ExPLoRA: Parameter-Efficient Extended Pre-Training to Adapt Vision Transformers under Domain Shifts

Samar Khanna [1] **Medhanie Irgau** [1] **David B. Lobell** [1] **Stefano Ermon** [1]

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 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. 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% 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.

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; Chen et al., 2020; Caron et al., 2020; 2021; Grill et al., 2020; 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; Liu et al., 2023; Qiu et al., 2023; Zhang et al., 2023b; Chavan et al., 2023; Lialin et al., 2023) 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, 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., 2018a; Aghajanyan et al., 2020), or that efficient updates can be achieved by exploiting matrix structure (Liu et al., 2023; Qiu et al., 2023).

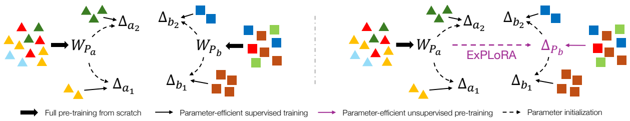

Figure 1: Consider two (fairly different) image domains, a and b. **Left**: the traditional paradigm of pre-training from scratch on each different domain to yield W_{P_a} and W_{P_b}, and then fine-tuning on the target datasets i to yield $\Delta_{a_i}, \Delta_{b_i}$, for domains a and b, respectively.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

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

In this work, we challenge this paradigm (Figure 1), asking whether pre-training from scratch on each new domain is strictly necessary, since doing so is expensive (in compute and time) and precludes knowledge transfer from natural images. We introduce ExPLoRA, which generalizes vision

[1] Stanford University. Correspondence to: Samar Khanna <samarkhanna@cs.stanford.edu>.

Published at ICML 2024 Workshop on Foundation Models in the Wild. Copyright 2024 by the author(s).

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 large, natural-image datasets. Selectively unfreezing 1-2 transformer blocks, we tune remaining weights with LoRA and continue unsupervised pre-training on the new domain. Subsequently fine-tuning using linear probing or LoRA on this new domain for supervised learning outperforms prior state-of-the-art (SoTA) approaches while training less than 5%-10% of the original weights. On satellite imagery, we demonstrate more than 7% improvement in linear probing top-1 accuracy over prior SoTA fully pre-trained and fine-tuned techniques. We conduct an extensive study on RGB, temporal, multi-spectral satellite images, and medical and wildlife imagery from WILDS (Koh et al., 2021), either matching or outperforming prior methods that use full-rank pre-training from scratch. Our contributions include:

1. Introducing ExPLoRA, a novel parameter-efficient method that extends unsupervised pre-training on target domains, achieving SoTA supervised-learning performance using a fraction of the original ViT weights.
2. Conducting a comprehensive case study on satellite imagery, outperforming existing techniques on datasets like fMoW. We also demonstrate generalization to other domains such as wildlife and medical imagery.

2. Background

MAE and DinoV2 Both the masked-autoencoder (MAE) (He et al., 2022) and DinoV2 (Oquab et al., 2023) are effective self-supervised learning techniques for ViTs. MAE uses an asymmetrical encoder-decoder architecture on images $\mathbf{x} \in \mathbb{R}^{C\times H\times W}$. Image patches are masked, and visible patches are fed to the ViT encoder $\mathcal{L}$. A smaller ViT decoder $\mathcal{L}_D$ reconstructs masked patches, aiming to minimize the mean-squared error on visible pixels. 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 loss from Dino (Caron et al., 2021) and a patch-based loss from iBOT (Zhou et al., 2021)

LoRA Low-rank adaptation (LoRA) (Hu et al., 2021) updates a set of unsupervised pre-trained weights to supervised fine-tuned weights via:

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

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 $r \ll \min(k_1, k_2)$.

3. Problem Setup

Consider a set of image domains $\mathcal{D} = \{1, 2, \dots\}$ and an associated data distribution for each domain $p_d(\mathbf{x})$, where $d \in \mathcal{D}$ and images $\mathbf{x} \in \mathbb{R}^{C_d\times H_d\times W_d}$, with channel, height, and width indexed by d. Let $D_P, D_F \subset \mathcal{D}$ be a set of domains representing the pre-training and fine-tuning data distributions $p_{D_P}(\mathbf{x})$ (eg: internet-scale natural image data) and $p_{D_F}(\mathbf{x})$ (eg: satellite data), respectively. Next, the fine-tuning joint distributions for each domain $d_F \in D_F$ are $p_{d_F}(\mathbf{x}, \mathbf{y})$, where $\mathbf{y}$ is the supervised-learning label.

We then assume access to the following: (i) pre-trained weights W_{D_P}, indexed by the collection of pre-training domains D_P, which have already been obtained via unsupervised learning (ii) samples from $p_{D_F}(\mathbf{x})$ representing unlabeled images from a new, different domain D_F (iii) a collection of target datasets $d_F \in D_F$ from the new domain D_F, sampled from distributions $p_{d_F}(\mathbf{x}, \mathbf{y})$. Thus, we would like to learn optimal weights W_{d_F} in a parameter-efficient manner for each supervised-learning dataset while leveraging knowledge stored in W_{D_P}.

Our goal is to learn fine-tuned weights W_{d_F} as follows:

$$W_{d_F} \approx W_{D_P} + \Delta_{D_F} + \Delta_{d_F} \tag{2}$$

where $\Delta_{d_F} \in \mathbb{R}^{k_2\times k_1}$ is an update matrix obtained via PEFT *supervised* learning (eg: LoRA) for the final domain d_F, and $\Delta_{D_F} \in \mathbb{R}^{k_2\times k_1}$ is an update matrix learned for the collection of fine-tuning domains D_F. Our key requirements for Δ_{D_F} are: (i) Δ_{D_F} must be learned via *unsupervised* pre-training on $p_{D_F}(\mathbf{x})$ (ii) Δ_{D_F} must only require learning a fraction of the k_1k_2 parameters that form W_{D_P}

Note that successfully learning Δ_{D_F}, would obviate the vast computing resources that are otherwise necessary to fully train foundation models for the new domain. Importantly, we emphasize learning Δ_{D_F} in an *unsupervised* manner so that the resulting model $W'_{D_F} = W_{D_P} + \Delta_{D_F} \approx W_{D_F}$ retains the benefits of pre-trained foundation models W_{D_F}, such as feature extraction, effective linear-probing, and generalization to further downstream tasks.

4. Method

To learn Δ_{D_F}, we propose ExPLoRA (i.e. **Ex**tended **P**re-training with **LoRA**). Let $\mathcal{L} = \{\mathtt{1}, \dots, \mathtt{L}\}$ denote the set of all $\mathtt{L}$ ViT blocks (or layers). For ViT-Large (ViT-L), $\mathcal{L} = \{1, \dots, 24\}$. The ExPLoRA approach is as follows:

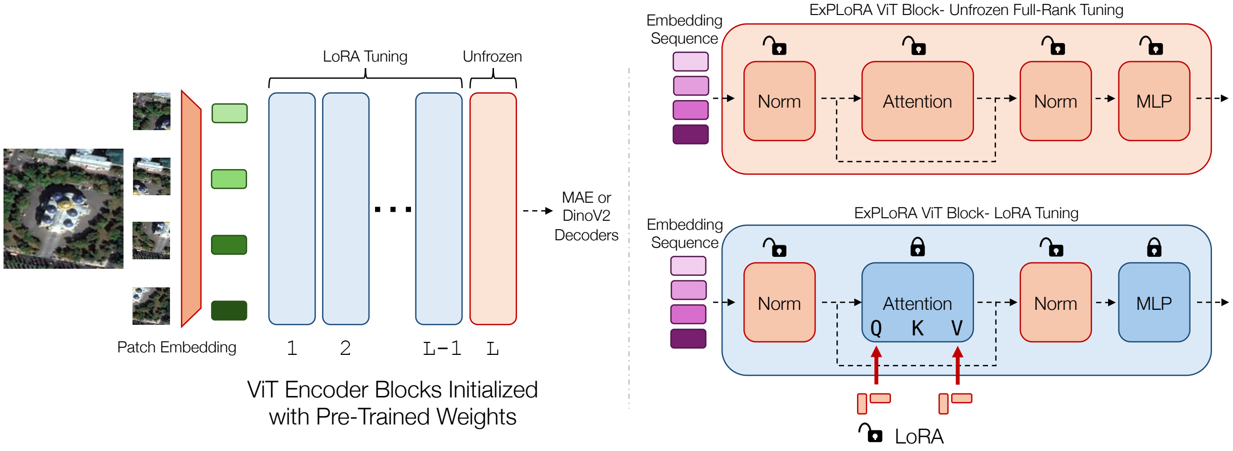

Figure 2: An overview of ExPLoRA. The set $\mathcal{L}$ of $\texttt{L}$ ViT blocks is partitioned into two sets: $\mathcal{U}$, which denotes blocks whose parameters are completely unfrozen, and $\mathcal{L} \setminus \mathcal{U}$ which denotes blocks that undergo LoRA tuning (only on the Q, V attention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

(i) Initialize the ViT with W_{D_P} from a large pre-training dataset (eg: DinoV2 or MAE weights).

(ii) Unfreeze all parameters of a subset of blocks $\mathcal{U} \subset \mathcal{L}$. Typically, $\mathcal{U} = \{\texttt{L}\}$ or $\mathcal{U} = \{\texttt{1,L}\}$.

(iii) For blocks $\mathcal{L} \setminus \mathcal{U}$, freeze all parameters and use LoRA with rank r on the Q, V weights of the attention layer.

(iv) Train the resulting model on *unlabeled* images $\mathbf{x} \sim p_{D_F}(\cdot)$ of the new domain D_F, with the same unsupervised objective as what was used for W_{D_P}.

In terms of notation, D-ExPLoRA-[$\texttt{L}$]-r32 would then denote a ViT initialized with DinoV2 weights (as opposed to M, which is MAE weights), where $\mathcal{U} = \{\texttt{L}\}$, and LoRA rank 32 is used on the Q, V matrices of every attention layer in $\mathcal{L} \setminus \mathcal{U}$. In this way, Δ_{D_F} requires 5-10% of the original parameters of the ViT while learning unsupervised representations for Δ_{D_F} that match or outperform learning W_{D_F} through full pre-training from scratch.

5. Experiments

We first conduct a case study on satellite imagery (Section 5.1), including an ablation study in Section 5.1.1. We then further evaluate on downstream tasks in Appendices C.1 and C.2 and Section 5.2. Training details including hyperparameter and compute configurations are mentioned in Appendix D. Results on the WiLDS benchmark are given in Section 5.2 and Appendix C.3.

5.1. Case Study: Satellite Imagery

We begin by examining satellite images largely because of the proliferation in works developing foundational models for satellite imagery via pre-training from scratch(Cong et al., 2022; Ayush et al., 2021; Reed et al., 2023; Tang et al., 2024) and since they represent a significant domain shift from natural images.

Dataset We use 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).

We compare our results in Table 1 against prior fully pre-trained SoTA foundation models as well as PEFT applied on pre-trained ViTs. Our results demonstrate that D-ExPLoRA-[$\texttt{L}$]-r32 is SoTA in terms of fMoW-RGB average accuracy at 79.09%, outperforming techniques that require fully pre-training ViTs on fMoW while using 5% of the original ViT encoder parameters. We then investigate linear-probing in Table 2, which entails training a linear head on extracted features from the frozen backbone, serving as a desirable metric of the quality of extracted embeddings. Our results demonstrate an improvement of over ↑7.3% in top 1 average accuracy over prior SoTA methods, demonstrating that ExPLoRA learns robust unsupervised representations for its target domain without expensive from-scratch pre-training. Importantly, ExPLoRA outperforms domain-specific prior SoTA methods (rows 1-4), as well as DinoV2, suggesting successful transfer learning on the target domain by leveraging pre-trained knowledge from natural images.

5.1.1. Ablation Study

Our ablation study (Table 3) on fMoW-RGB linear-probing determines whether our proposed configuration is optimal. A natural question is whether the performance improvement stems primarily from unfreezing blocks, or from LoRA-tuning the ViT. In row 1, we unfreeze $\texttt{L,L-1}$ (with no LoRA) and compare with ExPLoRA-$\texttt{L}$-r8 in row 6. Unfreezing an extra block consumes almost double the number of parameters without the same improvement in perfor-

Model	PEFT	Pre-train	Fine-tune	Top 1 Acc.
		#Params	#Params	
GASSL (Ayush et al., 2021)	Full	23.6M	23.6M	71.55
ScaleMAE (Reed et al., 2023)	Full	303.3M	303.3M	77.80
SatMAE (Cong et al., 2022)	Full	303.3M	303.3M	77.78
MAE (He et al., 2022)	Full	-	303.3M	76.91
SatMAE (Cong et al., 2022)	LoRA-r8	303.3M	0.8M	76.10
MAE (He et al., 2022)	LoRA-r8	-	0.8M	76.21
MAE (He et al., 2022)	BOFT-b2m8	-	0.9M	72.40
DinoV2 (Oquab et al., 2023)	LoRA-r8	-	0.8M	78.08
M-ExPLoRA-[L]-r32	LoRA-r8	15.7M	0.8M	76.42
D-ExPLoRA-[L]-r32	LoRA-r8	15.7M	0.8M	**79.09**

Table 1: Results on fMoW-RGB (validation). The "Pre-train #Params" and "Fine-tune #Params" refer to the trainable parameters required on the *new* domain, i.e. satellite images.

Method	Top 1 Acc.
GASSL	68.32
SatMAE	65.94
ScaleMAE	67.30
CrossScaleMAE	69.20
DinoV2	67.60
DinoV2†	69.00
D-ExPLoRA-[L]-r64	**75.77**
D-ExPLoRA-[L]-r64†	**76.53**

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.

Blocks	LoRA	Norm	LoRA	Num.	Top 1
Unfrozen	Rank	Unfrozen	Layers	Params	Acc.
[L-1,L]	0	✓	[]	25.3M	74.42
[]	256	✓	[Q,V]	25.9M	74.82
[]	128	✓	All	33.1M	55.03
[1]	32	✓	[Q,V]	15.7M	73.39
[L]	32	✗	[Q,V]	15.6M	75.14
[L]	8	✓	[Q,V]	13.4M	75.23
[L]	32	✓	[Q,V]	15.7M	75.44
[L]	64	✓	[Q,V]	18.7M	**76.53**

Table 3: Ablation study using DinoV2-ExPLoRA, measuring linear-probing accuracy on fMoW-RGB. If the LoRA rank is > 0, LoRA is only used on the frozen ViT blocks. All results are obtained by using concatenated features from the last 4 ViT blocks.

Method	PEFT	Top 1 Acc.
ConnectLater (Qu & Xie, 2024)	Full	93.90
ICON	Full	90.10
DinoV2	Lin. Probe	93.27
DinoV2	LoRA-r8	92.97
D-ExPLoRA-[L]$-r32$	Lin. Probe	**94.41**
D-ExPLoRA-[L]$-r32$	LoRA-r8	94.21

Table 4: Results on the validation set of Camelyon17

mance. Thus, simply increasing the number of unfrozen blocks is not as effective as ExPLoRA, and will also sharply decrease the parameter-efficiency.

Next, we see that high LoRA ranks used on all ViT layers (i.e. all attention and MLP matrices) significantly harms learning (row 3). In fact, it is much less effective than using just LoRA-r256 on the Q, V matrices of all $\mathcal{L}$ blocks (row 2). However, both rows 2 and 3 are much less parameter-efficient than ExPLoRA (rows 4-6). The choice of $\mathcal{U}$ matters as well. As seen in row 4 vs row 7, for the DinoV2 objective, $\mathcal{U} =$ [1] is not as effective as $\mathcal{U} =$ [L], ceteris paribus. We also notice a slight drop in accuracy from leaving the normalization layers across the ViT frozen, seen in row 5.

Lastly, we investigate the impact of LoRA rank on ExPLoRA. Changing the rank from 8 to 32 has a small improvement, but changing from 32 to 64 brings about a much larger improvement, 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. 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_P to those representing D_F, which then enables the unfrozen ViT block to effectively compress data from the new domain.

5.2. WiLDS Datasets

We also test ExPLoRA on the WILDS (Koh et al., 2021) benchmark, specifically the Camelyon17 (Bandi et al., 2018) and iWildcam (Beery et al., 2020) datasets, representing domain transfers to medical imagery (Section 5.2) and wildlife imagery (Appendix C.3), 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 4 demonstrate improved performance over domain-specific methods as well as DinoV2, once again successfully bridging the domain gap.

6. Conclusion

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.

References

Aghajanyan, A., Zettlemoyer, L., and Gupta, S. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. *arXiv preprint arXiv:2012.13255*, 2020.

Ayush, K., Uzkent, B., Meng, C., Tanmay, K., Burke, M., Lobell, D., and Ermon, S. Geography-aware self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10181–10190, 2021.

Bandi, P., Geessink, O., Manson, Q., Van Dijk, M., Balkenhol, M., Hermsen, M., Bejnordi, B. E., Lee, B., Paeng, K., Zhong, A., 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.

Beery, S., Cole, E., and Gjoka, A. The iwildcam 2020 competition dataset. *CoRR*, abs/2004.10340, 2020. URL https://arxiv.org/abs/2004.10340.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Chavan, A., Liu, Z., Gupta, D., Xing, E., and Shen, Z. One-for-all: Generalized lora for parameter-efficient fine-tuning. *arXiv preprint arXiv:2306.07967*, 2023.

Chen, X., Fan, H., Girshick, R., and He, K. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

Christie, G., Fendley, N., Wilson, J., and Mukherjee, R. Functional map of the world. In *CVPR*, 2018.

Cong, Y., Khanna, S., Meng, C., Liu, P., Rozi, E., He, Y., Burke, M., Lobell, D., and Ermon, S. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems*, 35:197–211, 2022.

Darcet, T., Oquab, M., Mairal, J., and Bojanowski, P. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*, 2023.

Delattre, S. and Fournier, N. On the kozachenko–leonenko entropy estimator. *Journal of Statistical Planning and Inference*, 185: 69–93, 2017.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M., and Lempitsky, V. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Helber, P., Bischke, B., Dengel, A., and Borth, D. 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.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Hu, W., Niu, G., Sato, I., and Sugiyama, M. Does distributionally robust supervised learning give robust classifiers? In *International Conference on Machine Learning*, pp. 2029–2037. PMLR, 2018.

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., Casas, D. d. l., Hanna, E. B., Bressand, F., et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

Kang, G., Jiang, L., Yang, Y., and Hauptmann, A. G. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4893–4902, 2019.

Khanna, S., Liu, P., Zhou, L., Meng, C., Rombach, R., Burke, M., Lobell, D. B., and Ermon, S. Diffusionsat: A generative foundation model for satellite imagery. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=I5webNFDgQ.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pp. 5637–5664. PMLR, 2021.

Lacoste, A., Luccioni, A., Schmidt, V., and Dandres, T. Quantifying the carbon emissions of machine learning. *CoRR*, abs/1910.09700, 2019. URL http://arxiv.org/abs/1910.09700.

Larsson, G., Maire, M., and Shakhnarovich, G. Fractalnet: Ultra-deep neural networks without residuals. *arXiv preprint arXiv:1605.07648*, 2016.

Li, C., Farkhoor, H., Liu, R., and Yosinski, J. Measuring the intrinsic dimension of objective landscapes. *arXiv preprint arXiv:1804.08838*, 2018a.

Li, H., Pan, S. J., Wang, S., and Kot, A. C. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5400–5409, 2018b.

Lialin, V., Muckatira, S., Shivagunde, N., and Rumshisky, A. Relora: High-rank training through low-rank updates. In *Workshop on Advancing Neural Network Training: Computational Efficiency, Scalability, and Resource Optimization (WANT@NeurIPS 2023)*, 2023.

Liu, W., Qiu, Z., Feng, Y., Xiu, Y., Xue, Y., Yu, L., Feng, H., Liu, Z., Heo, J., Peng, S., et al. Parameter-efficient orthogonal finetuning via butterfly factorization. *arXiv preprint arXiv:2311.06243*, 2023.

Ma, J., He, Y., Li, F., Han, L., You, C., and Wang, B. Segment anything in medical images. *Nature Communications*, 15(1): 654, 2024.

Man, X., Zhang, C., Feng, J., Li, C., and Shao, J. W-mae: Pre-trained weather model with masked autoencoder for multivariable weather forecasting. *arXiv preprint arXiv:2304.08754*, 2023.

Mañas, O., Lacoste, A., Giro-i Nieto, X., Vazquez, D., and Rodriguez, P. 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.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Oren, Y., Sagawa, S., Hashimoto, T. B., and Liang, P. Distributionally robust language modeling. *arXiv preprint arXiv:1909.02060*, 2019.

Pu, G., Jain, A., Yin, J., and Kaplan, R. Empirical analysis of the strengths and weaknesses of peft techniques for llms. *arXiv preprint arXiv:2304.14999*, 2023.

Qiu, Z., Liu, W., Feng, H., Xue, Y., Feng, Y., Liu, Z., Zhang, D., Weller, A., and Schölkopf, B. Controlling text-to-image diffusion by orthogonal finetuning. *Advances in Neural Information Processing Systems*, 36:79320–79362, 2023.

Qu, H. and Xie, S. M. Connect later: Improving fine-tuning for robustness with targeted augmentations. *arXiv preprint arXiv:2402.03325*, 2024.

Reed, C. J., Gupta, R., Li, S., Brockman, S., Funk, C., Clipp, B., Keutzer, K., Candido, S., Uyttendaele, M., and Darrell, T. 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.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. 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.

Singhal, P., Walambe, R., Ramanna, S., and Kotecha, K. Domain adaptation: challenges, methods, datasets, and applications. *IEEE access*, 11:6973–7020, 2023.

Sun, B., Feng, J., and Saenko, K. Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

Tang, M., Cozma, A., Georgiou, K., and Qi, H. Cross-scale mae: A tale of multiscale exploitation in remote sensing. *Advances in Neural Information Processing Systems*, 36, 2024.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7167–7176, 2017.

Van Etten, A., Lindenbaum, D., and Bacastow, T. M. Spacenet: A remote sensing dataset and challenge series. *arXiv preprint arXiv:1807.01232*, 2018.

Zhang, J., Zhou, Z., Mai, G., Mu, L., Hu, M., and Li, S. Text2seg: Remote sensing image semantic segmentation via text-guided visual foundation models. *arXiv preprint arXiv:2304.10597*, 2023a.

Zhang, Q., Chen, M., Bukharin, A., He, P., Cheng, Y., Chen, W., and Zhao, T. Adaptive budget allocation for parameter-efficient fine-tuning. In *The Eleventh International Conference on Learning Representations*, 2023b.

Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., and Kong, T. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021.

Zhou, Y., Chia, M. A., Wagner, S. K., Ayhan, M. S., Williamson, D. J., Struyven, R. R., Liu, T., Xu, M., Lozano, M. G., Woodward-Court, P., et al. A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7981): 156–163, 2023.

Appendix

We include supplementary material in the following sections.

A. Expanded Related Work

Visual Foundation Models Visual foundation models(VFMs), such as DinoV2 or masked autoencoders (MAE), have demonstrated remarkable performance across downstream tasks such as classification or semantic segmentation (Bommasani et al., 2021; Oquab et al., 2023; He et al., 2022). However, there has also been a rise in domain-specific VFMs (Cong et al., 2022; Zhou et al., 2023; Ma et al., 2024; Man et al., 2023; Zhang et al., 2023a), like SatMAE, which is designed to handle temporal or multi-spectral satellite imagery(Cong et al., 2022). With these models containing 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 to various downstream tasks, mitigating the prohibitive costs associated with full model tuning by updating only a fraction of the parameters. For example, 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; Lialin et al., 2023; Pu et al., 2023). Lialin et al. (2023) introduce LoRA for pre-training via the sum of multiple low-rank matrices, but require full parameter tuning as a "warm start". Some approaches use multiplicative orthogonal weight updates to frozen weights, effectively retaining pre-training knowledge (Qiu et al., 2023; Liu et al., 2023). While our ExPLoRA method can be configured with LoRA fine-tuning for downstream tasks, it supplements existing PEFT methods rather than replacing them, particularly in the case of unsupervised domain adaptation.

Domain Adaptation The central problem in domain adaptation is managing the distribution shift with respect to training and testing data. Domain adaptation approaches have explored this issue from several perspectives(Singhal et al., 2023). Discrepancy-based methods minimize the difference between feature distributions of the source and target domains using discrepancy metrics(Gretton et al., 2012; Kang et al., 2019; Sun et al., 2016) for domain loss. Adversarial methods aim to amplify domain confusion while simultaneously being rigorously trained to recognize and distinguish between different domains (Ganin et al., 2016; Li et al., 2018b; Tzeng et al., 2017). Group DRO aims to minimize the loss in the worst-case domain within the context of domain adaptation, specifically addressing subpopulation shift, where data distributions differ but may have some overlap (Hu et al., 2018; Oren et al., 2019).

B. Method: Further Details

In this section, we expand on Section 4 and provide details for each of the pre-training and fine-tuning configurations.

ExPLoRA for DinoV2 For our experiments, we use the DinoV2 ViT-L model as W_{D_P}, without registers (Darcet et al., 2023). We unfreeze the 24th block of the ViT, and use LoRA only on the query Q and value V matrices of each attention layer in all other blocks. We also unfreeze the normalization layers (which require very few parameters) throughout the network. We train each of the Dino, iBOT, and Koleo (Delattre & Fournier, 2017) inear heads fully, without any frozen parameters.

ExPLoRA for MAE We initialize both the MAE encoder and decoder with pre-trained weights. We only unfreeze the last block of the ViT encoder, and tune the Q, V matrices of each attention layer in all other blocks (including all blocks of the decoder) with LoRA.

For the multi-spectral MAE introduced in (Cong et al., 2022) we need to unfreeze the patch embedding layers for each group of channels (as these cannot be initialized from W_{D_P} since W_{D_P} only considers 3 channel RGB inputs). We then find that when using ExPLoRA, unfreezing blocks `1` and `L` is necessary to successfully achieve domain transfer.

Fine-tuning While fine-tuning DinoV2, we discard the linear heads used for the pre-training loss components (Section 2), and load all other model weights. Similarly, for MAE, we discard the decoder weights. We then initialize a linear head for classification, or decoder for segmentation, either of which is fully trainable. Freezing all ViT encoder weights, we then use LoRA-$r8$ on the Q, V matrices of every attention layer. We find that the drop-path augmentation (Larsson et al., 2016) is especially useful for fine-tuning, and use a value of 0.2 for the ViT-L models.

C. Additional Experiments

We expand on Section 5 with results on downstream satellite-image datasets in Appendices C.1 and C.2 as well as iWildcam (Beery et al., 2020) in Appendix C.3.

C.1. Multi-Spectral Satellite Images

Dataset 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 aim to assess the feasibility of domain transfer from natural images to multi-spectral, low-resolution satellite images. This presents a significant challenge compared to fMoW-RGB, as none of the natural image datasets in D_P include sensor information beyond visible light RGB bands. We utilize the group-channel ViT-L SatMAE model introduced in (Cong et al., 2022), initializing it with MAE self-supervised weights from the natural image domain. Since the patch embedding layers differ from those of MAE, we unfreeze and train them from scratch during ExPLoRA pre-training, resulting in minimal overhead to the parameter count.

Model	Backbone	PEFT	Pre-train #Params	Fine-tune #Params	Top 1 Acc.
ImgNet-Supervised	ResNet152	Full	60.3M	60.3M	54.46
MAE (He et al., 2022)	ViT-L	Full	-	303.3M	51.61
SatMAE (Cong et al., 2022)	ViT-L	Full	303.3M	303.3M	**61.48**
MAE (He et al., 2022)	ViT-L	LoRA-r8	-	0.8M	46.97
SatMAE (Cong et al., 2022)	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 5: 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.

From Table 5, we observe the challenge of domain transfer from natural images to multi-spectral satellite images, as discussed in Section 4. Even fully fine-tuning from MAE weights results in nearly a 10% drop in accuracy (row 2). LoRA tuning solely from MAE weights (row 4) performs even worse. Ablating by initializing with MAE weights and unfreezing only 4 transformer blocks during pre-training (row 6) is insufficient to bridge the domain gap. Notably, with ExPLoRA, unfreezing the first and last transformer blocks yields surprisingly good results, surpassing even fully pre-training from scratch (when using LoRA for fine-tuning). This underscores ExPLoRA’s ability to bridge wide domain gaps while utilizing only a fraction (in this case, around a tenth) of the original parameters.

C.2. Additional Satellite Datasets

fMoW-Temporal Also sourced from fMoW-RGB (Christie et al., 2018), each input is a sequence of up to 3 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 the dataset’s training images (without labels) with $\mathcal{U} = [\mathtt{L}]$ and LoRA rank 32. Our LoRA-tuned model then outperforms the domain-specific SatMAE for PEFT (Table 6), demonstrating successful transfer learning at a fraction of the pre-training parameters used by temporal SatMAE, which was fully pre-trained and fully fine-tuned on this dataset.

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 5. 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 7). This demonstrates the successful use of our extended pre-trained model on further downstream datasets.

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

Method	PEFT	Top 1 Acc.
GASSL (Ayush et al., 2021)	Full	74.11
SatMAE (Cong et al., 2022)	Full	79.69
MAE (He et al., 2022)	LoRA-r8	69.30
SatMAE (Cong et al., 2022)	LoRA-r8	75.27
M-ExPLoRA-[L]-r32	LoRA-r8	**75.98**

Table 6: fMoW-Temporal validation set results

Method	PEFT	Top 1 Acc.
SeCo (Mañas et al., 2021)	Full	93.14
SatMAE (Cong et al., 2022)	Full	98.98
SatMAE (Cong et al., 2022)	LoRA-r8	**98.73**
DinoV2 (Oquab et al., 2023)	BOFT-b8m2	96.60
M-ExPLoRA-[1,L]-r32	LoRA-r8	98.54

Table 7: EuroSAT validation set results

Method	PEFT	mIoU
GASSL (Ayush et al., 2021)	Full	78.51
SatMAE (Cong et al., 2022)	Full	78.07
ScaleMAE (Reed et al., 2023)	Full	**78.90**
DinoV2 (Oquab et al., 2023)	Lin. Probe	76.21
DinoV2 (Oquab et al., 2023)	LoRA-r8	76.69
D-ExPLoRA-[L]-r64	LoRA-r8	76.69

Table 8: SpaceNet validation set results

Method	PEFT	Top 1 Acc.
GASSL (Ayush et al., 2021)	Full	57.63
SatMAE (Cong et al., 2022)	Full	**71.77**
SatMAE (Cong et al., 2022)	LoRA-r8	69.45
MAE (He et al., 2022)	LoRA-r8	70.36
DinoV2 (Oquab et al., 2023)	LoRA-r8	**70.40**
D-ExPLoRA-[L]-r32	LoRA-r8	**70.40**

Table 9: NAIP validation set results

pre-train on the training set. A significant portion of the data consists of images with extensive black regions, indicating areas without meaningful visual information. Considering this limitation and the small dataset size, it is not clear whether additional pretraining 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 8).

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 9, 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.

C.3. Additional WILDS datasets

Lastly, we evaluate ExPLoRA on wildlife images from Beery et al. (2020). Note that the jump from natural images to medical images (Section 5.2) is larger than for wildlife images, likely because natural image datasets such as ImageNet(Deng et al., 2009) already contain many images of animals. However, there are little to no medical images such as images of cell tissue in the pre-training datasets for natural image models.

Method	PEFT	Top 1 Acc.
DinoV2 (Oquab et al., 2023)	Lin. Probe	66.04
DinoV2 (Oquab et al., 2023)	LoRA-r8	67.10
D-ExPLoRA-[L]-r32	Lin. Probe	62.95
D-ExPLoRA-[L]-r32	LoRA-r8	**68.07**

Table 10: Results on the validation set of iWildcam

iWildcam The iWildcam classification requires identifying one of 182 animal species given an image. We pre-train on the training set of the iWildcam classification task, finding that this outperforms pre-training on the extra-unlabeled set. In Table 10, 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.

D. Training Details

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

D.1. fMoW RGB

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. For D-ExPLoRA configurations, we use the default setting as used in the code provided by (Oquab et al., 2023) (including for the masking probabilities, relative loss weighting, data augmentations and transforms etc.) Similarly, we use the same settings for M-ExPLoRA pre-training as described in (He et al., 2022; Cong et al., 2022).

We use a single NVIDA-RTX 6000 Ada GPU, or 4 NVIDIA-RTX A4000 GPUs for all our experiments, on an academic GPU cluster. For all D-ExPLoRA configurations, we use a batch size of 32, a base learning rate of 5e-3, no weight decay, and a warmup and decaying cosine learning rate scheduler. We train for 200,000 iterations.

For M-ExPLoRA configurations, we use an effective batch size of 1024 (through gradient accumulation), a base learning rate of $1.5e-4$, no weight decay, and a warmup and decaying cosine scheduler, with a warmup of 1 epoch, and a total training time of 200 epochs.

Fine-tuning We use a base learning rate of 1e-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. We only use the drop-path augmentation, doing away with mixup and cutmix

Our GPU requirements for fine-tuning are the same as in pre-training.

Linear probing We adapt the code provided in (Oquab et al., 2023) for linear probing, with a batch size of 256 and a collection of different learning rates: $[1e-4, 1e-3, 5e-3, 1e-2, 2e-2, 5e-2, 0.1]$. We evaluate both probing on average pooled features as well as on the [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.

D.2. fMoW Sentinel

Pre-training 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 D.1, except that we use a base learning rate of 4.5e-4 for pre-training and train for 50 epochs (given the larger dataset size) on 4 NVIDIA RTX A4000 GPUs for 80 hours.

D.3. Downstream datasets

Hyperparameter and training configuration details are the same as in Appendix D.1 if the images are RGB, and the same as in Appendix D.2 if the images have more channels or are temporal.

D.4. 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]

[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/

E. Limitations and Future Work

While effective, there are many aspects of ExPLoRA that deserve further study. The strategy of unfreezing a small number of blocks 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 be better suited to work with ExPLoRA during pre-training. Further work to evaluate ExPLoRA for natural language domains would be valuable, as would an investigation into whether we can do away entirely with unfreezing a transformer block.

F. 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. Many researchers in academia or smaller companies are then completely dependent on the resources of such organizations in order to leverage ML for their own research or use-cases. Techniques such as PEFT can alleviate this dependence and enable those with much fewer computational resources to perform investigations and customize models for their own use-cases. We hope that ExPLoRA further enables ML practitioners and users to tailor foundation models for their own needs while requiring comparatively few resources. Our hope in doing so is to accelerate the deployment and use of ML for important domains such as sustainability or medicine.

For satellite images, for example, automated analyses that accurately characterize activity on the planet can guide an array of social, economic and environmental policies. Manually curating such observations is time-consuming and expensive, but pre-training foundation models on such data carries its own costs and environmental impact (see Appendix G). Thus, we provide a cheaper and more effective way to distill knowledge from foundation models that were already trained on natural images. Insights gained from such a model can further aid researchers and policymakers at a fraction of the cost, enabling more flexible uses of foundation models towards downstream datasets and tasks.

G. 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 11.

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 11: 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 11.