S2MAE: A Spatial-Spectral Pretraining Foundation Model for Spectral Remote Sensing Data

Xuyang Li[1,2] Danfeng Hong[1,2*] Jocelyn Chanussot[3]
[1]Aerospace Information Research Institute, Chinese Academy of Sciences, 100094 Beijing, China
[2]School of Electronic, Electrical and Communication Engineering,
University of Chinese Academy of Sciences, 100049 Beijing, China
[3]Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France
lixuyang23@mails.ucas.ac.cn, hongdf@aircas.ac.cn, jocelyn.chanussot@inria.fr

Abstract

In the expansive domain of computer vision, a myriad of pre-trained models are at our disposal. However, most of these models are designed for natural RGB images and prove inadequate for spectral remote sensing (RS) images. Spectral RS images have two main traits: (1) multiple bands capturing diverse feature information, (2) spatial alignment and consistent spectral sequencing within the spatial-spectral dimension. In this paper, we introduce Spatial-SpectralMAE (S2MAE), a specialized pre-trained architecture for spectral RS imagery. S2MAE employs a 3D transformer for masked autoencoder modeling, integrating learnable spectral-spatial embeddings with a 90% masking ratio. The model efficiently captures local spectral consistency and spatial invariance using compact cube tokens, demonstrating versatility to diverse input characteristics. This adaptability facilitates progressive pretraining on extensive spectral datasets. The effectiveness of S2MAE is validated through continuous pretraining on two sizable datasets, totaling over a million training images. The pre-trained model is subsequently applied to three distinct downstream tasks, with in-depth ablation studies conducted to emphasize its efficacy.

1. Introduction

Spectral imaging, with its ability to capture a diverse spectrum of spectral information, significantly enhances the precision and recognition of objects and scenes beyond the capabilities of RGB data alone. This has positioned multi/hyperspectral (MS/HS) remote sensing data as a preferred and vital component in numerous Earth Observation (EO) applications [19]. These applications encompass various domains such as land use/land cover mapping, ecosystem monitoring, weather forecasting, energy resource development, biodiversity conservation, and geological exploration.

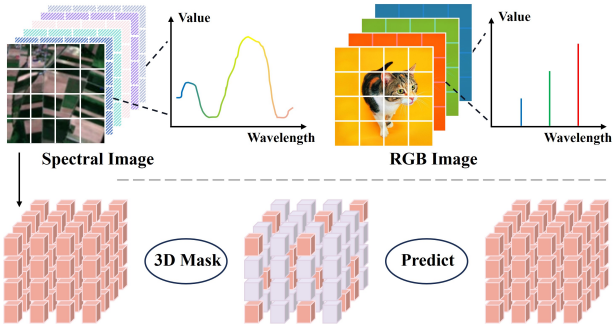

Figure 1. Spectral RS data displays spatial invariance and channel-wise continuity. The 3D random mask helps in learning local spatial-spectral correlations by enabling reconstruction.

In the realm of RS imagery, a wealth of open-resource images is readily accessible, yet a significant portion remains unlabeled. Existing algorithms and models tend to underutilize these expansive datasets, primarily relying on the limited labeled data available. However, the process of labeling such data is resource-intensive, time-consuming, and often financially burdensome. To unlock the full potential of these resources, there is a critical need for the development and implementation of self-supervised or unsupervised methods driven by data.

A wave of pioneering self-supervised methods has emerged within the RS community [1, 4, 15, 17, 18, 26–28, 30, 33, 34, 40, 42, 43]. Wang *et al.* [39] trained a plain vision transformer on RS RGB images and developed rotated varied-size window attention for fine-tuning the model. Mall *et al.* [26] improved SeCo [27] by designing the CACo loss to better utilize contrastive learn-

*Corresponding author.

ing (CL) for mining temporal invariance in RS data. However, most existing approaches primarily focus on RGB data [1, 27, 28, 32, 39, 42], neglecting rich spectral information in RS. SatMAE [4] is a masked autoencoder (MAE) model tailored for MS images, utilizing the group mask strategy with group embedding to pretrain a model. While pioneering, SatMAE's group mask design (see Fig. 3) falls short in three aspects: (1) inadequate interaction between groups, impeding spectral sequencing comprehension (e.g., between RGB and Red Edge), (2) limited band combinations in grouping; for example, SatMAE divided 10 channels into 3 groups, limiting adaptability to varying channel numbers and (3) extra group inductive bias due to specific band combinations. These shortcomings drive us to ponder: Can MAE models exploit local spectral continuity in spectral data with variable band counts to learn strong representations and reduce inductive bias?

To address these issues, we propose Spatial-SpectralMAE (S2MAE), an extension of MAE in characterizing spectral images using a 3D masking approach. The 3D masking method, first introduced in [9], evaluates MAE's efficacy in videos. Despite its ability to learn robust representations with minimal biases, it tends to focus more on local environmental details than on the subject's movement (critical in video analysis) due to random masking and varying object dynamics across frames. In contrast, localized information is pivotal for spectral images due to their low resolution (e.g., 10m, 20m). Also, spectral images do not vary in spatial dimensions, with each channel embodying unique spectral reflectance for distinct characterization information (see Fig. 1). Consequently, utilizing 3D masking in spectral data, integrating local spectral continuity and spatial invariance via small tensor cubes, is expected to be more effective than in videos. This method has proven effective for hyperspectral image classification [20, 34], and we think it will be valuable for all spectral RS data on more tasks. Furthermore, leveraging the advantages of transformers, S2MAE efficiently manages diverse input image traits such as size, resolution, and channels, enabling progressive pretraining across various spectral RS datasets. It should be noted that our extended version, i.e., SpectralGPT, with more advanced design, more general EO applications, and more analysis and discussion, can be found in [19].

In overview, our contributions encompass:

(1) We've devised Spatial-SpectralMAE (S2MAE), a general self-supervised framework for spectral imagery, utilizing 3D masked transformers with a 90% mask ratio. It overcomes the limitations of SatMAE, enhancing encoder capabilities to learn strong representations through local spectral continuity and spatial invariance for spectral images of an arbitrary number of bands.

(2) We adopt a progressive pretraining approach for S2MAE, leveraging two Sentinel-2 datasets: fMoW-Sentinel [4] and BigEarthNet [35]. These datasets exhibit differences not only in image sizes and geographical coverage but also accumulate an extensive training set exceeding a million images in total.

(3) S2MAE and existing foundational models are assessed across three distinct downstream tasks, including single/multi-label classification, and change detection. Additionally, validation through numerous ablation studies is conducted, complemented by factors such as mask ratio, model scale, decoder depth, patch size, and other relevant aspects.

2. Related Work

Self-supervised learning. In the domains of Natural Language Processing (NLP) and Computer Vision (CV), various Pretrained Foundation Models (PFMs) have gained prominence, including BERT [7], GPT [31], LLaMA [38] series in NLP, and MoCo [13], MAE [14], DINO [41], SAM [22] in CV. These models are shaped by two complementary self-supervised learning techniques: contrastive learning (CL) and masked language/image modeling (MLM/MIM). In the Computer Vision field, CL aims to capture invariance across a batch of images by considering identical and different images with data augmentation. On the other hand, MIM aims to unveil spatial correlations by reconstructing masked patches within a single image. Research consistently favors CL over MIM in linear probing evaluations, but MIM excels in fine-tuning assessments. Innovative methods like SiameseIM [36], MimCo [44], and SiamMAE [10] integrate CL and MIM to achieve robust data invariance, enhancing the efficacy and resilience of these foundational models.

Masked Autoencoders. Masked autoencoders (MAE) [14] are special frameworks for MIM. The architecture involves an asymmetric design where the encoder operates on partially unmasked tokens and a lightweight decoder reconstructs the masked tokens. MAE has demonstrated state-of-the-art performance in various vision benchmarks, leading to its extension in numerous follow-up works across different data modalities. MultiMAE [2] extends MAE to handle diverse input modalities, adjusting the training objective to predict multiple outputs. For video data, VideoMAE [37] innovates by introducing a video tube masking and reconstruction pretext task. MAE-ST [9] enhances MAE by randomly masking 3D patches in video data, enabling strong representations without spacetime biases. This approach is also applicable to handling spectral data. In this work, we randomly mask spectral-spatial agnostic patches and utilize the plain MAE framework for pretraining.

Progressive Pretraining. Progressive pretraining is a technique that enhances the generalization of models, initially developed for NLP [11, 23]. This methodology has

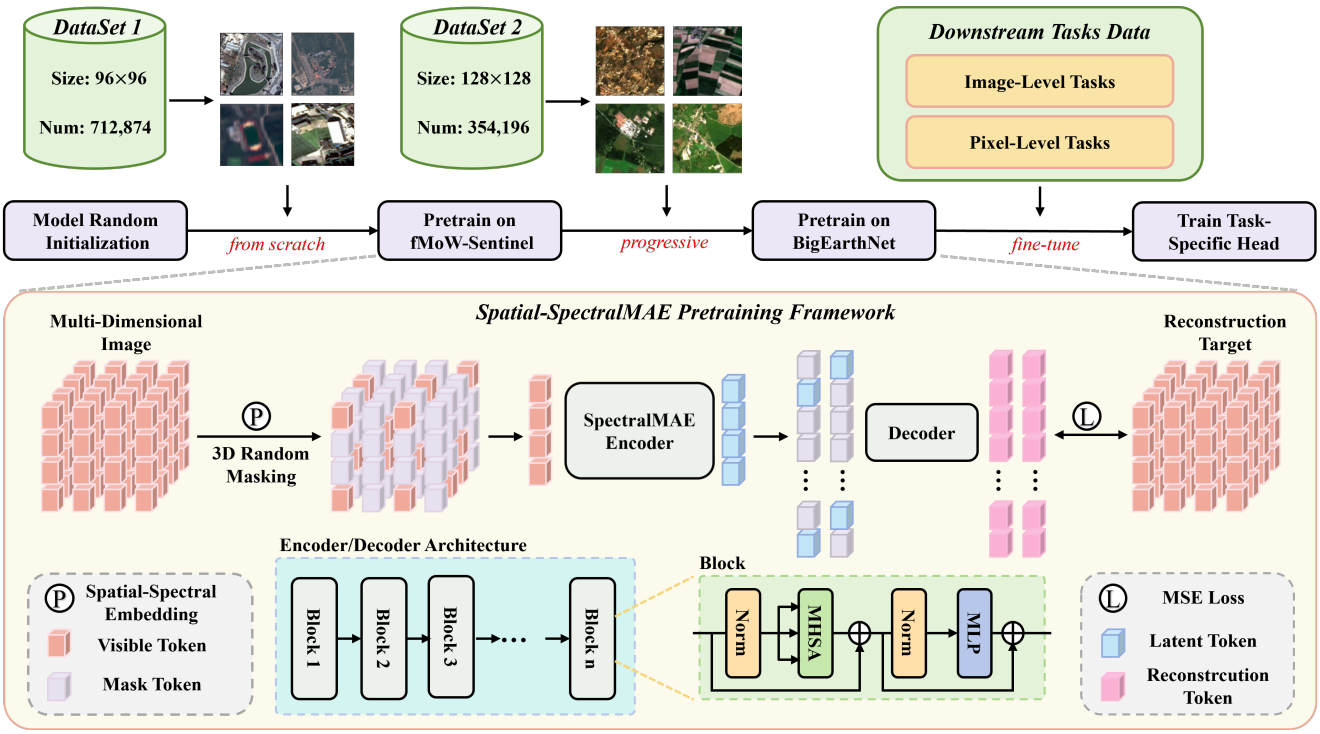

Figure 2. An illustrative workflow of the proposed S2MAE foundation model, which consists of three components: initial pretraining from scratch on one dataset (e.g., fMoW, with 712,874 images), progressive pretraining on more datasets (e.g., BigEarthNet, with 354,196 images), and fine-tuning for downstream tasks. In the pertaining phase, our S2MAE starts to train the model from scratch with a random initialization. Subsequently, the model undergoes progressive training using data with varying image sizes and geographic regions. S2MAE is constructed following the MAE architecture [14] and incorporates 3D masking, where 90% of the patches are masked. For downstream tasks, such as single classification, multi-label classification, and change detection, the pre-trained S2MAE is connected with the task-specific head networks to be trained and then performs fine-tuning.

also found applications in vision tasks; for instance, [21] utilizes BYOL-style continual pretraining for medical image segmentation, and [32] adopts a hierarchical strategy. In the domain of RS, GFM [28] employs the model pretrained on ImageNet [6] as an auxiliary distillation objective, effectively combining concepts from MIM and CL. Nevertheless, leveraging only one general pretraining framework to transfer model weights from a global dataset to a specific regional dataset in the spectral RS domain remains uncharted territory. Exploiting the adaptability of the 3D transformer, we take a progressive pretraining approach by pretraining S2MAE on two distinct spectral datasets. These datasets exhibit variations not only in image size and quantity but also in the geographical regions they cover.

3. Methodology

3.1. Method Overview of S2MAE

S2MAE is an extension of MAE. It employs a designed 3D transformer consisting of a random masking strategy, a plain ViT [8] encoder module, and a lightweight ViT decoder. After pretraining, we only use the encoder as a backbone for downstream tasks. Fig. 2 illustrates the framework of S2MAE. In detail, the implementation process of S2MAE can be broken down into the following steps:

Patchify. Given a spectral image $\boldsymbol{x} \in \mathbb{R}^{H\times W\times C}$, we partition it into non-overlapping 3D tensor patches along both the spatial and spectral dimensions. Each patch has a size of $p \times p \times k$, where p and k are the patch sizes in spatial and spectral dimensions, respectively. Using these settings, we have $\frac{H}{p} \times \frac{W}{p} \times \frac{C}{k}$ patches, denoted as $\boldsymbol{x} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_i, ..., \boldsymbol{x}_{\frac{H}{p}\times\frac{W}{p}\times\frac{C}{k}}\}$. All patches are then flattened and mapped with a trainable linear projection $\boldsymbol{E}_s$, combined with separable positional embeddings $\boldsymbol{E}_{spectral}$ and $\boldsymbol{E}_{spatial}$. We denote the integration of $\boldsymbol{E}_{spectral}$ and $\boldsymbol{E}_{spatial}$ as $\boldsymbol{E}_{pos}$. Thus, the data in the i-th patch can be expressed as $\tilde{\boldsymbol{x}}_i = \boldsymbol{E}_s\boldsymbol{x}_i + \boldsymbol{E}_{pos}$. Notably, the separate spatial-spectral position embedding utilized in this context differs from the vanilla embedding employed in SpectralGPT [19]. This approach serves two key purposes. Firstly, it prevents an unwarranted increase in the size of positional embeddings, particularly in a 3D context, as proposed in [9]. Secondly, it aids in decoupling spatial-spectral features

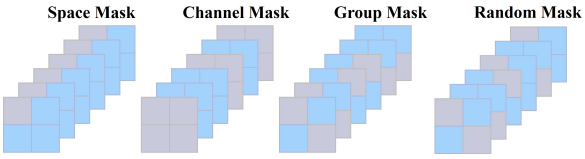

Figure 3. Four strategies for handling multi-dimensional spectral data through masked image modeling.

indirectly, thereby influencing model performance distinctively from SpectralGPT and yielding divergent results in downstream and ablation studies. Under specific hyperparameter configurations, the model may exhibit slightly superior performance relative to SpectralGPT.

Masking. Next, a masking operation is performed on these patches to identify visible (or unmasked) and masked patches,e.g., $\boldsymbol{x}_{vis}$ and $\boldsymbol{x}_{mask}$.

$$[\boldsymbol{x}_{vis}, \boldsymbol{x}_{mask}] = \mathbb{M} \odot \tilde{\boldsymbol{x}}, \tag{1}$$

where $\mathbb{M} \in \{0,1\}^{\frac{H}{p} \times \frac{W}{p} \times \frac{C}{k}}$ is a patch-wise binary mask indicating which patches should be masked, i.e., all data in the patch are set to zero. Only the visible patches are sent into the to-be-learned encoder.

Encoder. The encoder f_{en} is implemented using ViT, where each visible patch is processed through a series of transformer blocks. Thus, the encoder output in the i-th patch can be expressed as $\boldsymbol{z}_i = f_{en}(\boldsymbol{x}_{vis})$.

Decoder. The input to the decoder, denoted by g_{de}, is a complete set of tokens that includes the encoded visible patches and mask tokens (e.g., $\boldsymbol{z}_m$). The encoded features, which are the latent representations from the encoder, and the mask tokens are used as inputs and combined with positional embeddings to the lightweight ViT decoder. The output can be expressed as $\hat{\boldsymbol{x}} = g_{de}([\boldsymbol{z}_{vis}, \boldsymbol{z}_m] + \boldsymbol{E}_{pos})$, where $\boldsymbol{z}_{vis}$ is the encoded representations of visible patches.

Loss. The utilized loss function is Mean Squared Error (MSE) loss, and we solely calculate the loss of the masked patches, i.e., $\mathcal{L} = \frac{1}{mask} \sum_{i \in mask} (\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i)^2$.

3.2. Progressive Pretraining Procedure

Since S2MAE is adaptable to diverse input image sizes, we employ a progressive pretraining strategy by incorporating diverse spectral RS datasets. Specifically, we utilize the trainset of fMoW-sentinel and BigEarthNet datasets to validate the effectiveness of the progressive pretraining approach. Notably, using this approach, various RS data can be incorporated into pretraining without integrating them into a unified dataset.

Stanford University researchers curated the fMoW-S2 [4] dataset using geo-coordinates and timestamps from fMoW [3] for Sentinel-2 image time series. The fMoW-S2 dataset mirrors fMoW labels and consists of Sentinel-2 spectral images (B1-12 and B8A). It has 882,779 images divided into training (712,874), validation (84,939), and test (84,966) sets, each averaging around 45 pixels in height and 60 pixels in width. For more details, visit the fMoW-S2 dataset website [1]. BigEarthNet [35] dataset has 125 Sentinel-2 tiles, capturing data from June 2017 to May 2018 in ten European countries. It comprises 590,326 12-band images across 19 classes for multi-label classification. About 12% of images affected by snow, clouds, or shadows were removed. The dataset has 354,196 training patches and 118,065 validation patches. The combined total of two datasets exceeds one million images.

A random-initialized vanilla ViT is employed alongside S2MAE for initial pretraining on fMoW-S2 with the image size of $96 \times 96 \times 12$. Following this, the pre-trained S2MAE model is continued for pretraining on BigEarthNet, where the image size is $128 \times 128 \times 12$. To adapt the model to our spectral image data, we employ a patch size of $8 \times 8 \times 3$. To distinguish between different stages, the model pre-trained solely on the fMoW-S2 dataset is denoted as S2MAE, whereas the model pre-trained on both datasets in a progressive way is represented as S2MAE*.

4. Experiments

In this section, we outline the implementation of the pretraining procedure and proceed to evaluate our model's performance through three downstream tasks. Additionally, we present various ablation studies for a comprehensive analysis. For visual reconstruction results, see Fig. 4. More reconstructed spectral image results, along with extensive information regarding training settings for downstream tasks, are available in the supplementary materials.

4.1. Pretraining Implement Details

Utilizing the computational power of 8 NVIDIA GeForce RTX 4090 GPUs, we implement the AdamW optimizer[25] with a foundational learning rate of 2×10^{-4}, coupled with a half-cycle cosine decay schedule [24]. To ensure robustness, we adopt a 3D masking ratio of 90%, facilitating effective training. The model undergoes a pretraining regimen, encompassing 200 epochs on the fMoW-S2 dataset. Following this initial phase, the model's training proceeds with an additional 100 epochs on the BigEarthNet dataset.

4.2. Downstream Tasks Experiments

Single-label classification on EuroSAT. The EuroSAT [16] dataset comprises 27,000 Sentinel-2 images, collected from 34 European countries. These images are categorized into 10 distinct land use classes. Each image maintains a size of 64×64 pixels and encompasses 13 spectral bands. For consistency with previous data processing, band B10

[1] https://purl.stanford.edu/vg497cb6002

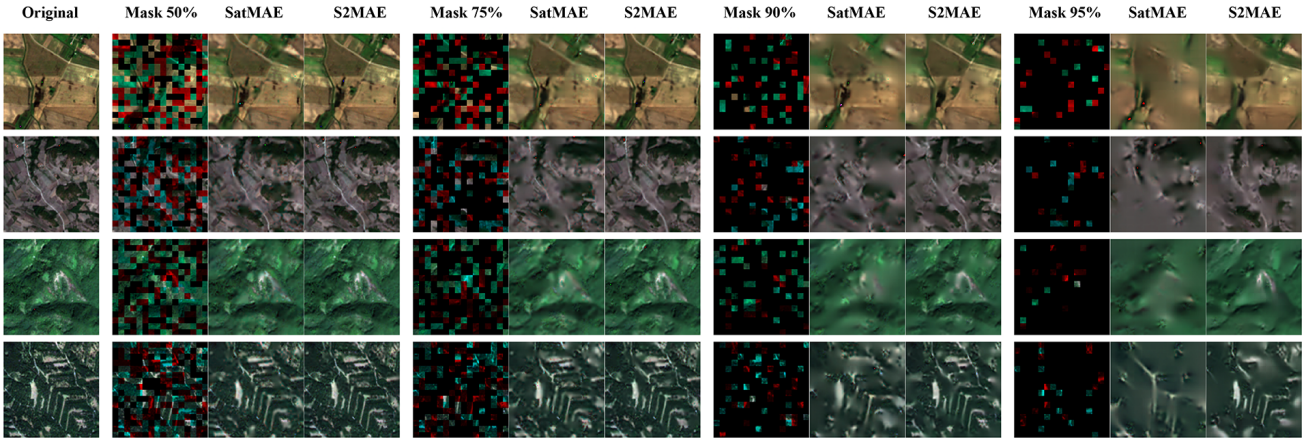

Figure 4. Visual comparison from the nature-color (RGB) image reconstruction perspective between SatMAE and S2MAE with varied masking ratios of 50%, 75%, 90%, and 95%, respectively. By masking out a greater number of patches, the reconstructed images exhibit noticeable differences from the originals (e.g., 50% vs. 95%), which is expected. It is worth noting that S2MAE holds stronger reconstruction capability (*cf. SatMAE*), even if the masking rate has reached over 90%, showing its powerful reasoning performance.

is excluded from all images, and the dataset adopts the train/validation splits recommended in [29].

For this task, the pre-trained model's encoder serves as the backbone, and its output is subject to an average pooling layer to generate predictions. The pre-trained model is finetuned on the EuroSAT dataset, spanning 150 epochs with a batch size of 512. This process employs the AdamW optimizer with a base learning rate of 2×10^{-4}, alongside data augmentations consistent with previous work [14], including weight decay (0.05), drop path (0.1), reprob (0.25), mixup (0.8), and cutmix (1.0). The training objective involves minimizing the cross-entropy loss. In Tab. 1, we present a comparative analysis of S2MAE against alternative pretraining models, reporting the highest Top1 accuracy on the validation set. Our results highlight the efficacy of the proposed approach, achieving an impressive accuracy of 99.16%. Furthermore, when the model undergoes pretraining on both datasets, a noteworthy performance boost is observed. This underscores the advantage of leveraging diverse data sources for improving model performance.

Multi-label classification on BigEarthNet. BigEarthNet is introduced in Sec. 3.2; this versatile dataset not only serves for pretraining but also plays a role in the finetuning phase. By previous research [4, 27], we finetune our model using only a 10% subset of the training set. We default to reporting mean average precision (mAP) results for evaluating the performance on this task. We adopt the train/validation splits recommended in [29]. It should be noted that most existing methods, including pre-trained foundation models, usually use all images for training in the BigEarthNet dataset, while our proposed method only uses 10% of training samples and achieves higher classification performance.

Method	Pretrained Dataset	Acc. (%)
ResNet50[12]	ImageNet-1k	96.72
SeCo[27]	SeCo	97.23
ViT[8]	From scratch.	98.73
ViT[8]	ImageNet-22k	98.91
SatMAE[4]	fMoW-S2	99.09
S2MAE	fMoW-S2	99.16
S2MAE*	fMoW-S2+BigEarthNet	**99.19**

Table 1. Quantitative results of SOTA pre-trained foundation models for the single-label RS scene classification task in terms of accuracy on the EuroSAT dataset.

Method	Pretrained Dataset	mAP
ResNet50[12]	ImageNet-1k	80.06
ViT[8]	From scratch.	80.15
SeCo[27]	SeCo	82.82
ViT[8]	ImageNet-22k	84.67
SatMAE[4]	fMoW-S2	84.93
S2MAE	fMoW-S2	85.59
S2MAE*	fMoW-S2+BigEarthNet	**87.41**

Table 2. Quantitative results of SOTA pre-trained foundation models for the multi-label RS scene classification task in terms of mean average precision (mAP) on the BigEarthNet dataset.

Aligning with most of the settings applied in the EuroSAT fine-tuning experiments, except for an increased learning rate of 2×10^{-4} and the training epochs of 40 epochs. Given the multi-label classification nature of this task, our training objective involves the multi-label soft margin loss. Tab. 2 presents a comparative analysis of our pre-trained model against other proposed pre-trained mod-

els. When compared to ViT pre-trained on ImageNet-22k and SatMAE, our model outperforms them by 0.92% and 0.66% in terms of mAP, respectively. The introduction of additional pretraining data, BigEarthNet, leads to a significant performance boost, with the model achieving an impressive 87.41% mAP. This substantial improvement can be attributed to two key factors. Firstly, the model's initial pretraining on BigEarthNet, even without labels, equips it with a strong grasp of the dataset's distribution, accelerating convergence during fine-tuning and enhancing classification accuracy. Secondly, the adoption of the MIM method as a pretext task, coupled with a substantial data scale, necessitates alignment with the training strategy, emphasizing the significance of the random masking framework and a 90% masking ratio to facilitate more robust representation learning. Furthermore, as our evaluation focuses on a multi-label classification task and employs only 10% of the training data, the results underscore the superior generalization capabilities of our proposed model.

Change Detection The Onera Satellite Change Detection (OSCD) dataset [5] includes 24 pairs of Sentinel-2 images (2015-2018). There are 14 training and 10 evaluation images, with 13 spectral bands at resolutions of 10m, 20m, and 60m. Labels denote pixel-level urban changes.

On the OSCD dataset, we perform image cropping, generating patches of 128×128 pixels with a 50% overlap rate, and apply random flips and rotations as data augmentation techniques. Model training spans 50 epochs. The optimizer and loss function remain consistent with those employed in the EuroSAT experiment. We evaluate model performance in terms of precision, recall, and F1 score, with quantitative results presented in Tab. 3 and qualitative results illustrated in Fig. 5. Notably, our proposed model achieves the highest F1 score, surpassing the second-best model by a substantial margin of 0.52%. Though our model excels in F1 score and recall, it has a low precision among evaluated models. This is due to data imbalance, favoring recall over precision. The ViT architecture's complexity requires ample data to combat overfitting. With only 14 training and 10 testing images in OSCD, overfitting and limited responsiveness to out-of-domain data may occur. Addressing this may involve more fine-tuning data or reducing the model's complexity.

Method	Pretrained Dataset	Precision	Recall	F1
ResNet50[12]	ImageNet-1k	**65.42**	38.86	48.10
SeCo[27]	SeCo	57.71	49.23	49.82
ViT[8]	From scratch.	56.71	47.52	51.71
ViT[8]	ImageNet-22k	52.09	52.37	52.23
SatMAE[4]	fMoW-S2	55.18	50.54	52.76
S2MAE	fMoW-S2	53.89	55.87	53.28
S2MAE*	fMoW-S2+BigEarthNet	54.90	**56.81**	**54.26**

Table 3. Quantitative results of SOTA pre-trained foundation models for the RS change detection task in terms of precision, recall, and F1 score on OSCD. The best result is shown in bold.

4.3. Ablation Studies on S2MAE

In the pretraining stage, we conduct a thorough study of factors affecting downstream task performance. For a rigorous assessment, all ablation models undergo fine-tuning on a 10% subset of the BigEarthNet dataset using mAP measurement. ViT-B is chosen as the backbone. Except for the data scale ablations, the models undergo 200 epochs of pretraining on the fMoW-S2 dataset.

Positional embedding. We explored different embedding measures for S2MAE, as shown in Tab. 4a. The results revealed that embedding has a minimal impact on model performance. We adopted learnable spatial-spectral embedding to mitigate the growth in positional embeddings' size, as suggested in [9].

Decoder depth. Tab. 4b systematically examines the impact of decoder depth on model performance, following the principles of MIM methods where the pre-trained encoder serves as the backbone for downstream tasks while discarding the decoder component. Notably, the results reveal that a shallow decoder configuration is ill-suited for spectral model pretraining. This observation aligns with the hypothesis that spectral images, characterized by high dimensionality and complexity, require a decoder with enhanced capacity, consistent with prior findings in the field [9].

Reconstruction. Tab. 4c analyzes the impact of reconstruction targets on spectral RS images—comparing normalized, standardized, and raw data without such transformations. Normalization scales data to [0, 1], while standardization gives a mean of 0 and a standard deviation of 1. The study shows minimal performance difference between normalization and standardization. However, models pre-trained on raw data significantly underperform, likely due to the inherent nature of spectral images where spectral values are numerically large and vary between bands. Pretraining on raw data might require a longer schedule to converge and match the performance of models pretrained on normalized data. The study suggests employing a semantically meaningful target in a specific representation space could potentially enhance model performance.

Patch size. In Tab. 4d, larger patch sizes are shown to consistently decrease model performance, aligning with previous research [4]. This effect is attributed to ViT architecture characteristics: larger patch sizes, like 16×16, reduce fine-grained spatial information as the model processes fewer patch tokens. This decline in spatial detail negatively impacts overall model performance. Notably, the pre-trained model consistently improves mAP regardless of patch size, showcasing its performance enhancement across various configurations. Remarkably, a patch size of 8×8 yields superior recognition performance compared to

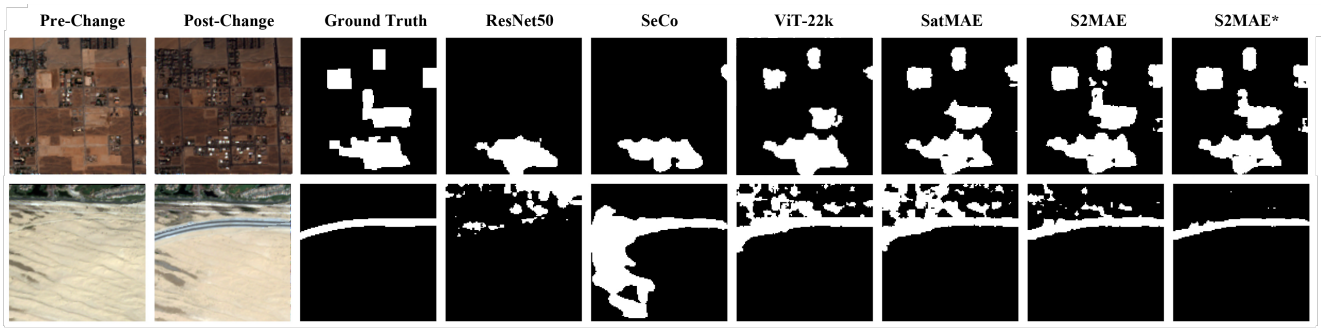

Figure 5. Visual results obtained by successively using different pre-trained foundation models, i.e., ResNet50, SeCo, ViT-22k, SatMAE, and our proposed S2MAE for the downstream change detection task on the OSCD dataset.

Embedding	Method	mAP
Vanilla	sin-cos	85.57
Spatialspectral	sin-cos	**85.62**
Spatialspectral	learnable	85.59

(a) Positional Embedding

Blocks	mAP
2	84.91
4	**85.59**
8	85.47

(b) Decoder Depth

Case	Scale	mAP
Without norm	-	84.86
Normalization	[0,1]	**85.59**
Standardization	[-1,1]	85.52

(c) Reconstruction Target

Init. Weights	Patch Size	mAP
Random	16	70.68
S2MAE	16	78.42
Random	8	80.15
S2MAE	8	**85.59**

(d) Patch Size

Ratio	mAP
25%	82.81
50%	84.32
75%	84.94
90%	**85.59**

(e) Masking Ratio

Pretrained Dataset	mAP
From scratch.	80.15
BigEarthNet	83.11
fMoW-S2	85.59
fMoW-S2+BigEarthNet	**87.41**

(f) Data Scale

Table 4. Ablation Analysis of the proposed S2MAE foundation model in terms of positional embedding, decoder depth, reconstruction target, patch size, masking ratio, and data scale, respectively. The best result is shown in bold.

16×16, highlighting the pre-trained model's versatility and efficacy.

Masking ratio. Tab. 4e highlights a key finding: higher masking ratios lead to improved model performance. Unlike the typical 75% masking ratio for RGB images, spectral RS images benefit from a masking ratio of 90% or higher. This aligns with [9]'s hypothesis linking masking ratio to data information redundancy. Spectral RS images have more redundancy, necessitating a higher masking ratio for effective model learning. A 90% masking ratio also enhances pretraining efficiency, reducing memory complexity and speeding up training—an advantageous practical outcome for model pretraining.

Data scale. In Tab. 4f, we analyze the impact of pretraining data, specifically focusing on two datasets: fMoW-S2 and BigEarthNet, both with a standardized input image size of 96×96. The findings highlight how dataset scale and distribution significantly affect model pretraining. fMoW-S2 proves superior to BigEarthNet in pretraining, attributed to its larger dataset and broader geographic coverage. Notably, the concept of continual pretraining, combining both datasets, results in higher mAP scores. This improvement is partially due to the transition from 96×96 images in fMoW-S2 pretraining to 128×128 images in BigEarthNet pretraining, illustrating the positive impact of increasing image size and data scale on overall model efficacy.

Pretraining Schedule. In Fig. 6, we present the fine-tuning results for models trained with varying pre-training epochs. Notably, the models pretrained for just 50 epochs exhibit significant performance gains compared to those trained from scratch. The observed trend in the figure indicates that the models continue to benefit from longer pre-training epochs, suggesting that extended training can further enhance performance.

Model scale. Tab. 5 compares fine-tuning results of ViT-B, ViT-L, and ViT-H, showcasing the mAP performance. ViT-B, equipped with 12 transformer layers with 86 million parameters, achieves an mAP of 85.59, surpassing scratch training by 5.44. ViT-L, featuring 24 layers and 307 million parameters, outperforms ViT-B with an mAP of 86.49, surpassing scratch training by 4.11. ViT-H, with 32 layers and 632 million parameters, achieves an mAP of 88.84, and remarkably, using S2MAE* pre-trained weights, ViT-H attains a state-of-the-art mAP of 90.72, outperforming models

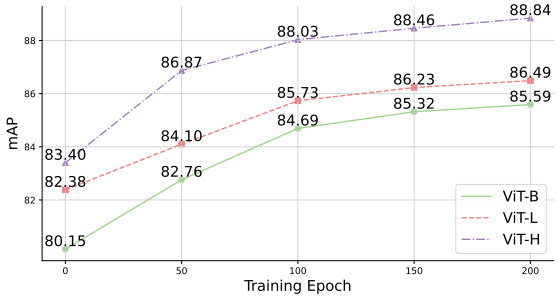

Figure 6. Pretraining schedule. Evaluation of S2MAE performance on BigEarthNet Classification for ViT-B, ViT-L, and ViT-H models. Longer training leads to improved performance.

Network Scale	Params	Pretained Weights	mAP
ViT-Base	86M	Random Init.	80.15
		S2MAE	85.59
		S2MAE*	**87.41**
ViT-Large	307M	Random Init.	82.38
		S2MAE	86.49
		S2MAE*	**88.51**
ViT-Huge	632M	Random Init.	83.40
		S2MAE	88.84
		S2MAE*	**90.72**

Table 5. Performance comparison using different pre-trained models across three ViT-based network scales (i.e., base, large, and huge) on the BigEarthNet dataset.

trained with the entire train set. These findings emphasize effective pretraining strategies and the suitability of larger ViT models for high-accuracy tasks.

5. Conclusion

In this paper, we introduced S2MAE, an extension of MAE for spectral RS imagery pretraining to investigate if the vanilla MAE can effectively learn robust representations by capturing local spectral consistency with minimal inductive bias. S2MAE incorporates a 3D transformer architecture, employing a random masking strategy and integrating learnable spectral-spatial embeddings. Our key observations include: (1) a crucial role of a high masking ratio (90%) for effective pretraining, particularly for highly redundant spectral images; (2) the importance of aligning the masking strategy with spectral properties, where 3D random masking proves more suitable for spectral data; (3) the enhanced performance of progressively pre-trained models using diverse RS datasets. This work aspires to contribute valuable insights to the domain of self-supervised learning in spectral RS imagery.

Limitations. S2MAE utilizes a 3D masking strategy to accentuate local spectral consistency. Nevertheless, there exists a discernible gap in the examination of masking strategies designed specifically for longer spectral sequences. Focusing on the reconstruction target of information in the spectral sequence dimension, may yield richer representations. Despite the computational advantages of a 90% masking ratio during pretraining, the complexity persists at a high level for downstream tasks. Looking ahead, we anticipate the exploration of more effective approaches, such as linearized self-attention, to augment the capabilities of S2MAE in subsequent research endeavors.

Acknowledgement. This work was supported by the National Key Research and Development Program of China under Grant 2022YFB3903401, the National Natural Science Foundation of China under Grant 42241109 and Grant 42271350, and by the MIAI@Grenoble Alpes (ANR-19-P3IA-0003).

References

[1] Kumar Ayush, Burak Uzkent, Chenlin Meng, Kumar Tanmay, Marshall Burke, David Lobell, and Stefano Ermon. Geography-aware self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10181–10190, 2021. 1, 2

[2] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. In *European Conference on Computer Vision*, pages 348–367. Springer, 2022. 2

[3] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6172–6180, 2018. 4

[4] Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems*, 35:197–211, 2022. 1, 2, 4, 5, 6

[5] Rodrigo Caye Daudt, Bertr Le Saux, Alexandre Boulch, and Yann Gousseau. Urban change detection for multispectral earth observation using convolutional neural networks. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 2115–2118. Ieee, 2018. 6

[6] 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. 3

[7] 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, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. 2

[8] 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. *ICLR*, 2021. 3, 5, 6

[9] Christoph Feichtenhofer, Yanghao Li, Kaiming He, et al. Masked autoencoders as spatiotemporal learners. *Advances in Neural Information Processing Systems*, 35:35946–35958, 2022. 2, 3, 6, 7

[10] Agrim Gupta, Jiajun Wu, Jia Deng, and Li Fei-Fei. Siamese masked autoencoders. *arXiv preprint arXiv:2305.14344*, 2023. 2

[11] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020. 2

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5, 6

[13] 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*, pages 9729–9738, 2020. 2

[14] 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. 2, 3, 5

[15] Xin He, Yushi Chen, Lingbo Huang, Danfeng Hong, and Qian Du. Foundation model-based multimodal remote sensing data classification. *IEEE Transactions on Geoscience and Remote Sensing*, 2023. 1

[16] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. 4

[17] Danfeng Hong, Bing Zhang, Hao Li, Yuxuan Li, Jing Yao, Chenyu Li, Martin Werner, Jocelyn Chanussot, Alexander Zipf, and Xiao Xiang Zhu. Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks. *Remote Sensing of Environment*, 299:113856, 2023. 1

[18] Danfeng Hong, Chenyu Li, Bing Zhang, Naoto Yokoya, Jon Atli Benediktsson, and Jocelyn Chanussot. Multimodal artificial intelligence foundation models: Unleashing the power of remote sensing big data in earth observation. *The Innovation Geoscience*, 2(1):100055, 2024. 1

[19] Danfeng Hong, Bing Zhang, Xuyang Li, Yuxuan Li, Chenyu Li, Jing Yao, Naoto Yokoya, Hao Li, Pedram Ghamisi, Xiuping Jia, Antonio Plaza, Paolo Gamba, Jon Atli Benediktsson, and Jocelyn Chanussot. Spectralgpt: Spectral remote sensing foundation model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. DOI:10.1109/TPAMI.2024.3362475. 1, 2, 3

[20] 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. 2

[21] András Kalapos and Bálint Gyires-Tóth. Self-supervised pretraining for 2d medical image segmentation. In *European Conference on Computer Vision*, pages 472–484. Springer, 2022. 3

[22] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 2

[23] Zihan Liu, Genta Indra Winata, and Pascale Fung. Continual mixed-language pre-training for extremely low-resource neural machine translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2706–2718, Online, 2021. Association for Computational Linguistics. 2

[24] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2016. 4

[25] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018. 4

[26] Utkarsh Mall, Bharath Hariharan, and Kavita Bala. Change-aware sampling and contrastive learning for satellite images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5261–5270, 2023. 1

[27] Oscar Manas, Alexandre Lacoste, Xavier Giró-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. 1, 2, 5, 6

[28] Matías Mendieta, Boran Han, Xingjian Shi, Yi Zhu, and Chen Chen. Towards geospatial foundation models via continual pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16806–16816, 2023. 1, 2, 3

[29] Maxim Neumann, Andre Susano Pinto, Xiaohua Zhai, and Neil Houlsby. In-domain representation learning for remote sensing. *arXiv preprint arXiv:1911.06721*, 2019. 5

[30] Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K. Gupta, and Aditya Grover. Climax: A foundation model for weather and climate. In *ICML 2023*, 2023. 1

[31] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2

[32] Colorado J Reed, Xiangyu Yue, Ani Nrusimha, Sayna Ebrahimi, Vivek Vijaykumar, Richard Mao, Bo Li, Shanghang Zhang, Devin Guillory, Sean Metzger, et al. Self-supervised pretraining improves self-supervised pretraining. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2584–2594, 2022. 2, 3

[33] 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*, pages 4088–4099, 2023. 1

[34] Linus Scheibenreif, Michael Mommert, and Damian Borth. Masked vision transformers for hyperspectral image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2165–2175, 2023. 1, 2

[35] Gencer Sumbul, Marcela Charfuelan, Begüm Demir, and Volker Markl. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 5901–5904. IEEE, 2019. 2, 4

[36] Chenxin Tao, Xizhou Zhu, Weijie Su, Gao Huang, Bin Li, Jie Zhou, Yu Qiao, Xiaogang Wang, and Jifeng Dai. Siamese image modeling for self-supervised vision representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2132–2141, 2023. 2

[37] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022. 2

[38] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. 2

[39] Di Wang, Qiming Zhang, Yufei Xu, Jing Zhang, Bo Du, Dacheng Tao, and Liangpei Zhang. Advancing plain vision transformer towards remote sensing foundation model. *IEEE Transactions on Geoscience and Remote Sensing*, 2022. 1, 2

[40] Haotian Yan, Sundingkai Su, Ming Wu, Mengqiu Xu, Yihao Zuo, Chuang Zhang, and Bin Huang. Seamae: Masked pre-training with meteorological satellite imagery for sea fog detection. *Remote Sensing*, 15(16):4102, 2023. 1

[41] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 2

[42] Tong Zhang, Peng Gao, Hao Dong, Yin Zhuang, Guanqun Wang, Wei Zhang, and He Chen. Consecutive pre-training: A knowledge transfer learning strategy with relevant unlabeled data for remote sensing domain. *Remote Sensing*, 14 (22):5675, 2022. 1, 2

[43] Man Zhou, Jie Huang, Keyu Yan, Danfeng Hong, Xiuping Jia, Jocelyn Chanussot, and Chongyi Li. A general spatial-frequency learning framework for multimodal image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1

[44] Qiang Zhou, Chaohui Yu, Hao Luo, Zhibin Wang, and Hao Li. Mimco: Masked image modeling pre-training with contrastive teacher. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4487–4495, 2022. 2