Towards Geospatial Foundation Models via Continual Pretraining

Matías Mendieta[1*] Boran Han[2] Xingjian Shi[3] Yi Zhu[3] Chen Chen[1]
[1] Center for Research in Computer Vision, University of Central Florida
[2] Amazon Web Services [3] Boson AI
matias.mendieta@ucf.edu boranhan@amazon.com xshiab@connect.ust.hk
yi@boson.ai chen.chen@crcv.ucf.edu

Abstract

Geospatial technologies are becoming increasingly essential in our world for a wide range of applications, including agriculture, urban planning, and disaster response. To help improve the applicability and performance of deep learning models on these geospatial tasks, various works have begun investigating foundation models for this domain. Researchers have explored two prominent approaches for introducing such models in geospatial applications, but both have drawbacks in terms of limited performance benefit or prohibitive training cost. Therefore, in this work, we propose a novel paradigm for building highly effective geospatial foundation models with minimal resource cost and carbon impact. We first construct a compact yet diverse dataset from multiple sources to promote feature diversity, which we term GeoPile. Then, we investigate the potential of continual pretraining from large-scale ImageNet-22k models and propose a multi-objective continual pretraining paradigm, which leverages the strong representations of ImageNet while simultaneously providing the freedom to learn valuable in-domain features. Our approach outperforms previous state-of-the-art geospatial pretraining methods in an extensive evaluation on seven downstream datasets covering various tasks such as change detection, classification, multi-label classification, semantic segmentation, and super-resolution. Code is available at https://github.com/mmendiet/GFM.

1. Introduction

The significance of geospatial technologies has progressively increased for various applications worldwide. Progress in this domain can substantially improve our ability to understand the earth and how we interact with it. With the rising popularity of foundation models in vision and natural language, researchers have begun to investigate applying such principles to the geospatial domain in order to enhance the suitability of deep learning models in downstream tasks [29, 28, 9, 2]. In the literature, various works have explored two prominent approaches for introducing pretrained foundation models in geospatial applications. The first obvious approach is to leverage existing foundation models from the natural image domain, like those trained on the large-scale ImageNet-22k dataset [11]. In practice, this is done by *directly finetuning publicly-available ImageNet pretrained models on the downstream tasks*. This approach has the advantage of being straight-forward, as ImageNet models can be simply downloaded from many open-source model zoos, and has been shown to be effective [29, 30]. However, due to the domain gap between natural images

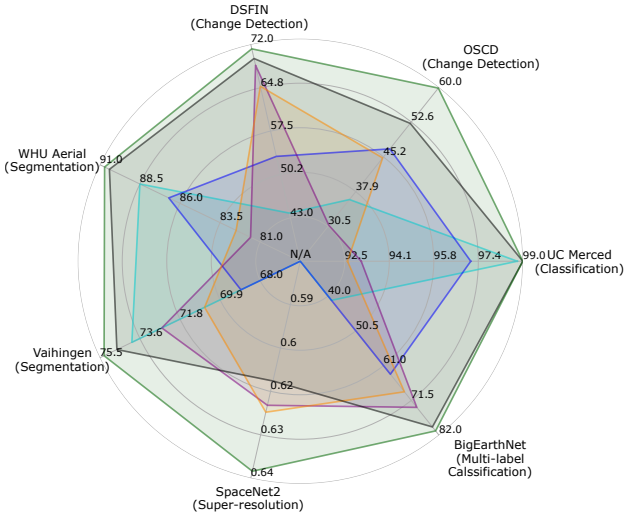

Figure 1. Our geospatial foundation model (GFM) achieves favorable performance on a broad set of tasks in comparison to other state-of-the-art geospatial pretraining methods (SeCo [28], SatMAE [9]) and ImageNet supervised pretraining baselines. Legend is as follows. Cyan: ImageNet-1k Supervised (ResNet50), Blue: SeCo [28], Purple: ImageNet-22k Supervised (ViT), Orange: SatMAE [9], Gray: ImageNet-22k Supervised (Swin), Green: GFM (ours).

*Work done as an intern at Amazon Web Services

and remote sensing, this approach is not optimal for geospatial data, and still leaves performance gains on the table.

In recent years, a second approach has gained significant traction, where researchers aim to pretrained models specific to the geospatial domain [28, 2, 9, 38]. These methods typically *train a network from scratch on a large corpus of remote sensing imagery* to learn in-domain representations transferable to downstream tasks. Unfortunately, this can require a significant amount of data and training time to achieve good performance, especially when employing large state-of-the-art (SOTA) transformer models. For instance, the current SOTA in geospatial foundation models, SatMAE [9], requires 768 hours on a V100 GPU for training a vision transformer [14]. This has substantial cost associated with producing the model, not just in terms of time and computation but also environmentally, with a total estimated carbon footprint of 109.44 kg CO_2 equivalent. Additionally, the final performance of such models are not consistently better across various tasks than simply utilizing publicly-available ImageNet pretrained models (Section 4), despite the high resource expense.

In this work, we propose to investigate a different paradigm for producing more effective geospatial foundation models with substantially less resource costs. First, we begin with a discussion on pretraining data selection, and ultimately construct a concise yet diverse collection of data from various sources to promote feature diversity and effective pretraining. Second, rather than following the aforementioned typical approaches, we investigate the potential of ***continual pretraining for the geospatial domain*** from readily-available ImageNet models. Continual pretraining has been practiced in the NLP domain with success in various works [16, 17, 26]. In this paradigm, existing foundation models are further improved for a specific domain or task through a secondary pretraining stage. This new single model can now be fine-tuned on the various downstream tasks in that domain. In principle, we reason that such a paradigm has the potential to boost performance by utilizing large-scale ImageNet representations as a base on which stronger geospatial foundation models can be built. Furthermore, such natural image models are constantly being improved and released by the general computer vision community, providing a consistent source of better baseline models. Therefore, an approach that could enable the geospatial domain to leverage these improvements with minimal resource needs and carbon footprint paves the way for continual, sustainable benefits for the geospatial community.

However, when we initially experiment with the standard continual pretraining formulation, we find it provides only marginal benefits (Section 3.2). Instead, we discover that utilizing ImageNet representations as an auxiliary distillation objective during pretraining leads to a stronger geospatial foundation model. Building upon this principle, we propose a multi-objective continual pretraining paradigm that significantly enhances performance while requiring minimal resources. Our approach leverages ImageNet's powerful representations to facilitate and expedite learning, while also enabling the acquisition of valuable in-domain features via self-supervised learning on geospatial data. Furthermore, our proposed Geospatial Foundation Model (GFM) exhibits strong performance, surpassing previous state-of-the-art (SOTA) methods across a diverse range of downstream tasks (Section 4). Our contributions are as follows:

- We investigate a novel paradigm for creating highly effective geospatial models with minimal resource costs. Our methodology begins with data selection and construction of a compact yet diverse dataset from multiple sources to promote feature diversity and enhance pretraining effectiveness, which we term GeoPile. We further explore the potential of continual pretraining from ImageNet models, but find it is not satisfactory in its standard formulation.

- Therefore, to achieve better performance with minimal resource needs, we propose a multi-objective continual pretraining paradigm. Our design is surprisingly simple yet effective, constructed as a teacher-student strategy with both a distillation objective and self-supervised masked image modeling. This approach allows GFM to leverage the strong representations of ImageNet to guide and quicken learning, while simultaneously providing the freedom to learn valuable in-domain features.

- We evaluate our GFM approach, as well as several baseline and SOTA methods, on 7 datasets covering important geospatial applications such as change detection, classification, multi-label classification, semantic segmentation, and super-resolution. Overall, our GFM performs favorably over previous methods (as shown in Figure 1).

2. Related Work

Geospatial Pretraining. Various works have experimented with employing supervised or self-supervised pretraining paradigms in the geospatial domain. The classical work of [29], and more recent paper [38], investigate supervised pretraining on individual datasets of various sizes. Interestingly, these still often found the ImageNet pretrained models to perform very well, particularly with vision transformers [14, 25]. Other works have explored self-supervised learning paradigms for remote sensing, primarily focused on contrastive methods. [28] and [2] employ a MoCo [7] style objective using spatially aligned but temporally different images as the positive pairs. [23] and [20]

also utilize a MoCo-inspired objective, but specify a cropping procedure to generate positives and negatives within and across images. [37] employs a colorization objective on Sentinel-2 imagery utilizing the various spectral bands. Most recently, SatMAE [9] explores the use of masked image modeling to train a large ViT model. This work is similar in some respect to ours, as we also train a transformer model with an MIM objective. However, we find that SatMAE often does not perform better than the off-the-shelf ImageNet-22k pretrained ViT (Section 4). This indicates both the difficulty of building strong geospatial pretrained models from scratch and highlights the potential usefulness of leveraging continual pretraining instead, as we investigate in this work.

Masked Image Modeling. Masked image modeling (MIM) has been proposed in various forms in recent years, and has recently been found to be particularly effective in the natural image domain, surpassing many contrastive works and being shown to be friendlier to downstream optimization [41, 18, 44, 3, 40] In general, the goal is to learn from data in a self-supervised manner by asking the model to generate pixel values for intentionally-withheld regions in an image. [32] is an early work with an aim of learning strong visual representations through inpainting masked regions. In [6], Chen et. al train a large transformer to predict pixels autoregressively. After the introduction of vision transformers (ViT) [14], many works continued to improve various MIM variants. [3] and [44] take inspiration from BERT [12] in natural language processing, and tokenize the image patches with either a pretrained model or jointly trained online tokenizer, with the objective being to reconstruct at a token-level rather than raw pixels. Recently, [41] and [18] show that a masked image modeling task of simply regressing directly on the image pixels is sufficient and effective. In this work, we leverage the framework from [41], as it is compatible with hierarchical transformer architectures [25].

In this work, we develop our pretraining objective based on a masked image modeling approach like [41, 18]. Exploration of the masked image modeling framework in geospatial applications is still in its early stages, and could help alleviate some concerns with contrastive approaches in this domain. Particularly, the choice of augmentations with contrastive methods can be quite difficult, as common selections such as greyscale, color jitter and others that heavily affect the intensity of the image can instill undesirable invariances [29]. On the other hand, MIM objectives like [41, 18] rely only on simple spatial augmentations such as flipping and cropping. Furthermore, a common remote sensing application is that of change detection, which requires a model to detect changes in two images from the same location but at different times. In order to still be effective on this task, works that use contrastive approaches on temporal positives introduce various design choices. For instance, SeCo [28] creates multiple feature subspaces during pretraining, each one invariant to a separate form of augmentation. [1] also employs temporal positives, but instead chooses the sampling locations for the pretraining data to ensure that image pairs contain primarily natural illumination and viewing angle variant, without major changes such as new urban developments.

Continual Pretraining. Continual pretraining has been primarily introduced in the natural language domain [16, 17, 26], in order to improve large language models (LLM). [16] illustrates the viability of two additional stages of pretraining, using in-domain data (domain-adaptive), and then even further using task-specific data (task-adaptive). [17] proposes a continual training paradigm for enabling temporal reasoning abilities to pretrained language models. [26] focus on using continual pretraining to enable mixed language neural machine translation. In the vision domain, [22] employs a BYOL [15] style continual pretraining paradigm for 2D medical image segmentation. [34] explores a hierarchical pretraining approach for task adaptation. However, they primarily focus on adapting to a specific downstream task at a time, employing three training stages on top of an existing pretrained model for each task individually. In contrast, we employ one efficient in-domain pretraining setting that can generalize to many downstream tasks, as illustrated in Section 4. Furthermore, rather than directly loading the pretrained weights from existing models as initialization, we find instead that leveraging the representations as an auxiliary distillation objective during the pretraining process enables learning stronger representations.

3. Methodology

In the following sections, we discuss the pretraining data selection (Sec, 3.1), investigate vanilla continual pretraining (Sec. 3.2), and present our GFM method (Sec. 3.3).

3.1. Pre-training Data Selection

A particularly common choice of source data among geospatial contrastive pretraining works is Sentinel-2 imagery [28, 1, 37] due to its large corpus of available data and ease of access. Therefore, to begin our study, we first gather a pretraining dataset of 1.3 million Sentinel-2 images using the sampling technique from [28]. After gathering the Sentinel-2 data, we employ it to pretrain a Swin-B [25] model with the masked image modeling (MIM) objective from [41]. We then finetune and evaluate this model on a wide variety of downstream datasets to get a broad understanding of its performance potential in many tasks (see Section 4 for task details). For a comparison, we finetune the ImageNet-22k pretrained Swin-B from the official Swin Transformer repository [25] on all downstream tasks as a

Figure 2. We visualize some example images from the pretraining datasets with Sentinel-2 (left) and GeoPile (right). Sentinel-2 has noticeably much lower feature diversity within a single image and across images than that of our GeoPile pretraining dataset.

baseline. In order to compare these models across all tasks, we introduce an average relative performance metric (ARP) in which we take the relative difference on each task with respect to the ImageNet-22k baseline, and then average that difference:

$$\text{ARP}(M) = \frac{1}{N}\sum_{i=1}^{N}\frac{\text{score}(M, \text{task}_i) - \text{score}(\text{baseline}, \text{task}_i)}{\text{score}(\text{baseline}, \text{task}_i)}. \quad (1)$$

Here "baseline" is the Swin-B model pretrained on ImageNet-22k, as mentioned above. M denotes the model for performance evaluation, and N is the number of tasks. There are 7 tasks used in Section 4 covering important geospatial applications such as classification, multi-label classification, semantic segmentation, change detection, and super-resolution. The reported ARP value is scaled by 100 to show as a percentage.

We compare these two models in Table 1. Interestingly, we find that the Sentinel-2 model performs poorly on downstream tasks compared to the ImageNet-22k baseline. To investigate further, we visualize multiple samples from Sentinel-2 in the left columns of Figure 2. Upon inspection, we note that the feature diversity within a single image and across images of Sentinel-2 is perceivably low. To further quantify this suspicion, we calculate the average image entropy over a randomly sampled set of 3000 images from the collected Sentinel-2 data as well as the typical ImageNet dataset as a baseline. Overall, the Sentinel images have an average entropy of 3.9 compared to 5.1 of ImageNet. Such an evaluation provides insights into the potential pitfalls of Sentinel-2 data in pretraining transformers. For MIM objectives, training data with a substantially lower entropy can make for an easier reconstruction task, since masked regions may be more similar to their neighbors. Therefore, the network does not have to work as hard to fill in the blanks, limiting the learning potential. Overall, these result indicate that the noticeably narrow scope of features and limited per-sample information in Sentinel-2 data may be limiting the potential of the pretrained model.

Table 1. Dataset Analysis. To evaluate each method, we finetune the pretrained model on seven different tasks, outlined in Section 4 and report the ARP metric defined in Equation 1. We also report the training time in hours on a V100 GPU, as well as the carbon impact estimations[2] in kg CO_2 equivalent [24]. Overall, our collected GeoPile pretraining dataset significantly improves downstream performance. † indicates the vanilla continual pretraining approach of initializing the model with ImageNet-22k weights prior to conducting MIM training on GeoPile. To further improve the performance in an efficient manner, we introduce our continuous pretraining paradigm GFM.

Method	# Images	Epochs	ARP ↑	Time ↓	CO_2 ↓
ImageNet-22k Sup.	14M	-	0.0	-	-
Sentinel-2 [28]	1.3M	100	-5.83	155.6	22.2
GeoPile	600k	200	0.92	133.3	19.0
GeoPile†	600k	200	1.24	133.3	19.0
GeoPile†	600k	800	1.45	533.2	76.0
GFM	600k	100	3.31	93.3	13.3

Table 2. Breakdown of datasets in the GeoPile. We gather approximately 600k samples from a combination of labeled and unlabeled satellite imagery with various ground sample distances and scenes.

Dataset	# Images	GSD	# Classes
NAIP [31]	300,000	1m	n/a
RSD46-WHU [27]	116,893	0.5m - 2m	46
MLRSNet [33]	109,161	0.1m - 10m	60
RESISC45 [8]	31,500	0.2m - 30m	45
PatternNet [45]	30,400	0.1m - 0.8m	38

Therefore, we set out to collect a diverse geospatial pretraining dataset. Sourcing from both labeled and unlabelled data, we form a new pretraining dataset which we term GeoPile. The breakdown of GeoPile is shown in Table 2. For textural detail, we ensure a variety of ground sample distances (GSD), including images with much higher resolution than Sentinel-2 (which has a GSD of 10m). Furthermore, the selected labeled datasets encompass a wide variety of classes from general remote sensing scenes, ensuring visual diversity across samples. We calculate the average entropy of our GeoPile dataset, and find it to be 4.6, much higher than that of Sentinel-2. Furthermore, the textural and visual diversity is qualitatively evident in Figure 2. In Table 1, the enhancing effect of the data selection is clearly shown by the substantial performance increase.

3.2. Vanilla Continual Pretraining

Next, after establishing our pretraining data selection, we investigate an alternate pretraining paradigm that bridges the gap between the two common approaches mentioned

[2] CO_2 estimations were completed with mlco2.github.io from [24].

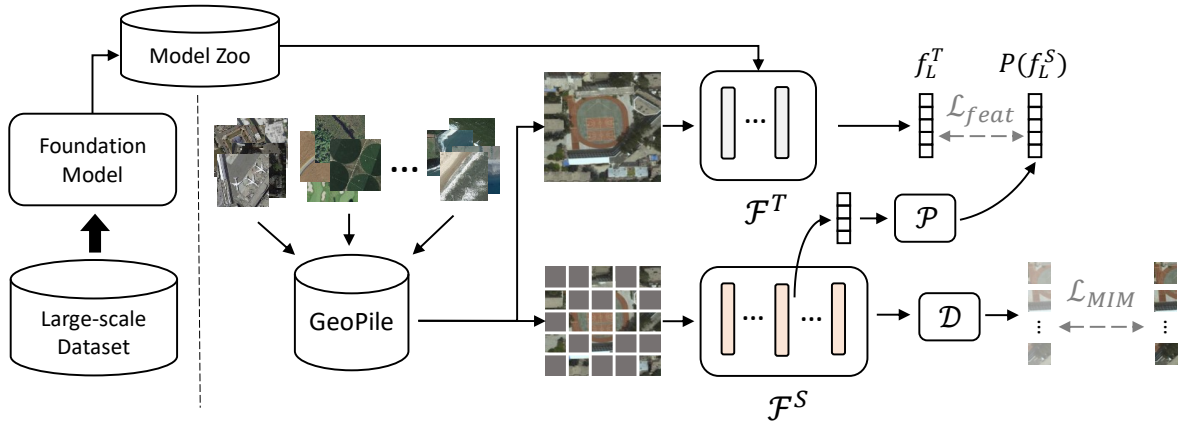

Figure 3. Our GFM continual pretraining pipeline, which leverages publicly-available large-scale models in concert with our compiled geospatial dataset and pretraining objective. First, we select a concise set of data from various sources, which we term GeoPile (Section 3.1). Next, we train GFM with our multi-objective continual pretraining approach. Our GFM framework is constructed as a teacher-student paradigm, with two parallel model branches. The teacher $\mathcal{F}^T$ is initialized with ImageNet-22k weights (top) and frozen during training. The student $\mathcal{F}^S$ is initialized from random initialization (bottom), and is trained to serve as the final geospatial foundation model. In a continual pretraining fashion, we leverage the intermediate features of an ImageNet-22k pretrained model to guide and quicken learning. Furthermore, we build in an MIM objective on the student branch to learn valuable in-domain features directly from the geospatial data.

in Section 1. Specifically, we investigate the potential of continual pretraining in the context of geospatial pretrained models. To do so, we first employ the vanilla continual pretraining approach; that is, using the ImageNet-22k weights as initialization prior to beginning the pretraining step with GeoPile. We find this to be helpful in improving performance over starting from scratch. This validates the possibility of continual pretraining as a beneficial paradigm to provide performance gain without additional resource costs. Nonetheless, the improvement is still limited, with $\sim$0.3% ARP increase over starting from scratch and $\sim$1.24% ARP over the baseline.

To further improve the performance of our pretrained model in comparison to the ImageNet-22k baseline, we increase the number of pretraining epochs in the next row of Table 1. While we are able to make improvements, this comes at the cost of substantially more computational cost and carbon footprint for marginal gain. Therefore, we ask the question: how can we significantly improve the performance further while maintaining minimal compute and carbon footprint overhead? To this end, we propose a simple and efficient approach for building geospatial pretrained models capable of strong downstream performance.

3.3. GFM Pretraining

A significant number of geospatial foundation model studies disregard the existing large-scale model representations. This is far from ideal, particularly for large transformer models known to require a vast amount of data and compute power to train. Instead, we reason that the valuable knowledge available in models like those trained on ImageNet-22k should be leveraged to produce strong performance with minimized overhead. To this end, we propose an unsupervised multi-objective training paradigm for effective and efficient pretraining of geospatial models, illustrated in Figure 3.

There are two main components in our framework. First, we randomly initialize an encoder $\mathcal{F}^S$ and decoder $\mathcal{D}$ set up for MIM as in [41]. During training, the input is randomly masked, and the network attempts to reconstruct the image at the output. This MIM objective is enforced with an L1 loss [41]:

$$\mathcal{L}_{MIM} = \frac{\|\mathbf{O}_\kappa - \mathbf{G}_\kappa\|_1}{N}, \tag{2}$$

where $\mathbf{O}_\kappa$ are the original pixel values from κ masked regions, $\mathbf{G}_\kappa$ are the generated reconstructions for those regions, and N is the total number of masked pixels.

For the continual pretraining of our framework, we initialize a second encoder branch $\mathcal{F}^T$ up to a chosen stage L and load the ImageNet-22k pretrained weights. This branch behaves as a form of teacher during the training process to the student branch ($\mathcal{F}^S$), which will serve as our final model. For the ImageNet teacher, we freeze the weights, to both ensure that the structured representations are maintained during the training process, and also reduce the computation required during optimization.

Rather than using the masked input as in the student branch, the teacher receives the unmasked image as input, and provides a feature output f_L^T at stage L. This feature has access to the full context of the input, enabling it to capture informative representations. We utilize this feature to guide the representations of the student, and form a secondary objective with the cosine similarity between branch

features:

$$\mathcal{L}_{feat} = -\frac{P(f_L^S)}{\left\|P(f_L^S)\right\|_2} \cdot \frac{f_L^T}{\left\|f_L^T\right\|_2}, \tag{3}$$

where f_L^S and f_L^T are the intermediate features of the student and teacher branches at stage L, and $\mathcal{P}$ is an linear projection layer. Therefore, the final loss during training is simply the summation of these objectives:

$$\mathcal{L} = \mathcal{L}_{MIM} + \mathcal{L}_{feat}. \tag{4}$$

This training paradigm enables an ideal two-fold optimization. Distillation from the intermediate features of the teacher ensure that the student can benefit from the teacher's diverse knowledge, learning more in less time. Furthermore, the student is simultaneously given freedom to adapt to in-domain data through its own pretraining objective, gathering new features to improve performance.

We analyze the ARP and resource cost of this approach in Table 1. Notably, our GFM is able to achieve better overall performance with substantially less computation and emissions impact compared to vanilla continual pretraining with the same dataset, illustrating that our multi-objective continual pretraining paradigm is an effective method for training these models. Comparatively, the SOTA geospatial pretrained method SatMAE [9] requires 768 hours on a V100 GPU and 109.44 kg equivalent CO_2 according to their reported results. Therefore, GFM enables more than 8× reduction in total training time and carbon impact. Moreover, we find that the performance of SatMAE is often not superior to the off-the-shelf ImageNet-22k pretrained ViT (Section 4). This implies that building powerful geospatial pretrained models from scratch is challenging and further underscores the benefits of utilizing continual pretraining instead. We show these results in the following section.

4. Experiments

To verify the effectiveness of our model in detail, we conduct experiments on seven geospatial datasets of various tasks including change detection (Section 4.1), classification (Section 4.2), segmentation (Section 4.3), and super-resolution (Section 4.4).

For pretraining, we employ 8 NVIDIA V100 GPUs with a batch size of 2048 (128 per GPU) and the image size of 192×192. All pretraining settings are the same as in [41]. For downstream tasks, 4 NVIDIA A10G GPUs are employed. During the pretraining stage, we utilize RGB bands as they are most commonly available among data sources and tasks. For downstream tasks with additional band inputs, we initialize the RGB patch embeddings with the pretrained weights and randomly initialize the remaining channels. Potentially improving performance even further though the employment of additional data modalities will be an intriguing avenue for future research. Additional training details for these tasks are provided in the *supplementary material*.

Table 3. Onera Satellite Change Detection Results

Method	Precision ↑	Recall ↑	F1 ↑
ResNet50 (ImageNet-1k) [19]	**70.42**	25.12	36.20
SeCo [28]	65.47	38.06	46.94
MATTER [1]	61.80	57.13	59.37
ViT (ImageNet-22k) [14]	48.34	22.52	30.73
SatMAE [9]	48.19	42.24	45.02
Swin (random)[25]	51.80	47.69	49.66
Swin (ImageNet-22k)[25]	46.88	59.28	52.35
GFM	58.07	**61.67**	**59.82**

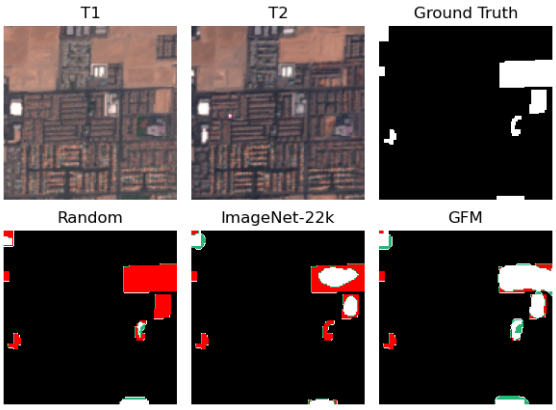

Figure 4. Qualitative results of downstream performance on OSCD comparing our GFM with ImageNet-22k and randomly initialized baselines. White, green, red colors show true positive, false positive, and false negative respectively.

Table 4. DSFIN Change Detection Results

Method	Precision ↑	Recall ↑	F1 ↑
ResNet50 (ImageNet-1k) [19]	28.74	**92.07**	43.80
SeCo [28]	39.68	81.02	53.27
ViT (ImageNet-22k) [14]	70.77	66.34	68.49
SatMAE [9]	70.45	60.29	64.98
Swin (random)[25]	57.97	62.06	59.94
Swin (ImageNet-22k)[25]	67.11	72.33	69.62
GFM	**74.83**	67.98	**71.24**

4.1. Change Detection

Change detection is a particularly important remote sensing task, helping us understand how humans interact with our planet over time, and natural phenomena that change our planet's landscape. We conduct experiments on both the Onera Satellite Change Detection (OSCD [5]) in Table 3 and DSIFN [43] in Table 4.

OSCD consists of 14 image pairs extracted from various regions around the world within a three year period of 2015

to 2018. The images are taken from Sentinel-2 with GSDs ranging from 10m to 60m, and split into 14 images for training and 10 for evaluation. The annotations indicate whether the change has occurred on a pixel level, and focus primarily on urban developments. Similarly, we also test our method on DSIFN dataset. This dataset contains high-resolution imagery, such as WorldView-3 and GeoEys-1 [43]. This dataset contains 3490 high resolution samples for training and 48 images for evaluation respectively. Every pair of images from a given location at two different timestamps will be fed into the swin encoder [25] for feature extraction. The difference between the features from each pair is computed and fed into an UPerNet [39] to generate the final binary segmentation masks [28, 4]. The encoder is initialized with the pretrained weights.

For both datasets, we report the precision, recall, and F1 score on the "change" class. As the results presented from OSCD (Table 3 and Figure 4) and DSIFN (Table 4), GFM shows a consistent improvement over the ImageNet-22k baseline across both datasets. Notably, SatMAE is able to improve over its ImageNet-22k baseline on OSCD, but lags behind on DSIFN. This further highlights the difficulty of training large vision transformers from scratch that can perform consistently across different GSDs.

4.2. Classification

Another common remote sensing application is that of classification. We evaluate two datasets common in the literature [28, 1]: UC Merced Land Use Dataset [42] and BigEarthNet [36]. The UC Merced Land Use Dataset is a classic dataset in the remote sensing field. It contains 21 classes, each with 100 images at 256x256 pixels and an approximate GSD of 1 foot. We split the data into train and validation according to [13]. BigEarthNet [36] (BEN) is a large-scale remote sensing dataset for multi-label classification. The data consist of 12-band Sentinel-2 images with sizes of 120x120, 60x60, and 20x20 pixels for the bands at 10m, 20m, and 60m GSDs, respectively. We employ the data split and 19 class evaluation as common in the literature [29, 28, 9].

In Table 5, we report the classification accuracy on UC Merced (UCM) and mean average precision results on BigEarthNet (BEN) for all methods. On UC Merced, we note the SeCo [28] pretrained model performs significantly worse than its ImageNet-1k pretrained counterpart with ResNet-50. These two datasets are very different in both classes, satellite source, and GSDs, and therefore having a diverse feature knowledge is imperative to maintaining performance despite these distinctions. Our model can provide robust performance in both cases by leveraging ImageNet representations and remote sensing data in its learning. Furthermore, one key motivation for training a geospatial foundation model is to improve the sample efficiency for downstream tasks. Notably, we find that our model maintains strong performance on BigEarthNet, even when only given 1% of the training data.

Table 5. UC Merced classification accuracy and BigEarthNet multi-label classification mean average precision results.

Method	UCM	BEN 10%	BEN 1%
ResNet50 (ImageNet-1k) [19]	98.8	80.0	41.3
SeCo [28]	97.1	82.6	63.6
ViT (ImageNet-22k)[14]	93.1	84.7	73.6
SatMAE [9]	92.6	81.8	68.9
Swin (random)[25]	66.9	80.6	65.7
Swin (ImageNet-22k) [25]	**99.0**	85.7	79.5
GFM	**99.0**	**86.3**	**80.7**

4.3. Segmentation

Segmentation is a popular remote sensing application for enabling automated extraction of building footprints or land cover mappings over wide regions. We therefore conduct experiments on this task on two different datasets. Vaihingen [35] is an urban semantic segmentation dataset collected over Vaihingen, Germany at a GSD of 0.9m. We employ the data split implemented in the MMSegmentation library [10] for our experiments, with 344 training and 398 for validation, all with an image size of 512x512 pixels. The WHU Aerial building [21] dataset is sampled over Christchurch, New Zealand at a GSD of 0.3m. Image tiles are provided at 512×512 pixels, split into 4736 for training and 2416 for evaluation.

We report the intersect of union (IoU) segmentation results for all methods in Table 6. ImageNet pretrained models are notably strong performers in all cases. On both datasets, SeCo lags substantially behind its ImageNet counterpart. Interestingly, SatMAE is able to bring improvement over ImageNet-22k on WHU, but fails to do so to a larger degree on Vaihingen. However, our approach is able to leverage the already strong ImageNet-22k representations and guide them towards the geospatial domain, resulting in overall improvement.

4.4. Super-resolution

In the previous experiments, we evaluated several common high-level tasks. Nonetheless, the low-level task of super-resolution is also important in the geospatial domain. For this task, we re-purpose the SpaceNet2 dataset, which contains 10,593 8-band images from four cities: Las Vegas, Paris, Shanghai, and Khartoum. The data is provided at both a GSD of 1.24m (multi-spectral, 162x162 pixels) and 0.3m (pan-sharpened multispectral, 650x650 pixels). We formulate a super-resolution task, taking as input the 1.24m multi-spectral images and generating the 0.3m pan-sharpened equivalent. We evaluate the super-resolution per-

Table 6. Results on the WHU Aerial and Vaihingen segmentation datasets. We finetune all methods for 40k iterations, and report the IoU for the building class on WHU and mean IoU (mIoU) across the 6 classes (impervious surface, building, low vegetation, tree, car, clutter) of Vaihingen.

Method	WHU Aerial	Vaihingen
ResNet50 (ImageNet-1k) [19]	88.5	74.0
SeCo [28]	86.7	68.9
ViT (ImageNet-22k) [14]	81.6	72.6
SatMAE [9]	82.5	70.6
Swin (random) [25]	88.2	67.0
Swin (ImageNet-22k) [25]	90.4	74.7
GFM	**90.7**	**75.3**

Table 7. SpaceNet2 Super-resolution Results. Notably, while SatMAE fails to enhance its baseline (ViT ImageNet-22k), our method exhibits substantial improvement over its respective baseline (Swin ImageNet-22k) in both PSNR and SSIM.

Method	PSNR ↑	SSIM ↑
ViT (ImageNet-22k)[14]	**23.279**	0.619
SatMAE [9]	22.742	0.621
Swin (random) [25]	21.825	0.594
Swin (ImageNet-22k) [25]	21.655	0.612
GFM	22.599	**0.638**

formance of our model and several baselines with the peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) in Table 7. The ViT-L ImageNet-22k model and our model are among the best in terms of PSNR and SSIM, respectively. Interestingly, SatMAE is not able to improve over its baseline. On the other hand, our method improves considerably over its ImageNet-22k baseline.

5. Ablation Studies

We perform multiple ablation studies on the choice of distillation stage, student initialization, training objectives, the pretraining dataset components. Further detailed results and discussions are provided in the *supplementary material*.

5.1. Distillation Stage

When implementing our feature map distillation objective, a natural question is at which point should the mapping take place. We experiment with different locations by stage in the Swin transformer and calculate the corresponding ARP in Figure 5. Overall, performing the distillation after Stage 3 yields the highest ARP. Hence, we employ this scheme for all downstream experiments. This result is also intuitively expected; distilling at Stage 3 gives a large portion of the model the supervisory signal from the teacher, while still allowing for purely domain-specific feature learning in the final layers.

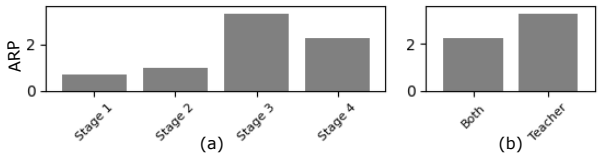

Figure 5. a) Distillation stage ablation results. b) Student initialization ablation results. "Both" indicates that the teacher and student branches are initialized with ImageNet weights prior to geospatial pretraining. "Teacher" indicates that just the teacher branch is initialized, as described in Section 3.3.

Table 8. GeoPile pretraining dataset ablation. We remove each dataset individually from GeoPile and report the number of images remaining and resulting ARP. The row "w/o curated datasets" removes all data other than NAIP imagery.

Data	# Images	ARP ↑
w/o WHU-RSD46	444,061	1.77
w/o MLRSNet	451,793	2.17
w/o Resisc45	529,454	1.57
w/o PatternNet	557,554	1.79
w/o curated datasets	300,000	0.53
w/o NAIP	260,954	1.50

5.2. Student Initialization

In our proposed framework, we maintain the teacher model frozen with ImageNet pretrained weights, and randomly initialize the student. Another alternative is to initialize the student also with ImageNet weights prior to beginning the geospatial pretraining process. However, as shown in Figure 5, this is not the most optimal option. Such initialization is unnecessary in our framework, since it already allows for seamless integration of ImageNet representations with valuable in-domain features. Forcibly doing so likely introduces too much bias towards the natural image representations. Therefore an unbiased student is most ideal and effective.

5.3. GeoPile Pretraining Dataset

To ablate components of the GeoPile, we remove each dataset individually to see its relative importance. Also, we compare using just the labeled data portion and using just the unlabeled NAIP imagery portion. As expected, using just data from labeled datasets gives better performance with less images than using just images gathered from just NAIP. The human-curated samples in these datasets are more likely to contain relevant objects and features, as they each correspond to a particular class of interest. Still, unlabeled data like NAIP can be sourced easily and with scale. Further scaling of both labeled and unlabeled portions could further improve performance; however, it will also increase the training time and sustainability impact. Therefore, we maintain GeoPile at approximately 600,000 images.

Table 9. Ablation results for the training objectives in GFM. For w/o teacher, we only conduct MIM with GeoPile. For w/o MIM, we simply perform the distillation objective from the ImageNet-22k model to our student model with GeoPile. We abbreviate the following for horizontal space: UC Merced (UCM), BigEarthNet (BEN), WHU Aerial (WHU), Vaihingen (Vai), SpaceNet2 (SN2).

Method	OSCD (F1)	DSFIN (F1)	UCM	BEN 10%	BEN 1%	WHU	Vai.	SN2 (PSNR)	SN2 (SSIM)
w/o teacher	57.3	67.65	98.8	**86.5**	80.0	90.5	74.0	22.509	0.631
w/o MIM	59.58	**71.86**	98.8	86.1	80.2	90.2	72.6	22.069	0.608
GFM	**59.82**	71.24	**99.0**	86.3	**80.7**	**90.7**	**75.3**	**22.599**	**0.638**

Table 10. Results for employing temporal pairs and datasets from SeCo [28] in our multi-objective pretraining framework. TP indicates that the teacher receives one image from a temporal pair, and the student receives the other. SI indicates that the same image is inputted to the teacher and student.

Dataset	Inputs	OSCD (F1)	DSFIN (F1)	UCM	BEN 10%	BEN 1%	WHU	Vai.	SN2 (PSNR)	SN2 (SSIM)
SeCo 100k [28]	TP	57.03	62.48	80.0	80.6	68.6	88.3	66.3	22.078	0.572
SeCo 100k [28]	SI	58.41	67.92	92.1	83.9	76.5	88.8	68.1	22.439	0.602
SeCo 1M [28]	SI	58.87	69.41	95.7	86.2	77.1	89.6	71.0	22.281	0.626
GeoPile	SI	**59.82**	**71.24**	**99.0**	**86.3**	**80.7**	**90.7**	**75.3**	**22.599**	**0.638**

5.4. Multi-objective Ablation.

To delve deeper into the evaluation of GFM's performance, we extend our analysis by conducting experiments in which we exclude the teacher component and MIM component individually, as detailed in Table 9. We find that training with the multi-objective approach is the best performer overall. This shows that the integrated distillation and MIM objectives within the GFM framework both contribute to producing a well-balanced mode for downstream tasks, and are important aspects of efficient and effective geospatial learning.

5.5. Temporal Pairs Experiment

Some works employ temporal pairs in the pretraining procedure [28, 2, 1], meaning two satellite images from the same spatial region but taken at different times. We also experiment with the use of temporal positives in our training paradigm using the dataset proposed in SeCo [28]. In this case, the teacher receives one image from a temporal pair, and the student receives the other. The temporal changes can possibly serve as a form of natural augmentation for the distillation objective. However, as shown in Table 10, we find that using temporal positives (TP) is worse than simply using the same image (SI) for both branches. Therefore, we simply use the same image for both branches for other experiments. We further scale up the data by employing the 1M sample Sentinel-based dataset from SeCo. Nonetheless, GeoPile proves to be more effective as a pretraining data source for our GFM.

6. Conclusion

In summary, this paper investigates an alternative paradigm from previous work towards producing better geospatial foundation models with substantially less resource cost. To this end, we first construct a concise yet diverse collection of data from various remote sensing sources for pretraining. Second, we propose a surprisingly simply yet effective multi-objective continual pretraining paradigm, in which we leverage the strong representations of ImageNet-22k to guide and quicken learning, while simultaneously providing the freedom to learn valuable in-domain features through self-supervised learning on geospatial data. We hope that our GFM approach will serve as an example to inspire other works in investigating efficient and sustainable methods for developing geospatial foundation models.

Broader Impact and Limitations. As the geospatial community continues to innovate, the resulting impact promises to positively benefit both the earth and society. Automating the process of extracting useful information from geospatial data can aid scientists, engineers, and others to make data-informed decisions on infrastructure advancement, food supply improvements, and natural disaster response. A potential limitation of our GFM approach is that it may still be somewhat constrained by the performance of the ImageNet-22k model. If perhaps a model was trained from scratch on an extremely large corpus of remote sensing data, the performance may eventually also lead to improved performance over ImageNet baselines. However, this would incur a substantial amount of training time and CO_2 impact. Furthermore, as mentioned in Section 1, natural image models are constantly being improved and released by the general computer vision community. Therefore, our approach enables the geospatial domain to effectively leverage these improvements for better in-domain performance with minimal carbon impact. We believe this is a sustainable way for the geospatial community to continually benefit from the most recent progress in computer vision, enabling a smarter, safer, and healthier planet.

References

[1] Peri Akiva, Matthew Purri, and Matthew J. Leotta. Self-supervised material and texture representation learning for remote sensing tasks. *CoRR*, abs/2112.01715, 2021. 3, 6, 7, 9

[2] Kumar Ayush, Burak Uzkent, Chenlin Meng, Kumar Tanmay, Marshall Burke, David B. Lobell, and Stefano Ermon. Geography-aware self-supervised learning. *CoRR*, abs/2011.09980, 2020. 1, 2, 9

[3] Hangbo Bao, Li Dong, and Furu Wei. Beit: BERT pre-training of image transformers. *CoRR*, abs/2106.08254, 2021. 3

[4] Rodrigo Caye Daudt, Bertr Le Saux, and Alexandre Boulch. Fully convolutional siamese networks for change detection. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 4063–4067, 2018. 7

[5] R. Caye Daudt, B. Le Saux, A. Boulch, and Y. Gousseau. Urban change detection for multispectral earth observation using convolutional neural networks. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, July 2018. 6

[6] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pre-training from pixels. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1691–1703. PMLR, 13–18 Jul 2020. 3

[7] Xinlei Chen, Haoqi Fan, Ross B. Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *CoRR*, abs/2003.04297, 2020. 2

[8] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, Oct 2017. 4

[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. *arXiv preprint arXiv:2207.08051*, 2022. 1, 2, 3, 6, 7, 8

[10] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020. 7

[11] 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

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[13] Ivica Dimitrovski, Ivan Kitanovski, Dragi Kocev, and Nikola Simidjievski. Current trends in deep learning for earth observation: An open-source benchmark arena for image classification, 2022. 7

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. 2, 3, 6, 7, 8

[15] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. *CoRR*, abs/2006.07733, 2020. 3

[16] Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. *CoRR*, abs/2004.10964, 2020. 2, 3

[17] Rujun Han, Xiang Ren, and Nanyun Peng. DEER: A data efficient language model for event temporal reasoning. *CoRR*, abs/2012.15283, 2020. 2, 3

[18] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. *CoRR*, abs/2111.06377, 2021. 3

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. 6, 7, 8

[20] Neal Jean, Sherrie Wang, Anshul Samar, George Azzari, David B. Lobell, and Stefano Ermon. Tile2vec: Unsupervised representation learning for spatially distributed data. *CoRR*, abs/1805.02855, 2018. 2

[21] Shunping Ji, Shiqing Wei, and Meng Lu. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Transactions on Geoscience and Remote Sensing*, 57(1):574–586, 2019. 7

[22] András Kalapos and Bálint Gyires-Tóth. Self-supervised pretraining for 2d medical image segmentation. *arXiv preprint arXiv:2209.00314*, 2022. 3

[23] Jian Kang, Ruben Fernandez-Beltran, Puhong Duan, Sicong Liu, and Antonio J. Plaza. Deep unsupervised embedding for remotely sensed images based on spatially augmented momentum contrast. *IEEE Transactions on Geoscience and Remote Sensing*, 59(3):2598–2610, 2021. 2

[24] Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*, 2019. 4

[25] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *CoRR*, abs/2103.14030, 2021. 2, 3, 6, 7, 8

[26] Zihan Liu, Genta Indra Winata, and Pascale Fung. Continual mixed-language pre-training for extremely low-resource neural machine translation. *CoRR*, abs/2105.03953, 2021. 2, 3

[27] Yang Long, Yiping Gong, Zhifeng Xiao, and Qing Liu. Accurate object localization in remote sensing images based on convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 55(5):2486–2498, 2017. 4

[28] Oscar Mañas, Alexandre Lacoste, Xavier Giró-i-Nieto, David Vázquez, and Pau Rodríguez. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. *CoRR*, abs/2103.16607, 2021. 1, 2, 3, 4, 6, 7, 8, 9

[29] Maxim Neumann, André Susano Pinto, Xiaohua Zhai, and Neil Houlsby. In-domain representation learning for remote sensing. *CoRR*, abs/1911.06721, 2019. 1, 2, 3, 7

[30] Keiller Nogueira, Otávio Augusto Bizetto Penatti, and Jefersson Alex dos Santos. Towards better exploiting convolutional neural networks for remote sensing scene classification. *CoRR*, abs/1602.01517, 2016. 1

[31] U.S. Department of Agriculture. National agriculture imagery program (NAIP). 4

[32] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. *CoRR*, abs/1604.07379, 2016. 3

[33] Xiaoman Qi, Panpan Zhu, Yuebin Wang, Liqiang Zhang, Junhuan Peng, Mengfan Wu, Jialong Chen, Xudong Zhao, Ning Zang, and P. Takis Mathiopoulos. Mlrsnet: A multi-label high spatial resolution remote sensing dataset for semantic scene understanding. *ISPRS Journal of Photogrammetry and Remote Sensing*, 169:337–350, 2020. 4

[34] 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. 3

[35] Franz Rottensteiner, Gunho Sohn, Jaewook Jung, Markus Gerke, Caroline Baillard, Sebastien Benitez, and Uwe Breitkopf. The isprs benchmark on urban object classification and 3d building reconstruction. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences I-3 (2012), Nr. 1*, 1(1):293–298, 2012. 7

[36] 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. 7

[37] Stefano Vincenzi, Angelo Porrello, Pietro Buzzega, Marco Cipriano, Pietro Fronte, Roberto Cuccu, Carla Ippoliti, Annamaria Conte, and Simone Calderara. The color out of space: learning self-supervised representations for earth observation imagery. *CoRR*, abs/2006.12119, 2020. 3

[38] Di Wang, Jing Zhang, Bo Du, Gui-Song Xia, and Dacheng Tao. An empirical study of remote sensing pretraining. *IEEE Transactions on Geoscience and Remote Sensing*, 2022. 2

[39] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *European Conference on Computer Vision*. Springer, 2018. 7

[40] Zhenda Xie, Zigang Geng, Jingcheng Hu, Zheng Zhang, Han Hu, and Yue Cao. Revealing the dark secrets of masked image modeling. *arXiv preprint arXiv:2205.13543*, 2022. 3

[41] 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. *CoRR*, abs/2111.09886, 2021. 3, 5, 6

[42] Yi Yang and Shawn Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, pages 270–279, 2010. 7

[43] Chenxiao Zhang, Peng Yue, Deodato Tapete, Liangcun Jiang, Boyi Shangguan, Li Huang, and Guangchao Liu. A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 166:183–200, 2020. 6, 7

[44] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan L. Yuille, and Tao Kong. ibot: Image BERT pre-training with online tokenizer. *CoRR*, abs/2111.07832, 2021. 3

[45] Weixun Zhou, Shawn Newsam, Congmin Li, and Zhenfeng Shao. Patternnet: A benchmark dataset for performance evaluation of remote sensing image retrieval. *ISPRS Journal of Photogrammetry and Remote Sensing*, 145:197–209, 2018. Deep Learning RS Data. 4