Masked Vision Transformers for Hyperspectral Image Classification

Linus Scheibenreif Michael Mommert Damian Borth
AIML Lab, School of Computer Science, University of St. Gallen
{firstname}.{lastname}@unisg.ch

Abstract

Transformer architectures have become state-of-the-art models in computer vision and natural language processing. To a significant degree, their success can be attributed to self-supervised pre-training on large scale unlabeled datasets. This work investigates the use of self-supervised masked image reconstruction to advance transformer models for hyperspectral remote sensing imagery. To facilitate self-supervised pre-training, we build a large dataset of unlabeled hyperspectral observations from the EnMAP satellite and systematically investigate modifications of the vision transformer architecture to optimally leverage the characteristics of hyperspectral data. We find significant improvements in accuracy on different land cover classification tasks over both standard vision and sequence transformers using (i) blockwise patch embeddings, (ii) spatial-spectral self-attention, (iii) spectral positional embeddings and (iv) masked self-supervised pre-training[1]*. The resulting model outperforms standard transformer architectures by* $+5\%$ *accuracy on a labeled subset of our EnMAP data and by* $+15\%$ *on Houston2018 hyperspectral dataset, making it competitive with a strong 3D convolutional neural network baseline. In an ablation study on label-efficiency based on the Houston2018 dataset, self-supervised pre-training significantly improves transformer accuracy when little labeled training data is available. The self-supervised model outperforms randomly initialized transformers and the 3D convolutional neural network by* $+7$-8% *when only* 0.1-10% *of the training labels are available.*

1. Introduction

Hyperspectral remote sensing provides measurements of the Earth's surface with high spectral resolution. This enables applications like the detection of specific material categories or agricultural parameters which often depend on fine-grained spectral reflectance patterns [15, 21]. In recent years, the availability of hyperspectral remote sensing data has strongly improved and the launch of the German hyperspectral Environmental Mapping and Analysis Program (EnMAP) mission in April 2022 made global hyperspectral data of high spectral and temporal resolution publicly available on a large scale [16]. In contrast to this trend, deep learning approaches for the analysis of hyperspectral remote sensing data are overwhelmingly developed on well-established benchmark datasets that are very small in comparison to commonly used datasets in other computer vision domains [10, 30]. To a large extent, this is due to the high acquisition cost of hyperspectral data itself and the corresponding labels for individual spectral sequences. This strongly limits the size of available labeled datasets and the development of deep learning approaches in the hyperspectral domain. In this work we aim to improve vision transformer architectures for the specific characteristics of hyperspectral data and to leverage the growing amount of freely available unlabeled hyperspectral remote sensing imagery for self-supervised pre-training of these models. We illustrate how to increase the performance while decreasing the amount of required labeled data for hyperspectral classification tasks.

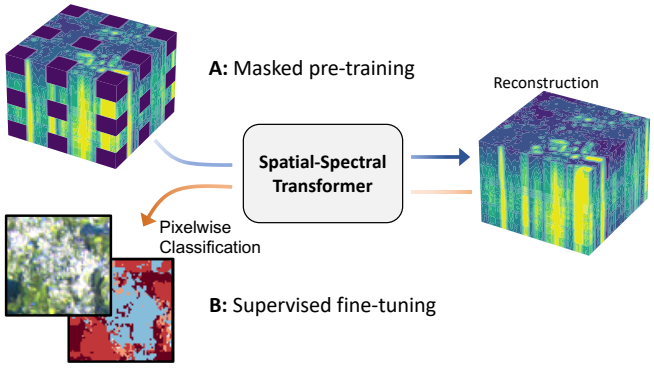

Figure 1. **A:** We propose the use of masked image modeling to pre-train spatial-spectral transformer networks on a large dataset of unlabeled hyperspectral EnMAP data. **B:** The pre-trained model can then be fine-tuned on small labeled datasets for supervised downstream tasks like land cover classification.

[1] Code available at github.com/HSG-AIML/MaskedSST

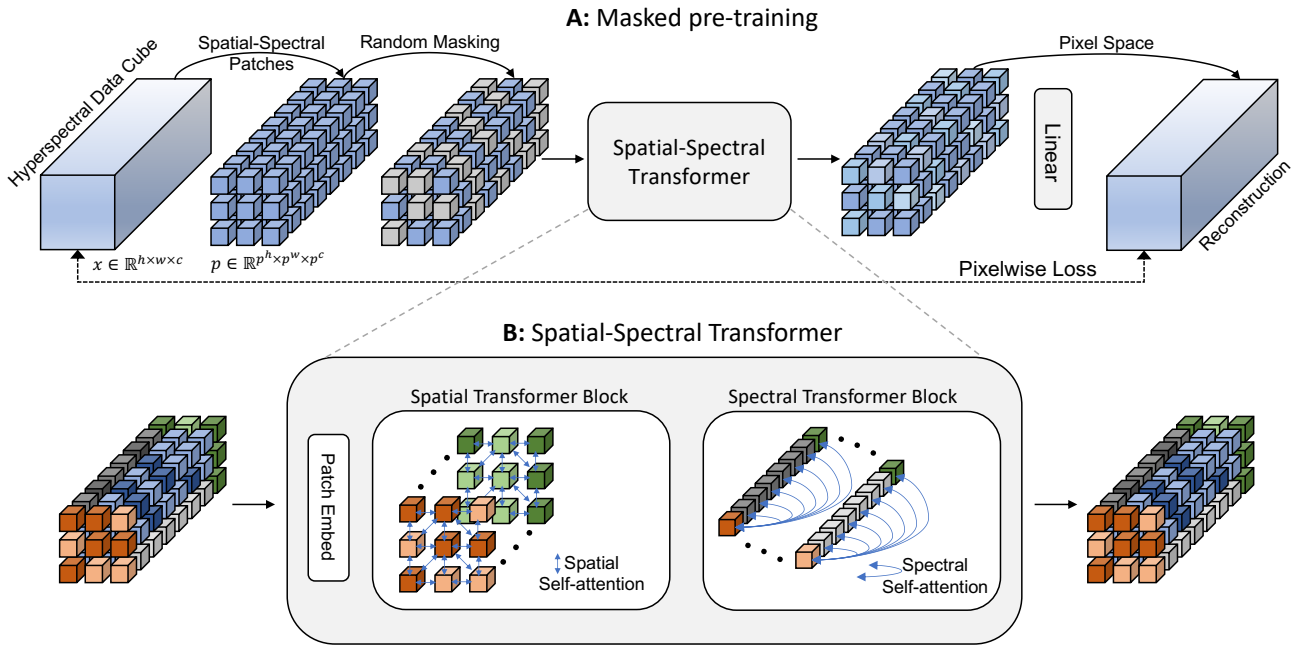

Figure 2. Overview of our proposed transformer model for hyperspectral data with spatial-spectral factorization within the masked self-supervised pre-training framework. **A:** The hyperspectral data cube is first divided into spatial-spectral patches $\mathbf{p} \in \mathbb{R}^{p^h \times p^w \times p^c}$. The patches are randomly masked, embedded and processed by the transformer, which sequentially applies self-attention spatially and spectrally between all embeddings. A linear layer maps representations of the masked patches back to pixel space to compute the reconstruction error. **B:** Our spectral-spatial transformer consists of a patch embedding layer and transformer blocks that apply self-attention among tokens with the same spectral or spatial index. The colors indicate token locations in the hyperspectral cube.

The contributions of this work can be summarized as follows:

- We collect a large scale unlabeled dataset of EnMAP observations over Europe and create a labeled dataset of Mexico City by matching EnMAP observations with land cover labels. Based on these datasets, our work provides a large scale evaluation of transformer models for hyperspectral data.
- We investigate different positional and spectral encoding schemes and show that block-wise embedding significantly improves the performance of transformers on hyperspectral data.
- To facilitate spatial-spectral learning with transformers, we utilize a spatial-spectral factorization scheme which greatly reduces the computational burden of the self-attention operation on high-dimensional hyperspectral data.
- We show that a self-supervised masked image modeling task for hyperspectral data improves model performance on downstream tasks, and can significantly improve label efficiency for transformer models.

2. Related Work

2.1. Hyperspectral Deep Learning

The high dimensionality and spectral correlation of hyperspectral data present unique challenges for machine learning methods. Accordingly, many machine learning techniques have been developed for common hyperspectral tasks such as dimensionality reduction [55], data fusion [51], unmixing [54] or classification [2]. In particular, deep learning approaches like fully connected [14], convolutional [26] (CNN), and recurrent neural networks [28] have been successfully applied on hyperspectral imaging data (see [2] for an overview). Hybrid transformer-CNN methods combine convolutional feature extractors with transformer networks (*e.g.*, [20,38,43]) to leverage the spatial inductive bias of CNNs in a transformer framework. Following the general trend in the deep learning field, pure transformer networks have also recently been developed for hyperspectral remote sensing imagery [22,31].

2.2. Vision Transformers

Transformer models are state-of-the art in natural language processing [11] (NLP) where their attention mechanism [42] models pairwise interactions between tokens, and allows them to capture long-range interactions. The trans-

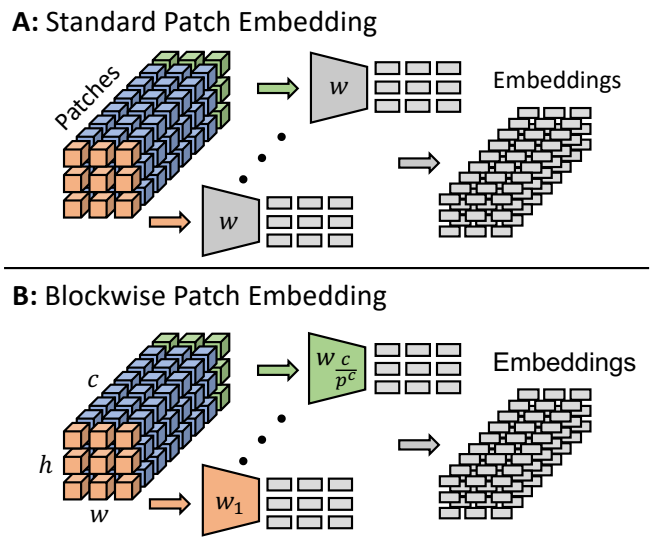

Figure 3. **A**: Standard patch-embedding approach for vision transformers. Patches are flattened and embedded using the same linear transform w. **B**: Blockwise spectral embedding for spatial-spectral patches. Each spectral interval is embedded with a specific linear transform w_b to account for the characteristics of the corresponding wavelength interval.

former approach has since also been successfully adapted for computer vision applications, where pre-trained transformers are now among the strongest general purpose backbones [12]. The standard vision transformer [12] (ViT) method first divides input images $\mathbf{x} \in \mathbb{R}^{h \times w \times c}$ into patches $\mathbf{p} \in \mathbb{R}^{n \times (p^h \times p^w \times c)}$ of patch size $p^h \cdot p^w$. This set of $n = (\frac{h}{p^h}) \cdot (\frac{w}{p^w})$ non-overlapping patches is linearly embedded to the transformer dimension d and summed with positional encodings for every patch. The resulting embeddings z are processed by the transformer encoder consisting of l layers of alternating multi-head self-attention [42] (MSA) and feed-forward (FF) blocks, both with layer normalization [3] (LN):

$$\begin{aligned} \mathbf{y}^l &= \mathrm{MSA}(\mathrm{LN}(\mathbf{z}^l)) + \mathbf{z}^l \\ \mathbf{z}^{l+1} &= \mathrm{FF}(\mathrm{LN}(\mathbf{y}^l)) + \mathbf{y}^l \end{aligned} \tag{1}$$

The major bottleneck in the application of (vision) transformers on high dimensional data is the quadratic complexity of the attention operation in the number of input tokens. A number of works try to improve the efficiency of transformers for large numbers of tokens by reducing the complexity of self-attention [7, 8], or by applying self-attention selectively rather than pairwise between all tokens [5, 23]. Most relevant for this work are transformer architectures for video data where different approaches to divide the self-attention operation along the temporal and spatial dimensions have been proposed [1, 5].

2.3. Self-supervised Learning

The goal of self-supervised learning (SSL) is to learn rich representations from unlabeled data. To that end, an artificial supervision signal is constructed from information that is inherent to the data sample. Models can then be trained to solve such a 'pretext-task' before the learned representations are transferred to different downstream tasks of interest. Common pretext-tasks include the prediction of relative rotation [13], solving of jigsaw puzzles [29] or image colorization [53]. More recently, contrastive learning has emerged as a powerful pre-training strategy [6,45]. This approach aims to solve an instance-wise classification problem between data samples with noise contrastive estimation [17]. The objective is to distinguish positive and negative pairs of data points, where the pairwise relationships are derived from inherent characteristics of the data samples rather than classical labels. The network thus learns to map positive pairs close to each other and far apart from negative samples in the representation space. Contrast can be defined on the image [6], patch [46], or pixel [47] level to control the granularity of resulting representations.

Masked Image Modeling Self-supervised learning through the prediction of masked data components is widely used in NLP and a central contributor to the success of transformer networks in this domain. The central idea is to replace a fraction of input tokens with a special mask token that has to be predicted by the transformer [11]. Following the success of such approaches with natural language, masked modeling approaches are now also used in the vision domain. These approaches closely follow the NLP approach by predicting masked tokens from adjacent visible tokens [4, 56], which requires a suitable tokenizer. Recent work has shown that in the vision domain token prediction can be substituted by directly regressing the values of masked pixel. This pre-training approach results in strong visual representations when combined with autoencoder networks [19] or by estimating pixel values from latent representations with a simple linear layer [48].

SSL in Remote Sensing Remote sensing offers large amounts of unlabeled data, which has been leveraged in a number of self-supervised learning strategies (see [44] for a review). Early approaches utilize hand-crafted pretext-tasks like inpainting and location prediction [40]. A number of methods have tailored the contrastive learning principle to the characteristics of remote sensing data by utilizing temporal information from consecutive overpasses [27, 52], multi-modal data from different sensors [33–35], or multi-spectral observations [37, 39] to define positive sample pairs. Masked image modeling approaches for remote sensing data utilize masked autoencoding [19] with multi-

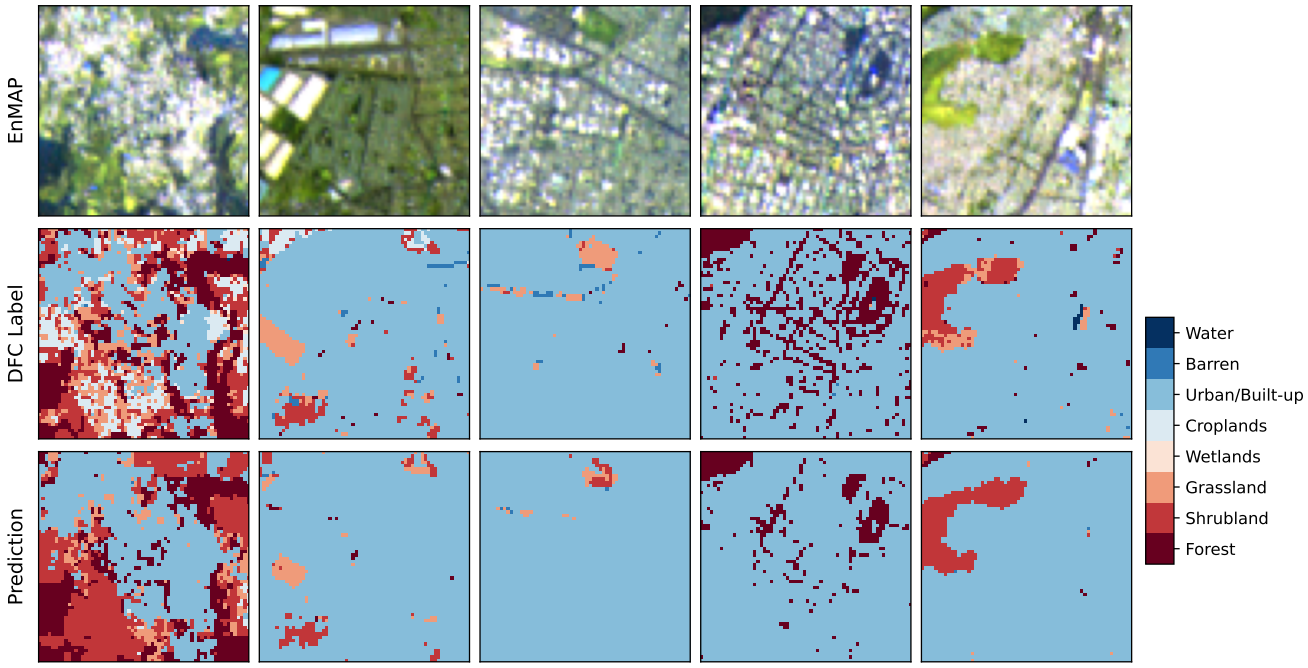

Figure 4. **Top:** Tiles from EnMAP L2 scenes over Mexico City. **Center:** Corresponding DFC2020 land cover labels. **Bottom:** Predicted land cover classes from the masked spatial-spectral transformer model.

spectral and multi-temporal data [9], or extend it to imagery of varying ground sampling distance [32]. In the hyperspectral domain, masked sequence modeling has been used to model the spectral signal [18], and within the masked autoencoding framework [24].

3. Method

This section introduces the proposed transformer model for hyperspectral data (3.1), the spatial-spectral patch embedding strategy (3.1.1) and how hyperspectral data is efficiently processed by factorizing self-attention spatially and spectrally (3.1.2). Finally, we present the masked pretraining scheme (3.2).

3.1. Transformer Architecture

This work adapts the vision transformer [12] architecture to hyperspectral imagery. Starting from a baseline transformer model, we successively add model components and adjust design choices to improve efficiency and performance on hyperspectral data. Our baseline **spectral model** processes the spectral sequence of individual pixel with an transformer encoder. Each pixel is divided into patches along the spectral dimension, resulting in $n = \frac{c}{p^c}$ blocks of size p^c, and then embedded with a shared linear transform. Learnable positional embeddings are added to the embedding sequence. As a spatial transformer baseline, we apply the original ViT [12] architecture on the RGB bands of hyperspectral data (**ViT-RGB**).

3.1.1 Spatial-Spectral Patch Embeddings

Spatial-Spectral Patches To incorporate the spatial context for the spectral sequence of each pixel, our **spatial-spectral model** divides the input image $\mathbf{x} \in \mathbb{R}^{h \times w \times c}$ into $n = (\frac{h}{p^h}) \cdot (\frac{w}{p^w}) \cdot (\frac{c}{p^c})$ patches of size $p^h \cdot p^w \cdot p^c$. This approach mirrors the spatial patching of the original ViT, but does not aggregate all spectral bands per location, thus retaining the hyperspectral 3D cube. This makes it possible to model both spectral and spatial relationships within the image using attention. Spatial-spectral patching increases the number of tokens by a factor of $\frac{c}{p^c}$ compared to spatial ViT patching and by $\frac{h}{p^h} \cdot \frac{w}{p^w}$ compared to the spectral transformer approach. Since the computational cost of self-attention is quadratic in the number of tokens, modeling all spatial-spectral relationships is practically infeasible for anything but very large spatial and spectral patch sizes $p^{\{h,w,c\}}$. We address this limitation in Section 3.1.2.

Blockwise Spectral Embedding Vision transformers create embeddings from patches through a learned linear transform that is shared between all patches (see Fig. 3 A). Unlike the spatial patches of ViT, which always represent the RGB intervals of the electromagnetic spectrum, our spatial-spectral patches represent multiple different spectral wavelength intervals for every spatial patch. To account for this diversity in the spectral signal, we propose a **blockwise spectral** embedding scheme that utilizes a separate

linear transform for each of the $\frac{c}{p^c}$ spectral blocks in the patched hyperspectral data (see Fig. 3 B). This approach is most similar to group embeddings which have been used for multi-temporal and multi-spectral remote sensing imagery [9].

Spectral Positional Embedding We investigate the utility of two different positional encoding techniques for the spatial-spectral embeddings: **Learnable positional embeddings** for every spatial-spectral patch that are optimized along with the transformer during model training. Alternatively, **spectral positional embeddings** explicitly encode spatial and spectral positional information of the hyperspectral data separately with fixed sine and cosine functions [9, 42] and the transformer dimensionality d.

$$\begin{aligned} \mathrm{PE}_{pos,2i} &= \sin(pos/10000^{2i/d}) \\ \mathrm{PE}_{pos,2i+1} &= \cos(pos/10000^{2i/d}) \end{aligned} \tag{2}$$

We allot one third of the embedding vector to the spectral positional embedding (*i.e.*, the encoding of the patch's index in the spectral sequence), and the remainder for the spatial embedding of horizontal and vertical position.

3.1.2 Spatial-Spectral Factorization

The spatial-spectral patch embedding strategy yields a large number of tokens for high dimensional hyperspectral data. This is a bottleneck for the attention operation, which has quadratic runtime in the number of tokens. To make training feasible, we resolve this limitation by factorizing the transformer model to sequentially process spatial and spectral relationships within the data (see Fig. 2 B). This approach is similar to separable convolutions in CNNs, where 2D and 1D convolutions are sequentially applied over and across feature maps [36]. This strategy reduces the computational load of self-attention from the squared product of the number of spatial and spectral patches to their squared sum, *i.e.*, $\mathcal{O}((\frac{h}{p^h} \cdot \frac{w}{p^w} \cdot \frac{c}{p^c})^2 \cdot d)$ to $\mathcal{O}((\frac{h}{p^h} \cdot \frac{w}{p^w} + \frac{c}{p^c})^2 \cdot d)$. In practice the number of required operations on our hyperspectral data changes by a factor of $\sim\frac{1}{75}$.

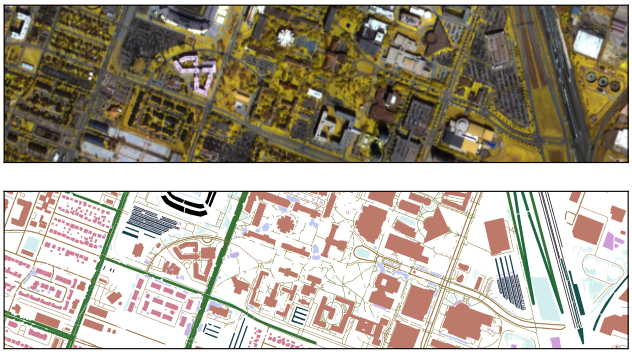

Figure 5. **Top:** RGB representation of the Houston2018 hyperspectral training set (bands 48, 32, 16). **Bottom:** Training labels for Houston2018 (20 classes, unlabeled pixels shown in white).

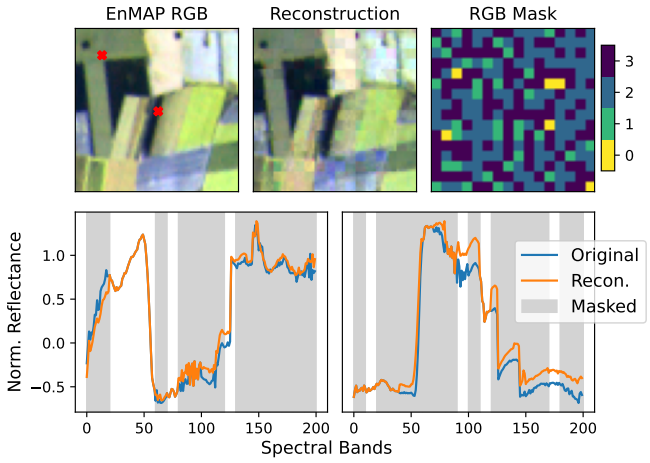

Figure 6. **Top:** Original RGB representation and reconstruction with 70% masking and mask patch size 4 after 200 training epochs. The heatmap indicates how many of the RGB bands were masked for each image patch. **Bottom:** Reconstruction along the spectral axis for the two pixels highlighted in red in the top-left image. Gray sections correspond to masked patches.

3.2. Masked Self-supervised Learning

The transformer models investigated in this work consist of a transformer encoder. During masked pre-training, we add an additional linear layer to map latent token representations from the transformer to pixel values, following the SimMIM method [48]. After patch embedding, a fraction of the embeddings is selected and replaced with a learnable `mask` token. The pre-training objective is to reconstruct the pixel values corresponding to the masked tokens (see Fig. 2). The reconstruction quality is measured by L1 loss, which is only evaluated for masked pixel tokens (see Fig. 6). Unlike similar approaches that utilize encoder-decoder architectures for masked pre-training [19], the small linear reconstruction head in this approach forces the encoder to focus its capacity on modeling the masked tokens, rather than leaving this task to the decoder. Masked sentence models [11] commonly mask 15% of tokens, while image [48] and video [41] models mask around 50% and 90%, respectively. We employ a blockwise-masking strategy (*i.e.*, by masking 4×4 windows of tokens instead of individual tokens) to prevent trivial solutions which are possible due to the high correlation of spectrally adjacent tokens.

4. Data

This work applies transformer models on hyperspectral remote sensing data. To that end, we utilize hyperspec-

	Model Components					Finetuned		Frozen	
Model Name	Spectral	Spatial	BPE	SPE	SSL	Acc. (%)	MAcc. (%)	Acc. (%)	MAcc. (%)
3D-CNN [25]	✓	✓				$\mathbf{83 \pm 0.3}$	$\mathbf{57 \pm 1.0}$	81 ± 0.4	54 ± 1.1
ViT-RGB [12]		✓				69 ± 0.5	20 ± 1.0	68 ± 0.3	16 ± 0.3
Transformer [42]	✓					77 ± 0.2	32 ± 0.4	72 ± 0.1	23 ± 0.2
Spectral T.	✓		✓			80 ± 0.1	38 ± 0.5	71 ± 1.0	27 ± 1.3
Masked Transformer	✓				✓	76 ± 0.2	29 ± 0.4	65 ± 0.0	14 ± 0.0
Masked Spectral T.	✓		✓		✓	81 ± 0.3	40 ± 0.7	78 ± 0.1	31 ± 0.1
SST	✓	✓				79 ± 0.1	38 ± 0.4	74 ± 0.4	32 ± 0.5
SST	✓	✓	✓			81 ± 0.1	40 ± 0.8	75 ± 1.0	27 ± 1.3
SST	✓	✓		✓		78 ± 0.2	33 ± 0.6	73 ± 0.6	24 ± 0.6
SST	✓	✓	✓	✓		82 ± 0.1	44 ± 0.3	76 ± 1.0	35 ± 1.0
Masked SST	✓	✓			✓	77 ± 0.5	31 ± 0.1	65 ± 0.0	14 ± 0.0
Masked SST	✓	✓	✓		✓	82 ± 0.1	42 ± 0.2	77 ± 0.1	29 ± 0.2
Masked SST	✓	✓		✓	✓	78 ± 0.4	32 ± 0.3	65 ± 0.0	14 ± 0.0
Masked SST	✓	✓	✓	✓	✓	82 ± 0.2	45 ± 0.6	79 ± 0.1	40 ± 0.1
MSST-Center	✓	✓	✓	✓	✓	82 ± 0.2	55 ± 0.5	$\mathbf{82 \pm 0.2}$	$\mathbf{55 \pm 0.2}$

Table 1. Hyperspectral classification performance of baselines and different transformer configurations on the EnMAP-DFC dataset. Columns 'Spectral' and 'Spatial' indicate whether the model utilizes spectral/spatial context. Please refer to Section 3.1.1 for details about blockwise patch embedding (BPE) and spectral positional encoding (SPE). SSL indicates that the model has been pre-trained on EnMAP data with the masked reconstruction task. SST refers to the spatial-spectral transformer model. 3D-CNN and MSST-Center provide predictions for the center pixel of a patch, the other methods for all pixels in the patch simultaneously. Finetuned results indicate performance after training all model parameters on labeled data, frozen indicates that only the classification head is trained on labeled data.

tral datasets from the Environmental Mapping and Analysis Program [16] and the IEEE GRSS Data Fusion Challenge (DFC) 2018 [49].

EnMAP The EnMAP satellite carries an imaging spectrometer that scans the Earth's surface with 224 spectral bands in the very-near infrared ($420 - 1000$nm) and shortwave infrared ($900 - 2450$nm) intervals [16]. The sensor has a spatial resolution of 30×30m and a 27-day revisit time. We collect a dataset consisting of 90 cloud-free EnMAP L2 scenes (orthorectified and atmospherically corrected) over Europe in Q4 2022. The EnMAP scenes are divided into non-overlapping 64×64 pixel tiles, and invalid atmospheric bands are removed (resulting in a total of 200 spectral bands). Our dataset consists of 19 792 tiles, for a total of more than 81M hyperspectral pixels.

EnMAP-DFC We create a labeled EnMAP dataset by matching two atmospherically corrected EnMAP L2 scenes over Mexico City with land cover data for the same region that was published for the IEEE GRSS DFC 2020 [50] (see Fig. 4). This dataset consists of 357 64×64 pixel tiles with pixel-wise labels for the classes `Forest`, `Shrubland`, `Grassland`, `Wetland`, `Cropland`,`Urban/Built-up`, `Barren` and `Water`. For our experiments, the data is randomly split into 286 training/validation tiles and 71 tiles for final testing. We note that some label noise is introduced due to the difference in labeling date (2020) and time of the EnMAP overflight in 2022 (see Fig. 4 top and center rows).

Houston2018 As a second labeled hyperspectral dataset, we use the Houston data from the IEEE GRSS DFC in 2018 [49]. This dataset consists of aerial imagery of the city of Houston (see Fig. 5), obtained with a hyperspectral instrument in the $380 - 1050$nm spectral range with 48 bands and 1m spatial resolution. The scene has 1202×4172 hyperspectral pixels, 590 149 of which are labeled into 20 fine-grained classes. We use the official train/test split of the dataset in our experiments (504 712 pixels for training and validation, 85 437 for testing).

Metrics We evaluate model performance for land cover classification on EnMAP-DFC and Houston2018 with accuracy and macro accuracy metrics. The standard accuracy measures the fraction of correctly classified samples over the entire dataset (see Eqn. 3). Macro accuracy provides the average of class-wise accuracies, which can deviate from accuracy on unbalanced datasets (see Eqn. 4). We report the average and standard deviation of each metric, computed over 5 training runs with different random seeds.

$$\text{Acc.} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{3}$$

where TP,TN,FP and FN correspond to true positive, true negative, false positive, and false negative, respectively.

$$\text{MacroAcc.} = \frac{\sum_{c \in \mathbb{C}} \text{Acc.}(\{x_i | y_i = c\}_i)}{|\mathcal{C}|} \quad (4)$$

where x_i is a data sample, y_i the corresponding class label, and $\mathcal{C}$ the set of all classes in the dataset.

5. Experiments & Results

5.1. Baselines

We use three different baseline models in this work. The **ViT-RGB model** directly applies the ViT [12] approach with dimensionality $d = 96$, 4 blocks, and 8 heads in the multi-head self-attention to the RGB bands of hyperspectral data. Accordingly, the data only consists of three spectral bands, which are aggregated during patch embedding. This model yields an accuracy of $69 \pm 0.5\%$ on the EnMAP-DFC dataset (see Table 1) and $19 \pm 1.4\%$ on Houston2018 (see Table 2). Our **spectral model** is a sequential transformer of the same size as the ViT-RGB. Hyperspectral pixel are processed individually, and self-attention acts between spectral tokens of width $p^c = 10$. Unlike the ViT-RGB, this approach can fully leverage the spectral information of the hyperspectral dataset, and improves the accuracy significantly to $77 \pm 0.2\%$ on EnMAP-DFC. On the Houston2018 dataset, the spectral model reaches an accuracy of $47 \pm 3.2\%$. As a convolutional baseline, we use an established **3D-CNN model** [25] with strong performance on hyperspectral datasets [2]. Using 3D convolutions, this model can incorporate both spectral and spatial information, yielding a performance of $83 \pm 0.3\%$ for EnMAP and $45 \pm 1.8\%$ on Houston2018. Unlike the presented transformer approaches (see Table 1), the 3D-CNN model only makes predictions for the center-pixel of every input patch. This improves performance but necessitates a sliding-window inference strategy to create pixelwise land cover maps, which strongly increases computational cost. We re-train our best performing model using this approach on the EnMAP data and adopt the same strategy on the Houston2018 dataset for comparability.

5.2. Spatial-Spectral Embedding

We extend the spectral transformer to deal with spatial-spectral signals by embedding the data along both spatial and spectral axes (see Fig. 3). This increases the number of embeddings by a factor of $\frac{h}{p^h} \cdot \frac{w}{p^w}$ compared to the spectral model. The spatial-spectral factorization strategy detailed in Section 3.1.2 allows our **spatial-spectral model (SST)** to efficiently process the increased number of tokens. The model consists of two stacked transformers (with $d = 96$, 4 transformer blocks and 8 heads) that sequentially process the tokens with $p^{h,w} = 1$ and $p^c = 10$ along the spatial and spectral dimension, respectively. This approach yields an accuracy of $79 \pm 0.1\%$ on EnMAP-DFC and serves as the basic backbone for the other presented transformer modifications. Adding the blockwise patch embedding (BPE) scheme allows the model to embed patches conditionally on their position along the spectral axis (see Fig. 3) and improves accuracy to $81 \pm 0.2\%$. We find that spectral positional embeddings (SPE) slightly harm the performance of the SST model ($78 \pm 0.2\%$), while the combination of BPE and SPE yields an improvement to $82 \pm 0.1\%$ on the EnMAP data. This model reaches an accuracy of 43 ± 2.4 on Houston2018.

Model	Acc (%)	MAcc (%)
3D-CNN [25]	45 ± 1.8	$\mathbf{45 \pm 1.0}$
Transformer [42]	33 ± 1.1	26 ± 1.0
ViT-RGB	19 ± 1.4	21 ± 1.4
Spectral T.	47 ± 3.2	43 ± 1.6
SST	43 ± 2.4	40 ± 1.7
Masked SST	$\mathbf{48 \pm 2.8}$	42 ± 1.2

Table 2. Land cover classification results for the Houston2018 dataset. SST corresponds to spatial-spectral transformer with BPE. The masked SST is pre-trained and fine-tuned on Houston2018 training data. All models besides the standard transformer are trained for center pixel prediction.

5.3. Masked Pre-training

Self-supervised masked modeling increases the data efficiency of transformer models for natural language [11] or image [48] applications. We pre-train our transformer configurations for hyperspectral data on the unlabeled EnMAP dataset with a masked pixel reconstruction strategy. The model is trained for 200 epochs to reconstruct the 70% of patches which were masked in a 4×4 blockwise fashion. This pre-training yields small improvements over training from scratch on the larger EnMAP-DFC dataset (*e.g.*, $+1\%$ accuracy for the masked SST with BPE). On the Houston2018 data, masked pre-training improves the SST model by $+5\%$ to $48 \pm 2.8\%$ accuracy. The combination of masked pre-training and BPE results in strong representations, as re-

	Dataset Fraction			
Model	0.1%	1%	10%	100%
3D-CNN [25]	28 ± 1.8	38 ± 1.3	42 ± 1.0	45 ± 1.8
Transf. [42]	10 ± 0.1	10 ± 0.2	17 ± 1.8	33 ± 1.2
ViT-RGB	14 ± 1.7	14 ± 1.2	17 ± 0.9	19 ± 1.4
Spectral T.	17 ± 1.7	34 ± 2.0	44 ± 2.9	47 ± 3.2
SST	27 ± 3.2	38 ± 1.7	43 ± 2.8	43 ± 2.4
Masked SST	$\mathbf{35 \pm 2.0}$	$\mathbf{46 \pm 3.1}$	$\mathbf{47 \pm 1.9}$	$\mathbf{48 \pm 2.8}$

Table 3. Land cover classification accuracy on Houston2018 for different training set sizes (100%: 504 712 labeled pixels). When labeled training data is scarce, the pre-trained transformer significantly outperforms the other models.

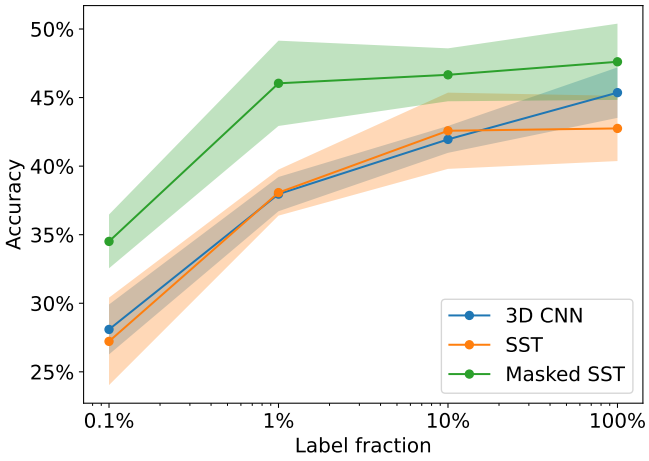

Figure 7. Performance of 3D-CNN baseline [25] and spatial-spectral transformer models trained on varying fractions of the Houston2018 dataset.

vealed by linear probing on the frozen transformer weights (see columns 'Frozen' in Table 1). The masked SST improves by $+3\%$ accuracy to $79 \pm 0.1\%$ and $+5\%$ in macro accuracy over the SST without pre-training on the EnMAP-DFC dataset.

Data Efficiency To investigate model performance on downstream applications with little labeled data, we train pre-trained and randomly initialized SST models with BPE and SPE on successively smaller portions of the Houston2018 dataset (see Table 3). Using as little as 0.1% of the Houston2018 training data ($\sim$504 pixels) results in an accuracy of $27 \pm 3.2\%$ for the randomly initialized SST model and $28 \pm 1.8\%$ for the baseline 3D-CNN [25]. The pre-trained SST model reaches an accuracy of $35 \pm 2.0\%$, which corresponds to an $+8\%$ increase that can be attributed to self-supervised pre-training (see Fig. 7). We observe a similar performance advantage for the self-supervised model when training on 1% and 10% of the Houston2018 training set (see Table 3). We note that the masked SST model outperforms the SST model without pre-training and the 3D-CNN with as little as 1% of the labeled training data.

6. Discussion

This work investigates the utility of masked hyperspectral image reconstruction for self-supervised learning of transformers. We pre-train different transformer model configurations on unlabeled data and evaluate them on the labeled EnMAP-DFC and Houston2018 datasets. A comparison of the baseline vanilla transformer and ViT-RGB methods reveals the high importance of spectral information for the EnMAP-DFC land cover classification task: ViT-RGB, which has access to larger spatial context but disregards spectral information beyond the RGB-bands, performs significantly worse than the standard transformer trained on the entire spectral sequence (-8% accuracy). Interestingly, both baseline transformer approaches lag significantly behind the convolutional 3D-CNN baseline [25]. We find that blockwise patch embedding is an important enhancement for spectral transformers and provides an implicit encoding of each token position in the spectral sequence. Despite EnMAP's high spectral resolution, the use of spatial context provides improvements in model performance. Incorporating spectral positional embeddings into the spatial-spectral transformer further boosts classification accuracy in our experiments when combined with blockwise patch embedding.

In order to leverage large unlabeled hyperspectral datasets and to boost the label efficiency of transformer models, we utilize masked data reconstruction as self-supervised pre-training task. Linear probing from the self-supervised representations indicates that masked hyperspectral image reconstruction yields meaningful representations that can achieve strong classification performance on EnMAP-DFC. We further conduct an ablation study on the label efficiency of our masked spatial-spectral transformer on the Houston2018 dataset. The pre-trained model can be fine-tuned with 1% of the labeled data to surpass the performance of the baseline models trained on 100% of the labeled data.

7. Conclusion

Our systematic evaluation of vision transformer models for hyperspectral remote sensing data reveals the benefits of different positional encoding schemes and the importance of modeling spatial-spectral interactions with self-attention. Factorizing self-attention between the spatial and spectral dimensions enables self-attention for high-dimensional hyperspectral data. We further showcase the potential of masked transformer pre-training and evaluate the resulting models with different amounts of labeled training data. The results of this study indicate that masked pre-training is highly effective to improve label efficiency of transformer models, and can also boost performance when a large number of labels is available. We believe that these results will be highly relevant for the hyperspectral remote sensing community as transformer networks continue to excel for vision tasks and more large unlabeled hyperspectral datasets start to become publicly available.

Acknowledgements

We thank the EnMAP mission team, the Hyperspectral Image Analysis Lab (University of Houston) and the IEEE GRSS IADF for providing the data used in this work as well as the anonymous reviewers for their helpful comments.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. ViViT: A Video Vision Transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6836–6846, 2021. 3

[2] Nicolas Audebert, Bertrand Le Saux, and Sébastien Lefèvre. Deep Learning for Classification of Hyperspectral Data: A Comparative Review. *IEEE Geoscience and Remote Sensing Magazine*, 7(2):159–173, 2019. 2, 7

[3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer Normalization. *arXiv preprint arXiv:1607.06450*, 2016. 3

[4] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *International Conference on Learning Representations*, 2022. 3

[5] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is Space-time Attention all you need for Video Understanding? In *ICML*, volume 2, page 4, 2021. 3

[6] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607. PMLR, 2020. 3

[7] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating Long Sequences with Sparse Transformers. *arXiv preprint arXiv:1904.10509*, 2019. 3

[8] Krzysztof Marcin Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Quincy Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. In *International Conference on Learning Representations year=2021*. 3

[9] 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 *Advances in Neural Information Processing Systems*. 4, 5

[10] 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. Ieee, 2009. 1

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019. 2, 3, 5, 7

[12] 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. *ArXiv*, abs/2010.11929, 2021. 3, 4, 6, 7

[13] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised Representation Learning by Predicting image Rotations. *arXiv preprint arXiv:1803.07728*, 2018. 3

[14] Kamlesh Golhani, Siva K Balasundram, Ganesan Vadamalai, and Biswajeet Pradhan. A Review of Neural Networks in Plant Disease Detection Using Hyperspectral Data. *Information Processing in Agriculture*, 5(3):354–371, 2018. 2

[15] Megandhren Govender, Kershani Chetty, and Hartley Bulcock. A Review of Hyperspectral Remote Sensing and its Application in Vegetation and Water Resource Studies. *Water Sa*, 33(2):145–151, 2007. 1

[16] Luis Guanter, Hermann Kaufmann, Karl Segl, Saskia Foerster, Christian Rogass, Sabine Chabrillat, Theres Kuester, André Hollstein, Godela Rossner, Christian Chlebek, et al. The EnMAP Spaceborne Imaging Spectroscopy Mission for Earth Observation. *Remote Sensing*, 7(7):8830–8857, 2015. 1, 6

[17] Michael Gutmann and Aapo Hyvärinen. Noise-Contrastive Estimation: A New Estimation Principle for Unnormalized Statistical Models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 297–304. JMLR Workshop and Conference Proceedings, 2010. 3

[18] Ji He, Lina Zhao, Hongwei Yang, Mengmeng Zhang, and Wei Li. HSI-BERT: Hyperspectral image classification using the bidirectional encoder representation from transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 58(1):165–178, 2019. 4

[19] 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*, pages 16000–16009, 2022. 3, 5

[20] Xin He, Yushi Chen, and Zhouhan Lin. Spatial-spectral Transformer for Hyperspectral Image Classification. *Remote Sensing*, 13(3):498, 2021. 2

[21] Uta Heiden, Karl Segl, Sigrid Roessner, and Hermann Kaufmann. Determination of Robust Spectral Features for Identification of Urban Surface Materials in Hyperspectral Remote Sensing Data. *Remote Sensing of Environment*, 111(4):537–552, 2007. 1

[22] Danfeng Hong, Zhu Han, Jing Yao, Lianru Gao, Bing Zhang, Antonio Plaza, and Jocelyn Chanussot. SpectralFormer: Rethinking Hyperspectral Image Classification with Transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–15, 2021. 2

[23] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross Attention for Semantic Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 603–612, 2019. 3

[24] Damian Ibanez, Ruben Fernandez-Beltran, Filiberto Pla, and Naoto Yokoya. Masked Auto-Encoding Spectral–Spatial Transformer for Hyperspectral Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022. 4

[25] Ying Li, Haokui Zhang, and Qiang Shen. Spectral–Spatial Classification of Hyperspectral Imagery with 3D Convolutional Neural Network. *Remote Sensing*, 9(1):67, 2017. 6, 7, 8

[26] Konstantinos Makantasis, Konstantinos Karantzalos, Anastasios Doulamis, and Nikolaos Doulamis. Deep Supervised Learning for Hyperspectral Data Classification through Convolutional Neural Networks. In *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 4959–4962. IEEE, 2015. 2

[27] Oscar Mañas, Alexandre Lacoste, Xavier Giro-i Nieto, David Vazquez, and Pau Rodriguez. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9414–9423, 2021. 3

[28] Lichao Mou, Pedram Ghamisi, and Xiao Xiang Zhu. Deep Recurrent Neural Networks for Hyperspectral Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3639–3655, 2017. 2

[29] Mehdi Noroozi and Paolo Favaro. Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles. In *European Conference on Computer Vision*, pages 69–84. Springer, 2016. 3

[30] ME Paoletti, JM Haut, J Plaza, and A Plaza. Deep Learning Classifiers for Hyperspectral Imaging: A Review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 158:279–317, 2019. 1

[31] Yuhao Qing, Wenyi Liu, Liuyan Feng, and Wanjia Gao. Improved Transformer Net for Hyperspectral Image Classification. *Remote Sensing*, 13(11):2216, 2021. 2

[32] Colorado J Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. Scale-MAE: A Scale-Aware Masked Autoencoder for Multiscale Geospatial Representation Learning. *arXiv preprint arXiv:2212.14532*, 2022. 4

[33] Sudipan Saha, Patrick Ebel, and Xiao Xiang Zhu. Self-supervised Multisensor Change Detection. *IEEE Transactions on Geoscience and Remote Sensing*, 2021. 3

[34] Linus Scheibenreif, Joëlle Hanna, Michael Mommert, and Damian Borth. Self-supervised Vision Transformers for Land-cover Segmentation and Classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1422–1431, 2022. 3

[35] Linus Scheibenreif, Michael Mommert, and Damian Borth. Contrastive Self-supervised Data Fusion for Satellite Imagery. In *International Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2022. 3

[36] Laurent Sifre and Prof Stéphane Mallat. Rigid-motion Scattering for Image Classification. *English. Supervisor: Prof. Stéphane Mallat. Ph. D. Thesis. Ecole Polytechnique*, 2, 2014. 5

[37] Vladan Stojnic and Vladimir Risojevic. Self-supervised Learning of Remote Sensing Scene Representations using Contrastive Multiview Coding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1182–1191, 2021. 3

[38] Le Sun, Guangrui Zhao, Yuhui Zheng, and Zebin Wu. Spectral–Spatial Feature Tokenization Transformer for Hyperspectral Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022. 2

[39] Aidan M Swope, Xander H Rudelis, and Kyle T Story. Representation Learning for Remote Sensing: An Unsupervised Sensor Fusion Approach. *arXiv preprint arXiv:2108.05094*, 2021. 3

[40] Chao Tao, Ji Qi, Weipeng Lu, Hao Wang, and Haifeng Li. Remote Sensing Image Scene Classification with Self-supervised Paradigm under Limited Labeled Samples. *IEEE Geoscience and Remote Sensing Letters*, 2020. 3

[41] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training. In *Advances in Neural Information Processing Systems*. 5

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017. 2, 3, 5, 6, 7

[43] Wenxuan Wang, Leiming Liu, Tianxiang Zhang, Jiachen Shen, Jing Wang, and Jiangyun Li. Hyper-ES2T: Efficient Spatial–Spectral Transformer for the Classification of Hyperspectral Remote Sensing Images. *International Journal of Applied Earth Observation and Geoinformation*, 113:103005, 2022. 2

[44] Yi Wang, Conrad M Albrecht, Nassim Ait Ali Braham, LiChao Mou, and Xiao Xiang Zhu. Self-supervised Learning in Remote Sensing: A Review. *IEEE Geoscience and Remote Sensing Magazine (GRSM)*, 2022. 3

[45] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised Feature Learning via Non-parametric Instance Discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018. 3

[46] Tete Xiao, Colorado J Reed, Xiaolong Wang, Kurt Keutzer, and Trevor Darrell. Region similarity representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10539–10548, 2021. 3

[47] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate Yourself: Exploring Pixel-level Consistency for Unsupervised Visual Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16684–16693, 2021. 3

[48] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663, 2022. 3, 5, 7

[49] Yonghao Xu, Bo Du, Liangpei Zhang, Daniele Cerra, Miguel Pato, Emiliano Carmona, Saurabh Prasad, Naoto Yokoya, Ronny Hänsch, and Bertrand Le Saux. Advanced Multisensor Optical Remote Sensing for Urban Land Use and Land Cover Classification: Outcome of the 2018 IEEE GRSS Data Fusion Contest. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(6):1709–1724, 2019. 6

[50] Naoto Yokoya, Pedram Ghamisi, Ronny Hänsch, and Michael Schmitt. 2020 IEEE GRSS Data Fusion Contest: Global Land Cover Mapping with Weak Supervision [tech-

nical committees]. *IEEE Geoscience and Remote Sensing Magazine*, 8(1):154–157, 2020. 6

[51] Naoto Yokoya, Claas Grohnfeldt, and Jocelyn Chanussot. Hyperspectral and Multispectral Data Fusion: A Comparative Review of the Recent Literature. *IEEE Geoscience and Remote Sensing Magazine*, 5(2):29–56, 2017. 2

[52] Yuan Yuan and Lei Lin. Self-supervised Pretraining of Transformers for Satellite Image Time Series Classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:474–487, 2020. 3

[53] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European Conference on Computer Vision*, pages 649–666. Springer, 2016. 3

[54] Xiangrong Zhang, Yujia Sun, Jingyan Zhang, Peng Wu, and Licheng Jiao. Hyperspectral Unmixing via Deep Convolutional Neural Networks. *IEEE Geoscience and Remote Sensing Letters*, 15(11):1755–1759, 2018. 2

[55] Wenzhi Zhao and Shihong Du. Spectral–Spatial Feature Extraction for Hyperspectral Image Classification: A Dimension Reduction and Deep Learning Approach. *IEEE Transactions on Geoscience and Remote Sensing*, 54(8):4544–4554, 2016. 2

[56] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. iBOT: Image BERT Pre-Training with Online Tokenizer. *International Conference on Learning Representations*, 2022. 3