mallya, D. Hoiem, N. K. Jha, and J. Kautz, "Dreaming to distill: Data-free knowledge transfer via deepinversion," in *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 8715–8724.

[47] G. Fang, K. Mo, X. Wang, J. Song, S. Bei, H. Zhang, and M. Song, "Up to 100x faster data-free knowledge distillation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, 2022, pp. 6597–6604.

[48] Z. Fang, J. Wang, X. Hu, L. Wang, Y. Yang, and Z. Liu, "Compressing visual-linguistic model via knowledge distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1428–1438.

[49] L. Magister, J. a. Mallinson, J. Adamek, E. Malmi, and A. Severyn, "Teaching small language models to reason," *arXiv preprint arXiv:2212.08410*, 2022.

[50] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. March, and V. Lempitsky, "Domain-adversarial training of neural networks," *Journal of machine learning research*, vol. 17, no. 59, pp. 1–35, 2016.

[51] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," in *International conference on machine learning*, Pmlr, 2018, pp. 1989–1998.

[52] M. Chen, H. Xue, and D. Cai, "Domain adaptation for semantic segmentation with maximum squares loss," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2090–2099.

[53] Y. Kim, D. Cho, K. Han, P. Panda, and S. Hong, "Domain adaptation without source data," *IEEE Transactions on Artificial Intelligence*, vol. 2, no. 6, pp. 508–518, 2021. DOI: 10.1109/TAI.2021.3110179.

[54] R. Dey, A. Matin, T. B. Faruk, S. Pallickara, and S. L. Pallickara, "DEEPSALT: Bridging laboratory and satellite spectra through domain adaptation and knowledge distillation for large-scale soil salinity estimation," in *Proceedings of the 2025 IEEE International Conference on Big Data*, 2025.

[55] A. Matin, P. Khandelwal, S. Pallickara, and S. L. Pallickara, "DISCERN: Leveraging knowledge distillation to generate high resolution soil moisture estimation from coarse satellite data," in *2023 IEEE International Conference on Big Data (BigData)*, 2023, pp. 1222–1229.

[56] J. Willard, X. Jia, S. Xu, M. Steinbach, and V. Kumar, "Integrating scientific knowledge with machine learning for engineering and environmental systems," *ACM Computing Surveys (CSUR)*, vol. 55, no. 4, pp. 1–37, 2022.

[57] A. Karpatne, G. Atluri, J. H. Faghmous, M. Steinbach, A. Banerjee, A. Ganguly, S. Shekhar, N. Samatova, and V. Kumar, "Knowledge-guided machine learning: Current trends and future prospects," *arXiv preprint arXiv:2403.00573*, 2024.

[58] P. Khandelwal, S. L. Pallickara, and S. Pallickara, "Deepsoil: A science-guided framework for generating high precision soil moisture maps by reconciling measurement profiles across in-situ and remote sensing data," in *Proceedings of the 32nd ACM International Conference on Advances in Geographic Information Systems*, 2024, pp. 233–246.

[59] P. Khandelwal, J. D. Niemann, D. J. Mulla, S. Pallickara, and S. L. Pallickara, "Subterra: Estimating soil moisture at root zone depths using science-guided learning," in *2025 IEEE Conference on Artificial Intelligence (CAI)*, IEEE, 2025, pp. 328–335.

[60] P. Khandelwal, A. Bachinin, T. B. Faruk, E. Lewark, S. L. Pallickara, and S. Pallickara, "Deepsoil: Physics-guided learning for high-resolution soil moisture mapping from surface to root-zone via multimodal sensing and process-based modeling," *ACM Transactions on Spatial Algorithms and Systems*, 2026.

[61] K. Bruhwiler, P. Khandelwal, D. Rammer, S. Armstrong, S. L. Pallickara, and S. Pallickara, "Lightweight, embeddings based storage and model construction over satellite data collections," in *2020 IEEE International Conference on Big Data (Big Data)*, IEEE, 2020, pp. 246–255.

[62] S. Armstrong, P. Khandelwal, D. Padalia, G. Senay, D. Schulte, A. Andales, F. J. Breidt, S. Pallickara, and S. L. Pallickara, "Attention-based convolutional capsules for evapotranspiration estimation at scale," *Environmental Modelling & Software*, vol. 152, p. 105 366, 2022.

[63] S. Mitra, D. Rammer, S. Pallickara, and S. L. Pallickara, "Glance: A generative approach to interactive visualization of voluminous satellite imagery," in *2021 IEEE International Conference on Big Data (Big Data)*, IEEE, 2021, pp. 359–367.

[64] A. Matin, R. Dey, T. B. Faruk, S. Pallickara, and S. L. Pallickara, "Knowledge-guided masked autoencoder with linear spectral mixing and spectral-angle-aware reconstruction," in *AAAI 2026 Workshop on Knowledge Graphs and Machine Learning (KGML)*, 2026.

[65] A. Bachinin, R. Dey, P. Khandelwal, S. Leuthold, M. F. Cotrufo, S. Pallickara, and S. L. Pallickara, "Science-informed multitask transformer for soil property prediction from ftir spectroscopy," in *2025 IEEE International Conference on eScience (eScience)*, IEEE, 2025, pp. 48–57.

[66] R. Dey, A. Matin, N. Orwick, Y. Zhang, S. Pallickara, and S. L. Pallickara, "When to trust, how to distill: Multi-foundation model guidance for lightweight, robust scientific time series forecasting," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, To appear, Jeju, South Korea: ACM, 2026.

[67] R. Dey, A. Bachinin, T. B. Faruk, A. Matin, M. Hong, Y. Zhang, S. Pallickara, and S. L. Pallickara, "XFormer: A multi-stage uncertainty-guided deep learning framework for time series extreme event forecasting," in *2026 IEEE Conference on Artificial Intelligence (CAI)*, IEEE, 2026.

[68] M. Raissi, P. Perdikaris, and G. E. Karniadakis, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear

partial differential equations," *Journal of Computational physics*, vol. 378, pp. 686–707, 2019.

[69] K. Kashinath, M. Mustafa, A. Albert, J.-L. Wu, C. Jiang, S. Esmaeilzadeh, K. Azizzadenesheli, R. Wang, A. Chattopadhyay, A. Singh, A. Manepalli, D. Chirila, R. Yu, R. Walters, B. White, H. Xiao, H. A. Tchelepi, P. Marcus, A. Anandkumar, P. Hassanzadeh, and Prabhat, "Physics-informed machine learning: Case studies for weather and climate modelling," *Philosophical Transactions of the Royal Society A*, vol. 379, no. 2194, p. 20 200 093, 2021.

[70] J. M. Bioucas-Dias, A. Plaza, N. Dobigeon, M. Parente, Q. Du, P. Gader, and J. Chanussot, "Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches," *IEEE journal of selected topics in applied earth observations and remote sensing*, vol. 5, no. 2, pp. 354–379, 2012.

[71] S. Ozkan, B. Kaya, and G. B. Akar, "Endnet: Sparse autoencoder network for endmember extraction and hyperspectral unmixing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 1, pp. 482–496, 2019.

[72] Y. Liu, L. Shu, T. Wang, C. Zhang, J. Gao, and Y. Zheng, "A survey on knowledge-guided machine learning," *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[73] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

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

[75] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[76] F. Tung and G. Mori, "Similarity-preserving knowledge distillation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1365–1374.

[77] M. Kottek, J. Grieser, C. Beck, B. Rudolf, and F. Rubel, "World map of the köppen-geiger climate classification updated," 2006.

[78] A. Matin, T. B. Faruk, S. Pallickara, and S. L. Pallickara, "HyperKD: Distilling cross-spectral knowledge in masked autoencoders via inverse domain shift with spatial-aware masking and specialized loss," in *Proceedings of the 12th IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, Best Paper Award, 2025.

[79] D. Gabor, *Theory of communication. part 1: The analysis of information. j inst electr eng-part iii: Radio commun eng 93 (26): 429–441*, 1946.

[80] Y. Meyer, *Wavelets and operators: volume 1*. Cambridge university press, 1992.

[81] R. Mehrotra, K. R. Namuduri, and N. Ranganathan, "Gabor filter-based edge detection," *Pattern recognition*, vol. 25, no. 12, pp. 1479–1494, 1992.

[82] J. Zujovic, T. N. Pappas, and D. L. Neuhoff, "Structural texture similarity metrics for image analysis and retrieval," *IEEE Transactions on Image Processing*, vol. 22, no. 7, pp. 2545–2558, 2013.

[83] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[84] J. Wei and X. Wang, "An overview on linear unmixing of hyperspectral data," *Mathematical Problems in Engineering*, vol. 2020, no. 1, p. 3 735 403, 2020.

[85] F. A. Kruse, A. B. Lefkoff, J. W. Boardman, K. B. Heidebrecht, A. T. Shapiro, P. J. Barloon, and A. F. H. Goetz, "The spectral image processing system (sips)—interactive visualization and analysis of imaging spectrometer data," *Remote Sensing of Environment*, vol. 44, no. 2–3, pp. 145–163, 1993. DOI: 10.1016/0034-4257(93)90013-N. [Online]. Available: https://www.sciencedirect.com/science/article/pii/003442579390013N.

[86] R. H. Yuhas, A. F. Goetz, and J. W. Boardman, "Discrimination among semi-arid landscape endmembers using the spectral angle mapper (sam) algorithm," in *Summaries of the Third Annual JPL Airborne Geoscience Workshop*, 1992, pp. 147–149. [Online]. Available: https://ntrs.nasa.gov/citations/19940012238.

[87] R. A. Freeze and R. Harlan, "Blueprint for a physically-based, digitally-simulated hydrologic response model," *Journal of hydrology*, vol. 9, no. 3, pp. 237–258, 1969.

[88] J. R. Nimmo, "The processes of preferential flow in the unsaturated zone," *Soil Science Society of America Journal*, vol. 85, no. 1, pp. 1–27, 2021.

[89] D. Entekhabi, E. G. Njoku, P. E. O'neill, K. H. Kellogg, W. T. Crow, W. N. Edelstein, J. K. Entin, S. D. Goodman, T. J. Jackson, J. Johnson, et al., "The soil moisture active passive (smap) mission," *Proceedings of the IEEE*, vol. 98, no. 5, pp. 704–716, 2010.

[90] Y. H. Kerr, A. Al-Yaari, N. Rodriguez-Fernandez, M. Parrens, B. Molero, D. Leroux, S. Bircher, A. Mahmoodi, A. Mialon, P. Richaume, et al., "Overview of smos performance in terms of global soil moisture monitoring after six years in operation," *Remote Sensing of Environment*, vol. 180, pp. 40–63, 2016.

[91] D. Phiri, M. Simwanda, S. Salekin, V. R. Nyirenda, Y. Murayama, and M. Ranagalage, "Sentinel-2 data for land cover/use mapping: A review," *Remote Sensing*, vol. 12, no. 14, p. 2291, 2020.

[92] S. S. Staff, "Gridded national soil survey geographic (gnatsgo) database for the conterminous united states," *United States Department of Agriculture, Natural Resources Conservation Service*, 2020.

[93] J. T. Abatzoglou, "Development of gridded surface meteorological data for ecological applications and modelling," *International Journal of Climatology*, vol. 33, no. 1, pp. 121–131, 2013.

[94] "NASA JPL(2013).nasa shuttle radar topography mission global 1 arc second number," Accessed 2023-06-05. DOI: https://doi.org/10.5067/MEaSUREs/SRTM/SRTMGL1N.003.

[95] J. E. Bell, M. A. Palecki, C. B. Baker, W. G. Collins, J. H. Lawrimore, R. D. Leeper, T. P. Meyers, D. A. Robinson, and G. L. Scott, "U.s. climate reference network soil moisture and temperature observations," *Journal of Hydrometeorology*, vol. 14, no. 3, pp. 977–988, 2013.

[96] K. Copenhaver, Y. Hamada, S. Mueller, and J. B. Dunn, "Examining the characteristics of the cropland data layer in the context of estimating land cover change," *ISPRS international journal of geo-information*, vol. 10, no. 5, p. 281, 2021.

[97] U. G. Survey, *Annual nlcd (national land cover database)—the next generation of land cover mapping*, 2025.

[98] P. K. Sharma, D. Kumar, H. S. Srivastava, and P. Patel, "Assessment of different methods for soil moisture estimation: A review," *J. Remote Sens. GIS*, vol. 9, no. 1, pp. 57–73, 2018.

[99] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004. DOI: 10.1109/TIP.2003.819861.

[100] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes (VOC) Challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010. DOI: 10.1007/s11263-009-0275-4.

[101] S. Bianco, R. Cadene, L. Celona, and P. Napoletano, "Benchmark analysis of representative deep neural network architectures," *IEEE access*, vol. 6, pp. 64 270–64 277, 2018.

[102] M. Litaor, M. Williams, and T. Seastedt, "Topographic controls on snow distribution, soil moisture, and species diversity of herbaceous alpine vegetation, niwot ridge, colorado," *Journal of Geophysical Research: Biogeosciences*, vol. 113, no. G2, 2008.