EXPLORA: PARAMETER-EFFICIENT EXTENDED PRE-TRAINING TO ADAPT VISION TRANSFORMERS UNDER DOMAIN SHIFTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Parameter-efficient fine-tuning (PEFT) techniques such as low-rank adaptation (LoRA) can effectively adapt large pre-trained foundation models to downstream tasks using only a small fraction (0.1%-10%) of the original trainable weights. An under-explored question of PEFT is in extending the pre-training phase without supervised labels; that is, can we adapt a pre-trained foundation model to a new domain via efficient self-supervised pre-training on this new domain? In this work, we introduce ExPLoRA, a highly effective technique to improve transfer learning of pre-trained vision transformers (ViTs) under domain shifts. Initializing a ViT with pre-trained weights on large, natural-image datasets such as from DinoV2 or MAE, ExPLoRA continues the unsupervised pre-training objective on a new domain, unfreezing 1-2 pre-trained ViT blocks and tuning all other layers with LoRA. We then fine-tune the resulting model only with LoRA on this new domain for supervised learning. Our experiments demonstrate state-of-the-art results on satellite imagery, even outperforming fully pre-training and fine-tuning ViTs. Using the DinoV2 training objective, we demonstrate up to 7.5% improvement in linear probing top-1 accuracy on downstream tasks while using <10% of the number of parameters that are used in prior fully-tuned state-of-the art approaches. Our ablation studies confirm the efficacy of our approach over other baselines, including PEFT and unfreezing more ViT blocks.

1 INTRODUCTION

Pre-training foundation models (Bommasani et al., 2021) for natural language (Brown et al., 2020; Chowdhery et al., 2023; Touvron et al., 2023; Jiang et al., 2024) and natural images (Oquab et al., 2023; He et al., 2022; Zhou et al., 2021; Rombach et al., 2022) has historically been computationally intensive, often limited to organizations with substantial resources. However, recent advancements in parameter-efficient fine-tuning (PEFT) techniques including low-rank adaptation (LoRA) and others (Hu et al., 2021; Zhang et al., 2023b; Chavan et al., 2023; Qiu et al., 2023; Liu et al., 2023; Jia et al., 2022) have sparked significant interest. These methods aim to adapt foundation models to downstream supervised-learning tasks using a small fraction (0.1%-10%) of the model's trainable weights, with many based on the hypothesis that the required weight updates to the pre-trained model have a "low intrinsic rank" (Hu et al., 2021; Li et al., 2018; Aghajanyan et al., 2020).

In this paper, we focus on visual foundation models (VFMs) such as DinoV2 or MAE (Oquab et al., 2023; He et al., 2022), which were trained on large-scale natural-image datasets. Despite the large investments in developing such models for natural images, they underperform when applied to other domains with visual data (e.g. medical or satellite images). For example, fine-tuning a model pre-trained on natural images on satellite image classification tasks is not as effective as fine-tuning one that was pre-trained on satellite images (Cong et al., 2022; Ayush et al., 2021). To bridge this gap, prevailing approaches invest similarly large levels of compute to pre-train VFMs on new domains, inspired by techniques developed for natural images (Cong et al., 2022; Reed et al., 2023; Tang et al., 2024; Khanna et al., 2024; Zhou et al., 2023; Moutakanni et al., 2024; Man et al., 2023).

In this work, we challenge this paradigm (fig. 1), asking whether pre-training from scratch on each new domain is strictly necessary, since doing so is expensive (in compute and time) and

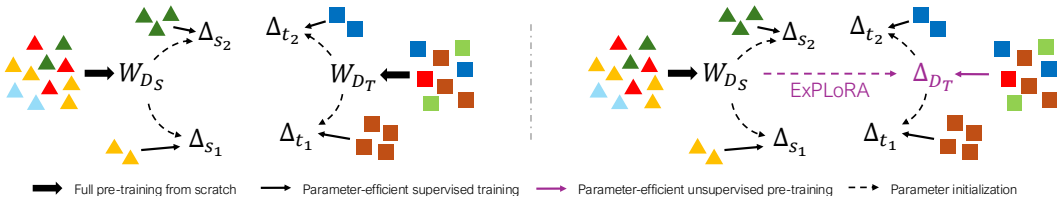

Figure 1: Consider two different image domains, D_S and D_T. **Left**: the traditional paradigm of pre-training from scratch on each domain to yield W_{D_S} and W_{D_T}, and then fine-tuning on target datasets i to yield $\Delta_{s_i}, \Delta_{t_i}$, for domains D_S and D_T, respectively. **Right**: our approach, which is to initialize with pre-trained weights from domain D_S and then learn unsupervised weights Δ_{D_T} for domain D_T in a parameter-efficient manner.

precludes knowledge transfer from natural images. Instead, we wish to more efficiently leverage the rich semantic information encoded in natural-image vision foundation models to adapt them to new domains. Our proposed solution addresses these concerns using PEFT techniques for domain adaptation via self-supervised learning.

We introduce **ExPLoRA**, which generalizes vision foundation models to new domains by extending the pre-training phase with parameter-efficient techniques. We initialize a vision transformer (ViT) (Dosovitskiy et al., 2021) with pre-trained weights from natural-image datasets such as MAE or DinoV2. Selectively unfreezing 1-2 transformer blocks, we tune remaining weights with LoRA and continue unsupervised pre-training on the new domain. Subsequently fine-tuning with linear probing or LoRA on this new domain for supervised learning outperforms prior state-of-the-art (SoTA) approaches while training under 6-10% of the original weights. On satellite imagery, for example, we demonstrate an 8% improvement in linear probing top-1 accuracy, and even an improvement over prior SoTA fully pre-trained and fine-tuned techniques. We conduct an extensive study on RGB, temporal, and multi-spectral satellite images, either matching or outperforming prior methods that fully pre-train from scratch. ExPLoRA also generalizes to different domains such as wildlife, medical, and agricultural imagery on the WILDS (Koh et al., 2021) benchmark. Our contributions include:

1. Introducing ExPLoRA, a novel parameter-efficient method that extends unsupervised pre-training on target domains, achieving state-of-the-art supervised-learning performance using a fraction of the original ViT weights (section 5).
2. Conducting a comprehensive case study on satellite imagery, showcasing improvements in linear probing top-1 accuracy and outperforming existing techniques on datasets like fMoW. We also demonstrate generalization to multiple other domains within WILDS (section 6).
3. Demonstrating ExPLoRA's efficacy via ablation studies and by analyzing the differences in local (eg: positional) and global (eg: class) information encoded in the patch representations output by each ViT block (section 6.3).

2 Related Work

VFMs VFMs such as DinoV2 or masked autoencoders (MAE) that pre-train with self-supervised learning (SSL) have demonstrated remarkable performance across downstream tasks such as classification or semantic segmentation Oquab et al. (2023); He et al. (2022); Grill et al. (2020); Chen et al. (2020). However, there has also been a rise in domain-specific VFMs (Cong et al., 2022; Reed et al., 2023; Moutakanni et al., 2024; Ma et al., 2024; Zhang et al., 2023a). For instance, SatMAE handles temporal or multi-spectral satellite image inputs. Since these models contain hundreds of millions of parameters, efficient adaptation to downstream tasks has become a key research focus.

PEFT PEFT methods have gained widespread adoption for efficiently adapting large models by updating only a fraction of parameters, mitigating the prohibitive costs of full model tuning. LoRA learns low-rank weight updates to frozen weights, while other methods modify the frequency or number of trainable parameters per layer (Hu et al., 2021; Zhang et al., 2023b; Chavan et al., 2023; Pu et al., 2023). Others use multiplicative orthogonal updates (Qiu et al., 2023; Liu et al., 2023) or inject adapter modules (Steitz & Roth, 2024; Yin et al., 2023; Chen et al., 2022; Yin et al., 2024; Lian et al., 2022), effectively retaining pre-training knowledge in frozen weights. Visual prompt tuning (VPT) methods concatenate learnable prompt tokens to image patch sequences, trading improved

fine-tuning performance with increased inference costs (Jia et al., 2022; Yoo et al., 2023; Pei et al., 2024; Han et al., 2023; Nie et al., 2023). ExPLoRA aims to supplement rather than replace these methods, and thus can be configured with any existing or future PEFT method for ViT fine-tuning.

Domain Adaptation Domain adaptation enables models trained on a source domain to perform well on a different but related target domain. Traditional transformer-based methods address this via domain alignment, discriminative feature learning, cross-attention with pseudo-labels (Sun et al., 2022; Chuan-Xian et al., 2022; Zhu et al., 2023), or adversarial learning with self-refinement (Yang et al., 2023; Xu et al., 2021), typically requiring either labeled target data or source domain labels. Recent work explores adapting ViTs through different means: e.g., continual pre-training via masked image modeling (Mendieta et al., 2023) and scaled LoRA adapters (Scheibenreif et al., 2024) for satellite imagery. ExPLoRA builds on this direction, enabling SSL directly on the target domain while using significantly fewer parameters. Further comparisons with related work are in appendix A.

3 Background

MAE The masked-autoencoder (MAE) (He et al., 2022) is an effective SSL technique for ViTs that uses an asymmetric encoder-decoder architecture on images $\mathbf{x} \in \mathbb{R}^{C \times H \times W}$, where patches are masked before being processed by the ViT encoder $\mathcal{L}$. The masked patches are then reconstructed by a smaller decoder $\mathcal{L}_D$, with both trained jointly using mean-squared error on the reconstructed visible pixels. While effective across domains (Cong et al., 2022; Bachmann et al., 2022), MAEs typically require full fine-tuning for downstream tasks, which makes them computationally expensive.

DinoV2 DinoV2 (Oquab et al., 2023) is a robust SSL method for ViTs. Unlike MAE, DinoV2 features have demonstrated strong zero-shot performance, enabling adaptation to downstream tasks even with a frozen ViT backbone. During pre-training, DinoV2 maintains two copies of a ViT encoder: the student (trainable) and the teacher, which is updated using an exponential-moving average of the student's parameters. The training objective incorporates a global, image-level loss from Dino (Caron et al., 2021), a patch-based loss from iBOT (Zhou et al., 2021), and regularizers including KoLeo (Delattre & Fournier, 2017) and Sinkhorn-Knopp centering (Caron et al., 2020).

LoRA Low-rank adaptation (LoRA) (Hu et al., 2021) assumes that the weight update to change a set of unsupervised pre-trained weights to supervised fine-tuned weights lives in a low-rank subspace,

$$W \approx W_0 + \Delta_W = W_0 + BA \tag{1}$$

where $W \in \mathbb{R}^{k_2 \times k_1}$ are the final, task-specific fine-tuned weights, $W_0 \in \mathbb{R}^{k_2 \times k_1}$ are the pre-trained weights, $\Delta_W \in \mathbb{R}^{k_2 \times k_1}$ is the weight update required to translate the pre-trained weights W_0 to the fine-tuned weights W. The key is that $\Delta_W = BA$ where $B \in \mathbb{R}^{k_2 \times r}$ and $A \in \mathbb{R}^{r \times k_1}$. That is, A and B form a low-rank factorization of Δ_W, where the rank $r \ll \min(k_1, k_2)$.

4 Problem Setup

Consider a set of image domains $\mathcal{D} = \{1, 2, \ldots\}$, where each domain $d \in \mathcal{D}$ is associated with a data distribution $p_d(\mathbf{x})$, and images $\mathbf{x} \in \mathbb{R}^{C_d \times H_d \times W_d}$ have domain-specific channel, height, and width. Let $D_S \subset \mathcal{D}$ represent a set of source domains (e.g., internet-scale natural image data) and $D_T \subset \mathcal{D}$ represent target domains (e.g., satellite imagery). The data from the source domains follow a distribution $p_{D_S}(\mathbf{x})$, and the target domain data come from $p_{D_T}(\mathbf{x})$. For some target domains $d_T \in D_T$, the joint distributions $p_{d_T}(\mathbf{x}, \mathbf{y})$ describe images $\mathbf{x}$ with associated supervised labels $\mathbf{y}$ used for downstream tasks. We then assume access to the following:

(i) W_{D_S}, pre-trained weights obtained via unsupervised pre-training on images from $p_{D_S}(\mathbf{x})$
(ii) $\mathcal{X}_{D_T} = \{\mathbf{x}_i\}_{i=1}^{N} \sim p_{D_T}(\mathbf{x})$, an unlabeled dataset of N images from new domains D_T
(iii) $\mathcal{Y}_{d_T} = \{\mathbf{x}_j, \mathbf{y}_j\}_{j=1}^{M_{d_T}} \sim p_{d_T}(\mathbf{x}, \mathbf{y})$ a labeled dataset of M_{d_T} images from domain $d_T \in D_T$

Our objective is to learn optimal weights W_{d_T} for each supervised-learning dataset $\mathcal{Y}_{d_T}$ in a parameter-efficient manner while leveraging the knowledge stored in W_{D_S}.

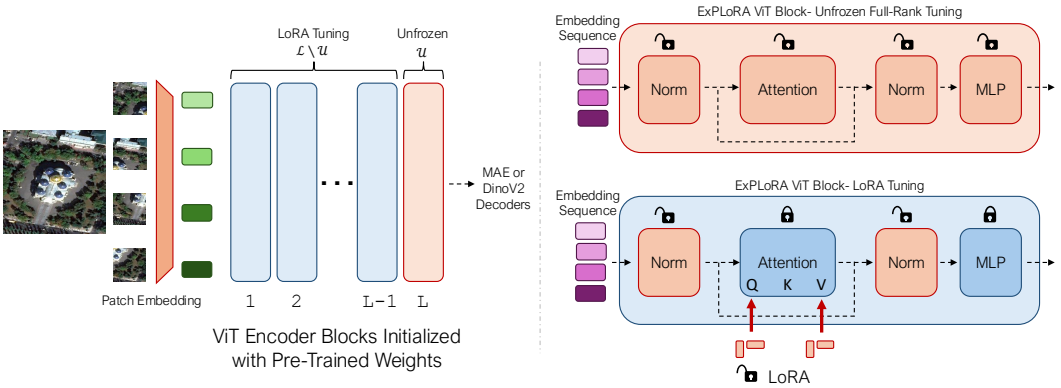

Figure 2: An overview of ExPLoRA. The set $\mathcal{L}$ of $\mathtt{L}$ ViT blocks is partitioned into two sets: $\mathcal{U}$, which denotes blocks whose parameters are completely unfrozen, and $\mathcal{L} \setminus \mathcal{U}$ which denotes blocks that undergo LoRA tuning (only on the Q, V attention matrices). Note that the normalization layers are always unfrozen across all blocks.

Algorithm 1 ExPLoRA

1: **Input:** W_{D_S} := pre-trained ViT with $\mathtt{L}$ layers $\mathcal{L} = \{\mathtt{1}, \ldots, \mathtt{L}\}$; $\mathcal{X}_{D_T}$:= unlabeled dataset
2: Initialize a frozen ViT with W_{D_S} from source domains D_S (e.g., DinoV2 or MAE weights).
3: Unfreeze all parameters of a subset of blocks $\mathcal{U} \subset \mathcal{L}$. (e.g., $\mathcal{U} = \{\mathtt{L}\}$ or $\mathcal{U} = \{\mathtt{1}, \mathtt{L}\}$).
4: Apply LoRA (with rank r) on Q and V weights in attention layers of frozen blocks in $\mathcal{L} \setminus \mathcal{U}$ and unfreeze normalization layers in these blocks.
5: Train all unfrozen parameters Δ_{D_T} on the unlabeled dataset $\mathcal{X}_{D_T}$ using the same unsupervised objective as what was used for W_{D_S} (e.g., DinoV2 or MAE).
6: **Output:** A new pre-trained model $W^*_{D_T} = W_{D_S} + \Delta_{D_T}$ for target domains D_T.

Traditionally, the approach (fig. 1) has been to begin pre-training from scratch on the new domains of interest in $\mathcal{X}_{D_T}$, and then fine-tune for each dataset $\mathcal{Y}_{d_T}$, representing the following:

$$W_{d_T} \approx W_{D_T} + \Delta_{d_T} \tag{2}$$

where W_{D_T} represents the weights learned from unsupervised pre-training on $\mathcal{X}_{D_T}$, and Δ_{d_T} are the weights learned from supervised fine-tuning on $\mathcal{Y}_{d_T}$. However, this method is computationally expensive: fully pre-training W_{D_T} from scratch for every new domain requires prohibitively large amounts of additional compute.

On the other hand, LoRA addresses this inefficiency in the following way:

$$W_{d_T} \approx W_{D_S} + \Delta_{d_T} = W_{D_S} + B_{d_T} A_{d_T} \tag{3}$$

The LoRA hypothesis is that the update Δ_{d_T} resides in a low-rank subspace when adapting pre-trained weights W_{D_S} to fine-tuned weights W_{d_T}. This hypothesis holds well when pre-training and fine-tuning distributions are similar, or where $d_T \in D_S$. However, when there is significant domain shift, such as between natural images and multi-spectral satellite data, the low-rank assumption often breaks down (see section 6.1.3).

Our goal is to learn W_{D_T} in a parameter-efficient manner to bridge the large domain shift to D_T while leveraging the knowledge encoded in W_{D_S}. We propose the following factorization of W_{d_T}:

$$W_{d_T} \approx W_{D_S} + \Delta_{D_T} + \Delta_{d_T} \tag{4}$$

where $\Delta_{D_T} \in \mathbb{R}^{k_2 \times k_1}$ is an additional update matrix learned from unsupervised pre-training on $\mathcal{X}_{D_T}$. Crucially, Δ_{D_T} requires only a fraction of the $k_1 k_2$ parameters of W_{D_S}, making it significantly more efficient than full-rank pre-training. The resulting model, $W^*_{D_T} = W_{D_S} + \Delta_{D_T} \approx W_{D_T}$, retains the benefits of unsupervised pre-trained VFMs, including strong feature extraction, effective linear probing, KNN classification, and generalization to downstream tasks.

5 METHOD

To learn Δ_{D_T}, we propose ExPLoRA (i.e. **Ex**tended **P**re-training with **LoRA**), a method that efficiently adapts pre-trained ViTs for new target domains D_T, described in algorithm 1.

In terms of notation, D-[L]-$r64$ refers to a ViT initialized with DinoV2 weights (denoted by D), where $\mathcal{U} = \{\texttt{L}\}$, and LoRA rank 64 is applied to the Q, V matrices of attention layers in $\mathcal{L} \setminus \mathcal{U}$. Thus, Δ_{D_T} comprises of all weights in $\mathcal{U}$, LoRA matrices in $\mathcal{L} \setminus \mathcal{U}$, and normalization layers. For $\mathcal{U} = \{\texttt{L}\}$, Δ_{D_T} consists of only 5% of the original ViT parameters. As we show in section 6, our extended pre-training approach can match or even outperform full pre-training on new domains from scratch.

ExPLoRA for DinoV2 We initialize a ViT-L with W_{D_S} from the DinoV2 ViT-L encoder, without registers (Darcet et al., 2023). Since the DinoV2 pre-trained checkpoints don't contain the Dino or iBOT linear heads, we initialize a shared Dino-iBOT linear head from scratch. This shared head is fully trained during extended pre-training, adding only a minimal number of trainable parameters.

ExPLoRA for MAE We initialize a ViT-L with W_{D_S} from the MAE ViT-L encoder. Since MAE provides the pre-trained decoder, we use these weights to initialize our MAE decoder He et al. (2022). During extended pre-training, in addition to the ExPLoRA recipe in algorithm 1, we apply LoRA with rank r' on the Q, V matrices of each attention layer in the frozen decoder. Note that the LoRA rank r' may differ from the LoRA rank r used in the ViT encoder (appendix B.4). All other decoder weights, apart from the layer-normalization layers, are kept frozen. No block is fully unfrozen in the MAE decoder, as it will be discarded after extended pre-training. This helps to minimize the number of additional parameters trained in the decoder.

ExPLoRA for Multi-Spectral Inputs For the multi-spectral ViT introduced by SatMAE we need to additionally unfreeze the positional encoding and the patch embedding weights for each group of channels. These cannot be initialized from W_{D_S}, as W_{D_S} is trained on RGB inputs, whereas multi-spectral inputs can have more or different channels. As part of Δ_{D_T} in algorithm 1, the positional encodings and patch embeddings for multi-spectral data are adapted during extended pre-training. Aside from this, the approach remains unchanged from that of DinoV2 or MAE described earlier.

Storage Considerations After running ExPLoRA, we receive a new unsupervised model $W^*_{D_T} = W_{D_S} + \Delta_{D_T}$ for the target domains D_T. Any components that are not part of the ViT encoder (eg: the Dino linear head or the MAE decoder) are discarded. Post-ExPLoRA, only Δ_T, consisting of 1-2 unfrozen ViT blocks, LoRA matrices, and layer-normalization weights, are stored for each D_T. Like LoRA and other PEFT methods, ExPLoRA significantly reduces additional storage requirements compared to fully training W_{D_T} from scratch.

Fine-Tuning post-ExPLoRA After extended pre-training with ExPLoRA, the output weights $W^*_{D_T}$ behave as any fully pre-trained model W_{D_T}. We can now use the ViT for feature extraction, PEFT, or fine-tuning as desired. For instance, we could initialize a linear head for classification or a decoder for segmentation, either of which is fully trainable. We can then freeze all ViT weights and apply LoRA-$r8$ on the Q, V matrices of the attention layers (or use another PEFT method). Lastly, we use supervised fine-tuning on each labeled dataset $\mathcal{Y}_{d_T}$ to train the unfrozen parameters Δ_{d_T}. This yields our final model W_{d_T} (eq. (4)), which can be used for classification, segmentation, detection etc.

6 EXPERIMENTS

Our experimental results consist of a case study on satellite imagery (section 6.1), with an ablation study in section 6.1.2 and analysis in section 6.3. We evaluate on multiple downstream tasks in sections 6.1.3, 6.1.4 and 6.2. Additional experiments and ablations are provided in appendix B and training hyperparameter and compute configurations are mentioned in appendix C. Our results achieve a new SoTA top 1 accuracy of 79.2% (↑1.4%) on the competitive fMoW-RGB benchmark, outperforming fully pre-trained and fine-tuned models while using 6% of the ViT encoder parameters. We also achieve a ↑8.2% improvement in linear probing accuracy on the same dataset. Across other satellite datasets, we match fully-pretrained prior state-of-the-art methods, and demonstrate competitive performance on WiLDS benchmark datasets as well.

6.1 CASE STUDY: SATELLITE IMAGERY

We examine satellite images given their importance towards multiple societal applications (section 7) and since they represent a significant domain shift from natural images. There is a large and growing body of research on developing foundational models for satellite imagery from scratch Ayush et al. (2021); Cong et al. (2022); Reed et al. (2023); Tang et al. (2024), thus presenting a good benchmark for ExPLoRA.

6.1.1 RGB SATELLITE IMAGES

Dataset We first consider the functional map of the world (fMoW) dataset of high-resolution satellite images, each paired with one of 62 classification labels (Christie et al., 2018). fMoW is used as a benchmark for many satellite-image foundation models (Cong et al., 2022; Reed et al., 2023).

Model	Arch.	PEFT	Pre-train #Params	Fine-tune #Params	Top 1 Acc.
ScaleMAE [49]	ViT-L	Full	303.3M	303.3M	77.80
SatMAE [17]	ViT-L	Full	303.3M	303.3M	77.78
SatMAE [17]	ViT-L	LoRA-$r8$ [28]	303.3M	0.8M	76.10
ScaleMAE [49]	ViT-L	LoRA-$r8$ [28]	303.3M	0.8M	78.01
GFM [41]	ViT-L	LoRA-$r8$ [28]	303.3M	0.8M	73.03
GDA [51]	ViT-L	GDA-$r16$ [51]	8.5M	8.5M	71.88
MAE [26]	ViT-L	LoRA-$r8$ [28]	-	0.8M	76.21
M-[L]-$r64$	ViT-L	LoRA-$r8$ [28]	18.7M	0.8M	76.55
DinoV2 [44]	ViT-L	LoRA-$r8$ [28]	-	0.8M	78.08
DinoV2 [44]	ViT-L	BOFT-$b2m8$ [37]	-	0.9M	72.40
DinoV2 [44]	ViT-L	VPT-100 [29]	-	0.4M	77.29
DinoV2 [44]	ViT-L	GVPT-100 [63]	-	0.4M	76.22
DinoV2 [44]	ViT-L	AdaLoRA-$r8$ [65]	-	1.2M	78.87
DinoV2 [44]	ViT-L	Adapter+ [53]	-	1.4M	78.16
DinoV2 [44]	ViT-L	SA^{2}VP [45]	-	1.1M	77.53
D-[L]-$r64$	ViT-L	SA^{2}VP [45]	18.7M	1.1M	78.51
D-[L]-$r64$	ViT-L	LoRA-$r8$ [28]	18.7M	0.8M	**79.15**

Table 1: Results on the fMoW-RGB validation dataset. The "Pre-train #Params" and "Fine-tune #Params" refer to the trainable parameters of the ViT encoder required on the *new* domain, i.e. satellite images. M-[L]-$r64$ and D-[L]-$r64$ refer to ExPLoRA models initialized with MAE and DinoV2 weights, respectively (section 5).

We compare our results in table 1 against both prior fully pre-trained SoTA foundation models as well as PEFT techniques applied on ViTs pre-trained with MAE and/or DinoV2 weights. Our results demonstrate that D-ExPLoRA-[L]-$r64$ is SoTA in terms of fMoW-RGB average accuracy at 79.15%. ExPLoRA outperforms techniques that require fully and/or continually pre-training ViTs on fMoW while using 6% of the original ViT encoder parameters. Further experiments with MAE are in B.7.

ExPLoRA-initializations with LoRA fine-tuning outperform other unsupervised initializations paired with PEFT techniques by 1-3%, including SoTA matrix-adaptation methods like AdaLoRA (Zhang et al., 2023b), BOFT (Liu et al., 2023), VPT approaches such as GVPT (Yoo et al., 2023) and SA^{2}VP (Pei et al., 2024), and adapter methods like Adapter+ (Steitz & Roth, 2024). We also outperform satellite image domain adaptation methods such as GFM (Mendieta et al., 2023) and GDA (Scheibenreif et al., 2024). Additionally, applying SA^{2}VP to ExPLoRA-initialized ViTs improves performance over a DinoV2 initialization by 1%, showcasing ExPLoRA's compatibility with other PEFT methods and its versatility as an initialization for new domains.

Using our strongest performing variant (i.e. ExPLoRA with DinoV2), we investigate linear-probing performance on fMoW-RGB compared with prior SoTA methods in table 2. Linear-probing represents freezing the backbone and then training a linear head on the features extracted from the frozen backbone, serving as a desirable metric of the quality of extracted embeddings. Our results demonstrate an improvement of over ↑8.2% in top 1 average accuracy over prior SoTA methods, demonstrating that ExPLoRA learns robust unsupervised representations for its target domain without requiring expensive from-scratch pre-training. Importantly, ExPLoRA outperforms domain-specific prior SoTA solutions (rows 1-4), as well as DinoV2, which suggests successful transfer learning on the target domain by leveraging knowledge from pre-training on natural images.

Method	Arch.	Top 1 Acc.
GASSL [2]	ResNet	68.32
SatMAE [17]	ViT-L	65.94
ScaleMAE [49]	ViT-B	67.30
CScaleMAE [55]	ViT-B	69.20
DinoV2 [44]	ViT-L	67.60
DinoV2† [44]	ViT-L	69.00
D-[L]-$r64$	ViT-L	**76.86**
D-[L]-$r64$†	ViT-L	**77.48**

Table 2: Linear-probing on fMoW-RGB. The first four rows fully pre-train on the dataset. † denotes concatenating features from the last 4 ViT blocks. All other rows use the features of the last ViT block.

Blocks Unfrozen	LoRA Rank	Norm Unfrozen	LoRA Layers	Num. Params	Top 1 Acc.
[L]	0	✓	[]	12.7M	74.83
[L-1,L]	0	✓	[]	25.3M	75.97
[]	256	✓	[Q,V]	25.9M	75.51
[]	128	✓	All	33.1M	55.03
[L]	64	✓	Mlp	16.5M	48.55
[1]	64	✓	[Q,V]	18.7M	75.97
[9]	64	✓	[Q,V]	18.7M	75.45
[L-1]	64	✓	[Q,V]	18.7M	77.40
[L]	0	✓	VPT-100	12.8M	70.14
[L]	64	✗	[Q,V]	18.6M	76.78
[L]	8	✓	[Q,V]	13.4M	76.31
[L]	32	✓	[Q,V]	15.7M	76.40
[L]	64	✓	[Q,V]	18.7M	**77.48**

Table 3: Ablation study using DinoV2-ExPLoRA, measuring linear-probing accuracy on fMoW-RGB. All results are obtained by using concatenated features from the last 4 ViT blocks.

6.1.2 Ablation Study

We perform an ablation study (table 3) on linear-probing performance for fMoW-RGB to determine whether our proposed configuration performs optimally. A natural question is whether the improvement in performance stems primarily from unfreezing blocks, or from LoRA-tuning the rest of the ViT. We investigate this by unfreezing blocks $\{\texttt{L}, \texttt{L-1}\}$ in row 2 (with no LoRA), and comparing that with ExPLoRA-L-$r8$ in row 10. As seen, unfreezing an extra block consumes almost double the number of parameters, but fails to yield the same improvement in performance $\downarrow 0.34\%$. Thus, simply increasing the number of unfrozen blocks will likely improve performance, but will not do so as effectively as ExPLoRA, and will also significantly and sharply decrease the parameter-efficiency.

Next, we investigate whether high LoRA ranks used on all ViT layers (i.e. all attention and MLP matrices, not just Q, V) is beneficial. Surprisingly, this significantly harms learning (row 4, 5). In fact, it is much less effective than using just LoRA-$r256$ on the Q, V matrices of all $\mathcal{L}$ blocks (row 3). However, both rows 3 and 4 are much less parameter-efficient than ExPLoRA (rows 6-8, 11-13).

The choice of $\mathcal{U}$ matters as well. As seen in rows 6-8, and 13, for the DinoV2 objective, $\mathcal{U} = \{\texttt{1}\}$ or $\mathcal{U} = \{\texttt{9}\}$ are not as effective as $\mathcal{U} = \{\texttt{L-1}\}$ or $\mathcal{U} = \{\texttt{L}\}$, ceteris paribus. To understand this result further, see section 6.3. We also notice a slight drop in accuracy from leaving the normalization layers across the ViT frozen, seen in row 10.

Lastly, we investigate the impact of LoRA rank on ExPLoRA. Changing the rank from 8 to 32 has a small improvement ($\uparrow 0.09\%$), but changing from 32 to 64 brings about a much larger improvement ($\uparrow 1.08\%$), with only a relatively small increase in trainable parameters. This demonstrates that higher ranks are necessary during pre-training for effective learning on the new domain. Further ablations on compute efficiency (B.2), data efficiency (B.3), MAE decoder rank (B.4), and ViT backbone size (B.5) are in appendix B.

6.1.3 Multi-Spectral Satellite Images

Dataset Next, we consider the fMoW-Sentinel dataset, a large dataset of Sentinel-2 images used in Cong et al. (2022). Each image consists of 13 spectral bands and is paired with one of 62 classes.

With fMoW-Sentinel, we evaluate transfer from natural images to multi-spectral, low-resolution satellite images - a harder task than fMoW-RGB due to the absence of non-RGB bands in D_S. We use the group-channel ViT-L from Cong et al. (2022), initialized with MAE. During algorithm 1, we additionally unfreeze only the patch embedding layers due to architectural differences.

Table 4 shows the challenge: fully fine-tuning from MAE drops accuracy by nearly 10% (row 2), LoRA tuning from MAE performs worse (row 4), and unfreezing four transformer blocks (row 6) fails to help. However, ExPLoRA with $\mathcal{U} = \{\texttt{1}, \texttt{L}\}$ outperforms even full pre-training from scratch for LoRA fine-tuning (row 5 vs. last row), demonstrating effective adaptation to a very different domain while using <10% of the parameters.

Model	Backbone	PEFT	Pre-train #Params	Fine-tune #Params	Top 1 Acc.
ImgNet-Supervised	ResNet152	Full	60.3M	60.3M	54.46
MAE [26]	ViT-L	Full	-	303.3M	51.61
SatMAE [17]	ViT-L	Full	303.3M	303.3M	**61.48**
MAE [26]	ViT-L	LoRA-r8	-	0.8M	46.97
SatMAE [17]	ViT-L	LoRA-r8	303.3M	0.8M	59.48
MAE-`[1,2,L-1,L]`	ViT-L	LoRA-r8	51.5M	0.8M	54.12
M-ExPLoRA-`[L]`-$r32$	ViT-L	LoRA-r8	16.2M	0.8M	51.84
M-ExPLoRA-`[1,L]`-$r32$	ViT-L	LoRA-r8	29.7M	0.8M	**60.15**

Table 4: Results on the fMoW-Sentinel validation set. The "Pre-train #Params" and "Fine-tune #Params" refer to the trainable parameters required on the *new* domain, i.e. multi-spectral satellite images. "MAE-`[1,2,L-1,L]`" refers to initializing the group-channel SatMAE model with MAE weights, unfreezing blocks 1,2,23,24 for ViT-L, and then continuing pre-training on fMoW-Sentinel.

Method	PEFT	Top 1 Acc.
GASSL [2]	Full	74.11
SatMAE [17]	Full	79.69
MAE [26]	LoRA-r8	69.30
SatMAE [17]	LoRA-r8	75.27
M-`[L]`-$r32$	LoRA-r8	**75.98**

Table 5: fMoW-Temporal validation set results

Method	PEFT	SpaceNet mIoU	Resisc45 Top 1 Acc.
SatMAE [17]	Full	78.07	94.80
ScaleMAE [49]	Full	**78.90**	95.70
DinoV2 [44]	LoRA-r8	76.69	97.60
D-`[L]`$-r64$	LoRA-r8	76.69	**97.65**
SatMAE [17]	Lin. Probe	50.89	88.30
ScaleMAE [49]	Lin. Probe	47.17	89.60
DinoV2 [44]	Lin. Probe	76.21	96.34
D-`[L]`$-r64$	Lin. Probe	76.34	**97.32**

Table 6: SpaceNet and Resisc-45 validation set results

6.1.4 Additional Satellite Datasets

We perform extensive experiments on downstream satellite datasets, with further results in B.1.

fMoW-Temporal Each input is a sequence of up to 3 fMoW-RGB (Christie et al., 2018) images of the same location, distributed temporally, and paired with one of 62 classes. Since the inputs are now temporal sequences, we initialize the temporal MAE architecture from Cong et al. (2022) with MAE weights, and pre-train on $\mathcal{X}_{D_T}$ with $\mathcal{U} = [\mathrm{L}]$ and LoRA rank 32. ExPLoRA then outperforms temporal SatMAE for PEFT (table 5), demonstrating successful transfer learning at a fraction of the pre-training parameters.

SpaceNet-v1 This dataset contains high resolution satellite images, each paired with a segmentation mask for buildings (Van Etten et al., 2018). The training and test sets consist of 5000 and 1940 images, respectively. For ExPLoRA, we pre-train on the training set. However, many images in the dataset contain extensive blacked-out regions, indicating limits of the visible region. Considering this limitation and the small dataset size, it is not clear whether additional pre-training is effective. We find that, despite this, ExPLoRA remains on par with the LoRA-tuned DinoV2 model and remains competitive with the fully pre-trained and fully fine-tuned domain-specific models (table 6).

RESISC-45 The RESISC-45 (Cheng et al., 2017) benchmark dataset consists of 31,500 satellite images of varying resolution (0.2m-30m GSD), with 45 classes. The data is split into 25,200 training and 6,300 validation images, as per Reed et al. (2023). In table 6, our D-ExPLoRA pre-trained on only high-resolution fMoW-RGB images achieves SoTA results of 97.32% on multi-resolution RESISC-45 images, with just linear-probing. Since we use the same pre-trained model as in the last row of table 1, we demonstrate successful transfer learning from ExPLoRA pre-training, without requiring any additional modifications for scale-aware representation learning (Reed et al., 2023).

6.2 WILDS Datasets

We test ExPLoRA on the WILDS (Koh et al., 2021) benchmark, specifically on Camelyon17 (Bandi et al., 2018), iWildcam (Beery et al., 2020) and GlobalWheat David et al. (2020; 2021) datasets, representing domain transfers to medical, wildlife, and agricultural imagery, respectively.

Method	PEFT	Top 1 Acc.
CLater [48]	Full	93.90
ICON	Full	90.10
DinoV2 [44]	Lin. Probe	93.27
DinoV2 [44]	LoRA-r8	92.97
D-[L]-$r32$	Lin. Probe	**94.41**
D-[L]-$r32$	LoRA-r8	94.21

Table 7: Classification results on the validation set of Camelyon17.

Method	PEFT	Top 1 Acc.
DinoV2 [44]	Lin. Probe	66.04
DinoV2 [44]	LoRA-r8	67.10
D-[L]-$r32$	Lin. Probe	62.95
D-[L]-$r32$	LoRA-r8	**68.07**

Table 8: Classification results on the validation set of iWildcam.

Method	Top 1 Acc.	AP@ 0.5:0.95	AR@ 0.5:0.95
ICON [32]	68.9	-	-
MAE [26]	82.5	53.8	58.7
DinoV2 [44]	82.3	52.1	57.1
D-[L]-$r64$	**82.7**	**54.5**	**59.2**

Table 9: Object detection results on the validation set of GlobalWheat. AP and AR stand for average precision and average recall, respectively.

Camelyon17 The WILDS Camelyon17 dataset consists of images of cancerous and non-cancerous cell tissue organized in labeled and unlabeled splits. We use the "train-unlabeled" split for pre-training ExPLoRA, and either use LoRA fine-tuning or linear probing on the training set of the labeled split. We report accuracy on the binary classification problem and compare with entries on the WILDS leaderboard which use unlabeled data. Our results in table 7 demonstrate improved performance over domain-specific methods as well as DinoV2, once again successfully bridging the domain gap.

iWildcam iWildcam classification requires identifying one of 182 animal species given an image. We pre-train on the training set, finding that this outperforms pre-training on the extra-unlabeled set. In table 8, we find an improvement over DinoV2 using LoRA-$r8$ PEFT. Surprisingly, the linear probing performance of the ExPLoRA suffers in comparison with DinoV2, suggesting possible loss in knowledge-transfer due to a small domain gap. Likely because natural image datasets W_{D_S} such as ImageNet (Deng et al., 2009) used for DinoV2 already contain many images of animals.

GlobalWheat The GlobalWheat dataset consists of a wheat head object detection task, where each image of a wheat field is associated with bounding boxes on the visible wheat heads David et al. (2020; 2021). ExPLoRA extends pre-training on the training set, and then we run fine-tuning using Detectron2 code for object-detection with ViTs (Wu et al., 2019). ExPLoRA outperforms both fully pre-trained baselines from the WILDS leaderboard and strong VFMs DinoV2 and MAE on top 1 accuracy, average precision, and average recall.

6.3 Analyzing ExPLoRA

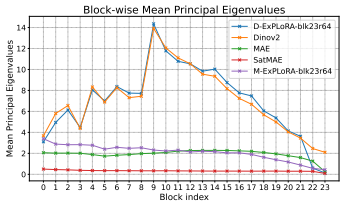

Figure 3: The mean of the principal components of the feature map outputted by each ViT block.

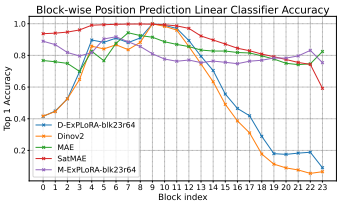

Figure 4: Linear probing each patch for position (local information), across all ViT blocks.

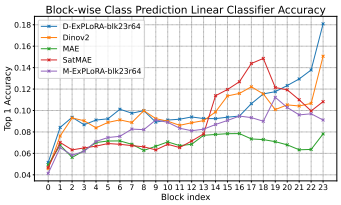

Figure 5: Linear probing each patch for classification (global information), across all ViT blocks.

The key design choice of ExPLoRA is to fully train a small subset $\mathcal{U} \subset \mathcal{L}$ of the ViT, while applying low-rank updates to the remaining frozen layers $\mathcal{L} \setminus \mathcal{U}$. For parameter-efficiency, we aim to keep $|\mathcal{U}| \ll |\mathcal{L}|$ and make and informed choice of which layers to unfreeze based on their potential to improve learning during extended pre-training.

We conduct an investigation on 5 models using a sample of $\mathcal{X}_{D_T}$. These models are DinoV2, D-ExPLoRA-[L]-$r64$, SatMAE, MAE, and M-ExPLoRA-[L]-$r64$. We do the following analyses: (i) PCA to measure the mean and variance of eigenvalues of patch feature vectors for each ViT block, in fig. 3 (ii) linear probing for local or global information (Darcet et al., 2023) by training logistic regression classifiers on each block's patch feature vectors, to predict either patch position (fig. 4) or image class (fig. 5).

Findings and Unfreezing Strategy for DinoV2 Our analysis reveals that the spectral properties of a block's feature map (fig. 3) and the ability to retrieve local information from its output patch tokens (fig. 4) are correlated. The classification accuracy for position and the mean of the principal eigenvalues peak in the middle-layers of the model, suggesting that the middle blocks capture local properties of patches (e.g., texture, relative position). Meanwhile, deeper blocks focus on global semantic understanding, as shown by increased classification accuracy for image class prediction in fig. 5. Combined, these results suggest that unfreezing deeper layers, such as $\mathcal{U} = \{\texttt{L}\}$, allows the model to better capture global features without overfitting to local details of images of D_T. This is empirically confirmed in table 3, where linear probing accuracy correlates inversely with the mean eigenvalue of each block (i.e., block 23 > block 22 > block 0 > block 9). The attention maps in fig. 8 further support this, showing that the deeper layers focus more clearly on central objects, while earlier layers (e.g., blocks 9, 10) exhibit more diffuse attention patterns spread around the border.

Findings and Unfreezing Strategy for MAE For MAE, we see a similar, but less pronounced trend. However, MAE is only trained for reconstruction, and so retains more local information across the ViT's layers. This is reflected by its lower patch-wise eigenvalues, higher localization accuracy, and lower global accuracies than Dino.

ExPLoRA's Impact D-ExPLoRA preserves local information in the middle layers but also improves localization accuracy in the last few layers. Importantly, it also enhances the global information contained in the patches for deeper model layers. This indicates a better understanding of the target domain, as seen in B.6, where ExPLoRA's attention highlights the central object more clearly.

7 Conclusion and Discussion

In this paper, we introduce ExPLoRA, a novel pre-training strategy to adapt pre-trained ViT foundation models for natural images to additional visual domains such as satellite imagery or medical data. We challenge the common paradigm of expensive pre-training from scratch for each new visual domain by offering a solution to transfer knowledge from foundation models that is both parameter-efficient and effective (even outperforming domain-specific foundation models). Our hope is that ExPLoRA enables further use of foundation models on domains other than natural images without requiring vast computational resources for pre-training.

While effective, there are many aspects of ExPLoRA that deserve further study. The strategy of fully training a small amount (or budget) of weights combines extremely well with PEFT techniques such as LoRA– we hope that future work investigates the reason behind this in further detail. Unresolved questions also include whether other parameter-efficient techniques might work better with ExPLoRA during pre-training. Further work to evaluate ExPLoRA for natural language would be valuable, as would an investigation into whether we can do away entirely with unfreezing a transformer block.

Broader Impact

As the scale of models and datasets grows exponentially, access to the computing power necessary to develop and make use of foundation models is increasingly restricted to the hands of a few organizations. This leaves many researchers in academia or smaller companies reliant on the resources of such organizations for ML research and applications. Techniques such as PEFT can alleviate this dependence and enable those with fewer computational resources to adapt, investigate, and customize models for their own needs. We hope that ExPLoRA furthers this goal, allowing ML practitioners to tailor foundation models with minimal compute, thus broadening access to powerful ML tools for critical fields like sustainability and medicine.

For example, automated analysis of satellite imagery can inform social, economic, and environmental policies, but manual curation is expensive, and pre-training models on such data has significant costs, both environmental and otherwise (see appendix D). ExPLoRA offers a more efficient way to distill knowledge from existing foundation models trained on natural images, sharply reducing costs while aiding researchers and policymakers and enabling flexible applications in downstream tasks.

REFERENCES

[1] Armen Aghajanyan, Luke Zettlemoyer, and Sonal Gupta. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. *arXiv preprint arXiv:2012.13255*, 2020.

[2] 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*, pp. 10181–10190, 2021.

[3] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. *arXiv preprint arXiv:2204.01678*, 2022.

[4] Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE transactions on medical imaging*, 38(2):550–560, 2018.

[5] Sara Beery, Elijah Cole, and Arvi Gjoka. The iwildcam 2020 competition dataset. *CoRR*, abs/2004.10340, 2020. URL https://arxiv.org/abs/2004.10340.

[6] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[8] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

[9] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

[10] Arnav Chavan, Zhuang Liu, Deepak Gupta, Eric Xing, and Zhiqiang Shen. One-for-all: Generalized lora for parameter-efficient fine-tuning. *arXiv preprint arXiv:2306.07967*, 2023.

[11] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[12] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*, 2022.

[13] 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, 2017.

[14] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

[15] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *CVPR*, 2018.

[16] Ren Chuan-Xian, Zhai Yi-Ming, Luo You-Wei, and Yan Hong. Towards unsupervised domain adaptation via domain-transformer. *arXiv preprint arXiv:2202.13777*, 2022.

[17] 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.

[18] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*, 2023.

[19] Etienne David, Simon Madec, Pouria Sadeghi-Tehran, Helge Aasen, Bangyou Zheng, Shouyang Liu, Norbert Kirchgessner, Goro Ishikawa, Koichi Nagasawa, Minhajul A Badhon, et al. Global wheat head detection (gwhd) dataset: A large and diverse dataset of high-resolution rgb-labelled images to develop and benchmark wheat head detection methods. *Plant Phenomics*, 2020.

[20] Etienne David, Mario Serouart, Daniel Smith, Simon Madec, Kaaviya Velumani, Shouyang Liu, Xu Wang, Francisco Pinto Espinosa, Shahameh Shafiee, Izzat SA Tahir, et al. Global wheat head dataset 2021: An update to improve the benchmarking wheat head localization with more diversity. *CoRR*, 2021.

[21] Sylvain Delattre and Nicolas Fournier. On the kozachenko–leonenko entropy estimator. *Journal of Statistical Planning and Inference*, 185:69–93, 2017.

[22] 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*, pp. 248–255. Ieee, 2009.

[23] 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.

[24] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33: 21271–21284, 2020.

[25] Cheng Han, Qifan Wang, Yiming Cui, Zhiwen Cao, Wenguan Wang, Siyuan Qi, and Dongfang Liu. Eˆ 2vpt: An effective and efficient approach for visual prompt tuning. *arXiv preprint arXiv:2307.13770*, 2023.

[26] 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*, pp. 16000–16009, 2022.

[27] 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.

[28] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[29] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pp. 709–727. Springer, 2022.

[30] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

[31] Samar Khanna, Patrick Liu, Linqi Zhou, Chenlin Meng, Robin Rombach, Marshall Burke, David B. Lobell, and Stefano Ermon. Diffusionsat: A generative foundation model for satellite imagery. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=I5webNFDgQ.

[32] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pp. 5637–5664. PMLR, 2021.

[33] Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. Quantifying the carbon emissions of machine learning. *CoRR*, abs/1910.09700, 2019. URL http://arxiv.org/abs/1910.09700.

[34] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Fractalnet: Ultra-deep neural networks without residuals. *arXiv preprint arXiv:1605.07648*, 2016.

[35] Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. *arXiv preprint arXiv:1804.08838*, 2018.

[36] Dongze Lian, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Scaling & shifting your features: A new baseline for efficient model tuning. *Advances in Neural Information Processing Systems*, 35:109–123, 2022.

[37] Weiyang Liu, Zeju Qiu, Yao Feng, Yuliang Xiu, Yuxuan Xue, Longhui Yu, Haiwen Feng, Zhen Liu, Juyeon Heo, Songyou Peng, et al. Parameter-efficient orthogonal finetuning via butterfly factorization. *arXiv preprint arXiv:2311.06243*, 2023.

[38] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024.

[39] Xin Man, Chenghong Zhang, Jin Feng, Changyu Li, and Jie Shao. W-mae: Pre-trained weather model with masked autoencoder for multi-variable weather forecasting. *arXiv preprint arXiv:2304.08754*, 2023.

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

[41] 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*, pp. 16806–16816, 2023.

[42] Théo Moutakanni, Piotr Bojanowski, Guillaume Chassagnon, Céline Hudelot, Armand Joulin, Yann LeCun, Matthew Muckley, Maxime Oquab, Marie-Pierre Revel, and Maria Vakalopoulou. Advancing human-centric ai for robust x-ray analysis through holistic self-supervised learning. *arXiv preprint arXiv:2405.01469*, 2024.

[43] Xing Nie, Bolin Ni, Jianlong Chang, Gaofeng Meng, Chunlei Huo, Shiming Xiang, and Qi Tian. Pro-tuning: Unified prompt tuning for vision tasks. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[44] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[45] Wenjie Pei, Tongqi Xia, Fanglin Chen, Jinsong Li, Jiandong Tian, and Guangming Lu. Sa^2vp: Spatially aligned-and-adapted visual prompt. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 4450–4458, 2024.

[46] George Pu, Anirudh Jain, Jihan Yin, and Russell Kaplan. Empirical analysis of the strengths and weaknesses of peft techniques for llms. *arXiv preprint arXiv:2304.14999*, 2023.

[47] Zeju Qiu, Weiyang Liu, Haiwen Feng, Yuxuan Xue, Yao Feng, Zhen Liu, Dan Zhang, Adrian Weller, and Bernhard Schölkopf. Controlling text-to-image diffusion by orthogonal finetuning. *Advances in Neural Information Processing Systems*, 36:79320–79362, 2023.

[48] Helen Qu and Sang Michael Xie. Connect later: Improving fine-tuning for robustness with targeted augmentations. *arXiv preprint arXiv:2402.03325*, 2024.

[49] 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*, pp. 4088–4099, 2023.

[50] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

[51] Linus Scheibenreif, Michael Mommert, and Damian Borth. Parameter efficient self-supervised geospatial domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27841–27851, June 2024.

[52] Shuchang Shen, Sachith Seneviratne, Xinye Wanyan, and Michael Kirley. Firerisk: A remote sensing dataset for fire risk assessment with benchmarks using supervised and self-supervised learning. In *2023 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pp. 189–196. IEEE, 2023.

[53] Jan-Martin O Steitz and Stefan Roth. Adapters strike back. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23449–23459, 2024.

[54] Tao Sun, Cheng Lu, Tianshuo Zhang, and Haibin Ling. Safe self-refinement for transformer-based domain adaptation, 2022. URL `https://arxiv.org/abs/2204.07683`.

[55] Maofeng Tang, Andrei Cozma, Konstantinos Georgiou, and Hairong Qi. Cross-scale mae: A tale of multiscale exploitation in remote sensing. *Advances in Neural Information Processing Systems*, 36, 2024.

[56] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[57] Adam Van Etten, Dave Lindenbaum, and Todd M Bacastow. Spacenet: A remote sensing dataset and challenge series. *arXiv preprint arXiv:1807.01232*, 2018.

[58] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

[59] Tongkun Xu, Weihua Chen, Pichao Wang, Fan Wang, Hao Li, and Rong Jin. Cdtrans: Cross-domain transformer for unsupervised domain adaptation. *arXiv preprint arXiv:2109.06165*, 2021.

[60] Jinyu Yang, Jingjing Liu, Ning Xu, and Junzhou Huang. Tvt: Transferable vision transformer for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 520–530, 2023.

[61] Dongshuo Yin, Leiyi Hu, Bin Li, and Youqun Zhang. Adapter is all you need for tuning visual tasks. *arXiv preprint arXiv:2311.15010*, 2023.

[62] Dongshuo Yin, Leiyi Hu, Bin Li, Youqun Zhang, and Xue Yang. 5%> 100%: Breaking performance shackles of full fine-tuning on visual recognition tasks. *arXiv preprint arXiv:2408.08345*, 2024.

[63] Seungryong Yoo, Eunji Kim, Dahuin Jung, Jungbeom Lee, and Sungroh Yoon. Improving visual prompt tuning for self-supervised vision transformers. In *International Conference on Machine Learning*, pp. 40075–40092. PMLR, 2023.

[64] Jielu Zhang, Zhongliang Zhou, Gengchen Mai, Lan Mu, Mengxuan Hu, and Sheng Li. Text2seg: Remote sensing image semantic segmentation via text-guided visual foundation models. *arXiv preprint arXiv:2304.10597*, 2023a.

[65] Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adaptive budget allocation for parameter-efficient fine-tuning. In *The Eleventh International Conference on Learning Representations*, 2023b.

[66] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021.

[67] Yukun Zhou, Mark A Chia, Siegfried K Wagner, Murat S Ayhan, Dominic J Williamson, Robbert R Struyven, Timing Liu, Moucheng Xu, Mateo G Lozano, Peter Woodward-Court, et al. A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7981):156–163, 2023.

[68] Jinjing Zhu, Haotian Bai, and Lin Wang. Patch-mix transformer for unsupervised domain adaptation: A game perspective. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3561–3571, 2023.

APPENDIX

We include supplementary material in the following sections.

A FURTHER CONTEXTUALIZATION WITH RELATED WORK

Here, we expand upon ExPLoRA's differences with recent related work.

A.1 COMPARISON WITH GEOSPATIAL DOMAIN ADAPTATION

Recent work has explored both continual pre-training (GFM) (Mendieta et al., 2023) and parameter-efficient domain adaptation (GDA) (Scheibenreif et al., 2024) for satellite imagery. We compare these approaches with ExPLoRA in table 10.

	GFM [41]	**GDA** [51]	**ExPLoRA** (Ours)
Parameter-efficient	✗	✓	✓
Training objective	MAE	MAE	Any
Architectural preservation	✓	✗	✓
Fine-tuning compatibility	Any PEFT	Only LoRA	Any PEFT

Table 10: Comparison of ExPLoRA with previous approaches to geospatial domain adaptation

ExPLoRA differs from these approaches in several key aspects. Unlike GFM which trains the full backbone, ExPLoRA achieves superior performance with only a fraction of trainable parameters. While GDA is also parameter-efficient, it requires non-mergeable scaling vectors that induce inference latency and modify the ViT, whereas ExPLoRA's LoRA adapters can be merged into the ViT's weights. Additionally, ExPLoRA extends beyond MAE architectures (supporting DinoV2 and others) and allows flexible configurations between pre-training and fine-tuning, including varying LoRA ranks or using different PEFT methods, which GDA doesn't support.

We also demonstrate ExPLoRA's broader applicability through experiments on larger datasets (fMoW-RGB, fMoW-Sentinel, which have 400k-800k images vs 90k images in FireRisk (Shen et al., 2023), the largest dataset used in GDA) and domains beyond remote sensing (i.e. WiLDS). Our analysis in section 6.3 provides insights into block-wise information encoding, offering practitioners a systematic approach for block selection during PEFT– a unique feature not present in prior work.

B ADDITIONAL EXPERIMENTAL RESULTS

We include further experimental results as a continuation of section 6.

B.1 RESULTS ON ADDITIONAL DOWNSTREAM DATASETS

NAIP We consider a land-cover classification dataset used in Ayush et al. (2021), where each of 244,471 training and 55,529 validation images are paired with one of 66 land cover classes obtained by the USDA's National Agricultural Imagery Program. In table 11, we first demonstrate similar performance between both natural-image backbones (rows 4 and 5), which surprisingly outperform SatMAE, which is pre-trained on fMoW-RGB. We use ExPLoRA to pre-train from DinoV2 to the training set of this dataset (without labels). Our results (row 6) demonstrate comparable performance, suggesting that for this dataset, domain-specific knowledge may not be highly relevant to successfully solve the task.

Method	PEFT	Top 1 Acc.
GASSL [2]	Full	57.63
SatMAE [17]	Full	**71.77**
SatMAE [17]	LoRA-r8	69.45
MAE [26]	LoRA-r8	70.36
DinoV2 [44]	LoRA-r8	**70.40**
D-`[L]`-$r32$	LoRA-r8	**70.40**

Table 11: NAIP validation set results

EuroSAT The dataset contains 27,000 13-band satellite images of 10 classes Helber et al. (2019), sourced from Sentinel-2. For ExPLoRA, we don't pre-train on the training set of this dataset, and instead use LoRA fine-tuning starting with the pre-trained weights learned in row 8 of table 4. We demonstrate improved performance over DinoV2, and match the performance achieved by the domain-specific SatMAE which was fully pre-trained on fMoW-Sentinel, and fully fine-tuned on EuroSAT (table 12). This demonstrates the successful use of our extended pre-trained model on further downstream datasets.

Method	PEFT	Top 1 Acc.
SeCo [40]	Full	93.14
SatMAE [17]	Full	98.98
SatMAE [17]	LoRA-r8	**98.73**
DinoV2 [44]	BOFT-b8m2	96.60
M-`[1,L]`$-r64$	LoRA-r8	98.54

Table 12: EuroSAT validation set results

B.2 The Importance of Extended Pre-training

To evaluate ExPLoRA's effectiveness, we analyze how its performance scales with computational resources. Specifically, we investigate two key questions: First, given a fixed compute budget, what is the optimal allocation between extended pre-training and fine-tuning? Second, for a fixed parameter budget, does investing compute in extended pre-training provide advantages over standard fine-tuning approaches?

We address these questions in fig. 6, focusing on DinoV2 models running on NVIDIA-A4000 GPUs. We evaluate D-ExPLoRA-[L]-$r64$ checkpoints at different stages of pre-training (50k, 100k, 150k, and 200k iterations), corresponding to 24, 48, 72, and 96 GPU-hours of extended pre-training respectively. Each checkpoint undergoes LoRA-$r8$ fine-tuning. We compare against three baselines: (i) Direct LoRA-$r8$ fine-tuning on DinoV2 weights (ii) Fine-tuning DinoV2 with block 24 unfrozen and LoRA-$r64$ (matching ExPLoRA's parameter budget) (iii) Fine-tuning DinoV2 with blocks 0, 1, 23, 24 unfrozen and LoRA-$r64$ (55.8M parameters vs ExPLoRA's 18.7M).

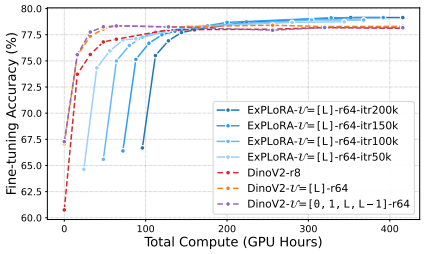

Figure 6: Fine-tuning accuracy versus total compute (measured in GPU-hours). Total compute includes both pre-training (if applicable) and fine-tuning phases.

Results in fig. 6 demonstrate that ExPLoRA's extended pre-training achieves a $\uparrow 0.9\%$ improvement in maximum top-1 fine-tuning accuracy within the same total compute budget (320 GPU hours). Notably, even increasing the parameter budget during fine-tuning fails to match this performance. While additional pre-training iterations beyond 50k improve initial fine-tuning accuracy, they have minimal impact on the final accuracy ceiling, highlighting ExPLoRA's computational efficiency.

B.3 Convergence and Data Efficiency

Another important question is on ExPLoRA's data efficiency- i.e. can ExPLoRA achieve good representations on the target domain without requiring many training iterations?

In fig. 7, we plot the linear-probing accuracy against the number of extended pre-training iterations for ExPLoRA (in blue). ExPLoRA improves quickly, requiring between 100-150k extended pre-training iterations to reach optimal performance. As discussed in section 6.1.2, unfreezing additional transformer blocks (in red) fails to achieve the same level of performance while requiring more parameters.

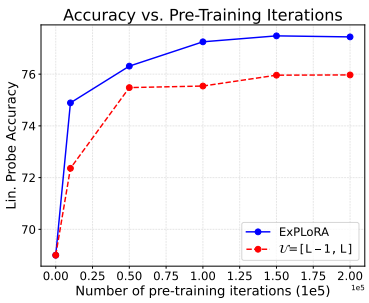

Figure 7: Lin. probe accuracy vs. number of training iterations.

One hypothesis for the effectiveness of pairing unfreezing blocks with LoRA tuning is that the low-rank updates to the ViT backbone "nudge" the sequence of embedded visual tokens from D_S to those representing D_T, which then enables the unfrozen ViT block to efficiently compress global information from the new domain.

B.4 Impact of MAE Decoder Rank

As outlined in section 5, we initialize the MAE decoder $\mathcal{L}_D$ with pre-trained weights W_{D_S} from He et al. (2022), keeping all decoder weights (except layer norm) frozen during extended pre-training on $\mathcal{X}_{D_T}$. We apply LoRA with rank r' to the Q, V weights of the attention layers in the decoder $\mathcal{L}_D$, while unfreezing 1-2 blocks $\mathcal{U}$ in the ViT encoder $\mathcal{L}$ and applying LoRA with rank r to the remaining layers $\mathcal{L} \setminus \mathcal{U}$ (algorithm 1).

Decoder Rank r'	Top 1 Acc.
8	59.75
16	59.77
32	**60.15**
64	59.21

Table 13: Ablation on M-$[1, \texttt{L}]$-$r64$ on the validation set of fMoW-Sentinel. Here, the LoRA rank used for the ViT-L encoder is fixed at $r = 64$, while the rank r' for MAE decoder is varied.

We evaluate ExPLoRA with M-$[1, \texttt{L}]$-$r64$ on fMoW-Sentinel, using a fixed encoder LoRA rank $r = 64$, unfreezing blocks $\mathcal{U} = \{1, \texttt{L}\}$, and varying the decoder rank r' We then fine-tune the resulting model with LoRA $r = 8$ and measure the highest top 1 accuracy on the validation set of fMoW-Sentinel. Results in table 13 show that increasing r' up to 32 improves fine-tuning performance, but performance declines when $r' = 64$, dropping $\downarrow 0.94\%$. This suggests that balancing the unfrozen parameters between the ViT encoder $\mathcal{L}$ (used for fine-tuning) and the MAE decoder $\mathcal{L}_D$ (discarded post pre-training) is crucial. Larger r' may improve the decoder's ability without benefiting the learned representations of $\mathcal{L}$. This issue doesn't arise in DinoV2, as the Dino-iBOT shared head is fully trained since it isn't provided by Oquab et al. (2023).

B.5 Impact of ViT backbone size

We also test the impact of the ViT backbone for ExPLoRA, varying the architecture for DinoV2 from ViT-B (86M, $\texttt{L} = 12$ layers, embedding dimension 768), ViT-L (303M parameters, $\texttt{L} = 24$ layers, embedding dimension 1024), and ViT-G (1100M parameters, $\texttt{L} = 40$ layers, embedding dimension 1280) for extended pre-training on fMoW-RGB. The ExPLoRA models we compare against are D-`[12]`-$r64$ for ViT-B, D-`[24]`-$r64$ for ViT-L, and D-`[32]`-$r32$ for ViT-G. We unfreeze the 12th, 24th, and 32nd layers for each of ViT-B, ViT-L, and ViT-G, picking these layers by extending the analysis from section 6.3 to ViT-B and ViT-G. We find that the 12th (last layer) for ViT-B and the 32nd (out of 40) layer for ViT-G output representations with low mean eigenvalues compared to other layers, thus presenting good candidates for unfreezing.

Method	Arch.	Top 1 Acc. Last 1/Last 4
DinoV2 [44]	ViT-B	63.62/65.90
DinoV2 [44]	ViT-L	67.60/69.00
DinoV2 [44]	ViT-G	70.07/70.36
D-`[12]`-$r64$	ViT-B	74.72/75.11
D-`[24]`-$r64$	ViT-L	76.86/77.48
D-`[32]`-$r32$	ViT-G	**77.29/77.79**

Table 14: Linear probing results on the validation set fMoW-RGB, where we vary the size of the ViT encoder $\mathcal{L}$ from ViT-B, ViT-L, and ViT-G. "Last 1/Last 4" refers to using the output representation from just the last 1 or the last 4 ViT layers, respectively.

In table 14, we see that as expected, ViT-G performs the best, but is only $\uparrow 0.31\%$ better in top 1 accuracy compared to ViT-L, while using many more parameters. On the other hand, we see the highest impact for ExPLoRA on ViT-B, where the top 1 accuracy improves by $\uparrow 9.21\%$ over the original DinoV2 ViT-B. These results further demonstrate the effectiveness and efficiency of ExPLoRA as a powerful technique to create unsupervised foundation models for new visual domains.

B.6 Attention Map Visualizations

To aid our analysis in section 6.3, we visualize attention scores for different ViT blocks across multiple models, including DinoV2, D-`[L]`-$r64$ (i.e. the last row of table 3), the second and third rows of table 3, MAE, SatMAE, and M-`[L]`-$r64$. These visualizations are shown in fig. 8 for 3 different images from the validation set of fMoW-RGB. Since our models are trained without registers, we truncate attention scores more than 5 standard deviations away from the mean, thus removing artifact attention scores with unusually high values on background patches (Darcet et al., 2023).

The visualizations in fig. 8 further support the analysis in section 6.3. For the Dino models, the attention scores of block 9-10 are diffuse and spread around the central object of the image, with quite a few border pixels highlighted. Conversely, the attention scores of the final layers are concentrated

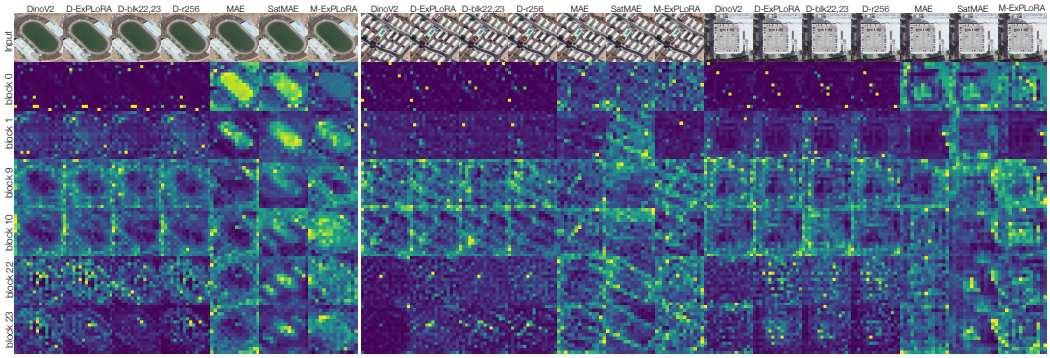

Figure 8: Attention maps visualized from the validation set of fMoW-RGB. The models considered, from left to right, are: DinoV2, D-ExPLoRA-`[L]`-$r64$, Dino with blocks 22,23 unfrozen during extended pretraining, Dino with LoRA-r256 during extended pre-training, MAE, SatMAE, and M-ExPLoRA-`[L]`-$r64$. We visualize the attention maps at the beginning, middle, and end blocks of the ViT-L.

more towards the central object. These visualizations further suggest that the middle layers focus on capturing local properties of the images such as texture, while the final layers capture global semantic information such as object-ness. Interestingly, the initial blocks for the Dino models display sparse attention patterns with spikes on seemingly random patches. This might suggest a form of caching to aid the computation of deeper layers that will extract local or global information.

For the MAE models, we see that the original MAE (pre-trained on natural images) seem to highlight more border pixels in the final layers of the ViT. Post extended pre-training with ExPLoRA, the final layers concentrate attention scores on the central object, more closely resembling the patterns of SatMAE (which was fully pre-trained on satellite images). ExPLoRA is thus able to successfully transfer knowledge from its initialized source-domain weights W_{D_S} to serve as a foundation model $W^*_{D_T}$ on the new target domain D_T.

B.7 Additional PEFT baselines for MAE

For completeness with table 1, we include PEFT methods used on MAE weights, which generally underperform compared with DinoV2. These results are in table 15.

Model	Arch.	PEFT	Pre-train #Params	Fine-tune #Params	Top 1 Acc.
SatMAE [17]	ViT-L	LoRA-$r8$ [28]	303.3M	0.8M	76.10
MAE [26]	ViT-L	LoRA-$r8$ [28]	-	0.8M	76.21
MAE [26]	ViT-L	DVPT-10 [29]	-	0.4M	72.35
MAE [26]	ViT-L	GVPT-100 [63]	-	0.4M	70.86
MAE [26]	ViT-L	SA^{2}VP [45]	-	1.1M	73.55
MAE [26]	ViT-L	AdaLoRA-$r8$ [65]	-	1.2M	75.25
MAE [26]	ViT-L	Adapter+ [53]	-	1.4M	74.10
M-`[L]`-$r64$	ViT-L	LoRA-$r8$ [28]	18.7M	0.8M	**76.55**

Table 15: MAE+PEFT results on the fMoW-RGB validation dataset, a continuation of table 1 The "Pre-train #Params" and "Fine-tune #Params" refer to the trainable parameters of the ViT encoder required on the *new* domain, i.e. satellite images. Since DinoV2 as an initialization outperformed MAE with all PEFT methods, we focus on DinoV2 in table 1 and include remaining PEFT with MAE experiments here for completeness.

C Training Details

In this section, we describe hyperparameters and hardware configurations used for our models.

C.1 Pre-Training

We use the ViT-Large architecture for all experiments. Since raw image sizes vary, the shorter image size is resized to 224 while preserving aspect ratio, and then a center crop is taken to yield images of

size $3 \times 224 \times 224$, representing the channels, height, and width. All pre-training is done on a single NVIDA-RTX 6000 Ada GPU, or 4 NVIDIA-RTX A4000 GPUs on an academic GPU cluster.

ExPLoRA with DinoV2 Most of the hyperparameters for D-ExPLoRA follow the defaults set by Oquab et al. (2023). That is, local (small) crops are between 5%-32% of the original image and are resized to 98x98 pixels, and global (large) crops are greater than 32% of the image and resized to 224x224 pixels. We share the parameters of the Dino-iBOT linear head (3 layers), with a bottleneck dimension of 256, a hidden dimension of 2048, and an output dimension of 65536, initialized from scratch. For Dino, we use Sinkhorn-Knopp (Caron et al., 2020) centering and Koleo (Delattre & Fournier, 2017) regularization with a weight of 0.1. For iBOT, we use masking ratios between 0.1 and 0.5 to mask half of the samples in the batch. The teacher model uses an initial EMA rate of 0.994, with a cosine warmup to 1.000 by the end of training. The teacher warmup and final temperatures are 0.04 and 0.07. The linear Dino-iBOT head is frozen for the first 3k training iterations. We train with the AdamW optimizer (no weight decay), with a base learning rate of 2×10^{-3} that is varied with a linear warmup and cosine decay schedule. Training is completed within 200,000 iterations, with a batch size of 32 and with 32 gradient accumulation steps (equalling an effective batch size of 1024), and with an epoch length set to 1000.

ExPLoRA with MAE Most of the hyperparameters we use for M-ExPLoRA pre-training follow those in He et al. (2022); Cong et al. (2022). We use an effective batch size of 1024 (through gradient accumulation), a base learning rate of 4.5×10^{-4}, no weight decay, and a warmup and decaying cosine scheduler, with a warmup of 1 epoch, and a total training time of 200 epochs. We use a masking ratio of 0.75 and we use the `norm_pix_loss` flag for the MSE loss.

C.2 PEFT Fine-Tuning

We fine-tune using 4 NVIDIA-RTX A4000 GPUs. We use a base learning rate of 10^{-3}, a cosine scheduler with warmup for 1 epoch, and train for 120 epochs. We use an effective batch size of 256, making use of gradient accumulation if the GPU cannot fit the full batch size in memory.

For data augmentations, we only use the drop-path augmentation (Larsson et al., 2016) at a rate of 0.2, with no dropout, mixup, or cutmix. We note that the original LoRA configuration outperforms other PEFT techniques when paired with the drop-path regularization technique. For example, we find that BOFT does not pair well with drop-path, instead performing most effectively with a custom multiplicative dropout technique (Liu et al., 2023). We include the result with the best hyperparameter configuration for each row in table 1.

C.3 Linear Probing

We use a single NVIDIA-RTX A4000 GPU for linear probing. We adapt the code provided by Oquab et al. (2023) for linear probing, with a batch size of 256 and a collection of different learning rates: $\left[1 \times 10^{-4}, 1 \times 10^{-3}, 5 \times 10^{-3}, 1 \times 10^{-2}, 2 \times 10^{-2}, 5 \times 10^{-2}, 1 \times 10^{-1}\right]$. We evaluate both probing on average pooled features as well as on the $[\texttt{CLS}]$ token, and also use output features from just the last block, or the last 4 blocks. All numbers reported represent the best validation set accuracy from the best performing configuration.

C.4 Multi-Spectral Images

We use the group-channel ViT-L architecture introduced in Cong et al. (2022). We don't use DinoV2 since there is no such architecture for DinoV2 pre-training. Input images are $13 \times 98 \times 98$, representing 13 multi-spectral bands. We follow the configuration in Cong et al. (2022) of dropping bands B1, B9, B10, and use the same grouping strategy. When loading MAE weights to the ViT-L encoder, the patch embeddings do not match and so the patch embedding and group channel encodings are trained from scratch. All other configuration details are the same as for M-ExPLoRA in appendix C.1, except that we use a base learning rate of 4.5×10^{-4} for pre-training and train for 50 epochs (given the larger dataset size) on 4 NVIDIA RTX A4000 GPUs for 80 hours.

Fine-tuning details are the same as in C.2.

C.5 DOWNSTREAM DATASETS

Hyperparameter and training configuration details are the same as in appendix C.1 if the images are RGB, and the same as in appendix C.4 if the images have more channels or are temporal.

C.6 DATASET LICENSES

The licenses for all datasets are included in the footnotes: fMoW[1], Sentinel-2[2], EuroSAT[3], SpaceNet[4], Camelyon17[5], iWildCam[6], GlobalWheat[7].

D ENVIRONMENTAL IMPACT

Following Cong et al. (2022), we compare the carbon footprint of pre-training using ExPLoRA with domain-specific solutions such as SatMAE. We use the carbon footprint calculator proposed by Lacoste et al. (2019). Our results are in table 16.

Method	fMoW-RGB		fMoW-Sentinel		fMoW-Temporal	
	GPU hours	kg CO_2 eq.	GPU hours	kg CO_2 eq	GPU hours	kg CO_2 eq.
SatMAE	768	109.44	576	82.08	768	109.44
ExPLoRA	**96**	**12.44**	**320**	**19.35**	**100**	**12.96**

Table 16: The estimated carbon footprint of pre-training on these datasets

Since we initialize with pre-trained weights on natural image domains, ExPLoRA is much less environmentally impactful while achieving similar or higher levels of performance. We achieve a 4x-8x reduction in total carbon emitted for each of the large pre-training satellite image datasets considered in table 16.

[1]fMoW license: `https://github.com/fMoW/dataset/raw/master/LICENSE`

[2]Sentinel-2 license: `https://scihub.copernicus.eu/twiki/pub/SciHubWebPortal/TermsConditions/Sentinel_Data_Terms_and_Conditions.pdf`

[3]EuroSAT license: `https://creativecommons.org/licenses/by/4.0/`

[4]SpaceNet v1 license: `http://creativecommons.org/licenses/by-sa/4.0/`

[5]Camelyon17 license:`https://creativecommons.org/publicdomain/zero/1.0/`

[6]iWildCam license:`https://cdla.dev/permissive-1-0/`

[7]GlobalWheat license:`https://opensource.org/licenses/MIT`