THESIS

TOWARDS GENERATING A PRE-TRAINING IMAGE TRANSFORMER FRAMEWORK FOR PRESERVING SPATIO-SPECTRAL PROPERTIES IN HYPERSPECTRAL SATELLITE IMAGES

Submitted by

Tanjim Bin Faruk

Department of Computer Science

In partial fulfillment of the requirements

For the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Fall 2024

Master's Committee:

Advisor: Sangmi Lee Pallickara
Co-Advisor: Shrideep Pallickara

M. Francesca Cotrufo

ABSTRACT

TOWARDS GENERATING A PRE-TRAINING IMAGE TRANSFORMER FRAMEWORK FOR PRESERVING SPATIO-SPECTRAL PROPERTIES IN HYPERSPECTRAL SATELLITE IMAGES

Hyperspectral images facilitate advanced geospatial analysis without the need for expensive ground surveys. Machine learning approaches are particularly well-suited for handling the geospatial coverage required by these applications. While self-supervised learning is a promising methodology for managing voluminous datasets with limited labels, existing encoders in self-supervised learning face challenges when applied to hyperspectral images due to the large number of spectral channels. We propose a novel hyperspectral image encoding framework designed to generate highly representative embeddings for subsequent geospatial analysis. Our framework extends the Vision Transformer model with dynamic masking strategies to enhance model performance in regions with high spatial variability. We introduce a novel loss function that incorporates spectral quality metrics and employs the unique channel grouping strategy to leverage spectral similarity across channels. We demonstrate the effectiveness of our approach through a downstream model for estimating soil texture at a 30-meter resolution.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my supervisor, Dr. Sangmi Lee Pallickara, for her unwavering support, guidance, and invaluable insights throughout the course of my research. Her expertise and encouragement have been instrumental in completing my thesis.

I am also sincerely grateful to my co-advisor, Dr. Shrideep Pallickara, and to my committee member, Dr. M. Francesca Cotrufo, for their constructive feedback and thoughtful suggestions. Their contributions have significantly enhanced the quality of my work.

Special thanks go to my colleague, Abdul Matin, whose guidance and engaging discussions pushed me in the right direction. I have truly benefited from his collaboration and his willingness to share knowledge.

Finally, I extend my heartfelt appreciation to my family members for their endless support and encouragement. Their love and patience have been my greatest source of strength.

This research was supported by the National Science Foundation (OAC-1931363, CNS-2312319), an NSF/NIFA Artificial Intelligence (AI) Institutes AI-CLIMATE Award [2023-03616], and the National Institute for Food and Agriculture (COL0-14021223).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

Prior to the advent of satellite technology, remote sensing was primarily confined to aerial photography or ground observations. Although these methods were effective, they had several limitations such as restricted coverage, limited spatial and spectral resolution, and infrequent revisit times [4, 5]. Satellite imagery provides a complementary view of the earth's surface that is often challenging or unattainable through traditional methods. For example, satellites can cover vast and remote areas that are inaccessible to aircraft or ground surveys [6]. They offer higher spatial resolution with frequent revisit times, making it possible to observe changes in various phenomena such as weather patterns, vegetation growth, and urban development [7]. Additionally, satellite sensors can capture a wide range of spectral bands—including visible, infrared, and microwave wavelengths—which enhances the detection and analysis of various surface materials and atmospheric conditions [8]. These advantages make satellite imagery a vital tool across diverse domains such as agriculture, environmental monitoring, disaster management etc.

The launch of Earth Resources Technology Satellite (ERTS) in 1972, later renamed Landsat 1, marked the beginning of a new era in remote sensing. Since then, numerous satellites have been deployed to orbit the Earth to gather vital remote sensing data [6, 9] and satellite technology has progressed a great deal in terms of data volume and data quality.

When Landsat 1 launched, it could only store 3.75 gigabytes onboard and had the largest recording capacity of any orbiting recorder at the time [10]. It also had a data downlink rate of 15 megabits per second [11]. In contrast, newer-generation satellites can store $1000\times$ [12] more data and have also achieved significantly higher transmission rates [13]. The implication is that they are storing more data than ever and also sending more data to Earth than ever. This poses a challenge for efficiently analyzing the vast amount of data. Furthermore, while Landsat 1 offered a spatial resolution of 80 and comprised only 4 bands [14], the introduction of a diverse array of new sensors has pushed spatial resolution down to as fine as 15cm and expanded the number of spectral bands

to as many as 250 [15]. This latter category of new-generation satellites, known as hyperspectral satellites, produces hyperspectral satellite imagery (HSI) and can capture continuous spectral bands with high-wavelength resolution, often exceeding 200 bands. In contrast, multispectral images typically capture fewer, spaced spectral bands (usually $3 - 11$ bands) [16]. Hyperspectral satellites can facilitate enhanced material identification and classification due to their unique spectral signatures [17] and can support precision agriculture by accurately mapping crop types and soil conditions [18, 19]. Examples of hyperspectral satellites include EnMAP (used in this study) [20], PRISMA [21], and Planet's Tanager-1, which is scheduled for launch in 2024 [22].

With the increasing number of satellite operations, the volume of data generated by these satellites has grown exponentially, creating substantial challenges for large-scale data analysis. To address these challenges, researchers have leveraged advanced machine-learning techniques, which can effectively transform voluminous satellite imagery data into valuable estimates of ground conditions [23–26].

The initial applications of machine learning in satellite imagery analysis primarily centered on classification and regression tasks. Among supervised learning methods, Support Vector Machines (SVM), Decision Trees, Random Forests, and *k*-Nearest Neighbor were commonly employed [27–29]. In the realm of unsupervised learning, *k*-means clustering was utilized for processing large datasets, while Principal Component Analysis (PCA) proved valuable for dimensionality reduction. These methods facilitated early achievements in land cover mapping, change detection, and crop identification [30, 31].

Despite their effectiveness, traditional machine-learning approaches had several limitations. One notable constraint was the requirement for manual feature engineering, a time-consuming process that relied heavily on domain expertise. Additionally, they could not fully capture the nuances in the satellite imagery, restricting their overall performance and accuracy [32, 33].

Deep learning techniques, when coupled with self-supervised learning approaches, effectively address these challenges by automatically extracting features and performing exceptionally well on large datasets. Several deep learning architectures such as Convolutional Neural Networks

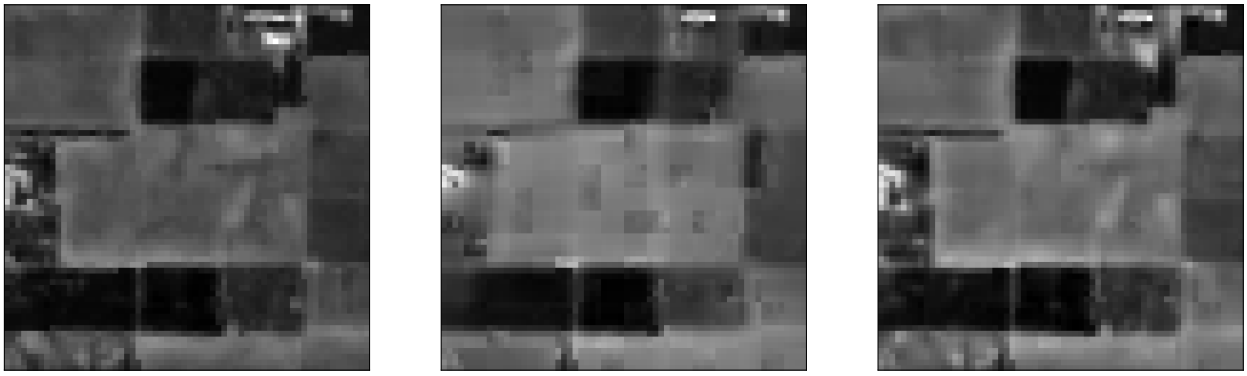

Figure 1.1: Input Reconstruction: Base Model vs Our Model

(CNNs), Autoencoders, Recurrent Neural Networks (RNNs), and Generative Adversarial Networks (GANs) have been successfully utilized for various remote sensing tasks [34, 35]. Moreover, self-supervised learning methods have demonstrated significant performance enhancements, particularly when models are pre-trained on extensive natural image datasets and subsequently fine-tuned for specific downstream tasks [36, 37]. The appeal of self-supervised learning lies in its ability to leverage large volumes of unlabeled data for model training, making it particularly valuable in domains where labeled data is scarce but unlabeled data is abundant, such as remote sensing, satellite imagery, and surveillance imaging.

However, applying deep learning techniques to HSI presents significant challenges, including the complexity of the feature space (also known as the curse of dimensionality) and computational requirements that are often impractical or prohibitively expensive. Existing self-supervised learning methods are predominantly designed for natural image datasets, such as ImageNet [38] and CIFAR-100 [39], which limits their direct applicability to large-scale regression models used in satellite image analysis.

In satellite imagery, each pixel value across channels is considered a critical observation for subsequent analysis, and the relationships among pixels differ significantly from those in natural images. The complexity is further amplified in HSI, where the number of spectral channels far exceeds the conventional three-band (RGB) structure of natural images. The majority of machine-learning models for satellite images have been predominantly developed for multispectral images

[40–42]. Understanding the complex relationships across the numerous bands in HSI remains challenging for achieving high accuracy in the final analytical products [43].

Figure 1.1 illustrates an example where a broadly accepted self-supervised approach [44] is employed to reconstruct HSI images and the model encounters difficulties in accurately reconstructing regions with higher spatial complexity.

1.1 Research Questions

Our study is motivated by the following research questions:

- **RQ1**: How can we design a dynamic masking strategy for the masked autoencoder framework to effectively utilize spatial variability in hyperspectral satellite images for improved encoder learning?
- **RQ2**: What channel grouping methodologies based on spectral similarity can be developed to capture complex spectral dependencies in hyperspectral imagery, and how do these methodologies impact the performance of the masked autoencoder framework?
- **RQ3**: How does accounting for spatial and spectral loss components within a masked autoencoder framework influence the quality of hyperspectral image reconstruction?

1.2 Approach Summary

In this study, we propose a novel pre-training masked autoencoder framework tailored for hyperspectral satellite images. Our framework dynamically adjusts the patch selection process to enhance the masking strategy using the spatial variability of the patches. It enables the encoder to learn from higher spatial variability during training. Our framework integrates spatial and spectral characteristics into a masked autoencoder (MAE) framework [44], extending the Visual Transformer (ViT) [1, 45]. Our framework clusters over 200 EnMAP HSI bands based on their spectral similarity, thereby capturing complex spectral dependencies within and across the channels in the

model. Our proposed approach shows strong performance in estimating soil texture at a 30-meter resolution.

1.3 Main Contributions

The contributions of our work can be summarized as follows:

1. A dynamic masking strategy for HSI for MAE pretraining, allowing for the adaptive adjustment of masking locations based on geospatial variability.

2. A channel grouping strategy for continuous spectral ranges that leverages the masking strategies to capture relationships both across groups of relevant hyperspectral bands and within each group.

3. A novel loss function for HSI that incorporates separate spatial and spectral components for higher-quality reconstruction.

1.4 Thesis Organization

The rest of the paper is organized as follows: Chapter 2 presents a comprehensive review of the literature and related work pertinent to our research. Chapter 3 elucidates our proposed methodology in detail, outlining the framework and component modifications. Chapter 4 discusses experimental results, including the findings and their implications. Chapter 5 examines a case study evaluating the effectiveness of our proposed method in a downstream task. Chapter 6 concludes the thesis, summarizing key findings, discussing limitations, and identifying future research directions.

Chapter 2

Related Work

In this chapter, we provide a concise overview of the application of machine learning techniques in the field of satellite imagery. We also conduct a systematic review of the state-of-the-art machine learning and deep learning methodologies relevant to hyperspectral satellite imagery.

2.1 Brief History of Computer Vision Models

The recent history of deep learning models in computer vision can be traced back to 2012 when AlexNet [46] won the ImagetNet competition. AlexNet demonstrated the power of deep convolutional neural networks (CNNs) and marked a turning point in the deep learning field. Later models found even better performance by increasing the number of layers, such as VGGNeT [47] (up to 19 layers) and ResNet [48] (over 100 layers). Furthermore, apart from increasing the number of layers, dense connectivity between the layers also proved to be fruitful, e.g. DenseNet [49]. Although these models showed strong performance in a supervised learning setup, the deep-learning community recognized the limitations caused by the scarcity of manually annotated data in specialized or emerging domains. This prompted a gradual shift towards developing models that are either less dependent on labeled data through semi-supervised learning or leverage unlabeled data through self-supervised learning. Self-supervised learning is particularly useful as vast amounts of unstructured and unlabeled data are available to train the models and can enable the models to learn meaningful representations from the data, which can then be applied to various downstream tasks.

2.2 Data Management

Data management is pivotal in scientific applications [50, 51], especially for spatiotemporally evolving phenomena [52]. Such data management frameworks include support for ad hoc queries [52], data sketches [53], and P2P grids [54]. Because data accesses are aligned with the

underlying data storage framework, how data is organized plays a key role in data access and overall completion times. Our proposed methodology does not place restrictions on the underlying data management frameworks.

2.3 Self-Supervised Learning

Self-supervised was introduced to reduce reliance on large annotated datasets and instead capture learning signals from the inherent structure of the training data with the help of various pretext tasks such as masked patches reconstruction and contrasting semantically related inputs. Following the development of several supervised learning models, self-supervised learning gained traction, particularly in the form of contrastive learning. Notable examples in this category include SimCLR [55] and MoCo [56] frameworks. These frameworks are usually paired with CNN-based models. In recent times, Vision Transformers (ViT) have emerged as powerful alternatives to CNNs, and have generally shown improved performance compared to CNN-based models, especially for large datasets.

2.3.1 Self-supervised Learning for Satellite Imagery

In the remote sensing field, the limited availability of human-annotated datasets poses a significant challenge for model training. To address this issue, self-supervised learning (SSL) can be leveraged to automatically extract rich semantic information from unstructured data, thereby eliminating the high costs associated with manual annotation processes. The latent features learned through SSL enable improved accuracy in downstream tasks.

Researchers have successfully applied the self-supervised learning technique in the remote sensing domain. Stojnić and Risojević [57] and Dimitrovski et al. [58] showed that self-supervised pre-training gives better results than using supervised pre-training on images with natural scenes and shows that this idea could also be easily extended to multispectral images.

Wang et al. [59] demonstrated the effectiveness of self-supervised learning on a large-scale, global, multimodal, and multiseasonal corpus of satellite imagery. Gao, Sun, and Liu [60] used

masked image modeling on remote sensing scene classification datasets which outperformed supervised state-of-the-art methods.

Spatiotemporal modeling efforts based on multispectral imagery include cloud removal [61], visualizations [62,63], and modeling the spatiotemporal dynamics of phenomena such as soil moisture [64]. Bruhwiler et al. [65] explore challenges in generating embeddings in multispectral satellite imagery. Unlike, the aforementioned studies, the crux of our effort focuses on hyper-spectral satellite imagery that are much higher dimensionality and introduces significant data management challenges.

2.4 Vision Transformer

The Transformer architecture, when first introduced, revolutionized the field of Natural Language Processing (NLP). Its ability to capture long-range dependencies made Transformer networks more effective than Recurrent Neural Networks (RNNs) which struggle to maintain context over long sequences. Although innovation led to breakthrough models in NLP such as BERT, T5, and GPT, researchers faced several challenges in adopting Transformers to computer vision tasks.

The primary difficulties in adapting transformers to vision were twofold. First, the difference in data formats: computer vision deals with either 2D or 3D data compared to the 1D sequences in text. Second, while CNNs possess inductive biases that make them effective for image processing, transformers lack these assumptions and rely on self-attention mechanisms.

Vision Transformers (ViT) introduced by Dosovitskiy et al. [1] addressed these challenges. ViT segments the images into fixed-size, non-overlapping patches and applies self-attention to these patch tokens. This approach enables ViTs to extract global features effectively. ViTs have shown competitive performance compared to CNNs on similar tasks.

MaskFeat [66] utilizes a vision transformer backbone to pre-train video and image models for predicting the Histogram of Oriented Gradients (HOG) feature representation, which can then be used for image classification. While MaskFeat directly utilizes HOG features for classification

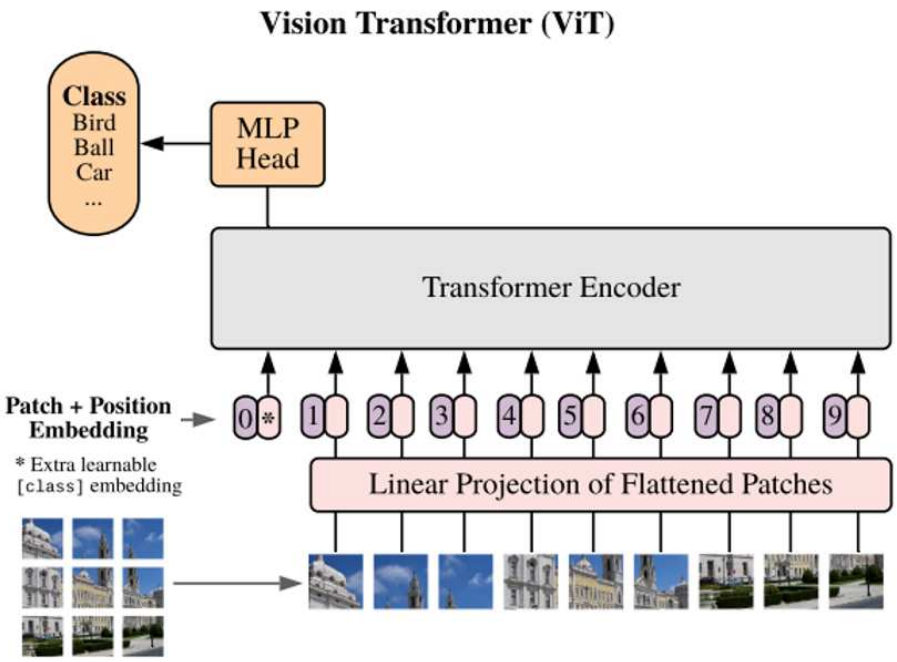

Figure 2.1: Vision Transformer Architecture [1]

on the ImageNet dataset, our proposed method capitalizes on HOG features to guide the masking strategy and is specifically designed for hyperspectral satellite images (HSI).

2.4.1 Vision Transformer for Satellite Imagery

Scheibenreif et al. [67] utilized a vision transformer with a self-supervised learning technique for land cover classification and segmentation tasks which outperformed supervised baselines. TSViT [68] employed a fully attentional model for general Satellite Image Time Series (SITS) processing based on the ViT. ViTs have also been with multispectral images for a variety of downstream tasks [69–71].

2.5 Masked Autoencoder

Following the success of ViTs, several self-supervised learning frameworks have been introduced, including DINO [72], iBOT [73], SimMIM [74], BEiT [75], and MAE. Among these frame-

works, Masked Autoencoders (MAEs) [44] have gained particular attention. MAEs mask a large portion of the input image and task the model with reconstructing the masked pixels. This approach is highly efficient and capable of producing high-quality image reconstructions. The robust performance of MAEs in datasets such as ImageNet underscores their potential in specialized domains. Examples include VideoMAE [76] with video tube masking and GMAE [77] for the domain of graphs.

SemMAE [78] uses a semantic-guided masking strategy that facilitates the learning of both local and global image structures and demonstrates its effectiveness on the ImageNet dataset. Recent work on the MAE masking strategy has also focused on using attention maps of pre-trained transformers. [79, 80]. However, these studies have evaluated their effectiveness only on traditional RGB-based image datasets. Our proposed method supplements the masking strategy based on HOG features and is suitable for HSI datasets. RMAE [81] enhances image representation learning by integrating region-based information, similar to words in natural language processing. RMAE progressively shifts from masking patches within each semantic region to masking entire regions. Our proposed method contrasts RMAE by selecting patches with higher information density and grouping spectral bands based on clustering algorithms. Similar to our proposed masking strategy which preferentially selects areas with higher spatial variability for encoder learning, AdaMAE [82] samples more tokens from the high spatiotemporal information regions, but it applies to the video domain.

2.5.1 Masked Autoencoders for Hyperspectral Satellite Imagery

SatMAE [40] presents a pre-training framework for temporal and multi-spectral satellite imagery based on Masked Autoencoder for multispectral images. SatMAE groups the multispectral bands with similar spatial resolution and wavelength characteristics. This approach enables the model to access information from unmasked regions in one group while those same regions may be masked in another. This significantly increases the model's learning ability. SatMAE applies this approach to images collected from the Sentinel-2 dataset, which comprised 10 spectral bands

after excluding bands with lower spatial resolution, which is considerably smaller than the $200+$ bands that HSI has to offer.

Cao et al. [83] propose a Transformer-based MAE using contrastive loss (TMAC) for hyperspectral image classification. TMAC has two branches – the MAE branch focuses on learning pixel-wise representations while the contrastive learning branch focuses on capturing the high-level, holistic features. Unlike TMAC, which utilizes dual branches, our proposed architecture enhances the masking strategy and modifies the loss function to effectively capture high-level features.

SpectralMAE [84] employs a masked autoencoder (MAE) specifically tailored for spectral dimensions for higher-quality spectral reconstruction. Compared to our proposed dynamic masking and reconstruction functionality, spectralMAE requires a fine-tuning phase that relies on fixed masking patterns and might have limited applicability beyond spectral reconstruction.

MaskedSST [85] provides training techniques for ViT to reduce computational costs and compares its performance against several models, including baseline ViT RGB and 3D-CNN models, trained on EnMAP and the Houston 2018 dataset. However, the overall model performance does not exceed the 3D-CNN performance.

SSMAE [86] combines transformer-based global feature extraction with lightweight CNNs for local feature enhancement. In contrast, our feature extraction is accomplished by a feature-guided masking strategy and a novel loss function.

SatMAE++ [87] extended SatMAE on various downstream classification tasks by introducing novel scaling strategies. Similarly, ScaleMAE [42] pretrains a masked autoencoder network at different scales for higher quality and accurate reconstruction. None of these efforts has been benchmarked using hyperspectral images. In contrast, our model evaluates its effectiveness on hyperspectral images collected from the EnMAP satellite, which has 218 bands.

Feature Guided Masked Autoencoder (FG-MAE), proposed by Wang et al. [88], reconstructs features such as Histogram of Oriented Gradients (HOG) and Normalized Difference Indices (NDI)

for multispectral images from masked input images. Our proposed work instead uses HOG for enhancing the masking strategy, thereby achieving higher-quality reconstruction.

2.6 Leveraging Autoencoder Learning on Satellite Imageries

Autoencoder-based approaches have been widely adopted for tasks involving high-dimensional hyperspectral data. For instance, Shi et al. [89] developed a 3D attention-based denoising network for hyperspectral images, which leveraged deep spatial-spectral features to reduce noise while maintaining high prediction accuracy, making it suitable for complex environmental applications such as soil property estimation. Additionally, He et al. [90] utilized convolutional autoencoders to restore hyperspectral images, emphasizing the effectiveness of latent feature extraction for downstream classification tasks like land-cover mapping. Similarly, Zhou et al. [91] proposed a stacked autoencoder for dimensionality reduction and feature learning, demonstrating its utility in hyperspectral data processing to enhance model efficiency. These studies highlight the value of using learned latent features from autoencoders to handle high-dimensional environmental datasets for downstream tasks such as the soil texture prediction task that we consider in this study.

Chapter 3

Methodology

In this chapter, we discuss how we develop our proposed feature-guided masking strategy. We then discuss the channel grouping strategies we explored. Finally, we describe the components of our proposed loss function.

3.1 Motivation

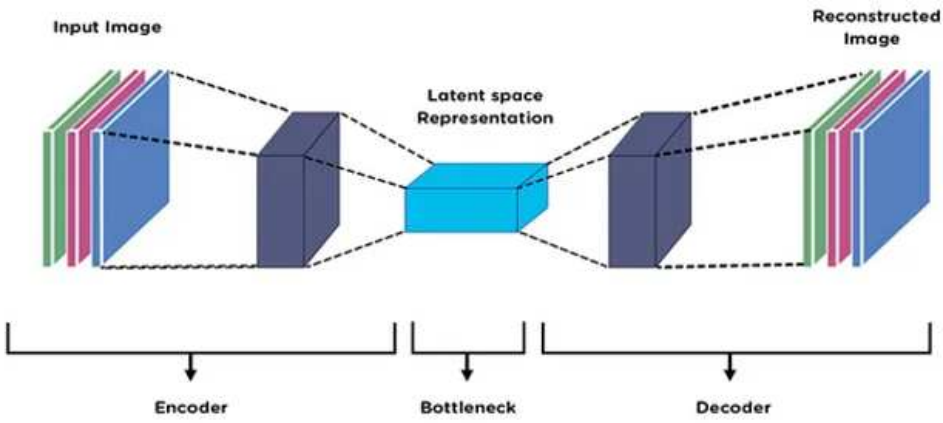

Figure 3.1: Autoencoder Architecture

Autoencoders have the general architecture of generating lower-dimensional embeddings for a set of data as illustrated in figure 3.1, which subsequently becomes useful for different downstream tasks [92]. This dimensionality reduction technique is particularly appealing for hyperspectral satellite images which have a large number of spectral bands [93]. The masked autoencoder architecture is a state-of-the-art deep learning framework that has also shown strong performance for a variety of tasks.

3.2 Background: Brief Introduction to Masked Autoencoders

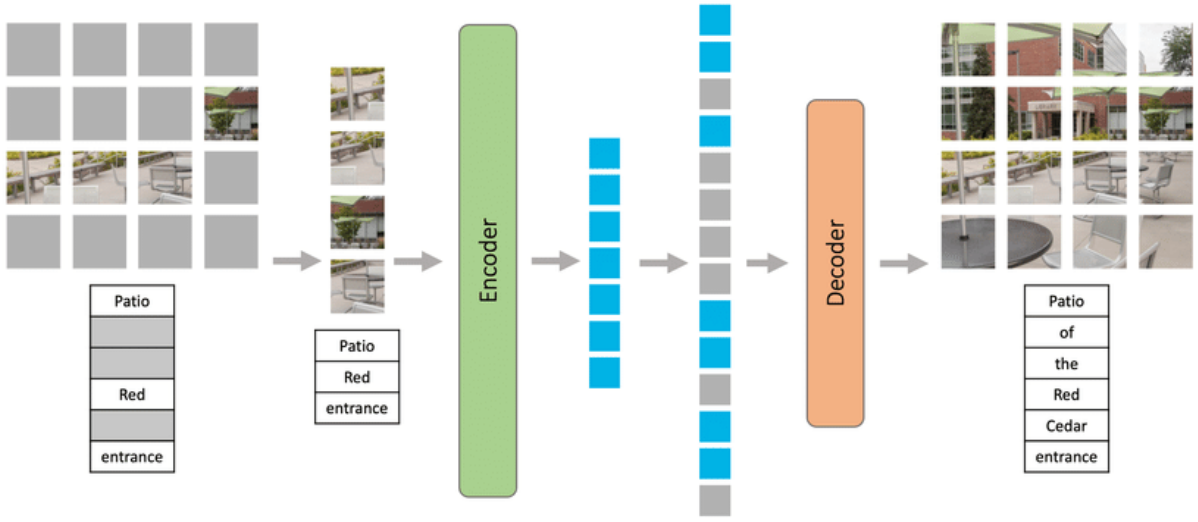

Figure 3.2: Masked Autoencoder Architecture [2]

Masked Autoencoders (MAEs) are autoencoders that use self-supervised techniques to learn latent representations from input images by masking a portion of the input data and then reconstructing the input data from the masked input by utilizing the learned latent representation. The images on which MAEs operate typically have the shape $C \times H \times W$, where C is the number of bands or channels in the image, H is the height of the image and W is the width of the image. MAEs resize the input image to create a set of non-overlapping square patches, each with shape $P \times P$. This results in N number of patches where $N = \frac{H \times W}{P^2}$. Similar to the input image, each patch has C channels and is represented by a flattened vector of length P^2C. Each of the flattened vectors is mapped to a lower dimension D with the help of a patch embedding function to create embedded tokens. As MAEs are based on Vision Transformer architecture and transformers inherently have no idea about position, positional encodings are added to the embedded tokens to keep track of the spatial information. A fraction r of the N tokens are randomly masked and the remaining visible tokens are fed as input to the encoder.

The decoder provides a learning signal to the encoder during the training process. For the encoder to provide higher-quality reconstruction, the encoder has to compress the most important features from the visible patches. This enables the encoder to learn rich and meaningful representation during the training process which becomes useful for various downstream tasks.

3.3 Network Architecture

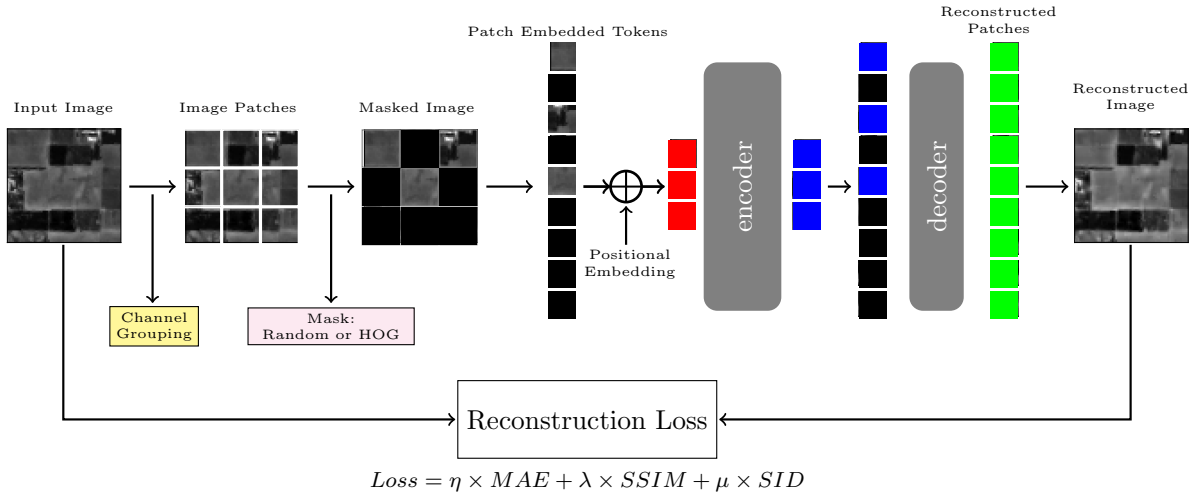

Figure 3.3: Network Architecture of Our Proposed Model. The Channel Grouping, Masking, and Loss Function are illustrated in the diagram.

Our proposed network architecture is based on the standard masked autoencoder architecture, but we propose improvements to three key areas - masking strategy, channel grouping strategy and the loss function, as illustrated in figure 3.3.

3.4 Feature Guided Masking Strategy

The random masking strategy utilized by MAEs to select only a portion of the embedded tokens for training reduces the computational cost significantly. This random masking strategy has been shown to reduce the pre-training time by a factor of 3 or more [94] and works well for images containing objects such as cats, birds, or vehicles. The authors of MAE demonstrated their

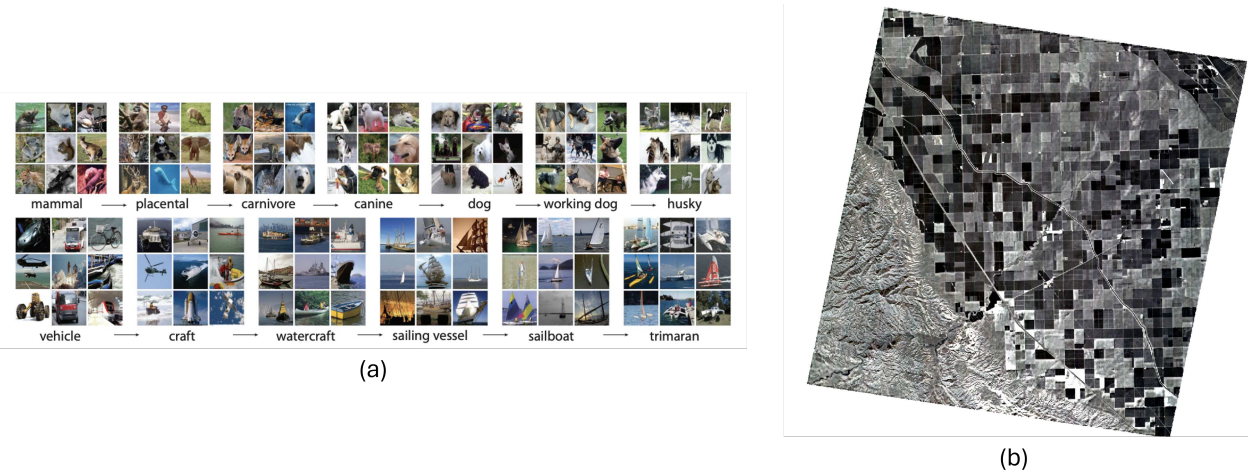

Figure 3.4: (a) ImageNet dataset (b) Satellite Imagery

model's effectiveness on the ImageNet [38] dataset. But random masking might not be well suited for satellite imagery which exhibits highly diverse landscapes compared to mostly homogenous patterns in images containing objects.

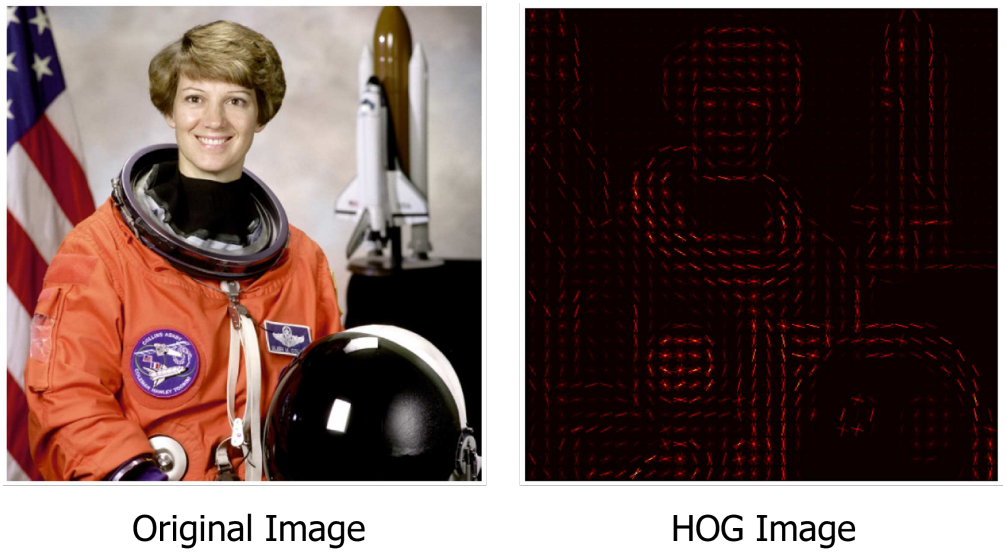

Figure 3.5: HOG Visualization

As an example, a 64×64 pixel cropped satellite image covers a region of approximately 2km $\times$ 2km and can encompass a complex array of topological features, including roads, buildings, agricultural fields, and mountains. Furthermore, the complexity of the landscape is not uniformly

distributed - while some areas may have more homogeneous characteristics (e.g., a large crop field), other areas may present greater diversity (e.g., a dense urban area) as constructed in figure 3.4. It is therefore essential for the model to perform coherently across the varied regions, regardless of how complex they might be, to support accurate downstream tasks.

We propose a masking strategy that emphasizes spatial variability across patches and guides the model to focus on areas with higher variability for better learning. Instead of assigning a uniform masking probability across the patches, we calculate the magnitude of spatial variability contained within the patches based on the Histogram of Oriented Gradients (HOG) [95]. HOG captures local shapes and appearances within each patch by quantifying the distribution of gradient orientations within that patch, as illustrated in figure 3.5. Utilizing HOG enables the model to prioritize complex regions during the mask creation step and enables the model to learn from higher variability patches more frequently.

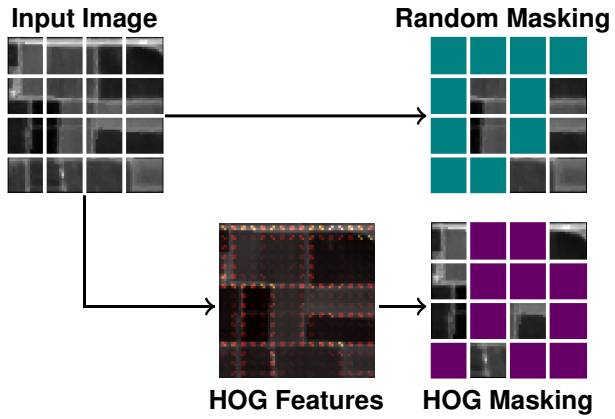

Figure 3.6: Contrasting *random masking* vs *HOG-based masking*. Our HOG-based masking preferentially selects areas with higher spatial variability for encoder learning.

Since HOG is typically applied on a single channel in an image, given a hyperspectral image of shape $C \times H \times W$ (where C is the number of channels and typically has a value more than 100), we apply Principal Component Analysis (PCA) to reduce the spectral dimensionality and identify the channel image that demonstrates the highest variance. We then extract the HOG feature values from each patch. We sort the patches based on their HOG feature values and split them into two

groups: $\alpha \times N$ patches with higher variability and $(1 - \alpha) \times N$ patches with lower variability, where alpha is an adjustable split ratio.

We assign two distinct local masking ratios to each of the two groups: beta for patches with higher variability and gamma for patches with lower variability. The local masking ratios are dynamically adjusted so that during training, more patches are selected from the lower variability group to be masked and the total number of masked tokens selected from the two groups maintain the specified (and configurable) masking ratio r.

More specifically, for a masking ratio r, which remains consistent throughout the entire training process, the following relationship between α, β and γ holds:

$$r = \alpha \times \beta + (1 - \alpha) \times \gamma \tag{3.1}$$

We adjust the local ratios (β and γ) in such a way that the following condition also holds:

$$\underbrace{\alpha \times \beta}_{\text{Higher Variability Group}} < \underbrace{(1 - \alpha) \times \gamma}_{\text{Lower Variability Group}} \tag{3.2}$$

The condition in 3.2 is maintained by setting $\alpha < 0.5$ and β to a much smaller value than γ. This allows the model to select more patches from the lower variability group and enables the encoder to focus on learning the areas with greater information density as the training progresses.

We illustrate the contrast between the random masking strategy and our proposed HOG-based masking strategy in figure 3.6.

3.5 Channel Grouping

Unlike RGB images, which have 3 channels, and multispectral satellite images which have less than 30 channels, HSI provides a large number of channels, typically more than 100. For example, the EnMAP dataset that we use in our experiments in chapter 4 has 218 channels per pixel. While leveraging this abundant spectral data effectively may present computational challenges, it

also opens up an opportunity to harness the number of channels to enhance the model's learning capabilities by grouping the channels based on some intelligent criteria.

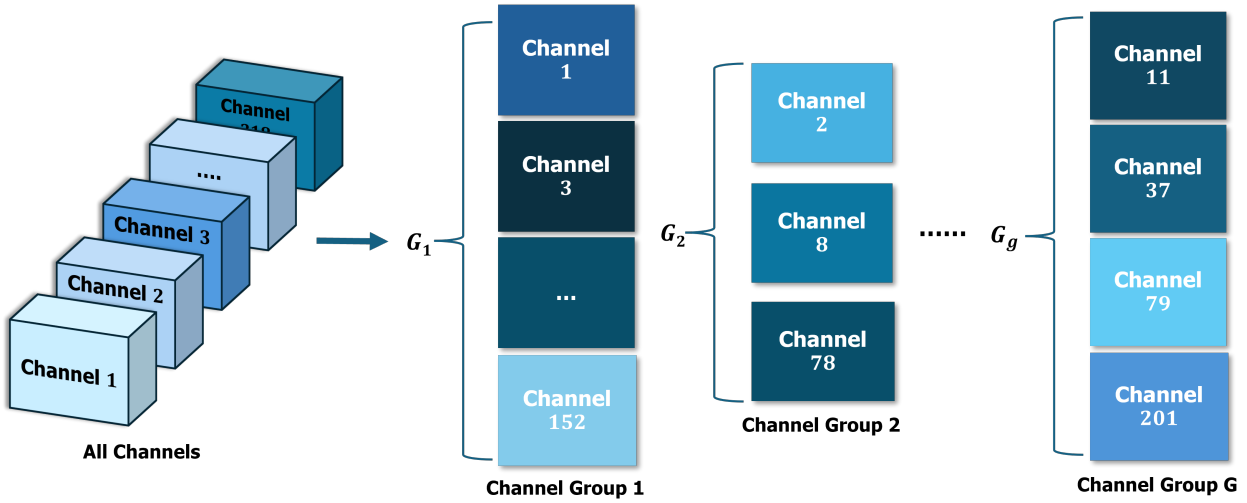

Figure 3.7: Illustration of Channel Grouping

Given an input image, the standard approach in MAE involves applying the same mask uniformly across all channels of the input image. While this strategy is effective on datasets like ImagetNet, which mostly contain RGB images, it underutilizes the large number of bands HSIs have to offer. If we group certain channels based on thoughtfully designed criteria, the benefits we can achieve are threefold:

1. **Enhanced Learning Through Diverse Masks:** The mask generated for each group will be different. As the number of groups increases, there is a higher likelihood that any given region will remain unmasked in at least one group, allowing the model to learn from those regions where it otherwise would not be able to learn if there was only one group, as illustrated in figure 3.8. This enables the model to reconstruct that region in groups where it is masked by applying the learning from groups where it is unmasked. Specifically for HSI, we have many bands or channels, so it is possible to create a large number of groups to facilitate higher-quality reconstruction.

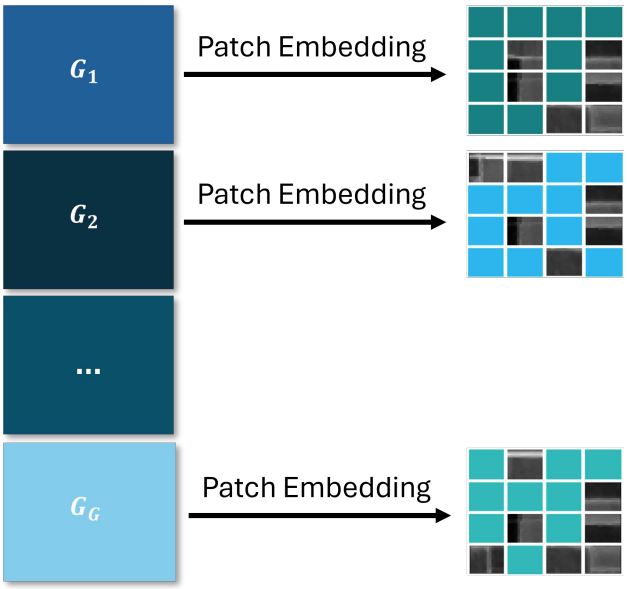

Figure 3.8: Separate Masks for Each Group, Enabled by Channel Grouping

2. **Efficient Capture of Spectral Correlations:** We can design the channel grouping criteria in such a way that bands with similar spectral characteristics are grouped together and the model focuses on learning the intrinsic spectral features within each group. Adjacent spectral bands often capture related information due to the continuous nature of the spectral signatures in HSI. By clustering these bands together, the model effectively learns the subtle variations and patterns in the image.

3. **Reducing Computational Overhead**: Instead of attempting to learn correlations across all 200+ channels simultaneously—which can be computationally intensive and prone to overfitting—channel grouping reduces the complexity associated with modeling inter-band relationships.

Therefore, we partition a hyperspectral image containing C channels into G unique groups, denoted as $g_1, g_2, g_3, \ldots, g_G$, as illustrated in figure 3.7. The number of channels $|g_i|$ in each group i equals the number of channels. In other words, we have $\sum_{i=1}^{G} |g_i| = C$. This grouping approach segments the input image I along its spectral axis, effectively making G separate images $I_1, I_2, \ldots, I_G$ each with dimensions $|g_i| \times H \times W$. Each image is first transformed into a set of

patches with dimensions $N \times P^2|g_i|$, where N represents the number of patches. For each group, a distinct patch embedding function is applied. Specifically, each group g_i is linearly projected to a lower-dimension D to create token embeddings with shape $N \times D$. The embedded tokens from all groups are concatenated to produce the final sequence with shape $GN \times D$.

In this study, we evaluated the following grouping techniques to effectively capture features within the channel groups while preserving valuable spectral information:

1. **Stacking**: All channels are combined into a single group.

2. **VNIR-SWIR**: Based on the Center Wavelength (CW) and Full Width at Half Maximum (FWHM) of the spectral bands (provided by the EnMAP Spectral Response Function), the spectral bands are divided into two groups: Visible and Near-infrared (VNIR) and Short-Wave Infrared (SWIR).

3. **Soil-Reflectance (SR)**: Utilizing spectral reflectance data of soil surfaces from the EnMAP Science Plan [96], we subdivide the channels into five distinct groups. We focus on the soil surface reflectance because it aligns with our downstream task.

4. ***k*-Means Clustering**: An unsupervised learning algorithm that assigns channels to clusters based on the mean of the nearest channel. This method is particularly effective for large datasets, similar to the ones found in HSI.

5. **Hierarchical Agglomerative Clustering (HAC)**: Similar to *k*-Means, HAC minimizes within-cluster distance using cluster linkage, e.g., Ward's method [97]. However, unlike k-Means, HAC does not assume spherical cluster shapes and can potentially capture more complex inter-channel relationships.

6. **Spectral Comparison Index (SCI)**: This technique groups channels based on the Mean Absolute Error between pairs of channels. SCI maximizes spectral similarity by comparing pairwise MAE values and provides accessible similarity scores.

3.6 Loss Function

During the training phase, Masked Autoencoders calculate the pixel-level loss between the original input image and the reconstructed image by utilizing the Mean Squared Error (MSE) loss. Traditional loss functions such as L_1 (Mean Absolute Error) or L_2 (Mean Squared Error) only measure loss in mathematical notations. They do not utilize any domain-specific knowledge associated with the input data and they also cannot capture structural mismatches as illustrated in figure 3.9.

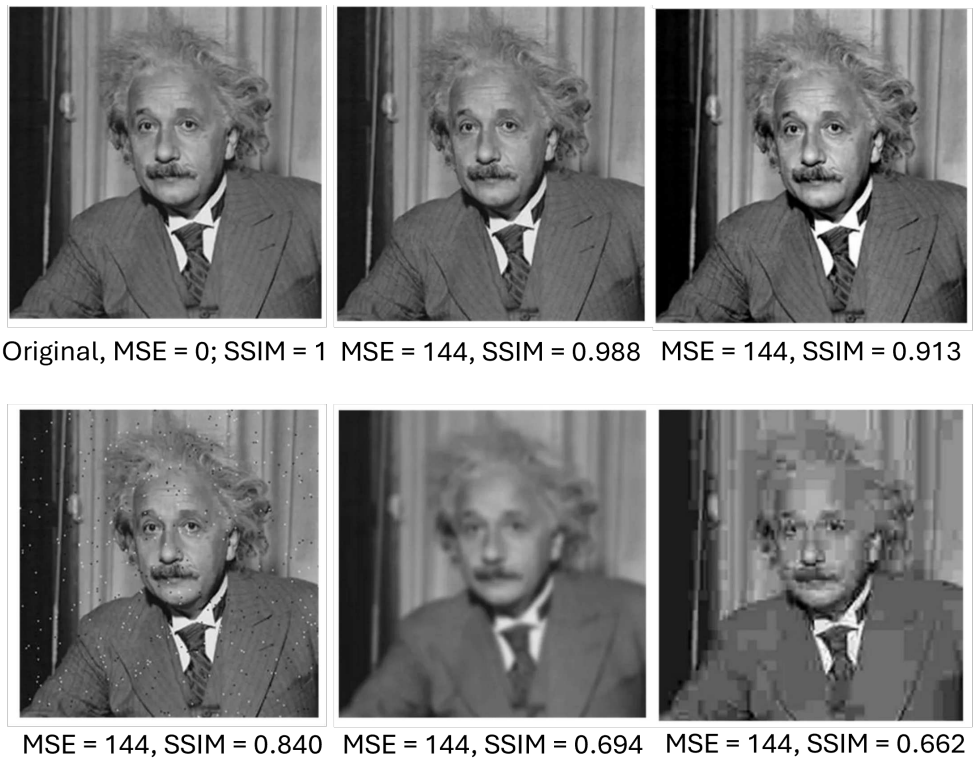

Figure 3.9: Traditional Loss Functions like MSE Cannot Capture Structural Mismatches [3]

$$\mathrm{MSE}(\mathrm{x}, \mathrm{y}) = \frac{1}{n} \sum_{i=1}^{n} (x_i - y_i)^2 \tag{3.3}$$

Although MSE, shown in Eq. (3.3), is frequently used in a wide range of deep-learning architectures, it is not specifically designed to capture structural mismatches [98]. However, in the case

of HSI, which consists of many spectral bands with high spatial variability, incorporating spatial and spectral characteristics in the loss function is crucial for effective model training.

We propose an enhanced loss function that utilizes the spatial and spectral characteristics of the input data to capture the nuances of both the intra and inter-channel dependencies and lead the model toward accurate reconstruction of the input data. Our proposed loss function combines Mean Absolute Error (MAE), shown in Eq. (3.4) with Structural Similarity Index (SSIM) [99], and Spectral Information Divergence (SID) [100].

$$\mathrm{MAE}(\mathrm{x}, \mathrm{y}) = \frac{1}{n} \sum_{i=1}^{n} |x_i - y_i| \tag{3.4}$$

SSIM, shown in Eq. (3.5) applies a sliding window technique to preserve the spatial structures local to each spectral band and uses human-perceivable structural information, e.g., luminance, texture, contrast, and structure to quantify image quality. SSIM uses the pixel mean of the input image (μ_x), the pixel mean of the predicted image (μ_y), the variance of the input image (σ_x) and the variance of the predicted image (σ_y).

$$\mathrm{SSIM}(\mathrm{x}, \mathrm{y}) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{3.5}$$

Owing to the large number of bands in HSI, complex patterns could emerge across this extensive spectral range. This necessitates the design of a loss function that captures the nuances of inter-channel dependencies for reconstructing spectral bands with high fidelity. While MAE focuses on the pixel space and SSIM focuses on the spatial patterns, SID, shown in (3.6), focuses on the spectral difference across the bands for each patch and prompts the model to learn spectral correlation during training. SID quantifies the divergence between the input and reconstructed spectra by first calculating the ratio of each corresponding spectral channel. It then applies a logarithmic transformation to these ratios to assess the information loss for each channel. These individual divergences are aggregated across all spectral channels, and the final loss is determined by taking

the average of these aggregated values. This averaged SID loss effectively measures the overall information discrepancy between the original and reconstructed spectra.

$$\mathrm{SID}(\mathrm{x}, \mathrm{y}) = \sum_{i=1}^{C} \left(x_i \cdot \log \left(\frac{x_i}{y_i} \right) - x_i + y_i \right) \tag{3.6}$$

Our proposed loss function is calculated as a weighted sum of MAE, SSIM, and SID and is represented as follows:

$$Loss = \eta \times MAE + \lambda \times SSIM + \mu \times SID \tag{3.7}$$

η, λ and μ are adjustable loss coefficients. Initially, more importance is given to MAE and then we gradually increase the weights of SSIM and SID coefficients during the later stages of the training.

Chapter 4

Experiments and Analysis

In this section, we describe the HSI dataset on which we conducted our experiments. We then specify details about our implementation and chosen evaluation metrics. Finally, we mention the experimental results and demonstrate the effectiveness of our proposed components through ablation studies.

4.1 Dataset

We collected hyperspectral images from the Environmental Mapping and Analysis Program (EnMAP) satellite launched in 2022. The spatial resolution of the images is 30m. The swath width of the EnMAP satellite is 30km, which refers to the earth's surface area covered by the satellite's sensors in a single pass.

EnMAP provides the hyperspectral images in the form of GeoTIFF files and each GeoTIFF file contains an area of approximately 30 km $\times$ 30 km. All images were collected from the region of California, U.S. between April 2022 and August 2024. We discarded images with high cloud and snow coverage and a strong presence of mountainous regions. Each GeoTIFF file contains 224 bands and the wavelengths range from 420 nm to 2450 nm.

After eliminating bands with no data values, some of the GeoTIFF files contained 218 bands, while others contained 219 bands. To maximize the size of our dataset, we selected the corresponding 218 bands from all GeoTIFF files. To facilitate model training, we created 64×64 pixels tiles from the original EnMAP images. Our entire dataset contains 22078 such tiles, categorized into a training dataset of 12466 tiles, a validation dataset of 2078 tiles, and a test data of 7534 tiles. We further divided the test dataset into two groups: Test dataset 1 with 6234 tiles and Test dataset 2 with 1300 tiles. Test dataset 2 contains *"unseen"* areas, which represent geospatial regions that the model did not encounter during training. Test dataset 2 is particularly important for assessing

the model's generalization ability, as spatially overlapping tiles often share similar topographical and environmental conditions due to their proximity.

4.2 Implementation Details

We implemented the Masked Autoencoder in PyTorch. We trained the models for 300 epochs with a variable learning rate, initialized with 10^{-3} with the Cosine Annealing scheme. The models are trained with Adam optimizer [101] with $\beta_1 = 0.9$ and $\beta_2 = 0.95$.

We use standard min-max normalization, randomly crop between 20% and 100% of the area of the tile, and perform a random horizontal flip on the tiles to augment the training dataset. We chose a batch size of 32, set the masking ratio to 75%, and used a patch size of 4 across all experiments. The training time of the proposed network is between $12 - 15$ hours with a single NVIDIA A100 80G Tensor GPU based on the model setup.

4.3 Evaluation Metrics

We used Mean Absolute Error (MAE), Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) to evaluate the reconstruction performance.

SSIM, shown in Eq. 3.5, is a metric used to assess the quality of images by comparing the similarity between a reference image and a distorted or altered image [102].

PSNR, shown in Eq. 4.1, measures the quality of reconstructed or compressed images compared to their original versions [103].

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{\text{MAX}^2}{\text{MSE}} \right) \tag{4.1}$$

The MSE for images with dimensions (C, H, W) is calculated as:

$$\text{MSE} = \frac{1}{C \cdot H \cdot W} \sum_{c=1}^{C} \sum_{h=1}^{H} \sum_{w=1}^{W} (X_{c,h,w} - Y_{c,h,w})^2 \tag{4.2}$$

Where C is the number of channels and H and W are the height and width of the images.

Table 4.1: Experimental results on the Test Dataset

Identifier	Masking Strategy	Grouping Strategy	Loss Function	Test Set 1			Test Set 2		
				MAE ↓	PSNR ↑	SSIM ↑	MAE ↓	PSNR ↑	SSIM ↑
1	Random	Stack	MAE	0.0143	33.36	0.8831	0.0152	33.06	0.8258
2	Random	Stack	MAE + SSIM	0.0106	35.66	0.8957	0.0114	35.75	0.8952
3	Random	Stack	MAE + SID	0.0153	32.66	0.8093	0.0166	32.37	0.7972
4	Random	Stack	MAE + SSIM + SID	0.0094	36.58	0.9119	0.0099	36.73	0.9106
5	HOG	Stack	MAE	0.0110	34.86	0.8844	0.0115	35.68	0.8839
6	HOG	Stack	MAE + SSIM	0.0104	35.75	0.8985	0.0109	35.99	0.8988
7	HOG	Stack	MAE + SID	0.0111	35.23	0.8844	0.0117	35.47	0.8843
8	HOG	Stack	MAE + SSIM + SID	0.0096	36.33	0.9094	0.0100	36.57	0.9094
9	Random	DBSCAN	MAE	0.0121	35.35	0.8807	0.0127	35.28	0.8890
10	Random	kMeans	MAE	0.0079	38.23	0.9318	0.0083	38.21	0.9339
11	Random	SCI	MAE	0.0093	37.22	0.9131	0.0094	37.11	0.9124
12	Random	kMeans	MAE + SSIM	**0.0047**	**42.41**	**0.9718**	**0.0050**	**42.33**	**0.9718**
13	Random	kMeans	MAE + SID	0.0070	38.64	0.9348	0.0074	38.52	0.9341
14	Random	kMeans	MAE + SSIM + SID	0.0049	42.05	0.8693	0.0053	41.99	0.9694
15	HOG	kMeans	MAE	0.0068	39.50	0.9300	0.0074	39.50	0.9280
16	HOG	kMeans	MAE + SSIM	0.0071	39.50	0.9300	0.0078	39.30	0.9276

4.4 Baseline Model

To benchmark our proposed masking and grouping strategies and loss function, we used a ViT-Base backbone with $D = 1024$, 24 transformer blocks, each with 16 attention heads. The decoder, considerably smaller than the encoder, has an embed dimension of 512, 8 transformer layers, and is also equipped with 16 attention heads per layer.

4.5 Model Performance

The results on partitions of the test dataset can be seen in table 4.1. From the results, it is easily visible that our proposed architecture performs significantly well across the board. For example, the HOG-based masking strategy achieves a 24.34% reduction in MAE loss, a 7.92% increase in PSNR, and a 6.57% increase in SSIM on test set 2 compared to the baseline model. Our proposed *k*-Means grouping strategy achieves a 45.39% reduction in MAE loss, a 15.58% increase in PSNR, and a 13.09% increase in SSIM on test set 2 compared to the baseline model. Our loss function components (SSIM and SID) achieve a 34.89% reduction in MAE loss, a 11.10% increase in PSNR, and a 10.27% increase in SSIM on test set 2 compared to the baseline model. Finally, our

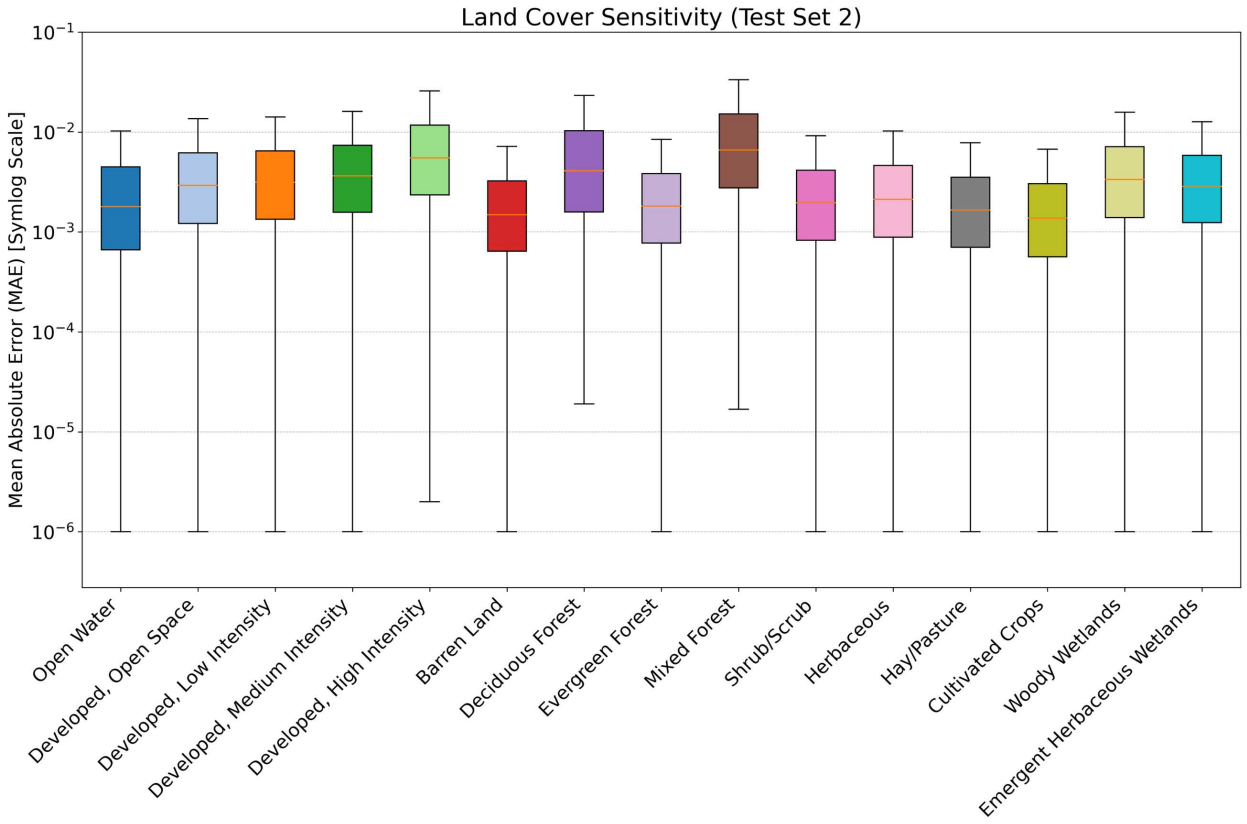

Figure 4.1: Land Cover Sensitivity on Test Set 2

best-performing model achieves a 65.13% reduction in MAE loss, a 27.01% increase in PSNR, and a 14.81% increase in SSIM on test set 2 compared to the baseline model.

4.6 Model Sensitivity Analysis

We performed a series of sensitivity analyses to evaluate the performance of our model across different land cover types, seasonal changes, and spectral bands.

4.6.1 Land Cover Type Sensitivity Analysis

We retrieved the National Land Cover Database (NLCD) for the state of California [104]. We have used the most recent NLCD dataset published in 2021 at a 30m resolution. We cropped the NLCD dataset to align with our EnMAP tiles and then extracted the corresponding NLCD

Table 4.2: Per Land Cover Type MAE on Test Dataset

	MAE	
Land Cover Type	Overlap	w/o Overlap
Open Water	0.0064	0.0044
Perennial Snow/Ice	N/A	N/A
Developed, Open Space	0.0054	0.0052
Developed, Low Intensity	0.0055	0.0055
Developed, Medium Intensity	0.0059	0.0063
Developed, High Intensity	0.0102	0.0103
Barren Land	0.0036	0.0033
Deciduous Forest	0.0051	0.0060
Evergreen Forest	0.0044	0.0033
Mixed Forest	0.0039	0.0088
Shrub/Scrub	0.0039	0.0036
Herbaceous	0.0038	0.0041
Hay/Pasture	0.0039	0.0037
Cultivated Crops	0.0028	0.0027
Woody Wetlands	0.0057	0.0061
Emergent Herbaceous Wetlands	0.0049	0.0049

values using geographic coordinates. Each pixel in an EnMAP tile is mapped to one of the 16 corresponding land cover types.

The boxplot of land cover types found in the NLCD dataset is shown in 4.1. The Mean Absolute Error values for each land cover type on both partitions of the test dataset are shown in 4.2. Our proposed model performs well across all the land cover categories.

4.6.2 Seasonal Sensitivity Analysis

We grouped our EnMAP tiles based on the timestamp of the parent GeoTIFF file. Each tile generated from a single GeoTIFF file is assigned the same timestamp. Instead of dividing the tiles based on seasons, we opted to divide them based on the months for finer granularity.

The boxplot of the months available in test set 2 is shown in figure 4.2. The y-axis is shown in a symmetrical log scale for better visualization.

The Mean Absolute Error values for each month on both partitions of the test dataset are shown in 4.3.

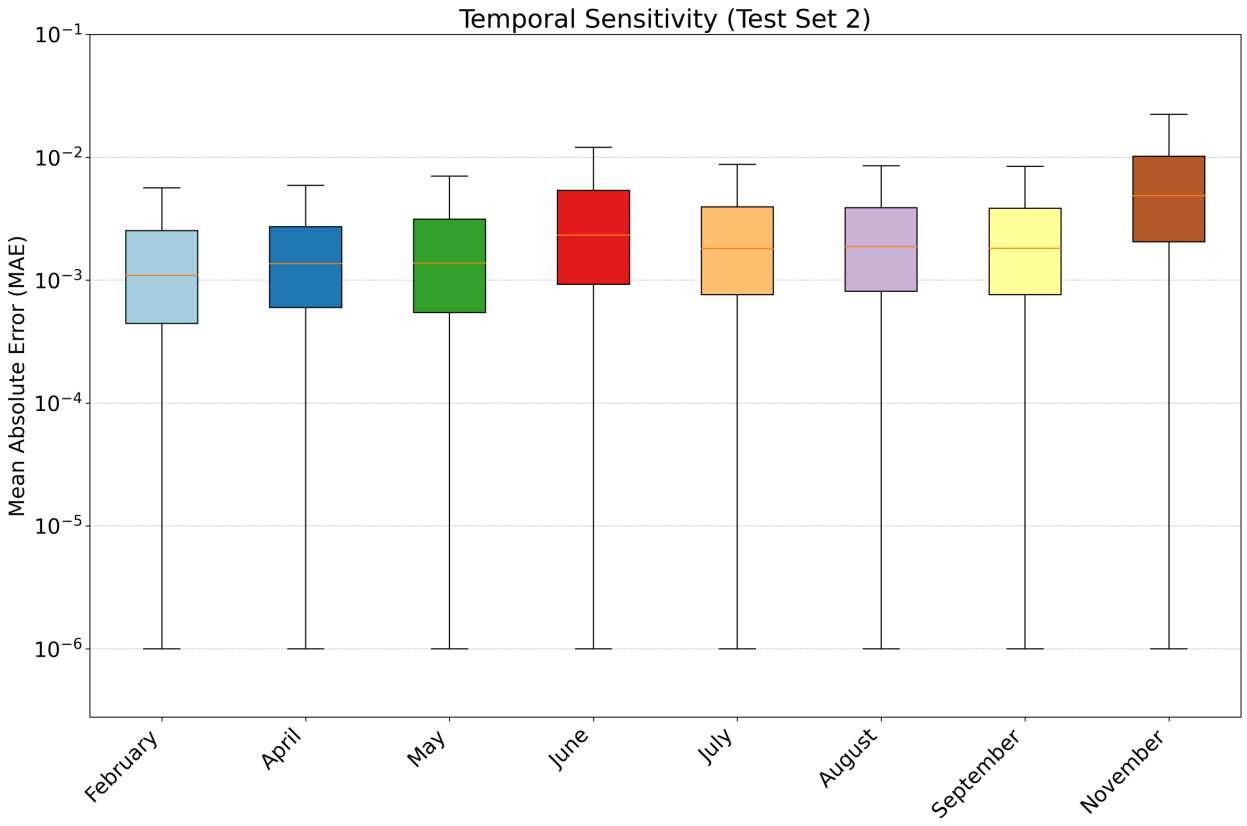

Figure 4.2: Temporal Sensitivity on Test Set 2

Table 4.3: Per Month MAE on Test Dataset

	MAE	
Land Cover Type	Overlap	w/o Overlap
January	N/A	N/A
February	0.0041	0.0023
March	N/A	N/A
April	0.0032	0.0022
May	0.0028	0.0027
June	0.0035	0.0049
July	0.0033	0.0035
August	0.0037	0.0034
September	0.0035	0.0035
October	0.0033	N/A
November	0.0030	0.0082
December	N/A	N/A

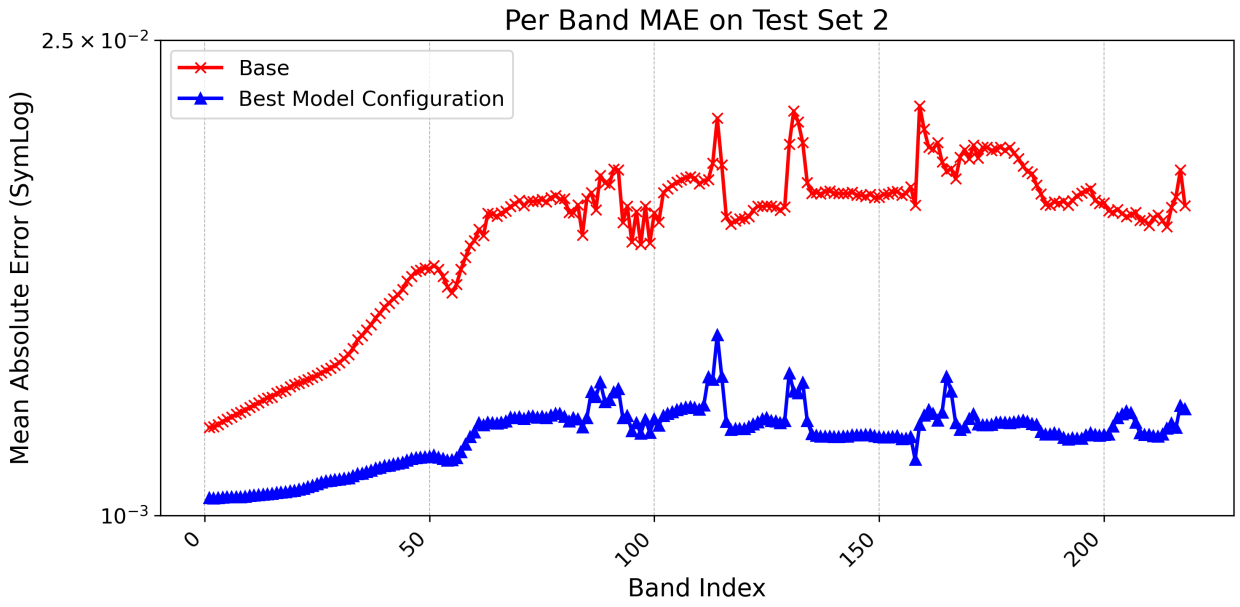

Figure 4.3: Per Band Mean Absolute Error on Test Set 2

4.6.3 Spectral Sensitivity Analysis

To evaluate how our model performed on each of the 218 bands on the test dataset, we calculated the Mean Absolute Error (MAE) for all bands and plotted the errors. The corresponding MAE for each band is illustrated in 4.3.

4.7 Ablation Study

We performed ablation studies to evaluate the effectiveness of our proposed refinements.

4.7.1 Masking Strategy

Table 4.4: Masking ablation study on Test Set 2

Masking	Grouping	Loss	MAE ↓	PSNR ↑	SSIM ↑
Random	Stack	MAE	0.0152	33.06	0.8258
HOG	Stack	MAE	**0.0115**	**35.68**	**0.8839**

Table 4.4 shows that our proposed HOG-based masking strategy performs significantly better than the random masking strategy for all evaluation metrics on test set 2 compared to the base model.

4.7.2 Channel Grouping Strategy

Table 4.5: Channel Grouping Ablation Study on Test Set 2

Masking	Grouping	Loss	MAE ↓	PSNR ↑	SSIM ↑
Random	Stack	MAE	0.0152	33.06	0.8258
Random	kMeans	MAE	**0.0083**	**38.21**	**0.9391**

Table 4.5 shows that our grouping strategy (*k*-Means) performs significantly better for all evaluation metrics on test set 2 compared to the base model.

4.7.3 Loss Function Components

Table 4.6 shows the performance of our loss function components. It is visible that SID performs poorly when coupled with MAE compared to SSIM coupled with MAE. But when all three components are coupled together, our proposed model performs significantly better than the baseline version.

Table 4.6: Loss Function Components Ablation Study on Test Set 2

Masking	Grouping	Loss	MAE ↓	PSNR ↑	SSIM ↑
Random	Stack	MAE	0.0152	33.06	0.8258
Random	Stack	MAE + SID	0.0166	32.37	0.7972
Random	Stack	MAE + SSIM	0.0114	35.75	0.8952
Random	Stack	MAE + SSIM + SID	**0.0099**	**36.73**	**0.9106**

Chapter 5

Case Study: Soil Texture Prediction

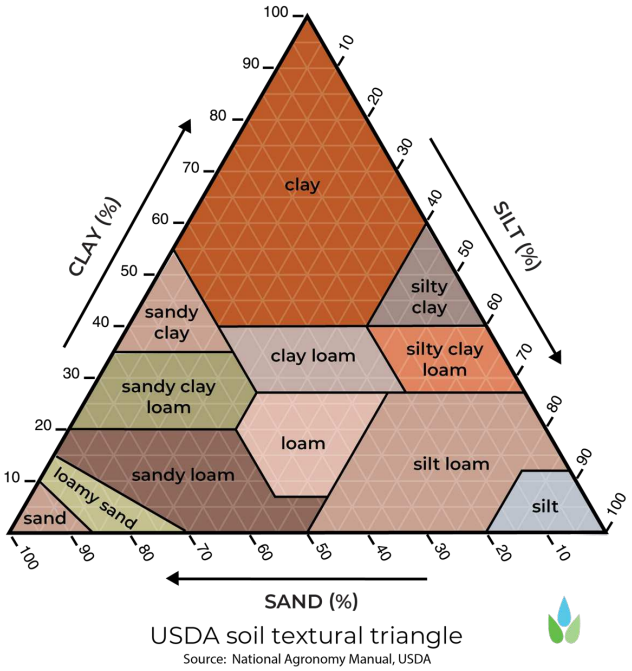

Figure 5.1: Soil Texture Triangle

To evaluate the effectiveness of our approach with downstream models, we conducted a case study to estimate soil texture using latent features generated by our proposed model. Soil texture is a critical determinant of soil health and productivity, influencing various soil properties and functions that encompass water retention, nutrient availability, and susceptibility to erosion. In this study, we employed the USDA's soil texture classification scheme, representing soil texture through proportions of sand, silt, and clay as illustrated in figure 5.1. Our downstream model estimates the proportions of sand and silt in the soil using EnMAP hyperspectral satellite images. For labeled samples, we utilized the POLARIS 30m probabilistic soil properties dataset across the contiguous United States, as published by the USDA [105].

Table 5.1: Test Error of the Downstream Task With Latent Features as Input From Different Versions of Our Proposed Model

Model	Mean Absolute Error
Baseline (Random Mask, No Grouping, MAE Loss)	0.0956
HOG Mask, No Grouping, MAE + SSIM + SID Loss	0.1033
HOG Mask, No Grouping, MAE Loss	0.1001
Random Mask, Grouping (*k*-Means), MAE Loss	0.0995
Random Mask, No Grouping, MAE + SSIM + SID Loss	0.0939
Random Mask, Grouping (*k*-Means), MAE + SSIM Loss	0.0863
Random Mask, Grouping (*k*-Means), MAE + SSIM + SID Loss	**0.0806**

We implemented VGG13 as a regression model to estimate the percentages of sand and silt. VGG13 is a variant of the VGG architecture [106] and its straightforward yet effective design, combined with its broader applications for computer vision tasks, aligns well with our objective of evaluating the effectiveness of our proposed model. The latent features, originally encoded with a dimensionality of $64 \times 64 \times 218$, were processed using our proposed model and served as input for our downstream model. The estimated targets, representing the percentages of sand and silt, are formatted as a $64 \times 64 \times 2$ matrix. The percentage of clay is calculated by subtracting the sum of the sand and silt percentages from the full percentage.

As illustrated in table 5.1, we compared the effectiveness of our proposed model in training VGG13 for soil texture estimation at a 30m resolution. The baseline model (MAE with ViT) does not incorporate any of our core refinements. The results indicate that the latent features generated by our proposed model improve accuracy by over 15% compared to the baseline model. This evaluation demonstrates that our model's output outputs effectively preserve geospatial characteristics, therefore enhancing the performance of the downstream model.

Regarding training efficiency, the use of our proposed model significantly accelerated the training process of the downstream model. Figure 5.2 presents the learning curve of the downstream model with and without our proposed model. While the base model begins to converge around the 50th epoch, the variations incorporating TerraMAE exhibit a more rapid convergence.

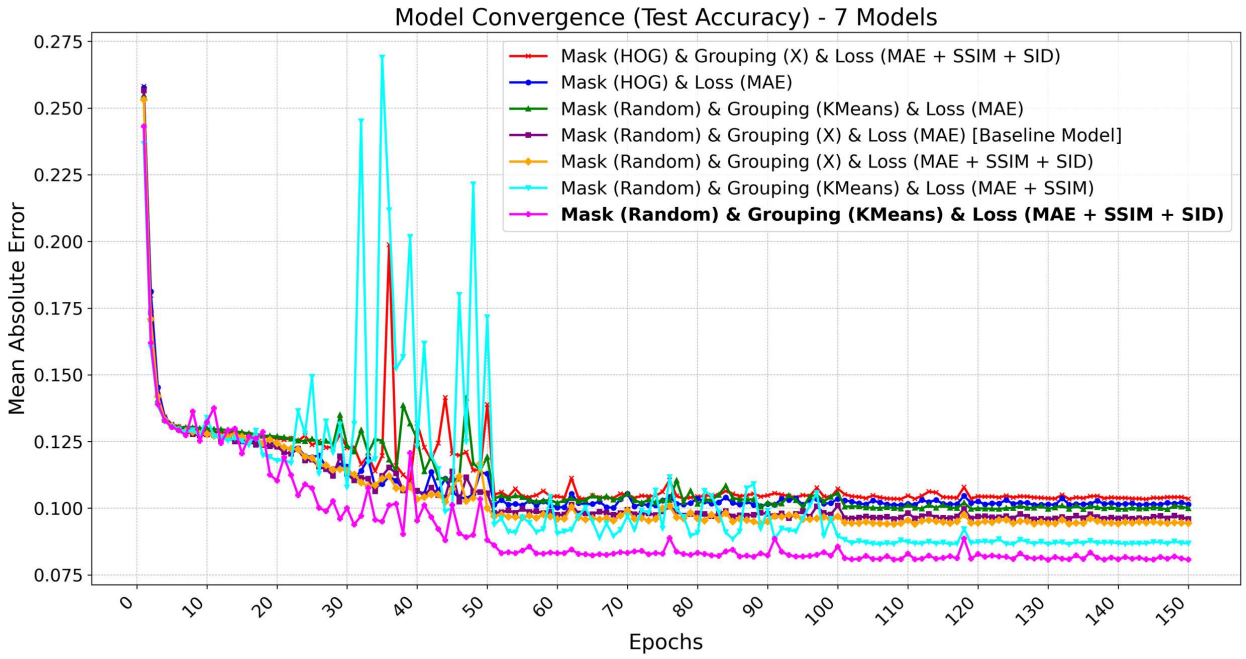

Figure 5.2: Test Errors of Downstream Models with Latent Features as Input from Different Versions of Our Proposed Framework

Chapter 6

Conclusion

Here, we presented a novel pre-training framework, which falls broadly within the category of Masked Encoders. Our proposed model is specifically designed for HSI characterized by a high number of channels (200+) and medium spatial resolution (30 meters). Our innovative masking strategy, which is based on the spatial variability of patches, along with a multi-part loss function, ensures that the training process is regulated by the geospatial and spectral characteristics of the areas scanned by hyperspectral satellites. Also, the relationships among channels are captured through spectral grouping, which takes into account spatial resolution and wavelength characteristics. This approach ensures that the relationships among channels, both within and across groups, are well-preserved. The effective capture of geospatial and spectral characteristics enhances the performance of the MAE framework and improves the outcomes for downstream tasks.

Bibliography

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020.

[2] Zhihong Chen, Yuhao Du, Jinpeng Hu, Yang Liu, Guanbin Li, Xiang Wan, and Tsung-Hui Chang. Multi-modal masked autoencoders for medical vision-and-language pre-training. In Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, and Shuo Li, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pages 679–689, Cham, 2022. Springer Nature Switzerland.

[3] ece.uwaterloo.ca. https://ece.uwaterloo.ca/~z70wang/research/ssim/. [Accessed 13-10-2024].

[4] M.S. Moran, Y. Inoue, and E.M. Barnes. Opportunities and limitations for image-based remote sensing in precision crop management. *Remote Sensing of Environment*, 61(3):319–346, 1997.

[5] Marion Pause, Christian Schweitzer, Michael Rosenthal, Vanessa Keuck, Jan Bumberger, Peter Dietrich, Marco Heurich, András Jung, and Angela Lausch. In situ/remote sensing integration to assess forest health—a review. *Remote Sensing*, 8(6), 2016.

[6] Alan S. Belward and Jon O. Skøien. Who launched what, when and why; trends in global land-cover observation capacity from civilian earth observation satellites. *ISPRS Journal of Photogrammetry and Remote Sensing*, 103:115–128, 2015. Global Land Cover Mapping and Monitoring.

[7] Michael A. Wulder, Nicholas C. Coops, David P. Roy, Joanne C. White, and Txomin Hermosilla. Land cover 2.0. *International Journal of Remote Sensing*, 39(12):4254–4284, March 2018.

[8] Alexander F.H. Goetz. Three decades of hyperspectral remote sensing of the earth: A personal view. *Remote Sensing of Environment*, 113:S5–S16, 2009. Imaging Spectroscopy Special Issue.

[9] Thomas Hiriart and Joseph H. Saleh. Observations on the evolution of satellite launch volume and cyclicality in the space industry. *Space Policy*, 26(1):53–60, 2010.

[10] Imaging the Past | Landsat Science — landsat.gsfc.nasa.gov. https://landsat.gsfc.nasa.gov/article/imaging-the-past/. [Accessed 13-10-2024].

[11] Landsat-1 to 3 - eoPortal — eoportal.org. https://www.eoportal.org/satellite-missions/landsat-1-3. [Accessed 13-10-2024].

[12] Erik Kulu. Planet Labs - Satellite Constellation - NewSpace Index — newspace.im. https://www.newspace.im/constellations/planet-labs. [Accessed 13-10-2024].

[13] https://www.facebook.com/48576411181. NASA's Laser Link Boasts Record-Breaking 200 Gbps Speed — spectrum.ieee.org. https://spectrum.ieee.org/laser-communications. [Accessed 13-10-2024].

[14] B.L. Markham, T. Arvidson, J.A. Barsi, M. Choate, E. Kaita, R. Levy, M. Lubke, and J.G. Masek. 1.03 - landsat program. In Shunlin Liang, editor, *Comprehensive Remote Sensing*, pages 27–90. Elsevier, Oxford, 2018.

[15] Shen-En Qian. *Overview of Hyperspectral Imaging Remote Sensing from Satellites*, chapter 2, pages 41–66. John Wiley & Sons, Ltd, 2022.

[16] Nathan Hagen and Michael W. Kudenov. Review of snapshot spectral imaging technologies. *Optical Engineering*, 52(9):090901–090901, September 2013.

[17] Mohammed Abdulmajeed Moharram and Divya Meena Sundaram. Land use and land cover classification with hyperspectral data: A comprehensive review of methods, challenges and future directions. *Neurocomputing*, 536:90–113, 2023.

[18] Billy G. Ram, Peter Oduor, C. Igathinathane, Kirk Howatt, and Xin Sun. A systematic review of hyperspectral imaging in precision agriculture: Analysis of its current state and future prospects. *Computers and Electronics in Agriculture*, 222:109037, 2024.

[19] Amanda Silveira Reis, Marlon Rodrigues, Glaucio Leboso Alemparte Abrantes dos Santos, Karym Mayara de Oliveira, Renato Herrig Furlanetto, Luís Guilherme Teixeira Crusiol, Everson Cezar, and Marcos Rafael Nanni. Detection of soil organic matter using hyperspectral imaging sensor combined with multivariate regression modeling procedures. *Remote Sensing Applications: Society and Environment*, 22:100492, 2021.

[20] Tobias Storch, Hans-Peter Honold, Sabine Chabrillat, Martin Habermeyer, Paul Tucker, Maximilian Brell, Andreas Ohndorf, Katrin Wirth, Matthias Betz, Michael Kuchler, Helmut Mühle, Emiliano Carmona, Simon Baur, Martin Mücke, Sebastian Löw, Daniel Schulze, Steffen Zimmermann, Christoph Lenzen, Sebastian Wiesner, Saika Aida, Ralph Kahle, Peter Willburger, Sebastian Hartung, Daniele Dietrich, Nicolae Plesia, Mirco Tegler, Katharina Schork, Kevin Alonso, David Marshall, Birgit Gerasch, Peter Schwind, Miguel Pato, Mathias Schneider, Raquel de los Reyes, Maximilian Langheinrich, Julian Wenzel, Martin Bachmann, Stefanie Holzwarth, Nicole Pinnel, Luis Guanter, Karl Segl, Daniel Scheffler, Saskia Foerster, Niklas Bohn, Astrid Bracher, Mariana A. Soppa, Ferran Gascon, Rob Green, Raymond Kokaly, Jose Moreno, Cindy Ong, Manuela Sornig, Ricarda Wernitz, Klaus Bagschik, Detlef Reintsema, Laura La Porta, Anke Schickling, and Sebastian Fischer. The enmap imaging spectroscopy mission towards operations. *Remote Sensing of Environment*, 294:113632, 2023.

[21] S. Cogliati, F. Sarti, L. Chiarantini, M. Cosi, R. Lorusso, E. Lopinto, F. Miglietta, L. Genesio, L. Guanter, A. Damm, S. Pérez-López, D. Scheffler, G. Tagliabue, C. Panigada,

U. Rascher, T.P.F. Dowling, C. Giardino, and R. Colombo. The prisma imaging spectroscopy mission: overview and first performance analysis. *Remote Sensing of Environment*, 262:112499, 2021.

[22] Planet Labs PBC. Tanager-1 is ready for launch: Planet's first hyperspectral satellite. https://www.planet.com/pulse/tanager-1-is-ready-for-launch-planets-first-hyperspectral-satellite/, 2024. Accessed: 2024-07-15.

[23] Jasmin Praful Bharadiya, Nikolaos Tzenios Tzenios, and Manjunath Reddy. Predicting crop yield using deep learning and remote sensing. *Journal of Engineering Research and Reports*, 24(12):29–44, Apr. 2023.

[24] Vijendra Kumar, Hazi Md. Azamathulla, Kul Vaibhav Sharma, Darshan J. Mehta, and Kiran Tota Maharaj. The state of the art in deep learning applications, challenges, and future prospects: A comprehensive review of flood forecasting and management. *Sustainability*, 15(13), 2023.

[25] Esther Rolf, Jonathan Proctor, Tamma Carleton, Ian Bolliger, Vaishaal Shankar, Miyabi Ishihara, Benjamin Recht, and Solomon Hsiang. A generalizable and accessible approach to machine learning with global satellite imagery. *Nature Communications*, 12(1), July 2021.

[26] S. Salcedo-Sanz, P. Ghamisi, M. Piles, M. Werner, L. Cuadra, A. Moreno-Martínez, E. Izquierdo-Verdiguier, J. Muñoz-Marí, Amirhosein Mosavi, and G. Camps-Valls. Machine learning information fusion in earth observation: A comprehensive review of methods, applications and data sources. *Information Fusion*, 63:256–272, 2020.

[27] Giorgos Mountrakis, Jungho Im, and Caesar Ogole. Support vector machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 66(3):247–259, 2011.

[28] M.A. Friedl and C.E. Brodley. Decision tree classification of land cover from remotely sensed data. *Remote Sensing of Environment*, 61(3):399–409, 1997.

[29] Gherardo Chirici, Matteo Mura, Daniel McInerney, Nicolas Py, Erkki O. Tomppo, Lars T. Waser, Davide Travaglini, and Ronald E. McRoberts. A meta-analysis and review of the literature on the k-nearest neighbors technique for forestry applications that use remotely sensed data. *Remote Sensing of Environment*, 176:282–294, 2016.

[30] Zhenhua Lv, Yingjie Hu, Haidong Zhong, Jianping Wu, Bo Li, and Hui Zhao. Parallel k-means clustering of remote sensing images based on mapreduce. In Fu Lee Wang, Zhiguo Gong, Xiangfeng Luo, and Jingsheng Lei, editors, *Web Information Systems and Mining*, pages 162–170, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[31] G.F. Byrne, P.F. Crapper, and K.K. Mayo. Monitoring land-cover change by principal component analysis of multitemporal landsat data. *Remote Sensing of Environment*, 10(3):175–184, 1980.

[32] Wei Han, Xiaohan Zhang, Yi Wang, Lizhe Wang, Xiaohui Huang, Jun Li, Sheng Wang, Weitao Chen, Xianju Li, Ruyi Feng, Runyu Fan, Xinyu Zhang, and Yuewei Wang. A survey of machine learning and deep learning in remote sensing of geological environment: Challenges, advances, and opportunities. *ISPRS Journal of Photogrammetry and Remote Sensing*, 202:87–113, 2023.

[33] Ying-Nong Chen, Kuo-Chin Fan, Yang-Lang Chang, and Toshifumi Moriyama. Special issue review: Artificial intelligence and machine learning applications in remote sensing. *Remote Sensing*, 15(3), 2023.

[34] Teja Kattenborn, Jens Leitloff, Felix Schiefer, and Stefan Hinz. Review on convolutional neural networks (cnn) in vegetation remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing*, 173:24–49, 2021.

[35] Emile Ndikumana, Dinh Ho Tong Minh, Nicolas Baghdadi, Dominique Courault, and Laure Hossard. Deep recurrent neural network for agricultural classification using multitemporal sar sentinel-1 for camargue, france. *Remote Sensing*, 10(8), 2018.

[36] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9(1), 2021.

[37] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):857–876, 2023.

[38] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[39] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. *University of Toronto*, 2009.

[40] Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David B. Lobell, and Stefano Ermon. Satmae: pre-training transformers for temporal and multi-spectral satellite imagery. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2024. Curran Associates Inc.

[41] Xuyang Li, Danfeng Hong, and Jocelyn Chanussot. S2mae: A spatial-spectral pretraining foundation model for spectral remote sensing data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24088–24097, June 2024.

[42] Colorado J Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. In

Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 4088–4099, October 2023.

[43] Mohamed Fadhlallah Guerri, Cosimo Distante, Paolo Spagnolo, Fares Bougourzi, and Abdelmalik Taleb-Ahmed. Deep learning techniques for hyperspectral image analysis in agriculture: A review. *ISPRS Open Journal of Photogrammetry and Remote Sensing*, 12:100062, 2024.

[44] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15979–15988, 2022.

[45] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zihang Jiang, Francis E. H. Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 538–547, 2021.

[46] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[47] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[48] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[49] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern*

Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, pages 2261–2269. IEEE Computer Society, 2017.

[50] Yogesh L. Simmhan, Sangmi Lee Pallickara, Nithya N. Vijayakumar, and Beth Plale. Data management in dynamic environment-driven computational science. In Patrick W. Gaffney and James C. T. Pool, editors, *Grid-Based Problem Solving Environments*, pages 317–333, Boston, MA, 2007. Springer US.

[51] Sangmi Lee Pallickara, Marlon Pierce, Qunfeng Dong, and ChinHua Kong. Enabling large scale scientific computations for expressed sequence tag sequencing over grid and cloud computing clusters. In *PPAM 2009 EIGHTH INTERNATIONAL CONFERENCE ON PARALLEL PROCESSING AND APPLIED MATHEMATICS Wroclaw, Poland*, 2009.

[52] Matthew Malensek, Sangmi Pallickara, and Shrideep Pallickara. Fast, ad hoc query evaluations over multidimensional geospatial datasets. *IEEE Transactions on Cloud Computing*, 5(1):28–42, 2017.

[53] Thilina Buddhika, Matthew Malensek, Sangmi Lee Pallickara, and Shrideep Pallickara. Synopsis: A distributed sketch over voluminous spatiotemporal observational streams. *IEEE Transactions on Knowledge and Data Engineering*, 29(11):2552–2566, 2017.

[54] Geoffrey Fox, Shrideep Pallickara, and Xi Rao. Towards enabling peer-to-peer grids: Research articles. *Concurr. Comput.: Pract. Exper.*, 17(7–8):1109–1131, June 2005.

[55] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

[56] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[57] Vladan Stojnic and Vladimir Risojevic. Self-supervised learning of remote sensing scene representations using contrastive multiview coding. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2021, virtual, June 19-25, 2021*, pages 1182–1191. Computer Vision Foundation / IEEE, 2021.

[58] Ivica Dimitrovski, Ivan Kitanovski, Nikola Simidjievski, and Dragi Kocev. In-domain self-supervised learning improves remote sensing image scene classification. *IEEE Geoscience and Remote Sensing Letters*, 21:1–5, 2024.

[59] Yi Wang, Nassim Ait Ali Braham, Zhitong Xiong, Chenying Liu, Conrad M. Albrecht, and Xiao Xiang Zhu. Ssl4eo-s12: A large-scale multimodal, multitemporal dataset for self-supervised learning in earth observation [software and data sets]. *IEEE Geoscience and Remote Sensing Magazine*, 11(3):98–106, 2023.

[60] Yuan Gao, Xiaojuan Sun, and Chao Liu. A general self-supervised framework for remote sensing image classification. *Remote Sensing*, 14(19), 2022.

[61] Paahuni Khandelwal, Samuel Armstrong, Abdul Matin, Shrideep Pallickara, and Sangmi Lee Pallickara. Cloudnet: A deep learning approach for mitigating occlusions in landsat-8 imagery using data coalescence. In *2022 IEEE 18th International Conference on e-Science (e-Science)*, pages 117–127, 2022.

[62] Saptashwa Mitra, Paahuni Khandelwal, Shrideep Pallickara, and Sangmi Lee Pallickara. Stash : Fast hierarchical aggregation queries for effective visual spatiotemporal explorations. In *2019 IEEE International Conference on Cluster Computing (CLUSTER)*, pages 1–11, 2019.

[63] Saptashwa Mitra, Daniel Rammer, Shrideep Pallickara, and Sangmi Lee Pallickara. Glance: A generative approach to interactive visualization of voluminous satellite imagery. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 359–367, 2021.

[64] 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 *ACM SIGSPATIAL '24, Atlanta, GA, USA*, 2024.

[65] Kevin Bruhwiler, Paahuni Khandelwal, Daniel Rammer, Samuel Armstrong, Sangmi Lee Pallickara, and Shrideep Pallickara. Lightweight, embeddings based storage and model construction over satellite data collections. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 246–255, 2020.

[66] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14648–14658, 2022.

[67] Linus Scheibenreif, Joëlle Hanna, Michael Mommert, and Damian Borth. Self-supervised vision transformers for land-cover segmentation and classification. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1421–1430, 2022.

[68] Michail Tarasiou, Erik Chavez, and Stefanos Zafeiriou. Vits for sits: Vision transformers for satellite image time series. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10418–10428, June 2023.

[69] Ryan Rad. Vision transformer for multispectral satellite imagery: Advancing landcover classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 8176–8183, January 2024.

[70] Aggrey Muhebwa, Gabriel Cadamuro, and Jay Taneja. Pixel perfect: Using vision transformers to improve road quality predictions from medium resolution and heterogeneous satellite imagery. *ACM J. Comput. Sustain. Soc.*, 1(1), September 2023.

[71] Yakoub Bazi, Laila Bashmal, Mohamad M. Al Rahhal, Reham Al Dayil, and Naif Al Ajlan. Vision transformers for remote sensing image classification. *Remote Sensing*, 13(3), 2021.

[72] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9630–9640. IEEE, 2021.

[73] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan L. Yuille, and Tao Kong. ibot: Image BERT pre-training with online tokenizer. *CoRR*, abs/2111.07832, 2021.

[74] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: a simple framework for masked image modeling. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9643–9653, 2022.

[75] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEit: BERT pre-training of image transformers. In *International Conference on Learning Representations*, 2022.

[76] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 10078–10093. Curran Associates, Inc., 2022.

[77] Sixiao Zhang, Hongxu Chen, Haoran Yang, Xiangguo Sun, Philip S. Yu, and Guandong Xu. Graph masked autoencoders with transformers, 2022.

[78] Gang Li, Heliang Zheng, Daqing Liu, Chaoyue Wang, Bing Su, and Changwen Zheng. Semmae: Semantic-guided masking for learning masked autoencoders. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 14290–14302. Curran Associates, Inc., 2022.

[79] Haijian Chen, Wendong Zhang, Yunbo Wang, and Xiaokang Yang. Improving masked autoencoders by learning where to mask. In *PRCV 2023*, 2023.

[80] He Zhu, Yang Chen, Guyue Hu, and Shan Yu. Information-density masking strategy for masked image modeling. In *2023 IEEE ICME*, 2023.

[81] Duy Kien Nguyen, Yanghao Li, Vaibhav Aggarwal, Martin R. Oswald, Alexander Kirillov, Cees G. M. Snoek, and Xinlei Chen. R-MAE: Regions meet masked autoencoders. In *The Twelfth International Conference on Learning Representations*, 2024.

[82] Wele Gedara Chaminda Bandara, Naman Patel, Ali Gholami, Mehdi Nikkhah, Motilal Agrawal, and Vishal M. Patel. Adamae: Adaptive masking for efficient spatiotemporal learning with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14507–14517, June 2023.

[83] Xianghai Cao, Haifeng Lin, Shuaixu Guo, Tao Xiong, and Licheng Jiao. Transformer-based masked autoencoder with contrastive loss for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–12, 2023.

[84] Lingxuan Zhu, Jiaji Wu, Wang Biao, Yi Liao, and Dandan Gu. Spectralmae: Spectral masked autoencoder for hyperspectral remote sensing image reconstruction. *Sensors*, 23(7), 2023.

[85] Linus Scheibenreif, Michael Mommert, and Damian Borth. Masked vision transformers for hyperspectral image classification. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2166–2176, 2023.

[86] Junyan Lin, Feng Gao, Xiaochen Shi, Junyu Dong, and Qian Du. Ss-mae: Spatial–spectral masked autoencoder for multisource remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–14, 2023.

[87] Mubashir Noman, Muzammal Naseer, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, and Fahad Shahbaz Khan. Rethinking transformers pre-training for multi-spectral satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27811–27819, June 2024.

[88] Yi Wang, Hugo Hernández Hernández, Conrad M Albrecht, and Xiao Xiang Zhu. Feature guided masked autoencoder for self-supervised learning in remote sensing. *arXiv preprint arXiv:2310.18653*, 2023.

[89] Qing Shi, Xuelong Tang, Rong Liu, and Liangpei Zhang. Hyperspectral image denoising using a 3d attention network. *IEEE Transactions on Geoscience and Remote Sensing*, 59(10):8297–8309, 2021.

[90] Wenjia He, Liangpei Zhang, and Huanfeng Shen. Convolutional stacked autoencoder for hyperspectral image restoration. *IEEE Transactions on Image Processing*, 30:3882–3896, 2021.

[91] Yifan Zhou, Xiaohui Ma, and Zhenlong Zhang. Stacked autoencoder for hyperspectral image dimensionality reduction. *Journal of Remote Sensing*, 44(2):345–359, 2020.

[92] Pengzhi Li, Yan Pei, and Jianqiang Li. A comprehensive survey on design and application of autoencoder in deep learning. *Applied Soft Computing*, 138:110176, 2023.

[93] Garima Jaiswal, Ritu Rani, Harshita Mangotra, and Arun Sharma. Integration of hyperspectral imaging and autoencoders: Benefits, applications, hyperparameter tunning and challenges. *Computer Science Review*, 50:100584, 2023.

[94] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, June 2022.

[95] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR'05*, 2005.

[96] S Chabrillat, Luis Guanter, Hermann Kaufmann, S Förster, Alison Beamish, Arlena Brosinsky, Hendrik Wulf, Saeid Asadzadeh, M Bochow, Niklas Bohn, et al. Enmap science plan. *EnMAP*, 2022.

[97] Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

[98] Zhou Wang and Alan C. Bovik. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE Signal Processing Magazine*, 26(1):98–117, 2009.

[99] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[100] Chein-I Chang. Spectral information divergence for hyperspectral image analysis. In *IEEE 1999 International Geoscience and Remote Sensing Symposium. IGARSS'99 (Cat. No.99CH36293)*, volume 1, pages 509–511 vol.1, 1999.

[101] Zijun Zhang. Improved adam optimizer for deep neural networks. In *2018 IEEE/ACM 26th international symposium on quality of service (IWQoS)*, pages 1–2. Ieee, 2018.

[102] Umme Sara, Morium Akter, and Mohammad Shorif Uddin. Image quality assessment through fsim, ssim, mse and psnr—a comparative study. *Journal of Computer and Communications*, 7(3):8–18, 2019.

[103] Quan Huynh-Thu and Mohammed Ghanbari. The accuracy of psnr in predicting video quality for different video scenes and frame rates. *Telecommunication Systems*, 49:35–48, 2012.

[104] Collin Homer, Chengquan Huang, Limin Yang, Bruce Wylie, and Michael Coan. Development of a 2001 national land-cover database for the united states. *Photogrammetric Engineering & Remote Sensing*, 70(7):829–840, 2004.

[105] Nathaniel W. Chaney, Eric F. Wood, Alexander B. McBratney, Jonathan W. Hempel, Travis W. Nauman, Colby W. Brungard, and Nathan P. Odgers. Polaris: A 30-meter probabilistic soil series map of the contiguous united states. *Geoderma*, 274:54–67, 2016.

[106] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.