Empirical Study of PEFT techniques for Crop Type Segmentation

Mohamad Hasan Zahweh[1,2], Hasan Nasrallah[2], Mustafa Shukor [3], Ghaleb Faour[2] and Ali J. Ghandour ⬤ [2*]

[1]   Lebanese University;
[2]   National Center for Remote Sensing - CNRS, Lebanon;
[3]   Sorbone University;
[*]   Corresponding author: Ali J. Ghandour, aghandour@cnrs.edu.lb;

Abstract: Parameter Efficient Fine Tuning (PEFT) techniques have recently experienced significant growth and have been extensively employed to adapt large vision and language models to various domains, enabling satisfactory model performance with minimal computational needs. Despite these advances, more research has yet to delve into potential PEFT applications in real-life scenarios, particularly in the critical domains of remote sensing and crop monitoring. In the realm of crop monitoring, a key challenge persists in addressing the intricacies of cross-regional and cross-year crop type recognition. The diversity of climates across different regions and the need for comprehensive large-scale datasets have posed significant obstacles to accurately identify crop types across varying geographic locations and changing growing seasons. This study seeks to bridge this gap by comprehensively exploring the feasibility of cross-area and cross-year out-of-distribution generalization using the State-of-the-Art (SOTA) wheat crop monitoring model. The aim of this work is to explore efficient fine-tuning approaches for crop monitoring. Specifically, we focus on adapting the SOTA TSViT model, recently proposed in CVPR 2023, to address winter wheat field segmentation, a critical task for crop monitoring and food security, especially following the Ukrainian conflict, given the economic importance of wheat as a staple and cash crop in various regions. This adaptation process involves integrating different PEFT techniques, including BigFit, LoRA, Adaptformer, and prompt tuning, each designed to streamline the fine-tuning process and ensure efficient parameter utilization. Using PEFT techniques, we achieved notable results comparable to those achieved using full fine-tuning methods while training only a mere 0.7% parameters of the whole TSViT architecture. More importantly, we achieved the claimed performance using a limited subset of remotely labeled data. The in-house labeled data-set, referred to as the Lebanese Wheat dataset, comprises high-quality annotated polygons for wheat and non-wheat classes for the study area in Beqaa, Lebanon, with a total surface of 170 km², over five consecutive years from 2016 to 2020. Using a time series of multispectral Sentinel-2 images, our model achieved a 84% F1-score when evaluated on the test set, shedding light on the ability of PEFT to drive accurate and efficient crop monitoring, designed mainly for developing countries characterized by limited data availability. We intend to publicly release the Lebanese winter wheat data set, code repository, and model weights.

Keywords: Remote sensing; crop monitoring; PEFT; transformers; crop type recognition

1. Introduction

Foundation models, mostly based on transformer architecture, have demonstrated impressive success in many domains and benchmarks. However, the exponential increase in the number of parameters makes training these models extremely expensive and challenging. This motivated researchers to propose more efficient approaches to fine-tune and adapt these models to downstream tasks.

Parameter-Efficient FineTuning (PEFT) encompasses a family of approaches that keep most of the parameters of pretrained models frozen and train only part or additional parameters. PEFT approaches have been extensively investigated and used in the NLP community and recently in the vision community. However, most of the work still considers general
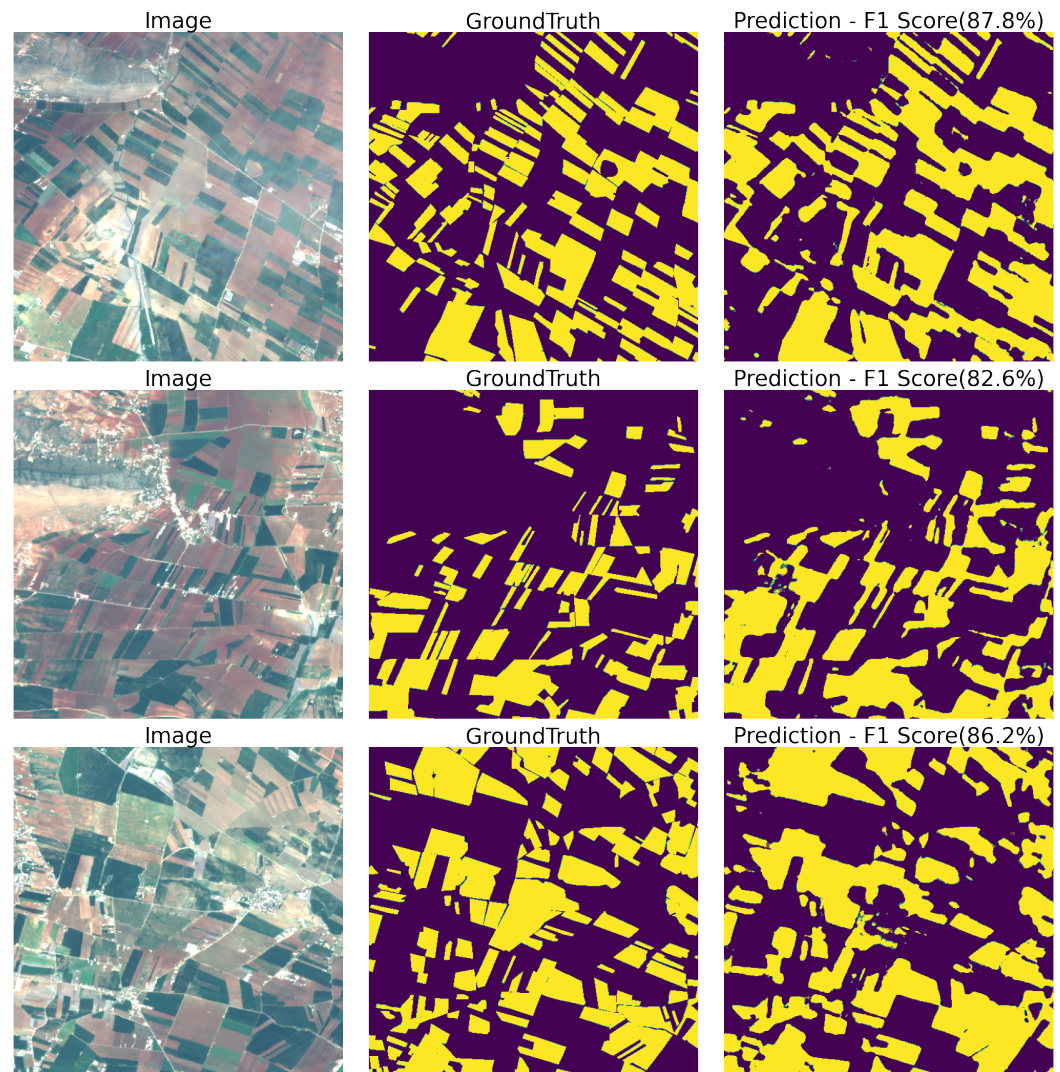
Figure 1. Best model predictions and F1-scores for some tiles versus the ground truth.

academic benchmarks and classical tasks. In this work, we explore efficient fine-tuning approaches for crop monitoring. Specifically, we investigated different PEFT techniques for crop monitoring. Specifically, we consider the recent state-of-the-art (SoTA), general satellite image time series model TSViT, and rather than fully fine-tune all model parameters, we fine-tune only a small number, using different techniques such as bias tuning, prompt tuning, adapters, or low rank adaptation. As shown in Figure 1, the predictions of our best model achieve high F1 scores. Our contributions are three-folds:

- We explore different PEFT techniques to efficiently adapt pre-trained models for crop type segmentation, and identify the most important parameters for each PEFT method to achieve the best performance.
- We experiment with different datasets from different countries like Germany and Lebanon and consider a realistic and challenging setup where we test the model on different year and regions.
- After an extensive set of experiments with the SoTA TSViT model, we produced an 85% F1-score on our test set with only 0.47% of the model parameters.

Method	Trainable Parameters (%)	F1-score
BitFit-Partial Bias	0.29	83.0
BitFit-Full Bias	0.54	83.9
VPT	0.29	83.5
LoRA	5.87	84.76
AdaptFormer	1.09	85.0
Head tune	0.05	56.0
Full finetuning	100	84.3

Table 1. Performance of various PEFT techniques and the baselines in term of F1-score and percentage of trainable parameters on Beqaa-Lebanon dataset. Adaptformer offers the best performance (even better than full fine-tuning) with an F1-score of 85%.

2. Framework

The Temporal Spatial Vision Transformer (TSViT) is a vision transformer that achieved SOTA for satellite images time series [1] in the PASTIS dataset [2]. The model consists of a temporal transformer followed by a spatial transformer. The architecture is based on common practices used for video processing, where the model used temporal encoding before spatial encoding.Given its state-of-the-art performance in crop field segmentation, we choose TSViT as the main model upon which we conduct our experiments and comparisons.

2.1. PEFT techniques

To avoid training all model parameters, especially large transformer models, great effort has been put into promoting efficient fine-tuning approaches [3]. Most of this work is focused on large-language models (LLMs) [4–9]. Some effort has been made to address vision model efficient tuning [10,11], however, very little work has considered PEFT techniques for remote sensing satellite images applications [12].

Parameter Efficient Fine Tuning (PEFT) techniques focus on training the least size of parameters in the most efficient way possible. These techniques achieve results equivalent to or better than full fine-tuning, reducing the cost of both training and storing the model with minimal loss (and sometimes non).

In our work, we investigate the effectiveness of the following PEFT techniques for the task of crop field segmentation given a time series of Sentinel-2 images: *(i)* BitFit, *(ii)* Visual Prompt Tuning, *(iii)* LoRA, *(iv)* Adapter.

BitFit [13] is a technique that focuses on training the Bias only, rather than the whole model. BitFit requires the least number of trainable parameters among other PEFT methods (note that head tuning requires fewer trainable parameters than BitFit).

Visual Prompt Tuning (VPT) [14] is a PEFT technique that adds additional prompt parameters concatenated to the input of the transformer that can help the model understand the different tasks and achieve better results. This prompt can be added to the first transformer encoder layer only, which is called **Shallow** prompt tuning, or added to each transformer encoder layer, which is known as **Deep** prompt Tuning. Furthermore, given that the TSViT architecture is uniquely characterized by the presence of two transformers, one focused on temporal information followed by another on spatial data, it opened up an opportunity to experiment with dual prompts.

LoRA [6] aims to train a smaller coinciding model with a low intrinsic rank r. Low-Rank Adaption (LoRA) freezes the original weight matrix W_0 and trains a separate update matrix that is formed by the product of two separate matrices A and B. In our LoRA experiments, we investigate the effect of the following hyperparameters of the additional modules: *(i) rt*: rank of the temporal transformer LoRA layers, *(ii) rs*: rank of the spatial transformer LoRA layers, and *(iii) rr*: rank of the rest of Lora layers.

Adapter [15] tuning is a method that aims to achieve efficient model adaptation by introducing two additional layers in the feed-forward steps. It can be added in series or in parallel with the feed forward layer.

Given the dual nature of the TSViT, our attention was primarily divided between two critical hyperparameters: the spatial adapter dimension and the temporal spatial dimension. The former pertains to the scale and complexity of spatial data processing, while the latter concerns the interpretation of temporal data. These dimensions essentially determine the granularity and depth of information processing within the transformer.

2.2. Baselines

Before the analytical comparison of our results, it is necessary to establish clear baselines. These baselines will serve as foundational benchmarks against PEFT. To ensure a logical comparison, we have incorporated a blend of traditional and specialized fine-tuning techniques. The techniques selected for our baseline are as follows:

- **Training from Scratch**: In this approach, the TSViT model was trained without using any prior weights. The goal of this strategy was to discern the innate potential of the TSViT architecture without the influence of fine-tuning.
- **Full Fine-Tuning**: Given a pre-trained model, we apply transfer learning by training all model parameters on a new data set using a low learning rate.It represents an aspirational benchmark: any PEFT technique that can perform as well or better than full fine-tuning would be deemed successful.
- **Head Fine-Tuning**: This technique introduces a layer to the front of the model, which then undergoes training. It is minimalist, targeting only the initial aspects of the model and setting the minimal performance expectation. Any PEFT technique that underperforms compared to this baseline would need re-evaluation.
- **Token Tuning**: TSViT stands out due to its reliance on temporal tokens for classification. This intrinsic characteristic opens up the possibility of a unique fine-tuning technique: Token Tuning. By manipulating the temporal tokens, one can effectively alter the output classes. It is a nuanced method, specifically for the TSViT model.

2.3. Datasets

We used two datasets in this study: *(i)* Beqaa-Lebanon that we prepared in-house for the scope of this work and *(ii)* Munich 480 dataset.

Beqaa-Lebanon data set was labeled for this project, where we annotated a weakly supervised data set of approximately 2,000 wheat parcels per year from 2016 to 2019. We then labeled an equivalent area of negative samples (non-wheat areas) per year to balance our data set. The total annotated area per year (from 2016 to 2019) is 170 km^2. Finally, we thouroughly annotated the year 2020 as a test set of surface area equal to 981 km^2, with about 16% positive samples (wheat).

The study area is divided into $24 * 24$ tiles as in the paper [1], which means that each input contains *T*24*24*C*, where *T* is the number of images that make up a time series of Sentinel-2 images for each wheat season, and *C* is the number of Sentinel-2 bands used. We set $T = 9$, corresponding to a series of images from November to July, and $C = 10$.

Munich 480 dataset [16] covers an area of $102km * 42km$ in Munich, Germany. It consists of 30 images per year split into $48 * 48$ tiles made of all 13 Sentinel-2 bands and covering 27 different crop classes.

3. Experimental Results

In this section, we will discuss the details of the implementation of the model, as well as the process of training or tuning the model using different techniques and baselines. In all our experiments, we used Adam optimizer, trained each model for 20 epochs, set batch size to 16 and kept a constant learning rate throughout the entire training process. All experiments were implemented using the Pytorch library on a single Nvidia GeForce TiTan XP 12GB GPU. But before that, we need first to introduce the embodied datasets.

	Temporal Dimension	Spatial Dimension	External Prompt	Deep Prompt	F1-Score(%)
Series 1	4	4	✗	✓	82.12
	8	8	✗	✓	82.88
	16	16	✗	✓	82.35
Series 2	4	4	✓	✓	82.16
	8	8	✓	✓	82.75
	16	16	✓	✓	**83.50**
Series 3	8	8	✓	✗	81.36
	4	4	✓	✗	79.00
Series 4	8	0	✓	✓	**83.50**
	0	8	✓	✓	69.73

Table 2. Visual Prompt Tuning Experimentation Table

In this subsection, we discuss and compare the results of different PEFT techniques on the Beqaa-Lebanon dataset. For comparison, F1-score is used as our main metric. Cross-validation was done after each epoch and the model with the best validation F1-score is saved.

3.1. BitFit and LoRA

BitFit is the simplest PEFT technique but showed promising results. As shown in Table 1, partial- and full-bias PEFT-ing of the TSVit model on the Beqaa-Lebanon dataset can achieve an F1 score of **83% and 83. 9%**, respectively. These results are comparable to full fine-tuning while only training **0.29% and 0.54%** of the model parameters, respectively.

For **LoRA**, we conducted four series of experiments to assess the influence of hyper-parameters discussed in Section 2.1. The best results are obtained when all intrinsic ranks are equal. This observation is best explained by the fact that the original and downstream tasks are the same here (crop-field segmentation), and thus the LoRA training should be balanced to match the flow of the model. The best results of LoRA tuning can be reached by balancing all modules' ranks where the highest F1-score of 84.9% is achieved with $rt = rs = rr = 4$. Detailed results are not included here for space limitations.

3.2. VPT and Adaptformer

Visual Prompt Tuning experiments delves further into the impact of various configurations. We conducted four series of experiments to assess the influence of different configuration settings, as shown in Table2.

- **Series 1**: External prompt is not used and the model is deep. The F1 score hovers around the low 82% range, regardless of the temporal and spatial dimensions.
- **Series 2**: We used an external deep prompt that shows outstanding performance, where the highest F1-score of 83.5% is achieved at dimensions $16 * 16$.
- **Series 3**: We used an external shallow prompt and witnessed a dip in performance, with F1 scores of 81.36% and 79% for dimensions $8 * 8$ and $4 * 4$, respectively. This suggests that although the shallow prompt might not be as effective as the deep prompt, it still outperforms head tuning, as shown in Table 1.
- **Series 4**: We experiment here with setting one dimension (temporal and spatial) at a time to zero while using an external deep prompt. An F1-score of 83.5% is achieved when the temporal dimension is set to 8 and the spatial dimension is equal to zero, but the performance drops significantly to 69.73% the other way around (temporal set to zero).

Series 1 and 2 showed that using the prompt inside the transformer's Q and K matrices only, which was proposed in the literature, does not have any advantage over adding it outside the transformer encoder layer. On the other hand, **Series 2 and 3** showed that deep prompts lead to better performance. **Series 4** concludes that prompt tuning of only
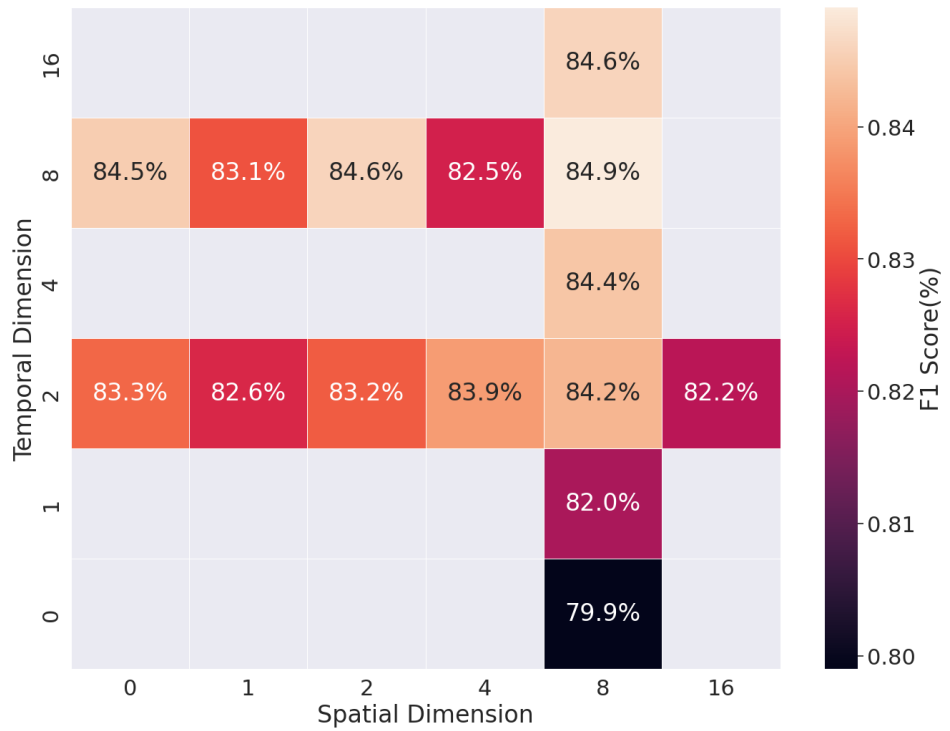
Figure 2. F1-scores results upon varying AdapterFromer's spatial and temporal dimensions on Beqaa-Lebanon dataset.

the temporal transformer is sufficient, and it is safe not to tune the spatial transformer.

Similarly to VPT, we test the influence of the **Adaptformer** spatial and temporal dimensions as shown in Figure 2. Spatial adapters showed negligible importance compared to temporal ones. A **84.5%** F1-score is attained when we set the temporal dimension to 8 and ignore the spatial dimension. On the contrary, we witness a **4.6%** drop in F1-score when doing the opposite (setting the temporal dimension to zero). AdaptFormer revealed the best performance among all PEFT techniques and the baselines with a F1-score of **84.9%** when the adapter's temporal and spatial dimensions were equal to 8.

We finally note that LoRA and Adaptformer were the only PEFT techniques that provided better F1-scores (84.76% and 85%, respectively) than full fining of the model while training only less than 1% of the model parameters. Also, as tabulated in Table 1, both Bitfit and VPT provided comparable but fewer results than full fine-tuning, while still having at least 27% better F1-score than the simple Head-tuning baseline. This shows that all considered PEFT techniques, when tuned carefully, are capable of reaching 98% of fully trained model performance.

We further investigated the effect of varying the learning rate over all PEFT methods. These experiments will not be shown here because of space limitations.

3.3. Munich 480 dataset

To further support our findings, we apply the following five training techniques in the Munich 480 dataset: *(i)* Head Tune, *(ii)* Partial Token Tune, *(iii)* Token Tune, *(iv)* AdaptFormer and *(v)* Training from scratch as shown in Table 3.

Partial and full token tuning provided better F1 scores (67.4% and 75.2%, respectively) than head tuning (63.9%). Adaptformer outperforms partial and full token tuning techniques with a margin greater than 10%. Moreover, Adaptformer only lays 4% behind full model training while only training 0.8%of the model parameters.

Training Technique	F1-score (%)	IoU (%)	Trainable Parameters (%)
AdaptFormer	84.7	74.3	0.808
Head Tune	63.9	48.4	0.079
Partial Token Tune	67.4	52.1	0.061
Full Token Tune	75.2	61.3	0.208
Training from scratch	88.9	80.7	100

Table 3. F1-scores of different investigated training scenarios in the Munich 480 dataset.

4. Conclusion

In this paper, we empirically studied the use of the parameter-efficient fine tuning (PEFT) technique to adapt the state-of-the-art TSViT model in different regions across the globe. We showed that PEFT methods can achieve accurate results with a size of about 1% of the model parameters. Our investigation provided several important findings on the effectiveness of applying parameter-efficient fine-tuning strategies in TSViT for winter wheat segmentation, where we were able to achieve an F1-score of 85% while training only 0.72% parameters. We analyzed each PEFT method's hyperparameters, and pointed out the best settings to achieve the highest possible individual performance.

Conflicts of Interest: "The authors declare no conflict of interest."

References

1. Tarasiou, M.; Chavez, E.; Zafeiriou, S. ViTs for SITS: Vision Transformers for Satellite Image Time Series, 2023, [arXiv:cs.CV/2301.04944].
2. Garnot, V.S.F.; Landrieu, L. Panoptic segmentation of satellite image time series with convolutional temporal attention networks. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 4872–4881.
3. Lialin, V.; Deshpande, V.; Rumshisky, A. Scaling down to scale up: A guide to parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.15647* **2023**.
4. Lester, B.; Al-Rfou, R.; Constant, N. The Power of Scale for Parameter-Efficient Prompt Tuning. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 3045–3059.
5. Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; Gelly, S. Parameter-efficient transfer learning for NLP. In Proceedings of the International Conference on Machine Learning. PMLR, 2019, pp. 2790–2799.
6. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* **2021**.
7. Li, X.L.; Liang, P. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190* **2021**.
8. Karimi Mahabadi, R.; Henderson, J.; Ruder, S. Compacter: Efficient low-rank hypercomplex adapter layers. *Advances in Neural Information Processing Systems* **2021**, *34*, 1022–1035.
9. Liu, H.; Tam, D.; Muqeeth, M.; Mohta, J.; Huang, T.; Bansal, M.; Raffel, C. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *arXiv preprint arXiv:2205.05638* **2022**.
10. Jia, M.; Tang, L.; Chen, B.C.; Cardie, C.; Belongie, S.; Hariharan, B.; Lim, S.N. Visual Prompt Tuning. In Proceedings of the European Conference on Computer Vision (ECCV), 2022.
11. Chen, S.; GE, C.; Tong, Z.; Wang, J.; Song, Y.; Wang, J.; Luo, P. AdaptFormer: Adapting Vision Transformers for Scalable Visual Recognition. In Proceedings of the Advances in Neural Information Processing Systems; Oh, A.H.; Agarwal, A.; Belgrave, D.; Cho, K., Eds., 2022.
12. Yuan, Y.; Zhan, Y.; Xiong, Z. Parameter-Efficient Transfer Learning for Remote Sensing Image-Text Retrieval. *IEEE Transactions on Geoscience and Remote Sensing* **2023**.
13. Zaken, E.B.; Ravfogel, S.; Goldberg, Y. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *arXiv preprint arXiv:2106.10199* **2021**.
14. Jia, M.; Tang, L.; Chen, B.C.; Cardie, C.; Belongie, S.; Hariharan, B.; Lim, S.N. Visual prompt tuning. In Proceedings of the European Conference on Computer Vision. Springer, 2022, pp. 709–727.
15. Chen, S.; GE, C.; Tong, Z.; Wang, J.; Song, Y.; Wang, J.; Luo, P. AdaptFormer: Adapting Vision Transformers for Scalable Visual Recognition. In Proceedings of the Advances in Neural Information Processing Systems; Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; Oh, A., Eds. Curran Associates, Inc., 2022, Vol. 35, pp. 16664–16678.
16. Rußwurm, M.; Körner, M. Multi-temporal land cover classification with sequential recurrent encoders. *ISPRS International Journal of Geo-Information* **2018**, *7*, 129.