ULISSE: Parameter-Efficient Adaptation of Earth Vision Models to Monitor Forest Disturbance in Sentinel-2 time series

Vito Recchia, Giuseppina Andresini, Annalisa Appice, Dino Ienco, Giuseppe Fiameni, Donato Malerba

ULISSE: Parameter-Efficient Adaptation of Earth Vision Models to Monitor Forest Disturbance in Sentinel-2 time series

Vito Recchia[a] (vito.recchia@uniba.it), Giuseppina Andresini[a,b] (giuseppina.andresini@uniba.it), Annalisa Appice[a,b] (annalisa.appice@uniba.it), Dino Ienco[c] (dino.ienco@inrae.fr), Giuseppe Fiameni[d] (gfiameni@nvidia.com), Donato Malerba[a,b] (donato.malerba@uniba.it),

[a] University of Bari Aldo Moro, Department of Informatics, via Edoardo Orabona, 4, Bari, 70125, Puglia, Italy

[b] Consorzio Interuniversitario Nazionale per l'Informatica - CINI, via Edoardo Orabona, 4, Bari, 70125, Puglia, Italy

[c] INRAE, UMR TETIS, University of Montpellier, Montpellier, France

[d] NVIDIA AI Technology Center, 2701 San Tomas Expressway, Santa Clara, 95050, California, US

Corresponding author at: Full address of the corresponding author, including the country name.

Giuseppina Andresini

University of Bari Aldo Moro, Department of Informatics, via Edoardo Orabona, 4, Bari, 70125, Puglia, Italy Tel: +39 080 5442407

Email: giuseppina.andresini@uniba.it

ULISSE: Parameter-Efficient Adaptation of Earth Vision Models to Monitor Forest Disturbance in Sentinel-2 time series

Vito Recchia[a], Giuseppina Andresini[a,b,*], Annalisa Appice[a,b], Dino Ienco[c,d], Giuseppe Fiameni[e], Donato Malerba[a,b]

[a] *University of Bari Aldo Moro, Department of Informatics, via Edoardo Orabona, 4, Bari, 70125, Puglia, Italy*
[b] *Consorzio Interuniversitario Nazionale per l'Informatica - CINI, via Edoardo Orabona, 4, Bari, 70125, Puglia, Italy*
[c] *INRAE, UMR TETIS, University of Montpellier, Montpellier, France*
[d] *INRIA, EVERGREEN, University of Montpellier, Montpellier, France*
[e] *NVIDIA AI Technology Center, 2701 San Tomas Expressway, Santa Clara, 95050, California, US*

Abstract

Europe is one of the most forest-rich regions in the world, with forestry mainly based on the management of coniferous trees. However, the spruce forest ecosystem is vulnerable to several disturbance agents. In particular, bark beetle outbreaks have been the scourge of spruce trees in the last decade, and they are expected to further intensify due to climate change, with significant adverse effects on forest ecosystems. Hence, the monitoring of forest disturbances caused by the rapidly escalating bark beetle outbreaks represents a significant ecological and forestry challenge. This monitoring is traditionally performed by foresters during field surveys. On the other hand, open Sentinel-2 images, available with the Copernicus mission and processed with sophisticated deep learning techniques, have been recently established as an alternative to field surveys performed by foresters to monitor various environmental phenomena such as bark beetle outbreaks. In particular, several deep learning approaches have been

*Corresponding author
Email addresses: vito.recchia@uniba.it (Vito Recchia), giuseppina.andresini@uniba.it (Giuseppina Andresini), annalisa.appice@uniba.it (Annalisa Appice), dino.ienco@inrae.fr (Dino Ienco), gfiameni@nvidia.com (Giuseppe Fiameni), donato.malerba@uniba.it (Donato Malerba)

recently proposed to map bark beetle tree dieback using Sentinel-2 images of forests. However, the current effectiveness of deep learning approaches, as a means to monitor bark beetle outbreaks in Sentinel-2 data is often limited by the reduced availability of ground truth information to supervise semantic segmentation models for this specific downstream task. In this study, we propose ULISSE, a deep learning semantic segmentation methodology for mapping forest tree dieback caused by bark beetle outbreak disturbances using Sentinel-2 image time series. ULISSE leverages a U-Net-like architecture with a multi-temporal encoder specifically designed to handle Sentinel-2 image time series. The framework integrates vision encoders pretrained on a large volume of Sentinel-2 images for land-cover classification. To capitalize on the representational capabilities of these pretrained encoders, we employ a Parameter-Efficient Fine-Tuning (PEFT) mechanism that adapts the multi-temporal encoder to our downstream segmentation task. This approach enables ULISSE to achieve high accuracy even with a limited amount of labelled data available at training time. Experimental results demonstrate the effectiveness of the proposed methodology across two case studies on mapping bark beetle disturbances in the Czech Republic and Romania study areas.

Keywords: Semantic Segmentation, Remote Sensing Vision Pretrained Models, Parameter-Efficient Fine-Tuning, Sentinel-2 Image time series Analysis, eXplainable Artificial Intelligence, Forest Disturbance Monitoring, Bark Beetle Outbreak Mapping

1. Introduction

Sentinel-2 is a high-resolution, multi-spectral satellite mission in the frame of the Copernicus programme,[1] operated by the European Union's Space program. The full mission specification was originally composed of two twin satellites (Sentinel-2A launched in June 2015, and Sentinel-2B launched in March 2017)

[1] `https://www.copernicus.eu/en`

flying in the same orbit, phased at 180°, and designed to have a revisit frequency of 5 days at the equator. The third Copernicus Sentinel-2 satellite (Sentinel-2C) was launched in September 2024 and, after a calibration time, it is expected to replace its predecessor, Sentinel-2A. Sentinel-2 data are freely available via the Copernicus Data Space Ecosystem [2]. The spectral band configuration of the Sentinel-2 mission is obtained with an optical instrument payload that samples 13 spectral bands, with four bands at 10m (B2, B3, B4, B8), six bands at 20m (B5, B6, B7, B8-A, B11 and B12) and three bands (B1, B9 and B10) at 60m spatial resolution, over an orbital swath width of 290 km. Nowadays, Sentinel-2 multispectral data are widely used in a plethora of Earth Observation applications, such as mapping changes in land cover (Andresini et al., 2023a; Abidi et al., 2024) and monitoring disturbances in forest areas (Reinosch et al., 2024).

In this study, we focus our attention on processing Sentinel-2 time series for mapping forest disturbance events caused by bark beetle outbreaks. Bark beetles are Scolytinae insects that create galleries in the phloem of their host spruce and disrupt the flow of nutrients in the tree. They represent one of the most aggressive natural disturbance agents across large areas of the European spruce forests. In particular, they are one of the major causes of forest tree dieback, particularly when their population density reaches the epidemic threshold. The timely and large-scale survey of bark beetle outbreaks is actually the main strategy against bark beetle disturbances. In fact, it allows forest managers to promptly plan forest sanitary cuts, which remain the most effective strategy to limit the spread of the disease outbreak. However, the survey activity is commonly based on terrestrial field work done with the assessment of foresters. This field work is laborious, time-demanding, and costly, particularly across large forests or forests with restricted accessibility (e.g., high-altitude forests). On the other hand, several recent studies have demonstrated that analyzing remote sensing imagery data via modern deep learning approaches can

[2] `https://dataspace.copernicus.eu/`

support forest managers by automating large-scale monitoring and mapping of forest tree dieback induced by bark beetle outbreaks.

Semantic segmentation is a Computer Vision task that learns a mapping model capable of assigning semantic labels to each pixel of a given image (Xiao et al., 2024). In general, the training phase of semantic segmentation methods uses full supervision by leveraging several images with their pixel-wise labelling maps as training data, in order to learn a mapping between the image content and its pixel-wise labels (Hao et al., 2020). Notably, in recent years, the performance of semantic segmentation methods for monitoring forest disturbance problems has been greatly enhanced through the use of sophisticated deep learning architectures (e.g., FCN, ViT, GCN, U-Net). These architectures have proven capable of achieving high segmentation accuracy by capturing spatial context characteristics within Sentinel-2 imagery (Thomas Ramos & Sappa, 2024).

In particular, the U-Net architecture, a widely adopted convolutional encoder-decoder network for imagery semantic segmentation, has been used in several recent studies to map forest tree dieback caused by bark beetle outbreaks across several European countries (Zhang et al., 2022; Andresini et al., 2024a,b). These research studies have highlighted that while deep learning models can achieve remarkable breakthroughs in semantic segmentation applications such as bark beetle outbreak mapping, their performance tightly depends on the quality and quantity of labelled training data.

In general, regarding the effectiveness of deep learning models developed for vision problems, a paradigm shift has been initiated by the recent emergence of large pretrained vision models (Xin et al., 2025). These state-of-the-art deep learning models are trained on massive image datasets with the intention of being reused across numerous downstream tasks. The underlying principle is that large pretrained vision models can be adapted to new applications through fine-tuning, rather than training models from scratch. This approach reduces development time and costs while mitigating challenges associated with limited labelled data and computational resources in downstream vision tasks. A notable

example is ImageNet (Deng et al., 2009), which has already served as a foundation for training a plethora of high-performance vision models. Following this research trend, recent advances in pretrained vision models have paved the way for a new era in the Earth Observation community (Lu et al., 2025) by enabling the development of models that learn generalized representations from satellite imagery at scale. Although established standards for supervised pretraining are emerging (Sumbul et al., 2021), with large pretrained vision models that have proven effective in several downstream tasks, these models still demand substantial computational and memory resources for conventional full fine-tuning strategies where every parameter of the model needs to be updated. Furthermore, when downstream labelled datasets are small, this adaptation strategy is prone to model overfitting, leading to poor generalization performance. Notably, obtaining large amounts of densely labelled data is challenging and expensive in many ecological scenarios, including bark beetle disturbance monitoring, where labels are commonly obtained through forester fieldwork. However, fieldwork based on human assessment of individual trees is labor-intensive and, for large forest areas, time-consuming and extremely costly (Bárta et al., 2021). In operational contexts, featured by a limited amount of labelled data, Parameter-Efficient Fine-Tuning (PEFT) methods have recently emerged as a solution to reduce the number of trainable parameters and computational overhead while aspiring to achieve performance comparable, or even better than, full fine-tuning on downstream tasks (Danish et al., 2026). Specifically, recent studies (Nwaiwu, 2025) have shown that PEFT methods can produce accurate predictive models even with limited labelled data. This is achieved by training only a small number of additional parameters, which reduces overfitting risk. Consequently, the general knowledge of pretrained vision models is preserved while enabling specific downstream task adaptation with the limited labelled data typically available in real-world ecological applications.

Based on the premises reported above, in this study, we introduce a deep learning semantic segmentation methodology, named ULISSE (U-net-based architecture Learned wIth Parameter-Efficient Fine-Tuning for Sentinel-2 imagery

timeSEries), for mapping bark beetle outbreaks in Sentinel-2 image time series of forest scenes. The idea of processing imagery time series for this task has matured in a few recent studies that have started the exploration of the potential of Sentinel-2 time series for mapping bark beetle outbreaks (Bárta et al., 2022; Jamali et al., 2023; Xu et al., 2024). However, these related studies consider temporal trends in pixel-level time series, giving up any potential information on the spatial context surrounding the pixel time series. Instead, information on the spatial context available surrounding a pixel allows semantic segmentation methods to boost accuracy performance in vision problems. Hence, our proposed methodology leverages the idea of learning a deep representation of the spatio-temporal information enclosed in Sentinel-2 image time series of forest scenes by resorting to a U-Net-like architecture. This architecture encloses a multi-temporal branch convolutional encoder, with one branch for each timestamp. This multi-temporal convolutional encoder is followed by a convolutional decoder. In addition, each convolutional encoder is a multispectral vision model, pretrained on a large amount of Sentinel-2 images for land cover classification. This multi-temporal branch encoder is fine-tuned on the downstream task using a PEFT mechanism, while the decoder is trained from scratch. In summary, the main characteristics of our proposed methodology are threefold: 1) It combines spatial and temporal information of Sentinel-2 image time series to improve accuracy in the considered semantic segmentation task. 2) It leverages a pretrained multispectral vision model, trained on a large amount of Sentinel-2 images, to design the multi-temporal encoder of the proposed semantic segmentation architecture. 3) It employs a PEFT mechanism to adapt the pretrained multispectral vision model, independently at each timestamp of the multi-temporal encoder, thereby reducing the number of trainable parameters.The experimental study provides an in-depth examination of the accuracy performance of the proposed methodology in two case studies regarding forest disturbances caused by bark beetles in Czech Republic and Romania.

The paper is organized as follows. Section 2 presents the related works. Section 3 describes the methodology used in this work. Section 4 illustrates

the results obtained in the evaluation study, while Section 5 discusses some of the remaining open challenges for evolving ULISSE from a proof-of-concept framework into a practical tool for large-area forest health monitoring. Finally, Section 6 draws conclusions and illustrates future developments.

2. Related work

This paper describes a deep learning semantic segmentation methodology to map forest disturbances caused by bark beetle outbreaks in Sentinel-2 time series. The proposed methodology is based on a multi-temporal, convolutional, encoder-decoder architecture for semantic segmentation and uses Parameter-Efficient Fine-Tuning mechanisms. Hence, the literature overview is organized into three fronts. Firstly, we focus on recent deep learning methods for semantic segmentation that have introduced seminal studies accounting for time series data (Section 2.1). Secondly, we analyze deep learning methods recently used for the semantic segmentation of bark beetle outbreaks in Sentinel-2 images (Section 2.2). Finally, we provide an overview of PEFT methods commonly employed to adapt pretrained vision models (Section 2.3).

2.1. Semantic segmentation

Semantic segmentation methods mainly train vision models to assign a label to each imagery pixel (Brar et al., 2025). Initial methods for semantic segmentation were mainly based on conventional thresholding, clustering, and morphological algorithms, while the most recent developments have been mainly dominated by deep learning methods since Shelhamer et al. (2017) proposed the use of Fully Convolution Networks (FCNs) for semantic segmentation problems. This network family integrates convolution layers followed by dropout layers to generate segmentation layouts and produce final pixel-wise classifications. Recent studies have extended FCNs to Pyramid Convolution Networks, which are able to learn rich contextual features to feed final semantic segmentation outcomes (Sang et al., 2020). Further studies train FCNs within Generative

Adversarial Networks to distinguish ground-truth segmentation maps of satellite images from maps produced by a segmentation arrangement (Mansourifar et al., 2022). More recently, few studies have explored the performance of Vision Transformers (ViTs) in semantic segmentation problems. These studies use self-attention mechanisms to obtain a deep representation of global relationships among imagery patches. However, as ViTs are not equipped with segmentation heads, mask transformers are introduced, to conduct image segmentation on ViT-based features (Kerssies et al., 2025). On the other hand, encoder-decoder architectures have been overwhelmingly successful in Computer Vision and remain among the most powerful deep learning architectures for semantic segmentation. In particular, U-Net is a popular family of convolutional encoder-decoder networks that is often used for semantic segmentation of Earth Observation imagery (Li et al., 2024; Dimitrovski et al., 2024). It uses fully convolution layers and skip connections to link feature maps from the convolutional encoder path to the corresponding convolutional decoder path.

With regard to the recent deep learning developments for satellite image time series classification, Pelletier et al. (2019) have described a Temporal Convolutional Neural Network (TempCNN) method, to train a classification model for multi-temporal Sentinel-2 images. In particular, TempCNN accounts for temporal features from Sentinel-2 image time series using a 1D convolutional operator, whereas the mapping task is performed by considering each pixel as an instance, without accounting for the spatial structure of the Sentinel-2 images. Differently, the Temporal-Spatial Vision Transformer (TSViT) method (Tarasiou et al., 2023), on a ViT architecture, is a proper semantic segmentation model, trained from scratch leveraging both spatial and temporal information, simultaneously. Specifically, it splits Sentinel-2 image time series into non-overlapping patches in space and time. These patches are tokenized for processing by a factorized temporal-then-spatial encoder. More recently, Szwarcman et al. (2024) have described Prithvi-EO-2.0. This is another spatio-temporal deep neural architecture based on a ViT pretrained with a masked autoencoder. In particular, Prithvi-EO-2.0 replaces the traditional ViT 2D positional embeddings with

3D patch embeddings to process spatio-temporal features. The model was pretrained accounting for six spectral bands only (i.e., Blue, Green, Red, Narrow NIR, SWIR 1, and SWIR 2). The pretrained model is adapted to downstream tasks with a full fine-tuning strategy. A ViT-based method, named SatMAE, is also described by Cong et al. (2022). This also integrates a pretrained ViT that can be adapted to any downstream task with a full fine-tuning strategy. In particular, the architecture of SatMAE is based on masked autoencoders trained on a sequence of non-overlapping imagery patches that are mapped into tokens containing both spatial and temporal embeddings. This method is formulated to be used with Sentinel-2 images in both classification and semantic segmentation tasks. However, the evaluation conducted in the referred study accounted for the temporal information in the classification task only, while it neglected any temporal information for the underlying semantic segmentation task. A further deep learning method formulated for handling spatio-temporal information of Sentinel-2 image time series is described by Cai et al. (2023). This method leverages a pretrained temporal encoder, named Exchanger. The Exchanger encoder is in charge of extracting the temporal feature embedding through a three-step process formulated with a collect–update–distribute paradigm. The Exchanger encoder is followed by the Mask2Former (Cheng et al., 2022) module, which includes masked attention, to extract localized features by constraining cross-attention within predicted mask regions. Extracted features are used to support semantic segmentation in land-cover segmentation tasks. A pretrained model, tailored for land-cover semantic segmentation, is made available for full fine-tuning adaptation for new downstream tasks. Finally, Zhu et al. (2025) have recently started the investigation of weakly supervised methods for semantic segmentation of multi-temporal Sentinel-2 images by introducing space-time perceptive clues to reduce the noise perturbation and rectify wrong semantic bias.

Several studies already integrate pretrained vision models into their proposed frameworks, but they rely on full fine-tuning adaptation strategies that are prone to overfitting in applications with limited labelled data. In contrast, our study

explores the use of PEFT mechanisms to adapt pretrained vision models for Sentinel-2 data with the aim of mitigating overfitting risks that may occur in real-world ecological downstream tasks. This represents the main motivation behind this research.

2.2. Semantic segmentation for bark beetle outbreak inventory

Although, several studies (Bárta et al., 2021; Candotti et al., 2022; Andresini et al., 2023b, 2024c) formulate the inventory of bark beetle outbreaks in Sentinel-2 images as a pixel-wise classification task and use traditional Machine Learning methods (e.g., Random Forest, XGBoost, Support Vector Machines) to support foresters in automatizing the inventory process, the latest research is rewarding deep learning methods for this task. For example, Zhang et al. (2022) proposes a U-Net architecture for mapping bark beetle outbreaks in Sentinel-2 images. The proposed architecture uses a Residual Network (ResNet) backbone as encoder, and introduces an attention mechanism in the decoding phase. Notably, in the field of remote sensing, the ResNet encoder is commonly used as the backbone of several U-Net architectures (Fan et al., 2022; Bhardwaj et al., 2025; Ramos & Sappa, 2025). More recently, Andresini et al. (2024b) have described a U-Net architecture for mapping bark beetle outbreaks in Sentinel-2 images. This model integrates an attention mechanism in the encoder, to amplify the crucial information, and uses a self-distillation approach in the decoder to transfer the knowledge within the U-Net architecture and obtain an ensemble-based classification of each pixel. The accuracy of this semantic segmentation method is evaluated on two case studies regarding the inventory mapping of forest tree dieback caused by insect outbreaks and wildfires, respectively. In addition, Andresini et al. (2024a) illustrates a multisensor data fusion schema to combine Sentinel-1 and Sentinel-2 data via fusion mechanisms in an underlying U-Net architecture trained for mapping bark beetle outbreaks in forest scenes. Recchia et al. (2024) describes an Attention-based CNN architecture trained for image classification and used with images of pixels obtained by seeing each imagery pixel within its surrounding, squared pixel neighborhood. On the other hand,

Pasquadibisceglie et al. (2025) introduces storytelling to obtain semantic stories of spectral-spatial features computed for Sentinel-2 imagery pixels. To this purpose, they use a Large Language Model, fine-tuned to the downstream task data, to perform the inventory of bark beetle outbreaks in forest scenes.

To the best of our knowledge, existing deep learning methods for mapping bark beetle outbreaks employ high spatial resolution images, with limited attention devoted to exploring Sentinel-2 image time series. Following recent literature, Sentinel-2 time series to map forest disturbances caused by bark beetle outbreaks are getting more and more attention Bárta et al. (2022); Andresini et al. (2024c); Östersund et al. (2024). However, in these pioneering studies, Sentinel-2 time series are processed to train a conventional pixel-wise classification model using standard Machine Learning techniques (i.e., Random Forest). Hence, these studies account for pixel-level temporal trends in Sentinel-2 time series, but neglect any potential knowledge derived from the spatial context of imagery pixels. The consideration of the spatial context information is one of the major reasons for the remarkable performance of vision models. These results motivated the idea developed in this study of exploring the performance of a deep learning architecture designed for mapping bark beetle outbreaks by considering Sentinel-2 time series.

2.3. Parameter-Efficient Fine-Tuning

Parameter-Efficient Fine-Tuning (PEFT) has recently emerged as a promising alternative to standard full fine-tuning strategies when adapting pretrained models to downstream tasks. When labelled data for a specific downstream task is limited, PEFT methods offer a practical solution by adjusting only a small number of additional parameters rather than updating the entire set of parameters given a large pretrained model. This strategy allows us to reduce both the computational resources required during fine-tuning and the risk of model overfitting (Han et al., 2024).

PEFT methods were firstly introduced in Natural Language Processing (NLP) to enable efficient transfer of Large Language Models across tasks (Houlsby

et al., 2019). For example, Zaken et al. (2022) propose a sparse fine-tuning method that modifies the bias parameters only, while keeping all other weights frozen. Low-Rank Adaptation (LoRA)(Hu et al., 2022), also originally developed for NLP, leverages low-rank decomposition to efficiently approximate weight updates. Specifically, LoRA learns a small set of additional parameters that are combined with the pretrained model weights to adapt this model to new downstream tasks. This low-rank adaptation strategy has inspired several extensions and has become one of the dominant PEFT paradigms, offering an excellent balance between performance, implementation simplicity, and reduced inference overhead (Yang et al., 2024).

In the field of Computer Vision, PEFT methods have only recently begun to gain attention after becoming well-established in NLP (Xin et al., 2024). In this regard, Aleem et al. (2024) introduces an extension of the LoRA method for Convolutional Neural Networks. Specifically, they design a novel approach that adds trainable low-rank decomposition matrices to convolutional layers, enabling adaptation of the underlying CNN backbone to new downstream classification tasks. The resulting framework is flexible enough to accommodate any modern pretrained Large Vision Model based on Convolutional Neural Network architectures. Recently, Mai et al. (2025) has performed an empirical study to evaluate the performance of LoRA with a ViT backbone adapted for downstream image classification tasks.

Liu et al. (2024b) proposes an extension of LoRA, referred to as Weight-Decomposed Low-Rank Adaptation (DoRA), which leverages pretrained parameter decomposition based on a Low-Rank Adaptation strategy. Specifically, DoRA decomposes the pretrained model's weights into two components—magnitude and direction—for fine-tuning, employing low-rank decomposition for direction updates with the primary objective of limiting the number of trainable parameters. Some experimental assessments have shown that this PEFT method demonstrates improved training stability while avoiding additional inference overhead compared to the original LoRA technique. As an alternative to LoRA, Qiu et al. (2023) proposes an orthogonal fine-tuning strategy, which is based on

pairwise neuron relationships on the unit hypersphere, and aims at preserving the geometric properties of the original model's parameters. Recently, Yuan et al. (2024a) has introduced a novel adaptation strategy, called Householder Reflection Adaptation (HRA), which bridges the gap between low-rank decomposition strategies and orthogonal adaptation. Given a large pretrained model, HRA fine-tunes its layers by multiplying each frozen set of parameters with an orthogonal matrix constructed from a chain of learnable Householder Reflections (HRs). This HR-based orthogonal fine-tuning is equivalent to an adaptive low-rank adaptation mechanism. Like the majority of PEFT methods, HRA seeks a compromise between the accuracy performance of the fine-tuned large model and the number of parameters updated during the refinement process.

Regarding the use of PEFT methods in Earth Observation tasks, Marti Escofet et al. (2026) have compared the performance of several PEFT methods, comprising LoRA, used in combination with some pretrained vision models in a water surface semantic segmentation downstream task. The evaluation, conducted using Sentinel-2 images acquired for several scenes at a single timestamp, shows that LoRA achieves the best performance among the considered approaches in the comparative study. This result underlines the potential of LoRA for the parameter-efficient fine-tuning of vision models especially tailored for satellite imagery data.

However, to the best of our knowledge, no previous study has explored the use of PEFT methods for a semantic segmentation downstream task involving Sentinel-2 time series data, which represents the novel contribution of this work.

3. Methodology

In this section, we describe the methodology proposed and evaluated in this study. Section 3.1 presents an overview of ULISSE. Section 3.2 introduces the U-Net-like architecture, while Section 3.3 describes the PEFT mechanisms associated with the encoder branches of ULISSE. Section 3.4 outlines the training procedure. Section 3.5 illustrates the whole pipeline, from data creation and

Table 1: List of main symbols

Symbol	Description
$\mathcal{D}$	labelled collection of Sentinel-2 image time series
$[\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_T}]$	Sentinel-2 image time series
M, N, C	Height, width and number of channels of a Sentinel-2 image
T	Length of a Sentinel-2 image time series
$\mathbf{Y}$	Ground truth binary map of a Sentinel-2 image time series
F	Semantic segmentation function
$\mathbf{W}_0^l$	Weight matrix computed at layer l
$\mathbf{E}^l$	Feature embedding matrix comouted at layer l
C_l	Number of channels at layer l
$\mathbf{A}^l, \mathbf{B}^l$	Low-rank matrices computed with LoRA/DoRA
r	Rank of $\mathbf{A}^l$ and $\mathbf{B}^l$
$\mathbf{m}^l$	Magnitude vector computed with DoRA
$\mathbf{H}_z^l$	Householder matrix computed with HRA

preparation through model training to the production of semantic segmentation maps. Table 1 introduces the main symbols of the notation used in the remaining part of this section.

3.1. Overview of ULISSE

Let us consider $\mathcal{D} = \{([\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_T}],\ \mathbf{Y})\}$, a labelled collection of Sentinel-2 image time series for non-overlapped scenes of a forest area. Each Sentinel-2 image time series $[\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_T}] \in \mathbb{R}^{M \times N \times C \times T}$ is a tensor of T timestamped images acquired at timestamps $t_1 < t_2 < \ldots < t_T$ with each image $\mathbf{X_i}$ characterized by M, N and C as the height, the width and the number of channels, respectively. Here, we consider multispectral bands acquired with a Sentinel-2 sensor for the C dimension. Each Sentinel-2 image time series is associated with a ground truth binary map $\mathbf{Y} \in \{healthy, damaged\}^{M \times N}$.

The architecture of ULISSE takes $\mathcal{D}$ as input and learns an output semantic

segmentation function $F\colon \mathbb{R}^{M\times N\times C\times T} \mapsto \{healthy, damaged\}^{M\times N}$, thanks to a U-Net-like architecture that consists of a multi-temporal encoder specifically designed to handle Sentinel-2 image time series, and a decoder. Each encoder is a pretrained ResNet-like model that is equipped with a PEFT mechanism. Figure 1 visually depicts the ULISSE framework, highlighting the multi-temporal encoder, the PEFT mechanism associated with each branch encoder, and the encoder-decoder architecture.

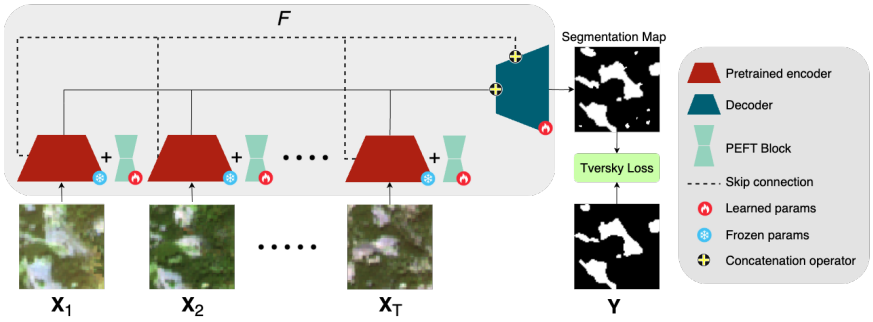


Figure 1: ULISSE schema to learn the semantic segmentation function F from a training set of labelled Sentinel-2 image time series $[\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_T}], \mathbf{Y}$

3.2. U-Net-like architecture for semantic segmentation of image time series

The architecture of ULISSE implements a U-Net-like network for the semantic segmentation of Sentinel-2 image time series of forest scenes. Generally, a U-Net architecture consists of: (1) a convolutional encoder that downsamples the original imagery feature space and reduces spatial resolution into an encoded feature embedding, (2) a convolutional decoder that upsamples the encoded feature embedding to the original input shape, and (3) several skip connections that directly link encoder blocks to decoder blocks in order to inject spatial details into the upsampling process. ULISSE extends the encoder-decoder structure of a conventional U-Net architecture by integrating a multi-temporal encoder equipped with T branches, to handle Sentinel-2 image time series.

Each encoder branch of ULISSE is a pretrained ResNet encoder. The ResNet is a Convolutional Neural Network, which has proven to achieve good accuracy

performance in several image classification tasks (Bohlol et al., 2025; Darwis et al., 2025). Pretrained versions of the ResNet encoder have already been used as backbones for several vision tasks, both in the general computer vision domain or in the remote sensing field (Yu & Geng, 2025; Huang et al., 2025). Specifically, a ResNet architecture is built by stacking multiple residual blocks to mitigate the performance degradation that may occur in a deep neural network by increasing the number of hidden layers. Each residual block includes residual connections, which are direct connections between non-consecutive layers in the deep neural network. These residual connections enable gradients to flow through the network by mitigating the vanishing gradients problem (Taghanaki et al., 2019).

In ULISSE, an aggregation strategy is used to map the output returned by each encoder branch into a single feature vector that is then used as input for the skip connections. This feature vector is obtained by applying the concatenation operator $\oplus$. This operator is commonly used in remote sensing applications to aggregate multi-source images (Zhou et al., 2023; Andresini et al., 2024a). In ULISSE, the concatenation step is performed before the skip connections of the U-Net architecture and produces a tensor by stacking the output vectors of all the parallel encoder branches. Let C_l denotes the channel size of the output returned at a layer l of each encoder branch. The application of the concatenation operator on layer l of all encoder branches returns an aggregated output with size $C_l \times T$. This output is subsequently reduced to size C_l through a 2D Convolution layer, ensuring that the outputs of the multi-branch encoder blocks are aligned with the number of channels in the corresponding decoder blocks so they can be used as input to the U-Net skip connection. The outputs of the concatenation step are used for both feeding the U-Net skip connections and creating the final encoder feature embedding outputted by the multi-branch encoder. This embedding is subsequently used as input to the decoder part of the encoder-decoder architecture.

The decoder part of ULISSE is implemented by progressively upsampling the spatial resolution of the final encoded feature embedding, to yield an output

with size $M \times N \times 1$. As the upsampling of the decoder is done symmetrically to the downsampling of the encoder, the decoder branch replicates the same block structure architecture as in the multi-branch encoder by following the opposite order. At the final layer of the decoder, a 1×1 Convolution is used to decrease the number of channels to 1 with the goal to produce the final pixel-level semantic segmentation map using the *Sigmoid* activation function.

3.3. PEFT of pretrained encoders

Each encoder branch of ULISSE is equipped with a small amount of additional parameters in order to implement a PEFT mechanism. In this work, we consider three alternative PEFT mechanisms, namely LoRA, DoRA and HRA, that are associated with the convolution layers of each encoder branch. In ULISSE, each temporal branch is equipped with independent PEFT parameters, with no parameter sharing across temporal branches.

Specifically, let us consider a convolution layer l so that $\mathbf{W}^l \in \mathbb{R}^{k_{M_l} \times k_{N_l} \times C_{l-1} \times C_l}$ denotes the weight matrix computed at l, where $k_{M_l} \times k_{N_l}$ is the kernel size, C_{l-1} and C_l are the number of input and output channels, respectively. The hidden layer representation h^l is defined as in the Equation 1:

$$h^l = \mathbf{W}^l * \mathbf{E}^{l-1} + \mathbf{b}^l, \tag{1}$$

where $*$ represents the convolutional operator, $\mathbf{E}^{l-1} \in \mathbb{R}^{M_{l-1} \times N_{l-1} \times C_{l-1}}$ is the three-dimensional feature embedding returned at layer $l-1$ and $\mathbf{b}^l$ is the bias vector. A standard convolutional layer returns a three-dimensional embedding $\mathbf{E}^l = \sigma_l(h^l)$ after the application of a non-linear activation function $\sigma_l \colon \mathbb{R}^{M_{l-1} \times N_{l-1} \times C_{l-1}} \mapsto \mathbb{R}^{M_l \times N_l \times C_l}$ so that $\mathbf{E}^l \in \mathbb{R}^{M_l \times N_l \times C_l}$. In the following, let $\mathbf{W}_0^l$ denote the pretrained weight matrix that is obtained at each layer l of the pretrained model placed in each encoder branch of the ULISSE architecture.

LoRA approximates the update of each pretrained weight matrix $\mathbf{W}_0^l$ through a low-rank decomposition involving two smaller matrices $\mathbf{A}^l$ and $\mathbf{B}^l$, with $\mathbf{A}^l \in \mathbb{R}^{k_{M_l} \times k_{N_l} \times C_{l-1} \times r}$, $\mathbf{B}^l \in \mathbb{R}^{1 \times 1 \times r \times C_l}$, and rank $r \leq min(C_l, C_{l-1})$. Hence, LoRA

reformulates Equation 1 as in Equation 2:

$$h^l = \mathbf{W}_0^l * \mathbf{E}^{l-1} + \frac{\alpha}{r}(\mathbf{B}^l * (\mathbf{A}^l * \mathbf{E}^{l-1})) + \mathbf{b}^l, \tag{2}$$

where α is a constant scaling factor. Notice that during the training of ULISSE, $\mathbf{W}_0^l$ is frozen, so it does not receive gradient updates, while $\mathbf{A}^l$ and $\mathbf{B}^l$ are trainable. The low-rank adaptation is implemented through sequential convolutions: first $\mathbf{A}^l$ reduces the channel dimension to rank r, subsequently $\mathbf{B}^l$ projects back to the output dimension C_l.

DoRA is a variant of LoRA, which also uses magnitude and direction components during the decomposition of each pretrained weight matrix $\mathbf{W}_0^l$ into $\mathbf{A}^l$ and $\mathbf{B}^l$. In particular, the magnitude is a learnable vector $\mathbf{m}^l \in \mathbb{R}^{C_l}$ that is initialized as $\mathbf{m}^l = \|\mathbf{W}_0^l\|_{C_l}$ with the column-wise norm computed over all input channels, while the direction is $\mathbf{W}_0^l$. Hence, DoRA reformulates Equation 1 as in Equation 3:

$$h^l = \left(\mathbf{m}^l \odot \frac{\mathbf{W}_0^l + \mathbf{B}^l * \mathbf{A}^l}{\|\mathbf{W}_0^l + \mathbf{B}^l * \mathbf{A}^l\|_{C_l}}\right) * \mathbf{E}^{l-1} + \mathbf{b}^l, \tag{3}$$

where $\odot$ represents element-wise multiplication broadcast across all dimensions. During the training of ULISSE, $\mathbf{W}_0^l$ is frozen, while $\mathbf{m}^l$, $\mathbf{A}^l$, and $\mathbf{B}^l$ are trainable.

Finally, HRA applies a sequence of Householder transformations to each pretrained weight matrix $\mathbf{W}_0^l$ along the output channel dimension C_l. Each Householder transformation z is defined by a learnable vector $\mathbf{v}_z^l \in \mathbb{R}^{C_l}$ (initialized randomly) that is normalized to obtain the vector $\mathbf{u}_z^l = \frac{\mathbf{v}_z^l}{\|\mathbf{v}_z^l\|_2}$. The Householder matrix is then obtained as $\mathbf{H}_z^l = \mathbf{I} - 2\mathbf{u}_z^l(\mathbf{u}_z^l)^\top$, where $\mathbf{I}$ is the Identity matrix. Hence, HRA reformulates Equation 1 as in Equation 4:

$$h^l = \left(\mathbf{W}_0^l \cdot \prod_{z=1}^{n} \mathbf{H}_z^l\right) * \mathbf{E}^{l-1} + \mathbf{b}^l, \tag{4}$$

where the product notation represents the sequential application of n Householder transformations on the pretrained weight matrix $\mathbf{W}_0^l$.

3.4. Training procedure

The training of the ULISSE architecture is conducted by using the selected PEFT mechanism to fine-tune each pretrained encoder branch, while training

the decoder from scratch, starting from a random initialization of the internal weights. ULISSE is trained following an end-to-end learning procedure in which each encoder branch is equipped with a PEFT module that learns its own parameters. The output of the encoder is feed-forward to a decoder module. During back-propagation, gradients computed from the decoder module are propagated from the decoder to each encoder branch to update each PEFT parameter, while the pretrained ResNet backbone encoders remain frozen.

As the downstream semantic segmentation task is highly imbalanced due to the fact that "damaged" class is extremely low-represented, the entire training process is conducted by minimizing the Tversky loss (Salehi et al., 2017). This is a loss function that is commonly used to handle a possible imbalance during model training without the need of applying any down-sampling of the majority class label "healthy" or over-sampling of the minority class label "damaged". The Tversky loss is defined as in Equation 5:

$$\mathcal{L}_{\text{Tversky}} = \frac{\mathsf{TDamaged}}{\mathsf{TDamaged} + \alpha\mathsf{FHealthy} + \beta\mathsf{FDamaged}}, \tag{5}$$

where TDamaged is the number of "damaged" pixels correctly predicted in the minority semantic label "damaged", FHealthy is the number of "damaged" pixels wrongly predicted in sthe majority semantic label "healthy", FDamaged is the number of "healthy" pixels wrongly predicted in the semantic label "damaged", while α and β are weight to control the trade-off between FHealthy and FDamaged. Both α and β are automatically optimized during the training stage as described in Section 4.2.

3.5. Outline of the whole pipeline

The complete pipeline, from data creation and preparation through model training to the production of semantic segmentation maps of testing scenes is shown in Figure 2.

In the training phase, Sentinel-2 image time series are processed with their associated ground truth masks, to learn a semantic segmentation model $F\colon \mathbb{R}^{M\times N\times C\times T} \mapsto \{healthy, damaged\}^{M\times N}$. Each timeserie includes T Sentinel-2 images, which

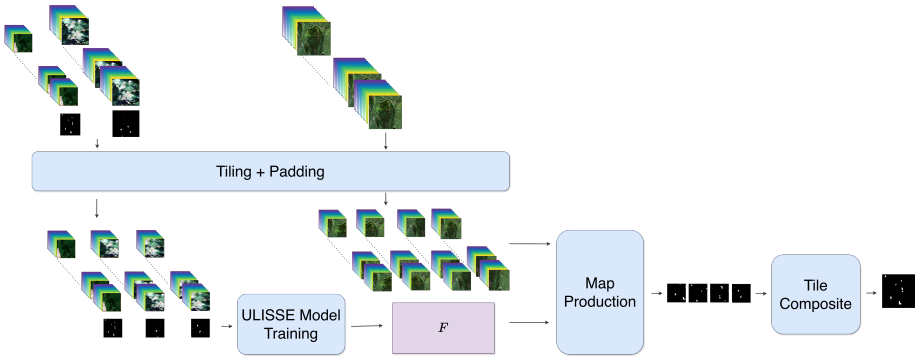


Figure 2: Whole ULISSE pipeline

are equally spanned in time. Each image covers C spectral channels, as described in Section 4.1.3. In addition, images in the same time series refer to the same scene, while different image time series refer to non-overlapping scenes with different spatial extents. So, to cope with the fixed spatial size requirements during the training phase of ULISSE, we use zero padding to obtain homogeneous, in terms of spatial extent, Sentinel-2 image time series tiles and associated ground truth masks of shape $M \times N$. The obtained collection of tiles and the associated ground truth masks is finally used as training data for the training phase of ULISSE to learn the mapping function F.

In the map production phase, ULISSE produces semantic segmentation maps for test scenes. As in the training phase, tiling and zero padding are applied to the Sentinel-2 image time series of test scenes to obtain time series tiles with shape $M \times N \times C \times T$. For each tile, ULISSE produces a semantic segmentation map of size $M \times N$. Finally, the areas originally padded are discarded, and all tile predictions are stitched together to generate the semantic segmentation map for the entire scene.

4. Evaluation study

In this section, we present both the case studies and the experimental scenario considered to evaluate the performance of ULISSE (Section 4.1). We describe the implementation details of ULISSE used in the evaluation (Section 4.2)

and we describe and discuss the results obtained by our framework and the competing approaches (Section 4.3). Finally, we conduct a post-hoc explainability analysis of the classification behaviour of our approach, ULISSE (Section 4.4).

4.1. Case study datasets and experimental scenario

We evaluated the performance of our proposed methodology in two case studies covering the mapping of the forest disease caused by bark beetle outbreaks in the Czech Republic and Romania, respectively. In both case studies, we considered scenes surrounding geo-referenced polygons of diseased forest areas, which were observed in September 2020, accessible through the DEFID2 database (Forzieri et al., 2023). Notably, the documentation of DEFID2 reports that the mortality severity of the recorded bark beetle outbreaks was high in the Czech Republic, with the percentage of contaminated trees between 80% and 100%, and medium in Romania, with a percentage of contaminated trees between 20% and 40%. This makes the semantic segmentation task addressed in this evaluation more challenging in Romania than in the Czech Republic.

4.1.1. DEFIDF2 forest scenes

Disease data for both case studies, the Czech Republic and Romania, were obtained from the DEFID2 dataset Forzieri et al. (2023). Forest scenes were generated from geo-referenced polygons of diseased forest areas by defining a buffer around them, which is then used to construct the training and testing scenes. After scene extraction, forest coverage is verified separately for each case study using the Forest Type 2018 product.[3]. This verification step ensures that the considered scenes are predominantly covered by forest: specifically, 93% and 86% of the total area for the Czech Republic and Romania datasets, respectively. The Czech Republic case study consists of a mosaic composed of 200 non-overlapping forest scenes, covering a total of 1,212,961 pixels at a spatial resolution of 10 meters. The size of scenes varies from 33×36 to 260×238 pixels,

[3] `https://doi.org/10.2909/77873ff3-4edf-48d4-94cd-c5b7b61da29e`

and the proportion of diseased forest surface per scene varies between 4.14% and 54.81%. The total percentage of diseased forest in the entire scene collection is 14.59%. The Romania case study comprises a mosaic of 124 non-overlapping forest scenes, covering 248,565 pixels at a spatial resolution of 10 meters. The size of scenes varies from 12×12 to 148×133 pixels, while the proportion of the diseased forest area per scene varies from 2.78% to 49.18% of the scene surface. The total percentage of diseased forest in the entire scene collection is 24.63%.

4.1.2. Sentinel-2 time window

In both datasets, the Sentinel-2 time series were obtained with a monthly temporal resolution from April 2020 to September 2020. The six-month time period was chosen because spruce bark beetle infestations are typically active in European countries from April to September, while bark beetles are dormant during autumn and winter. Similar to many other native insects, the European spruce bark beetle (Ips typographus) enters diapause during winter to adapt to the adverse environmental conditions (Hofmann et al., 2025). Diapause is a genetically programmed hibernation strategy that allows bark beetles to reduce their vital function activities, i.e., metabolism, mobility and reproduction. The inactivity of bark beetles during autumn and winter months has been recently examined by Dalponte et al. (2023), who has analysed pheromone trap data (number of insects caught) from a region in northern Italy between April 9, 2021, and October 13, 2021. Their analysis showed that the parental bark beetle generation became active in April 2021, while the first, second and third generations were active in May-June, July-August and August-September 2021, respectively. Forest dieback caused by the bark beetle outbreak was assessed in early autumn of the same year. Based on this pheromone trap analysis, Dalponte et al. (2023) subsequently examined the spectral separability of "damaged" forest areas from "healthy" areas during the period July-October 2021.

To set the observation period of this study, we also considered that in both case studies, the DEFIDF2 ground truth maps of the bark beetle outbreaks were assessed in September 2020, while the monitored areas were healthy in

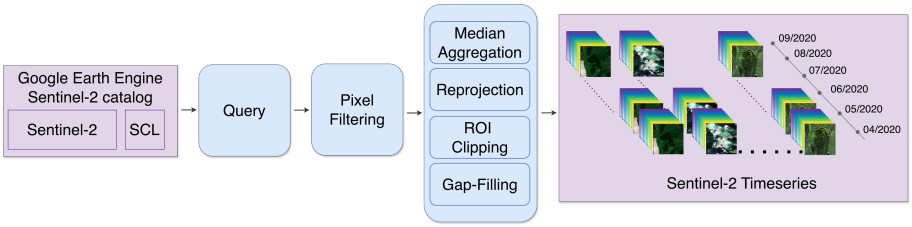


Figure 3: Workflow of Sentinel-2 dataset download and preparation

previous years. Additionally, Sentinel-2 images acquired between October and March may be highly affected by cloud cover and atmospheric phenomena, significantly limiting their utility.Hence, based on all these considerations, we chose to conduct our study considering a six-month time series spanning the period between April and September 2020, excluding Sentinel-2 images acquired before the beginning of the spring 2020 season.

4.1.3. Sentinel-2 time series dataset preparation

Figure 3 depicts the workflow to download and preprocess the Sentinel-2 image time series for each of the case studies. In the *Query* step, we retrieve the Sentinel-2 images for the area covered by each DEFIDF2 scene in the period between April 1st and September 20th, 2020. These images were downloaded from the *Sentinel-2 catalog* of ESA Copernicus open access hub, using the Google Earth Engine APIs[4]. This image catalog contains the multispectral radiometric data in the visible, near infrared, and short wave infrared parts of the spectrum for a total of thirteen spectral bands. These spectral data were acquired at their original spatial resolution (between 10 m and 60 m), through a constellation of two Copernicus satellites (Sentinel-2A launched in 2015 and Sentinel-2B launched in 2017) that, since their launch, covered the majority of the Earth's surface with an approximately revisit cycle of 5 days. In addition, all Copernicus Sentinel-2 images are atmospherically corrected by the Sentinel-2

[4]`https://developers.google.com/earth-engine`

Table 2: Selected Sentinel-2 bands with their spatial resolution and central wavelength

Band	**Band Name**	**Spatial Resolution**	**Central wavelength**
B2	Blue	10 m	490 nm
B3	Green	10 m	560 nm
B4	Red	10 m	665 nm
B5	Red-edge 1	20 m	705 nm
B6	Red-edge 2	20 m	740 nm
B7	Red-edge 3	20 m	783 nm
B8	Near Infrared	10 m	842 nm
B8A	Narrow Near Infrared	20 m	865 nm
B11	SWIR 1	20 m	1610 nm
B12	SWIR 2	20 m	2190 nm

Level 2A product generation and formatting tool (sen2cor v2.11) before being stored in the queried catalog.

In this study, we considered ten spectral bands and the Scene Classification Layer (SCL) band. Specifically, for each image, band B10 (SWIR – Cirrus) was removed since this band is reserved to atmospheric correction. In addition, following the recommendations of (Clasen et al., 2025), band B1 – Coastal Aerosol – and band B9 – Water Vapor – were also discarded, since at 60 m spatial resolution, and mainly used for cloud screening, atmospheric correction, and cirrus detection. The final list of considered Sentinel-2 bands is reported in Table 2.

The SCL band is produced via the Scene Classification Algorithm (Louis et al., 2016), which uses the reflectance properties of multi-spectral bands to identify three cloud classes (including cirrus) and six additional classes: shadows, cloud shadows, vegetation, non-vegetated areas, water, and snow. In this study, the SCL information was used in the *Pixel filtering* step, to filter out all pixels that were classified as noise, defective, dark, cloud, cloud shadow or thin cirrus.

In the *Median Aggregation* step, we aggregated the Sentinel-2 data of each scene at the monthly temporal resolution. This operation produced a single composite image for each month, resulting in a time series of six monthly Sentinel-2

images for each scene. The temporal aggregation was performed pixel-wise for each spectral band by computing the median of all valid band values, as determined by the SCL-based pixel classification, acquired in the considered month for a pixel.

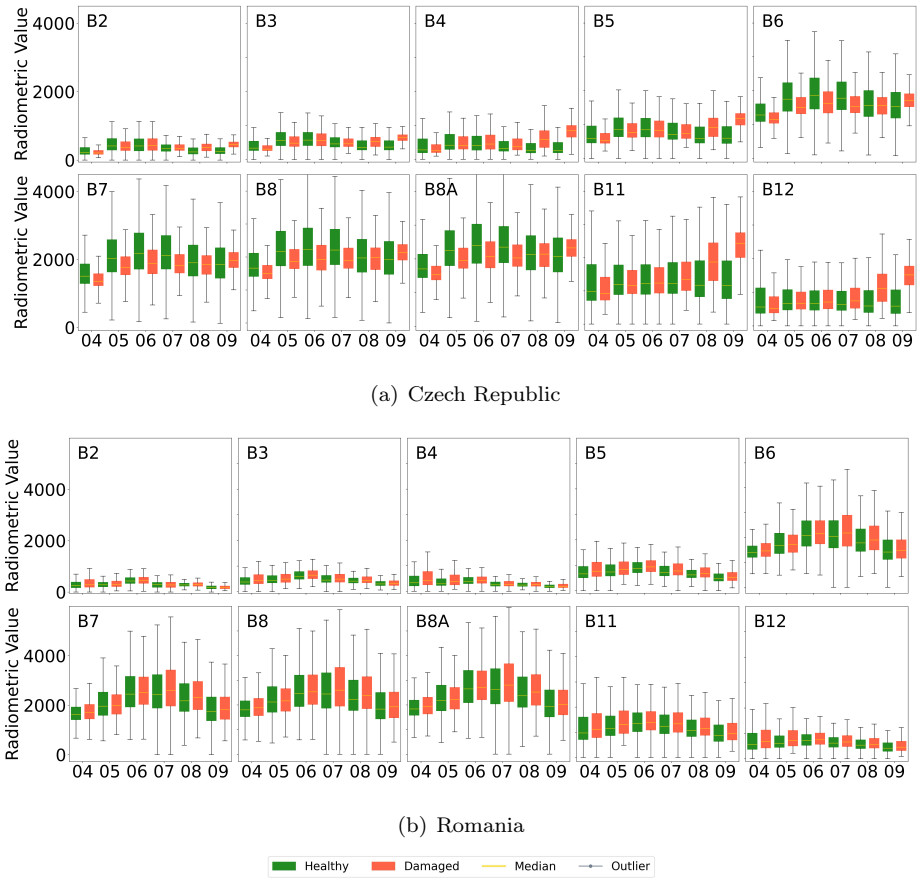
In the *Reprojection* step, we resampled each raster image at 10 m resolution and reprojected the image from the original Universal Transversal Mercator (that is in degrees), to the spatial reference system WGS84:EPSG3857 Web Mercator (that is in meters). The EPSG3857 reference system was selected as it is commonly used in Europe, to allow all scenes (both in the Czech Republic and Romania) to be in the same reference system. For the spatial resampling, we resampled the considered 20 m resolution bands of the Sentinel-2 spectrum (i.e., B5, B6, B7, B8, B11 and B12) to 10 m using the Nearest Neighbour interpolation. This step ensured that all considered bands had the same spatial resolution. The Nearest Neighbour method is the default choice with Google Earth Engine APIs. According to this method, the spectral values that are assigned to pixels in 10 m resolution are determined on the basis of the pixel values that surround its transformed position in the original 20 m resolution. As underlined by Lillesand et al. (2015), this up-sampling approach offers the advantage of computational simplicity while not altering the original spectral pixel values.

In the *ROI Clipping* step, we considered the region of interest (ROI) of each scene as it was retrieved from the scene boundary shapefile. The ROI of a scene was used to select the scene spectral data associated to the reference spatial extent.

Finally, in the *Gap-Filling* step, we performed missing value imputation, as this is a standard practice in the remote sensing (Inglada et al., 2017), to fill spectral missing information associated with invalid pixels. Any missing value in a spectral band was linearly interpolated using only the valid data (cloud-free, non-shadow, non-saturated pixels) from the multi-temporal information associated with the same band and pixel. The final Sentinel-2 image time series dataset, obtained through this workflow, contains a time series of six monthly

multispectral images at 10 m spatial resolution for each DEFIDF2 scene.

4.1.4. Sentinel-2 time series dataset inspection



(a) Czech Republic

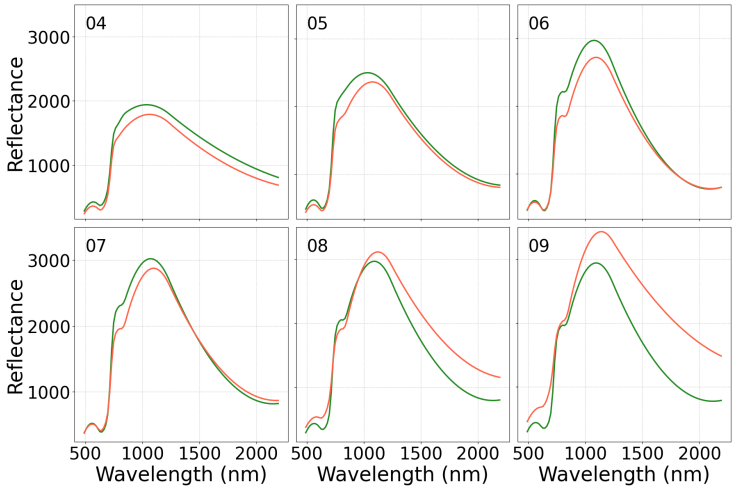
(b) Romania

Figure 4: Box plot distribution of the radiometric values of Sentinel-2 bands acquired monthly in April 2020 (04), May 2020 (05), June 2020 (06), July 2020 (07), August 2020 (08) and September 2020 (09) in Czech Republic (Fig. 4(a)) and Romania (Fig. 4(b)).

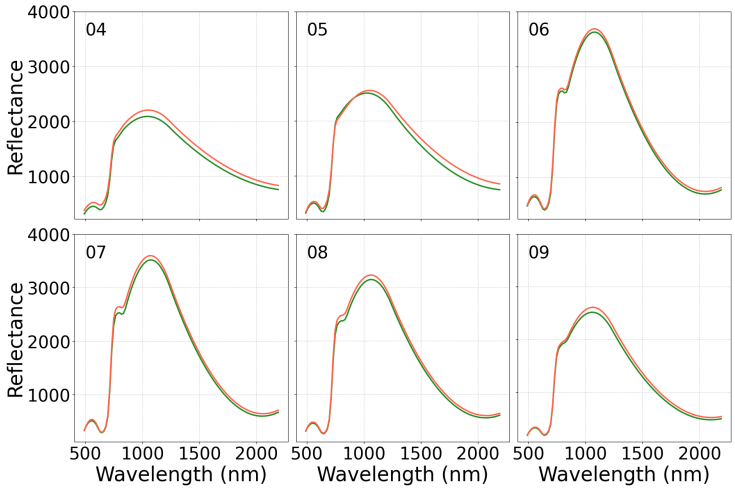
To analyse the spectral characteristics of forest disturbances caused by bark beetles, we examine the distribution of the radiometric values acquired for the Sentinel-2 bands in forest patches labelled as "healthy" and forest patches labelled as "damaged", for both case studies. With this regard, Figure 4 shows the box plots of the radiometric values of Sentinel-2 bands plotted per month, in both the Czech Republic and Romania, grouped with respect to semantic labels

"healthy" and "damaged". A difference in the box plots of the Sentinel-2 data acquired per band in the various months can be noted. In addition, box plots of the Czech Republic exhibit a greater divergence, on the B2, B3, B4, B5, B11 and B12 bands, between the two opposite semantic labels in the later months (August and September) than in the earlier months (April, May, June and July) of the considered time series. As the severity of the disease increases over time, this analysis suggests that these bands have the potential to contribute to the delineation of the diseased area when the stage of disease is more advanced. The box plots show that the divergence between the two opposite semantic labels is negligible in Romania, regardless of both bands and months. This, probably, depends on the fact that the severity achieved by the disease in Romania in September 2020 is moderate, so it is more complex to capture via remote sensing imagery, while the severity achieved by the disease in the Czech Republic in September 2020 is high, thus potentially easier to detect with Sentinel-2 imagery.

Similar conclusions can be drawn from the analysis of the spectral reflectance profiles associated with the two semantic labels. These profiles are shown in Figure 5, which depicts the monthly radiometric values of the Sentinel-2 bands for both case studies. The spectral reflectance profiles of the two semantic labels are well separated in the Czech Republic, while in Romania, they almost overlap. More specifically, in the Czech Republic, the spectral reflectance profile corresponding to the disturbance status is distinctly separated from the spectral reflectance profile of the healthy status from April to September 2020, particularly in the Red-edge wavelength region (B5-B7). Furthermore, in the same case study, the separation between the two spectral reflectance profiles increases markedly in the shortwave infrared (SWIR) region (B11-B12) during the late summer months (August and September). This behaviour is consistent with the analysis of forest disturbances' spectral characteristics discussed by Abdullah et al. (2019), who described how multiple forest stressors (including bark beetle infestation) alter the biophysical and biochemical properties of trees and, consequently, their spectral response. In particular, their study showed that



(a) Czech Republic



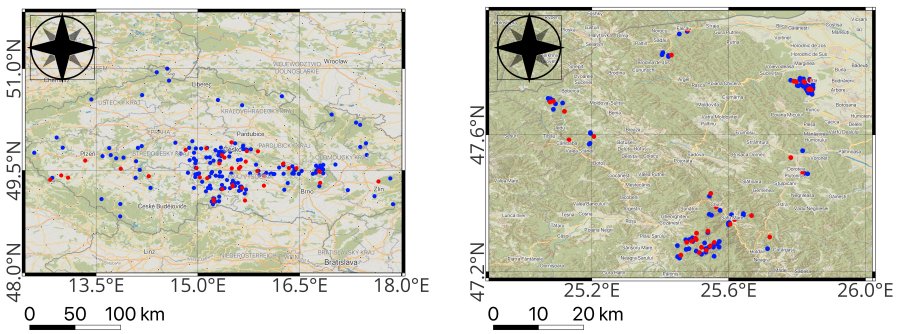
(b) Romania

Figure 5: Spectral reflectance profiles for the semantic labels: "healthy" and "damaged". The profiles are obtained averaging the spectral values of Sentinel-2 bands acquired monthly in April 2020 (04), May 2020 (05), June 2020 (06), July 2020 (07), August 2020 (08) and September 2020 (09), in Czech Republic (Fig. 5(a)) and Romania (Fig. 5(b)).

the Red-edge bands are more sensitive than other Sentinel-2 bands to diseased and insect attacks. In addition, the same study highlighted that near-infrared (NIR) and SWIR bands provide valuable information for assessing tree water content and nitrogen concentration, which can be used to detect vitality loss and cell structure degradation typically observed with reduced chlorophyll content and leaf water under diseased conditions. Notably, the importance of NIR and SWIR bands for monitoring foliar properties has recently been discussed also by Kluczek & Zagajewski (2025). Instead, Carletti et al. (2025) have recently assessed the importance of Red-edge regions for monitoring conifer forest dieback. In particular, they have examined the sensitivity of the Red-edge information to vegetative health and structure, photosynthetic activity, and, particularly, chlorophyll content, which is the main factor influencing the reflectance of vegetation in the visible and Red-edge regions. Although less pronounced, a slight separation between the spectral reflectance profiles of the two labels can also be observed in the Red-edge and SWIR regions of the Romania case study. Finally, the relevance of spectral information contained in Red-edge and SWIR regions for machine learning based detection of bark beetle outbreaks has been recently confirmed by Andresini et al. (2023b). Similarly, Kluczek & Zagajewski (2025) have recently shown that NIR and Red-edge regions of the spectrum are among the most important predictor spectral bands to correctly map bark beetle outbreaks with machine learning methods. In particular, their study resorts to Shapley values to show how Red-edge and SWIR bands have a premier role in enabling the classification model to recognize forest dieback caused by bark beetles.

4.1.5. Experimental scenario

The semantic segmentation models, considered in this study, were developed and evaluated by considering a random split of the scene mosaic of each dataset in a training scene set, used for models' development, and a testing scene set, used for models' evaluation. Specifically, in the Czech Republic, we used 160 random scenes (covering 1,014,708 pixels) as a training scene set, and the re-



(a) Czech Republic (b) Romania

Figure 6: Location of the 200 study scenes in the Czech Republic (Fig. 6(a)), and 124 study scenes in Romania (Fig. 6(b)). The blue circles denote the centroids of training set scenes, while red circles denote the centroids of testing set scenes.

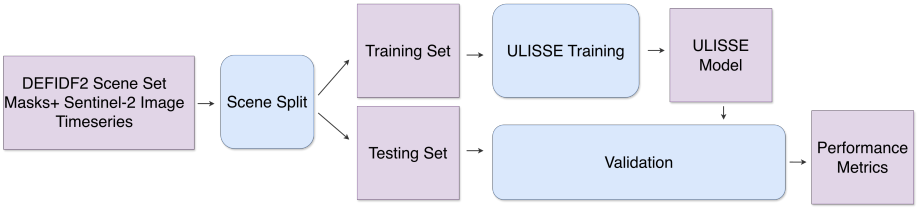


Figure 7: Experimental structure diagram

maining 40 scenes (covering 198,253 pixels) as a testing scene set. Instead, in Romania, we used 88 random scenes (covering 200,702 pixels) as a training scene set, and the left-out 36 scenes (covering 47,863 pixels) as a testing scene set. Figure 6 shows the maps of the scene locations, with their centroids in both the Czech Republic and Romania study sites, and their partitioning into training and testing scene sets. Figure 7 shows the diagram of the evaluation protocol adopted to assess the accuracy performance of ULISSE.

4.1.6. Performance metrics

We considered the Precision (P), Recall (R), F1 Score (F1) computed for the two semantic labels (i.e., "damaged" and "healthy"), as well as the Inter-

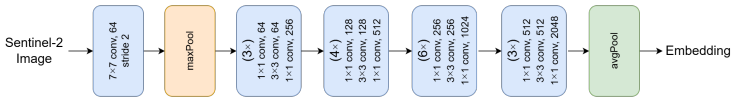
section over Union (IoU), to assess the performance of the forest disease maps of testing scenes produced with the considered semantic segmentation models for both case studies. These are standard metrics commonly used in semantic segmentation tasks and formulated as described in the following. Given a semantic label identified as Positive, and the opposite label identified as Negative, we denote TP – the number of pixels labelled with semantic label Positive and correctly predicted in semantic label Positive, TN – the number of pixels labelled with semantic label Negative and correctly predicted in semantic label Negative, FP – the number of pixels labelled with semantic label Negative and wrongly predicted in semantic label Positive, and FN – the number of pixels labelled with semantic label Positive and wrongly predicted in the semantic label Negative. The Precision is measured as $P = \frac{TP}{TP+FP}$, the Recall is measured as $R = \frac{TP}{TP+FN}$, and the F1 score is measured as the harmonic mean of Precision and Recall, that is, $F1 = \frac{2 \times P \times R}{P+R}$. Hence, in the following, P(D), R(D) and F1(D) denote Precision, Recall and F1 score computed considering the semantic label "damaged" as Positive and the semantic label "healthy" as Negative. On the other hand, P(H), R(H) and F1(H) denote Precision, Recall and F1 score computed considering the semantic label "healthy" as Positive and the semantic label "damaged" as Negative. Finally, the Intersection over Union score is the ratio between the intersected area and the combined area of prediction and ground "damaged" truth, that is, $\mathsf{IoU} = \frac{\mathsf{TP}}{\mathsf{TP+FP+FN}}$. For all these metrics, a higher score corresponds to a more accurate model.

4.2. Implementation details

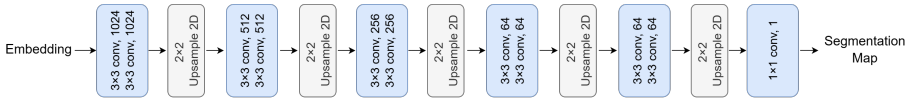
ULISSE was implemented in Python 3 using the Pytorch library.[5] Each encoder branch is a ResNet50 (i.e., a ResNet with 50 layers) with weights pretrained on BigEarthNet dataset and available online[6]. A description of the layer structure of the ResNet50 backbone is shown in Figure 8(a). Each residual block

[5]Data and code available at `https://github.com/s4rgax/ULISSE`

[6]`https://huggingface.co/BIFOLD-BigEarthNetv2-0/ResNet50-s2-v0.2.0`



(a) Encoder module (Backbone Resnet50)



(b) Decoder module

Figure 8: Layer structure of backbone encoder and decoder modules

of the pretrained ResNet50 consists of a stack of three Convolutional layers with 1×1, 3×3, and 1×1 kernel sizes, repeated multiple times. After the input layer, ResNet50 implements a Convolutional layer with a 7×7 kernel size, 64 filters, followed by a max pooling layer with a 3×3 kernel size. Then, it implements four bottleneck residual blocks with an increasing number of filters. More in detail, the first block, that is repeated 3 times, increases the filter number from 64 to 256, the second block, that is repeated 4 times, increases the filter number from 256 to 512, the third block, that is repeated 6 times, increases the filter number from 512 to 1024, and the fourth block, that is repeated 3 times, increases the filter number from 1024 to 2048. In addition, each block implements a residual connection. Finally, the ResNet50 model uses an Average Pooling layer to reduce the feature dimension before the final classification layer. The last layer of ResNet50, which was originally pretrained for image classification, is removed in each encoder branch of ULISSE. The Rectified Linear unit ($ReLu$) activation function is used for each hidden layer.

The layer structure of the decoder of ULISSE is shown in Figure 8(b). It is composed of four bottleneck upsampling blocks. The first upsampling block, that is repeated 3 times, decreases the filter from 2048 to 1024, the second upsampling block, that is repeated 6 times, decreases the filter from 1024 to 512, the third upsampling block, that is repeated 4 times, decreases the filter from 512 to 256, and the fourth upsampling block, that is repeated 3 times,

Table 3: Hyper-parameter search space

Hyper-parameter	Values
Mini batch size	{ 2^2, 2^3, 2^4, 2^5, 2^6}
r (LoRA, DoRA)	{ 2^2, 2^3, 2^4, 2^5, 2^6}
n (HRA)	{ 2^2, 2^3, 2^4, 2^5, 2^6}
Learning rate	$[10^{-6}, 10^{-3}]$
α	[0.1, 0.5]
β	1 - α

decreases the filter from 256 to 64. At the final layer, a 1×1 Convolution is used to decrease the number of filters from 64 to 1, and produce the semantic segmentation map, at pixel-level, using the *Sigmoid* activation function. The *ReLu* activation function is used for each hidden Convolutional layer of the decoder part. To implement the PEFT approaches we adopt the HuggingFace PEFT library[7].

ULISSE was trained with image tiles with size $224 \times 224 \times 10$ (M=224 and N=224) obtained from original scenes with a combination of tiling and zero padding operations as discussed above. The training was performed with a maximum number of epochs set to 200. Early stopping was used to retain the best semantic segmentation model. The Adam optimizer was adopted to update the model's weights. To select the hyperparameters of the final ULISSE architecture, we used the Hyperopt library by adopting 20% of the training set as validation set to automatically select the configuration that achieved the highest F1 measured on the validation set. The hyper-parameter search space explored in the evaluation conducted in this study is reported in Table 3.

[7] `https://github.com/huggingface/peft/`

4.3. Performance analysis

The evaluation assessment of our proposed methodology is organized as follows:

- We examine the effect of PEFT methods on the accuracy performance of the semantic segmentation maps produced in the model evaluation phase (Section 4.3.1).
- We perform an exploratory analysis of the relationship between the length of the time series processed and the accuracy performance of ULISSE (Section 4.3.2).
- We examine the sensitivity of the accuracy performance of ULISSE to the exclusion of an image from the processed imagery time series (Section 4.3.3).
- We explore the effect of the temporal aggregation strategy, adopted to combine the temporal information, on the accuracy performance of the model (Section 4.3.4).
- We examine the computational cost of ULISSE in both the train and test phases (Section 4.3.5).
- We conduct a comparative study, involving ULISSE and related semantic segmentation methodologies from the recent literature (Section 4.3.6).

4.3.1. PEFT analysis

To evaluate the effect of PEFT mechanisms on our method, we considered the following variants of ULISSE:

- RI: This is the Random-Init variant of ULISSE that gives up any PEFT mechanism, to train the model from scratch, starting from an initial random weights initialization.

Table 4: PEFT analysis – P(D), R(D), F1(D), P(H), R(H), F1(H) and IoU. The best results are in bold, while the runner-up results are underlined.

Dataset	**Metric**	**RI**	**FT**	ULISSE		
				LoRA	DoRA	HRA
Czech Republic	P(D)	76.09	74.35	<u>77.89</u>	**79.18**	77.29
	R(D)	71.57	**74.63**	<u>73.12</u>	71.58	72.42
	F1(D)	73.76	74.49	**75.43**	<u>75.19</u>	74.78
	P(H)	93.81	**94.40**	<u>94.15</u>	93.87	94.01
	R(H)	95.04	94.32	<u>95.43</u>	**95.85**	95.31
	F1(H)	94.42	94.36	<u>94.79</u>	**94.85**	94.65
	IoU	58.43	59.35	**60.55**	<u>60.24</u>	59.72
Romania	P(D)	48.56	50.68	52.42	**56.04**	<u>55.26</u>
	R(D)	**92.09**	<u>90.08</u>	87.55	78.84	76.28
	F1(D)	63.59	64.87	**65.57**	<u>65.52</u>	64.09
	P(H)	96.30	**95.61**	94.74	<u>91.95</u>	91.07
	R(H)	67.87	71.12	73.83	<u>79.63</u>	**79.66**
	F1(H)	79.63	81.57	82.98	**85.34**	<u>84.98</u>
	IoU	46.61	48.00	**48.78**	<u>48.72</u>	47.16

- FT: This is a full Fine-Tuning variant of ULISSE that gives up any PEFT method and uses the traditional full fine-tuning strategy to update all model's parameters, i.e., both encoder and decoder weights.
- LoRA, DoRA and HRA: These are three variants of ULISSE, which use LORA (Yu et al., 2023), DORA (Liu et al., 2024a) and HRA(Yuan et al., 2024b) as PEFT mechanisms, respectively.

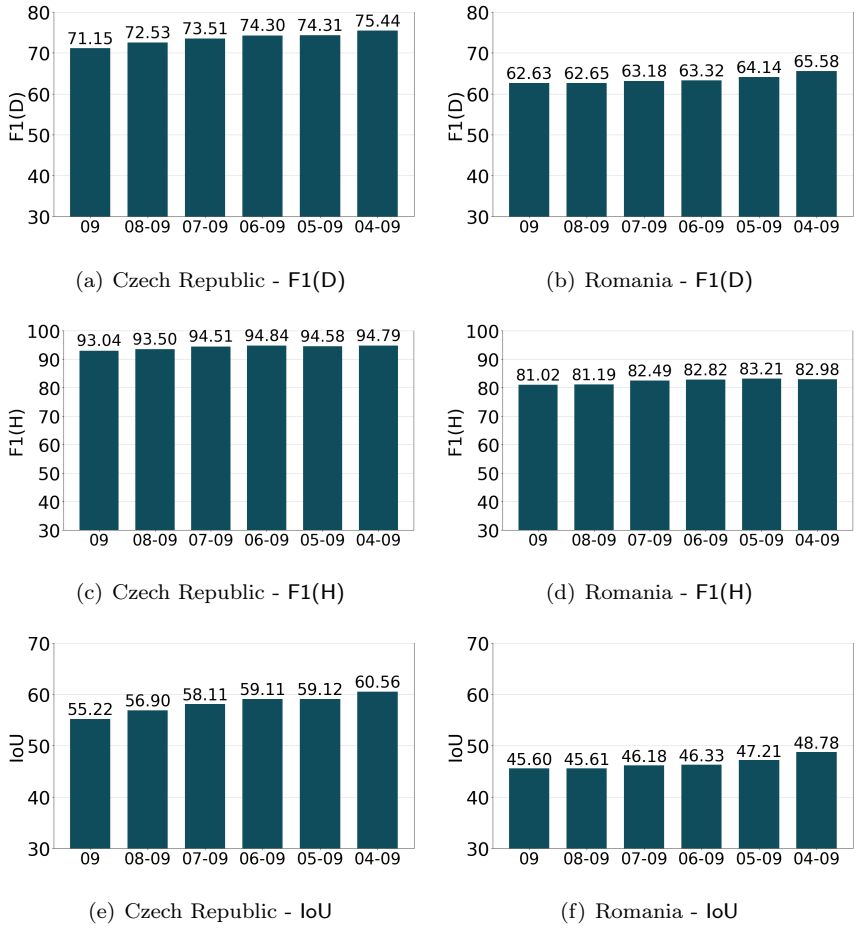
Table 4 reports the accuracy metrics – P(D), R(D), F1(D), P(H), R(H), F1(H) and IoU – which were measured in the evaluation phase conducted in the Czech Republic and Romania case studies. We note that all PEFT variants of ULISSE – LoRA, DoRA and HRA – outperform RI and FT variants in terms of F1(D),

F1(H) and IoU. More precisely, the use of a PEFT method commonly increases the Precision measured for label "damaged" (P(D)), while decreases the Recall measured for the same semantic label (R(D)). Accordingly, the use of a PEFT method commonly increases the Recall measured for label "healthy" (R(H)), but this improvement happens at the cost of a decrease of Precision measured for the same semantic label (P(H)). We recall that the considered case studies regard a semantic segmentation task where the cost of an alert regarding a false "damaged" area is higher for forest managers than the cost of a missing alert regarding a true "damaged" area. In fact, this type of mapping error would lead to postponing the sanitary cut intervention that forest managers commonly use to contain the spread of the disease. In contrast, alerts regarding false "damaged" areas are also problematic, as they may lead to unnecessary sanitary cut treatments with useless extra costs for the forest management. Based on this premise, a good trade-off between Precision and Recall for both semantic labels, and particularly in forest patches labelled as "damaged", is desirable. So, focusing the attention on the analysis of F1 scores, we noted that LoRA achieves the highest F1(D), as well IoU of the comparative study, while DoRA achieves the highest F1(H) of the comparative study in both case studies.

According to the analysis of results reported above, we select the LoRA variant, simply denoted as ULISSE in the following, to be considered in subsequent analyses. This variant achieves the best overall accuracy while optimizing predictions for label "damaged" in both case studies, while it still maintains reasonable performance for the prediction of label "healthy". This result is aligned with the conclusions drawn in the empirical study of Marti Escofet et al. (2026), who identify LoRA as the most effective PEFT method for fine-tuning a semantic segmentation model of water surface processing Sentinel-2 images of water scenes.

4.3.2. Sentinel-2 time series analysis

To assess the influence of the Sentinel-2 time series length on the accuracy performance of ULISSE, we considered six data scenarios, each defined by a



(a) Czech Republic - F1(D)

(b) Romania - F1(D)

(c) Czech Republic - F1(H)

(d) Romania - F1(H)

(e) Czech Republic - IoU

(f) Romania - IoU

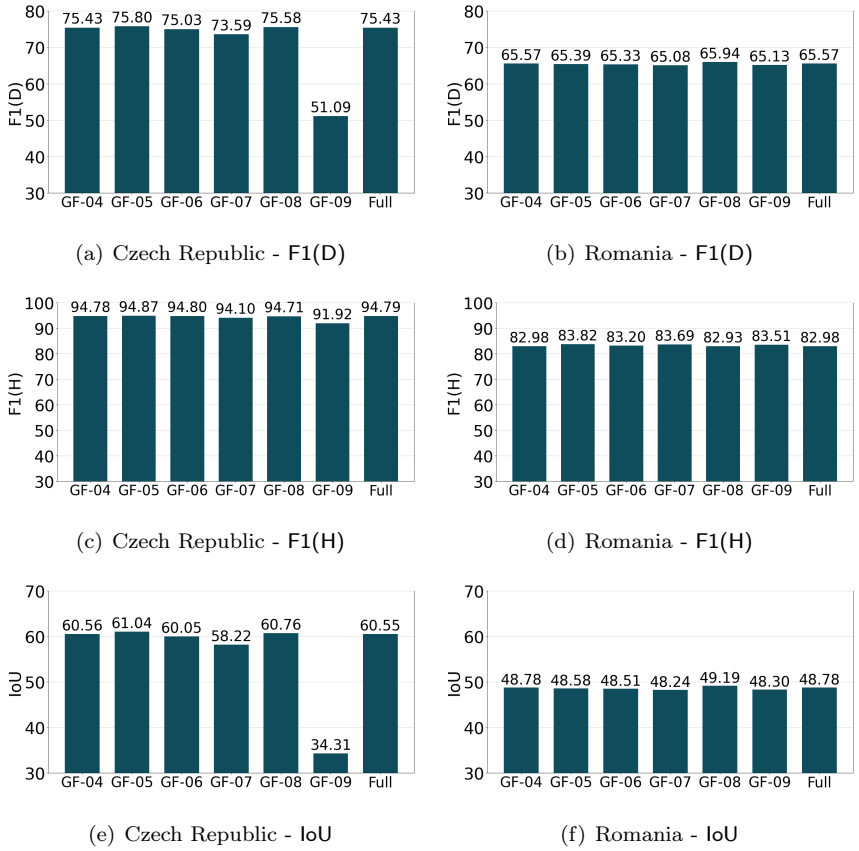
Figure 9: Sentinel-2 time series analysis. F1(D), F1(H) and IoU of ULISSE run with LoRA in the data configurations: 09 (September), 08-09 (August-September), 07-09 (July-September), 06-09 (June-September), 05-09 (May-September), and 04-09 (April-September) of Czech Republic (Figs. 9(a), 9(c) and 9(e)) and Romania (Figs. 9(b), 9(d) and 9(f))

starting and ending month for the analyzed period. For example, 04-09 represents the complete time series from April (04) through September (09), 07-09 covers the period from July (07) through September (09), and 09 uses only the September Sentinel-2 image, excluding any temporal information from the semantic segmentation process. Figure 9 shows the F1(D), F1(H) and IoU values measured on the semantic segmentation maps produced for the evaluation phase

of ULISSE in the considered data configurations of both the Czech Republic and Romania. These results show that temporal Sentinel-2 information contributes to improving the accuracy performance of the semantic segmentation maps, regardless of the considered case study and semantic label. Notably, the longer the time period used, the greater the gain in accuracy achieved. However, the improvement of the F1 score obtained by considering the temporal information is greater for the label "damaged" than for the label "healthy".

4.3.3. Monthly sensitivity analysis

In this Section, we investigate how the exclusion of images from the Sentinel-2 time series (e.g., due to persistent cloud cover throughout an entire month) affects the performance of ULISSE. To this end, we systematically exclude each of the six images comprising the time series data at the testing phase and employ linear interpolation-based gap-filling to impute each missing image. For example, GF-04 denotes the scenario in which the April image is excluded from all time series in the test set and interpolated via the gap-filling procedure. Full denotes the original test set without any gap-filled images. Figure 10 shows the F1(D), F1(H) and IoU values for ULISSE across both the Czech Republic and Romania study sites. The results demonstrated that, in general, the performance of ULISSE remains stable regardless of which image is excluded from the original time series. This suggests that the gap-filling procedure provides appropriate imputations for missing values without compromising the accuracy of ULISSE. In some cases, ULISSE even gains accuracy when excluding Sentinel-2 images from specific months. For instance, in the Czech Republic, the performance of ULISSE slightly improves when excluding the original Sentinel-2 images acquired in May. However, we also note that the accuracy performance decreases drastically in the Czech Republic when the gap-filling was performed on the September image, contrary to what is observed in the Romania case study. We hypothesize that the gap-filling is less effective for September images in the Czech Republic scenes because the mortality severity of recorded bark beetle outbreaks increased markedly during this month in this region. In contrast, for



(a) Czech Republic - F1(D)

(b) Romania - F1(D)

(c) Czech Republic - F1(H)

(d) Romania - F1(H)

(e) Czech Republic - IoU

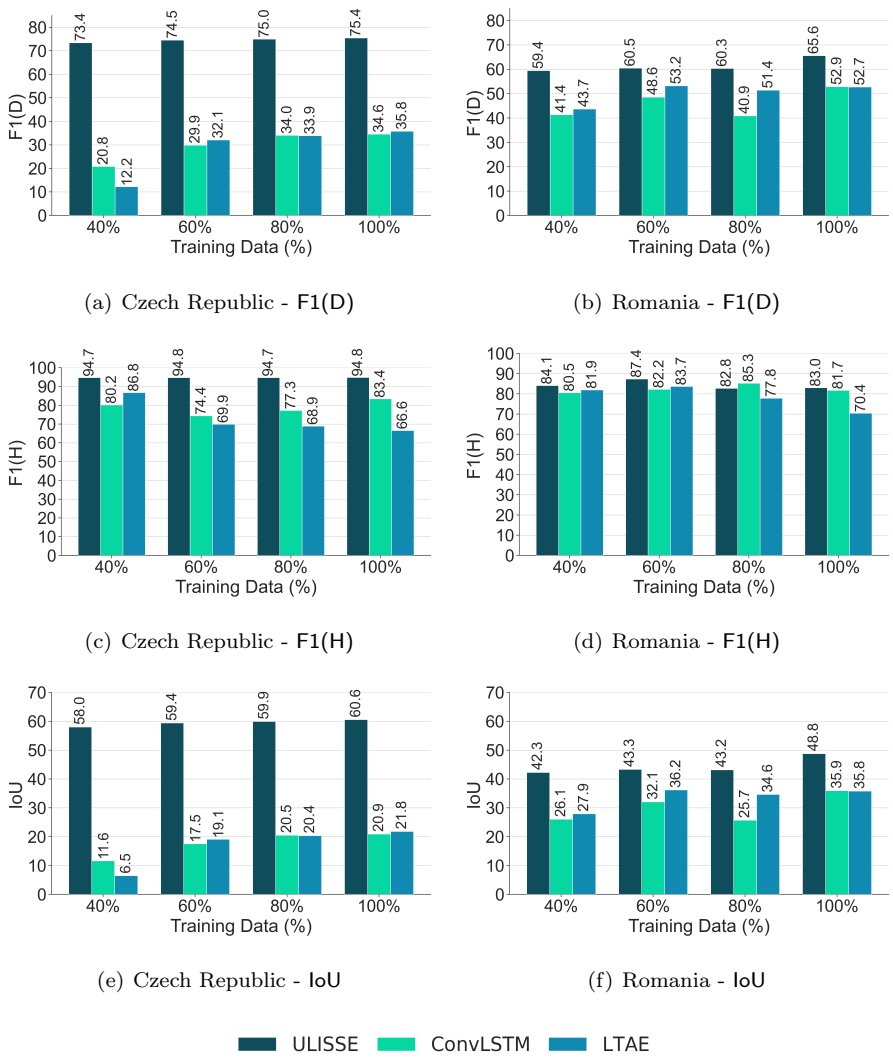
(f) Romania - IoU

Figure 10: Impact on ULISSE's performances of removing one month at a time from the temporal imagery sequence with missing values interpolation. F1(D), F1(H) and IoU values for gap-filling configurations GF-04 (April excluded), GF-05 (May excluded), GF-06 (June excluded), GF-07 (July excluded), GF-08 (August excluded), GF-09 (September excluded), and Full (original time series without excluded months) in Czech Republic (Figs. 10(a), 10(c) and 10(e)) and Romania (Figs. 10(b), 10(d) and 10(f)).

the Romania case study, the mortality severity observed in September remained consistent with the previous months.

4.3.4. Temporal aggregation strategy

To evaluate the temporal aggregation strategy employed by ULISSE, specifically, a multi-branch U-Net architecture with ResNet as the per-branch en-



(a) Czech Republic - F1(D)

(b) Romania - F1(D)

(c) Czech Republic - F1(H)

(d) Romania - F1(H)

(e) Czech Republic - IoU

(f) Romania - IoU

Figure 11: Comparison of ULISSE with alternative temporal aggregation strategy variants: ULISSE, ConvLSTM, and LTAE on Czech Republic and Romania datasets varying in percentage (40%, 60%, 80% and 100%) used as training data.

coder, followed by a concatenation aggregation mechanism, we compare our approach against two standard alternatives: i) a ConvLSTM-based temporal aggregation (Rußwurm & Körner, 2018), where embeddings from the different multi-branch encoders are aggregated via an LSTM recurrent neural network,

and ii) a transformer-based temporal aggregation, namely LTAE (Garnot & Landrieu, 2020) (lightweight temporal attention encoder), where embeddings from the different multi-branch encoders are aggregated via an attention mechanism. Specifically, we replace the original temporal aggregation strategy of ULISSE with the ConvLSTM-based approach (denoting this variant as ConvLSTM) and with the lightweight, transformer-based, temporal attention (denoting this variant as LTAE). Additionally, to further characterize the behaviour of the different temporal aggregation strategies, we vary the amount of training data considered, in terms of percentage of the original training dataset, starting from 40% up to 100% (the full training data) with a step of 20%. The results of this analysis are presented in Figure 11. We observe that ULISSE, equipped with its multi-branch U-Net architecture and concatenation aggregation mechanism, clearly outperforms both alternative temporal aggregation strategies (ConvLSTM and LTAE) across all performance metrics on both case studies. The performance gaps are generally more pronounced for the "damaged" label rather than the "healthy" label. Interestingly, when varying the amount of training data, ULISSE exhibits a clear, stable behaviour compared to the competing approaches. These results further underscore the quality and effectiveness of the proposed temporal aggregation strategy relative to standard strategies (e.g., recurrent neural and temporal transformers) prevalent in the remote sensing literature for multi-temporal data analysis. The improved performance of the original ULISSE temporal aggregation strategy is likely due to the skip connections inherent in the U-Net architectural design. These connections bridge encoder and decoder blocks directly, thereby injecting spatial details into the upsampling process. These spatial details are generally useful and particularly helpful for enhancing the detection and characterization of the low-represented "damaged" label, as they mitigate the model's bias towards the majority "healthy" label.

4.3.5. Computational cost analysis

In this section, we compare the full fine-tuning approach and ULISSE in terms of computational cost metrics during both the training and testing phases

on the Czech Republic and Romania case studies. For the training phase, we report the following metrics: giga floating-point operations (GFLOPs), peak GPU memory (in GB), peak power consumption (in Watts), and the number of trainable parameters (in millions) in the encoder. Note that the decoder is identical across all methods and is therefore excluded from this analysis. For the test stage, we consider peak GPU memory (in GB), peak power consumption (in Watts), and throughput measured as the number of 224× 224 images processed per second. For this analysis, we consider a worst-case scenario with a batch size of 64 for both training and testing phases. For ULISSE, we use the highest LoRA rank value from our experimental evaluation (rank = 64). The results are presented in Table 5. These results show that, at the training phase, ULISSE achieves a higher number of GFLOPs while limiting the number of trainable encoder parameters. Meanwhile, peak GPU memory and peak power consumption remain comparable between the two approaches. At the testing phase, both approaches achieve similar throughput, with ULISSE processing slightly more images per second. However, ULISSE requires considerably less GPU memory and consumes less energy than the full fine-tuning approach. This point is particularly important for deployment, as it makes ULISSE well-suited to run on less demanding hardware, such as devices equipped with smaller GPU cards (less than 8 GB in our case) and onboard systems where constraints on low energy consumption and limited hardware resources are critical.

4.3.6. Competing methods analysis

To assess the effectiveness of the proposed methodology, we compare the performance of ULISSE to several competing frameworks from the recent literature that have been deployed either to map bark beetle outbreaks in Sentinel-2 images or to handle Sentinel-2 time series in a general remote sensing mapping task. For this comparative study, we considered both the spatial version – ULISSE - 09, and the spatio-temporal version – ULISSE - 04-09 of our framework. To provide a comprehensive assessment of our framework and investigate the relative contributions of spatial and temporal information, we compare ULISSE against

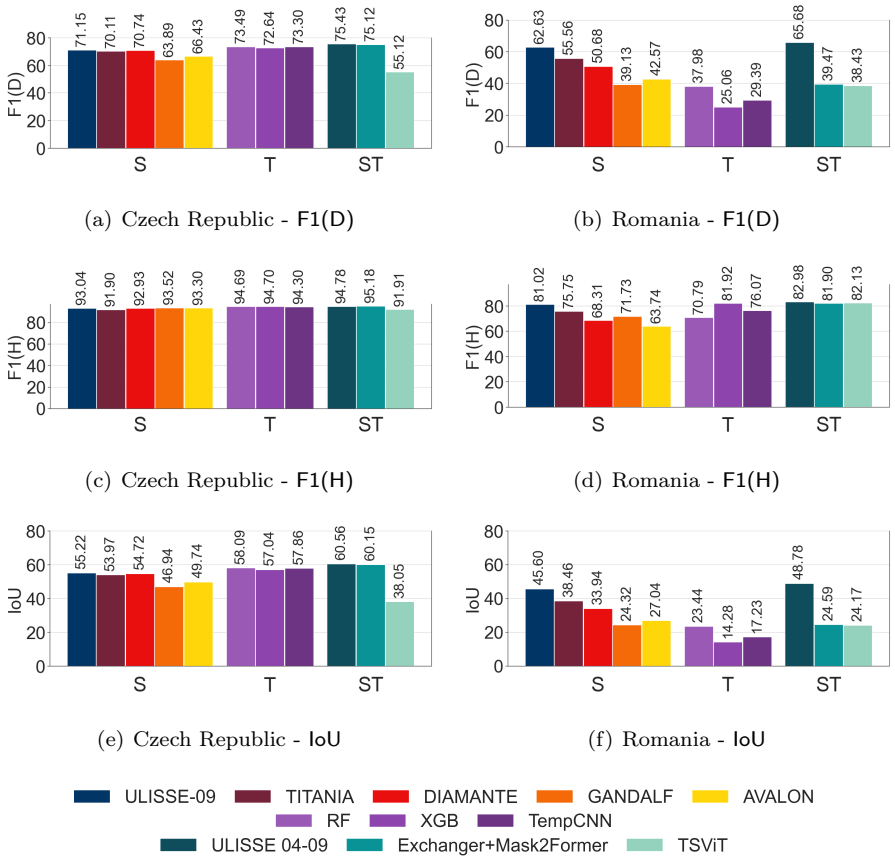
Table 5: Comparison between Full Fine-Tuning (FT) and ULISSE, in terms of resource consumption on Czech Republic and Romania datasets with batch size equals to 64 and rank value, for the LoRA approach, equals to 64. Best metric values per dataset are highlighted in bold.

	Czech Republic		**Romania**	
	FT	ULISSE	FT	ULISSE
Cost Metrics (Train)				
GFLOPs	2561.19	**3425.62**	2561.19	**3425.62**
Peak GPU Memory (GB)	**41.96**	45.65	**41.96**	45.65
Peak Power (W)	122.461	**94.74**	**112.64**	121.85
# Encoder Trainable Params (M)	141.18	**28.45**	141.18	**28.45**
Cost Metrics (Test)				
Peak GPU Memory (GB)	41.96	**7.64**	41.96	**7.64**
Peak Power (W)	122.461	**77.27**	112.64	**79.86**
Throughput (images/sec)	225.19	**228.99**	212.93	**229.97**

two categories of competing methods: those that exploit only one type of information (either spatial or temporal data configuration) and those that integrate both types of information simultaneously (spatio-temporal data configuration). Accordingly, we selected the following competing methods, for which the original publicly available implementation is available:

- TITANIA (Andresini et al., 2024b): This spatial method trains from scratch a U-Net model for the semantic segmentation of Sentinel-2 images. It integrates an attention mechanism into the U-Net architecture. In addition, it uses a self-distillation approach to transfer the knowledge within the U-Net model.

- DIAMANTE(Andresini et al., 2024a): This spatial method uses a deep data fusion strategy to train a multisensor U-Net architecture from scratch by



(a) Czech Republic - F1(D)

(b) Romania - F1(D)

(c) Czech Republic - F1(H)

(d) Romania - F1(H)

(e) Czech Republic - IoU

(f) Romania - IoU

Figure 12: Competing method analysis. F1(D), F1(H) and IoU of ULISSE, the spatial (S) competing methods: TITANIA, DIAMANTE, GANDALF and AVALON, the temporal (T) competing methods: RF, XGB and TempCNN, and the spatio-temporal (ST) competing methods: Exchanger+Mask2Former and TSViT. ULISSE was run with LoRA in both the purely spatial data configuration 09 (September), and in the spatio-temporal data configuration 04-09 (April-September) of both Czech Republic and Romania.

processing aligned Sentinel-1 and Sentinel-2 images.

- GANDALF(Pasquadibisceglie et al., 2025): This spatial method converts pixels of Sentinel-2 imagery into contextual stories and leverages a fine-tuning technique to adapt a text embedding model (Bert-medium) to the downstream task of imagery pixel classification.

- AVALON (Recchia et al., 2024): This spatial method transforms each pixel of a Sentinel-2 image into a pixel image that sees the pixel within its surrounding pixel neighborhood. It trains an Attention CNN architecture for classifying each image pixel.

- RF and XGB: These temporal methods leverage classification models trained with Random Forest (RF) and XGBoost (XGB), respectively, from the pixel-level stack of Sentinel-2 time series images. Several remote sensing studies on bark beetle mapping resorted to Random Forest (Bárta et al., 2021; Candotti et al., 2022; Andresini et al., 2024c) and XGBoost (Andresini et al., 2023b) to perform the inventory of bark beetle disturbances in Sentinel-2 images of forest areas.

- TempCNN (Pelletier et al., 2019): This is a temporal deep learning architecture that uses convolutions in the temporal dimension of Sentinel-2 pixel time series.

- Exchanger+Mask2Former (Cai et al., 2023): This is a spatio-temporal method for the semantic segmentation of Sentinel-2 time series data. It comprises: an Exchanger module that uses a hierarchical encoder to handle Sentinel-2 time series, and a Mask2Former module that is used for the semantic segmentation task. According to the suggestion of the authors, the pretrained semantic segmentation model, developed on the multi-temporal Sentinel-2 dataset PASTIS, can be adapted to any Sentinel-2 time series semantic segmentation task by adopting a full fine-tuning strategy.

- TSViT (Tarasiou et al., 2023): This is a spatio-temporal ViT-based semantic segmentation framework. It is designed to process Sentinel-2 image time series through Visual Transformers with a temporal-then-spatial attention mechanism.

All the competing methods used the same training-testing scene splits previously adopted for both the Czech Republic and Romania case studies. In

addition, the methods were run considering the parameter optimization procedures described in the reference papers[8]. For a fair comparison, we directly compare ULISSE - 09 with the spatial methods: TITANIA, DIAMANTE and GANDALF, while we compare ULISSE - 04-09 with the temporal methods: RF, XGB, TempCNN, and the spatio-temporal methods: Exchanger+Mask2Former and TSViT.

Figure 12 presents the comparative analysis results in terms of F1(D), F1(H), and IoU metrics across both the Czech Republic and Romania datasets. For the Czech Republic dataset, a clear trend emerges: temporal (T) approaches consistently outperform spatial (S) approaches across all metrics. In contrast, the Romania dataset does not exhibit such a discernible pattern. Regarding spatio-temporal (ST) approaches, which integrate both types of information, their performance varies across the two datasets. On the Czech Republic dataset, ST approaches achieve, generally, the best overall results. However, on the Romania dataset, only ULISSE demonstrates a clear and systematic improvement over single-information methods. This highlights the ability of our framework to effectively combine both information types compared to other ST approaches, such as Exchanger+Mask2Former and TsViT, which struggle to fully leverage the complementary information provided by the spatio-temporal data configurations. More precisely, ULISSE -09 outperforms TITANIA, DIAMANTE, GANDALF, and AVALON in the purely spatial data configuration, while ULISSE -04-09 generally surpasses all competing methods regardless of the data configuration employed. Two minor exceptions occur on the Czech Republic dataset, both concerning F1(H). In the purely spatial setting, GANDALF achieves a slightly higher F1(H) than ULISSE -09 (93.52 vs. 93.04). Similarly, in the spatio-temporal setting, Exchanger+Mask2Former marginally exceeds ULISSE -04-09 (95.18 vs. 94.78). Nevertheless, ULISSE consistently outperforms all competing methods

[8]We used implementations of RF and XGB available in scikit-learn with the Bayesian optimization of the parameters that both algorithms introduce to handle data imbalance in the learning stage.

in both F1(D) and IoU metrics.

First, we observe that all models achieved their highest absolute performance, in terms of both F1 and IoU, on the Czech Republic case study compared to the Romania case study. This difference can be easily attributed to two factors: i) the limited amount of labelled data associated with the Romania case study compared to the labelled data associated with the Czech Republic case study (250k vs 1.2M) and ii) the lower disease severity level in the Romanian dataset. Regarding this latter aspect, the distinction between healthy and damaged trees is more subtle and less pronounced in the Romanian case study since the disease is in an early stage, making accurate classification more complex. These two factors likely combine to make the Romania case study a challenging discrimination task. Second, we note that the performance gain between ULISSE and the second-best competitor is more pronounced in the Romania case study. This is likely because the PEFT-based strategy implemented by ULISSE effectively leverages the pretrained vision model, while models trained from scratch struggle to achieve satisfactory F1 and IoU performance with a limited amount of labelled data. This finding further supports the rationale behind ULISSE for reusing a pretrained backbone with a PEFT mechanism when only a limited number of labelled samples are available, compared to both standard full fine-tuning and learning from scratch paradigms. In contrast, in the Czech Republic case study, this gain is less pronounced, as semantic segmentation models trained from scratch can benefit from a larger training dataset.

4.4. Explanation insights

With the aim to make a step further in the analysis and characterization of our framework, ULISSE, here we investigate the contribution of each Sentinel-2 band, more precisely the corresponding time series, to the final classification. Specifically, we aim to explain the decision process behind our framework, ULISSE (run with LoRA), in the two case studies, by examining how each Sentinel-2 band contributed to the detected diseased forest areas in the testing scenes. To this purpose, we adopted the occlusion explanation approach

proposed by Covert et al. (2021), where Sentinel-2 band time series were systematically occluded by replacing their values with a zero-valued time series. We successively monitor the change in the output of the model. We decided to use a zero-valued time series for occlusion since zero is the value commonly employed to indicate cloudy, shadow, or defective Sentinel-2 acquisitions.

To explain the effect of a Sentinel-2 band, we computed the pixel-wise difference between the original model outcome, where all the bands are considered, and the model outcome when a time series band is occluded. Let us remind that our framework ULISSE employs a Sigmoid activation function for the final prediction. Accordingly, given any imagery pixel $\mathbf{x}$, if $Sigmoid(\mathbf{x}) \geq 0.5$ then pixel $\mathbf{x}$ is assigned to semantic label $+1$ ("damaged"), otherwise it is assigned to semantic label -1 ("healthy"). Notice that this binary assignment was decided following the principle, also guiding the model development, that the higher the Sigmoid value, the higher the probability that the considered pixel belongs to a diseased forest area. Conversely, the lower the Sigmoid value, the higher the probability that the considered pixel belongs to a healthy forest area. Hence, for each Sentinel-2 band B, let $Sigmoid^{\oplus B}(\cdot)$ denotes the Sigmoid value computed for each pixel with the ULISSE model when the time series of the Sentinel-2 band B is used, while $Sigmoid^{\ominus B}(\cdot)$ denotes the Sigmoid outcome obtained with the model when the time series of the Sentinel-2 band B is occluded. Based on these premises, we formulate the explanation score $xai()$, reported in Equation 6, to quantify the contribution of a Sentinel-2 band B on the semantic label predicted for a particular pixel $\mathbf{x}$. Specifically,

$$xai(\mathbf{x}, B) = y \times (Sigmoid^{\oplus B}(\mathbf{x}) - Sigmoid^{\ominus B}(\mathbf{x})), \tag{6}$$

where y represents the ground truth semantic label of pixel $\mathbf{x}$ (i.e., $y = -1$ if $\mathbf{x}$ belongs to an healthy forest patch, $y = +1$ if $\mathbf{x}$ truly belongs to a diseased forest patch). A high xai score indicates that a Sentinel-2 band contributes positively to the correct model decision on a pixel, while a low xai score indicates that a Sentinel-2 band contributes negatively to the correct model decision on a pixel. This interpretation of xai scores is based on the following considerations:

- If $y = +1$ and $Sigmoid^{\oplus B}(\mathbf{x}) \geq Sigmoid^{\ominus B}(\mathbf{x})$ then $xai(\mathbf{x}, B) \geq 0$. This positive score explains that the use of band B leads to an increase in the Sigmoid value computed for pixel $\mathbf{x}$, which may facilitate the correct assignment of pixel $\mathbf{x}$ to the semantic label "damaged" (+1).

- If $y = +1$ and $Sigmoid^{\oplus B}(\mathbf{x}) \leq Sigmoid^{\ominus B}(\mathbf{x})$ then $xai(\mathbf{x}, B) \leq 0$. This negative score explains that the use of band B leads to a reduction in the Sigmoid value computed for $\mathbf{x}$, which may cause a wrong assignment of pixel $\mathbf{x}$ to the semantic label "healthy" (-1).

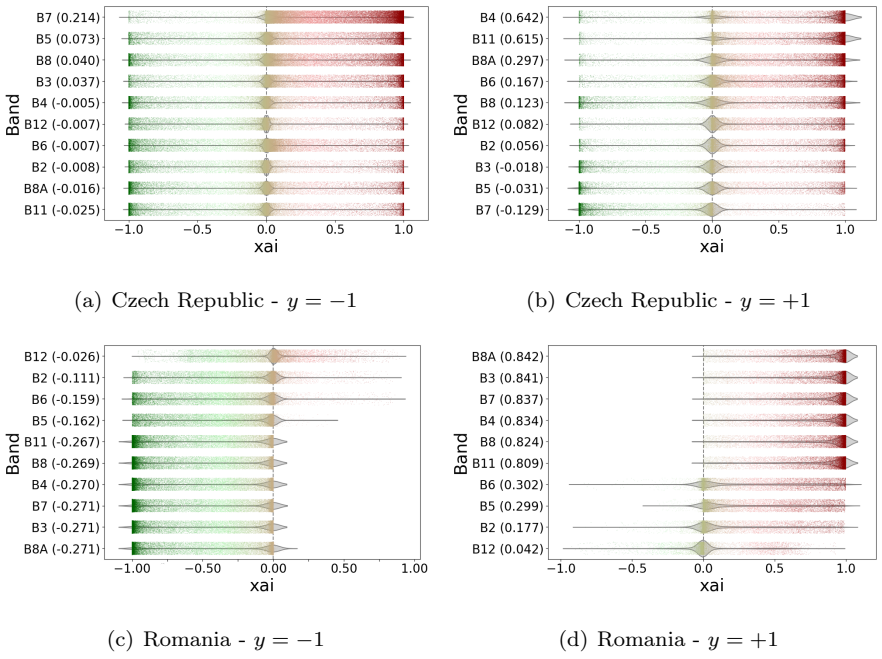
- If $y = -1$ and $Sigmoid^{\oplus B}(\mathbf{x}) \geq Sigmoid^{\ominus B}(\mathbf{x})$ then $xai(\mathbf{x}, B) \leq 0$. This negative score explains that the use of band B leads to an increase in the Sigmoid value computed for pixel $\mathbf{x}$, which may cause the wrong assignment of pixel $\mathbf{x}$ to the semantic label "damaged" (+1).

- If $y = -1$ and $Sigmoid^{\oplus B}(\mathbf{x}) \leq Sigmoid^{\ominus B}(\mathbf{x})$ then $xai(\mathbf{x}, B) \geq 0$. This positive score explains that the use of band B leads to a reduction in the Sigmoid value computed for $\mathbf{x}$, which may facilitate the correct assignment of pixel $\mathbf{x}$ to the semantic label "healthy" (-1).

We start with a global analysis of xai scores. In this regard, Figure 13 shows the beeswarm charts[9] of xai scores measured for each Sentinel-2 band, and for each imagery pixel of the maps produced in the Czech Republic and Romania, respectively. Pixels are grouped in separate charts (Figs. 13(a) and 13(b) for the Czech Republic, Figs. 13(c) and 13(d) for Romania) with respect to their ground truth semantic labels. In each chart, Sentinel-2 bands are ranked according to the average xai score measured for the plotted pixels in the considered band. These beeswarm charts deserve several considerations regarding differences or similarities in the decision behaviour of the semantic segmentation models adopted in the two case studies.

Regarding decisions on healthy semantic segments, the beeswarm charts re-

[9] `https://seaborn.pydata.org/generated/seaborn.swarmplot.html`



(a) Czech Republic - $y = -1$

(b) Czech Republic - $y = +1$

(c) Romania - $y = -1$

(d) Romania - $y = +1$

Figure 13: Beeswarm charts of *xai* scores (axis X) measured per Sentinel-2 band (axis Y) for all imagery pixels belonging to the healthy semantic segments ($y = -1$) and damaged semantic segments ($y = +1$) in the ground truth labeling maps of the testing scenes in Czech Republic (Figs. 13(a) and 13(b)) and Romania (Figs. 13(c) and 13(d)) case studies. In each plot, Sentinel-2 bands are sorted on axis Y (from top to bottom) according to the average *xai* score measured on testing pixels.

ported in Figures 13(a) and 13(c) show a clear difference in how the Sentinel-2 bands contribute to delineating the healthy forest patches in the Czech Republic and Romania. In general, the occlusion of B7, B5, B8 and B3 increases the Sigmoid values measured for healthy pixels in the Czech Republic, with the achievement of a positive *xai* score, on average, for these four bands (Fig. 13(a)). This means that each one of these bands makes it easier for the underlying semantic segmentation model to achieve the correct assignment of healthy pixels in the Czech Republic. Notably, this explanation insight is in line with the high value of Precision P measured in the evaluation of the accuracy performance of ULISSE in the Czech Republic (see configuration LoRA in Table 4).

On the other hand, the occlusion of any of the Sentinel-2 bands causes a general decrease in Sigmoid values measured for healthy pixels in Romania. This is underlined by the negative *xai* scores, on average, measured for all Sentinel-2 bands (Fig. 13(c)) with respect to the semantic label "healthy" in Romania. Once more, these negative scores are aligned with the low value of Precision P measured in Romania (see configuration LoRA in Table 4).

Instead, regarding decisions on damaged forest areas, the beeswarm charts reported in Figures 13(b) and 13(d) show that B4, B8, and B8A are in the top-5 bands, with positive *xai* scores, on average, in both case studies. Notably this is partially aligned with the study of Abdullah et al. (2019), who outlined that the Green and Red bands (B3 and B4) are commonly monitored in the context of the ecological process analysis, to examine the status of chlorophyll degradation and nitrogen deficiency, while the NIR bands (B8 and B8A), as well the SWIR bands (B11 and B12) commonly provide information to assess water content and nitrogen concentration in trees. In any case, we also note that the explanation analysis discloses the existence of some differences in how the semantics segmentation models behave in the two case studies. For example, all bands have a positive effect on correct decisions regarding the semantic label "damaged" in Romania, while B3, B5, and B7 have a negative effect on the decisions regarding the same label in the Czech Republic. On the other hand, B3, B5, and B7 have a positive effect on correct decisions regarding healthy forest areas in the Czech Republic. Conversely, B3 and B7, which have a negative effect on correctly recognizing the semantic label "damaged" in the Czech Republic, have a positive effect for the correct recognition of the same label in Romania, where they are in the second and third place in the positive *xai* ranking, respectively. In any case, these differences regarding the explanations of the decisions yielded in the two case studies are not surprising, considering that the mortality severity of the monitored disturbance was different in the considered case studies, i.e., "high" in the Czech Republic and "medium" in Romania. The explanation analysis confirms once more that the Romania case study exhibits a more challenging detection task than the one associated with the Czech Republic case

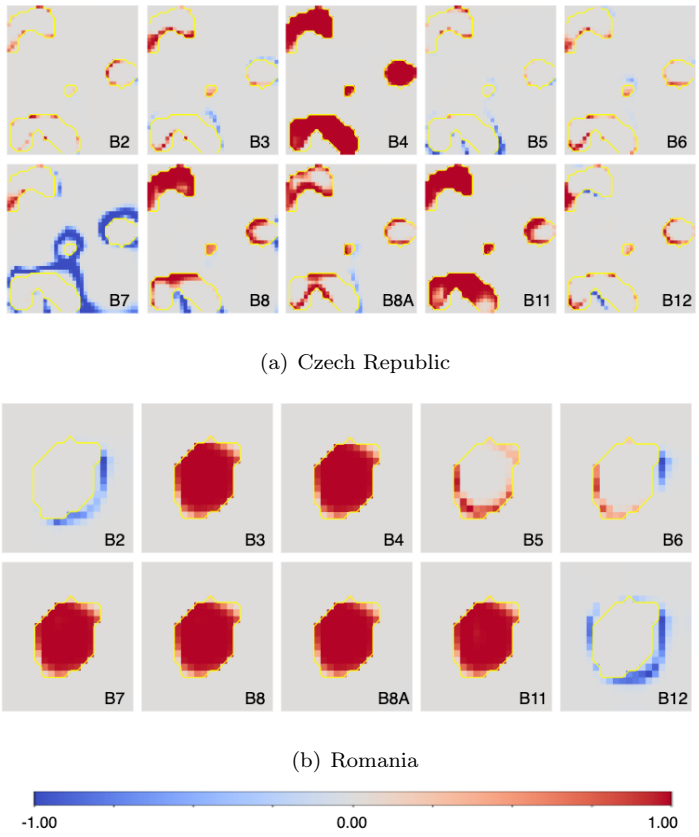
study. In fact, the analysis of the average xai score highlights that the considered semantic segmentation models struggle to correctly detect healthy areas in Romania.

To complete this explanation analysis, Figures 14(a) and 14(b) show explanation values measured for all Sentinel-2 bands across a testing scene of the Czech Republic and Romania, respectively. For both scenes, we decided to show explanation values obtained pixel-wise and masked according to the formula reported in Equation 7:

$$xai'(\mathbf{x}, B) = \begin{cases} 0 & \text{if } xai(\mathbf{x}, B) \leq 0 \\ y \times xai(\mathbf{x}, B) & \text{otherwise} \end{cases}. \tag{7}$$

xai' applies a zero-valued mask to all pixels for which the explained band has a negative effect on the correct decision (i.e., the decision that matches the ground truth). In addition, it ranges in [0,1] for pixels for which the considered band contributes positively to yield correct predictions for the semantic label "damaged", while it inverts the value in the range [-1,0] for pixels for which the considered band contributes positively to yield correct predictions for the semantic label "healthy". Notably, the obtained visual results, on the example scenes, confirm the conclusions drawn in the global explanation analysis

In particular, the example scene of the Czech Republic (Fig. 14(a)) shows that B4 and B11 emerge as prominent bands to correctly delimit diseased forest areas, while B7 and B5 emerge as prominent bands to correctly delimit healthy boundaries surrounding the diseased forest areas. Similarly, B8A, B3, B7, B4, B8 and B11 positively contribute to correctly delimiting the diseased forest areas in the considered scene of Romania (Fig. 14(b)). Also, this achievement supports, at the local level, the effectiveness of insights already derived from global explanations with regard to the effect of these bands on decisions of the semantic label "damaged" in Romania. However, it is also confirmed that bands B2 and B12 contribute to correctly delimiting the boundary of healthy forest patches surrounding the diseased area. Notably, these bands are listed in the study of Abdullah et al. (2019) who explored the relationships between the in-



(a) Czech Republic

(b) Romania

Figure 14: Explanation score maps, obtained according to the formulation of xai' using the zero masking, and computed for each Sentinel-2 band in a testing scene of Czech Republic (Fig. 14(a)) and Romania (Fig. 14(b)), respectively

formation conveyed in the Sentinel-2 bands and the spectral signs of bark beetle disturbances. In fact, according to their study, Red-edge bands (B5 and B7) are naturally involved in the model decisions' explanations as they are proven to be particularly sensitive to diseased and insect attacks. On the other hand, Green and Red bands (B3 and B4) are involved as they are commonly affected by the chlorophyll degradation and nitrogen deficiency, while both phenomena occur with forest dieback. Finally, NIR bands (B8 and B8A) and SWIR bands (B11 and B12) are involved in the model decisions' explanations as they are related to the water content and nitrogen concentration in trees, which are commonly

compromised by the bark beetle presence.

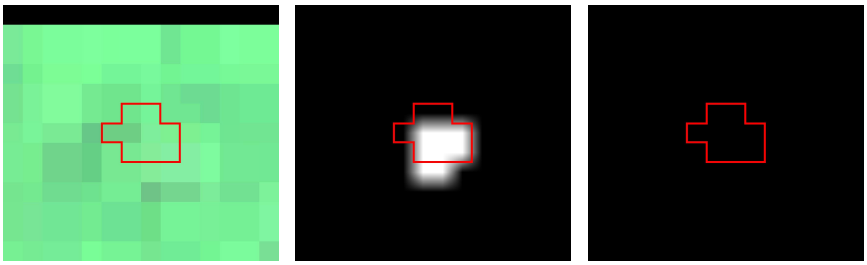
5. Discussion

To summarize, our research study has introduced a new methodological framework to map bark beetle damage via satellite image time series of Sentinel-2 imagery. Following a data-centric AI paradigm (Roscher et al., 2025), we have demonstrated how a publicly available pretrained convolutional model can be leveraged to accommodate a specific multi-temporal semantic segmentation task through a combination of modern parameter-efficient fine-tuning mechanisms and a multi-temporal U-Net architecture. We have shown that our framework is particularly effective in label-limited scenarios, such as the Romanian dataset, where recent competing, multi-temporal semantic segmentation approaches struggle to achieve satisfactory results. Despite these promising results, we acknowledge that our research has primarily focused on designing and assessing the methodological aspects of the proposed framework, while follow-up research in the ecological domain remains to be explored.

Wall-to-wall map generation and large-area processing: The mapping assessment in this study was conducted over spatially limited areas to provide a proof-of-concept for the mapping capabilities of ULISSE. Figure 15 depicts two examples of binary maps generated by ULISSE and the FT model on a scene from the Czech Republic (top) and a scene from the Romania case study (bottom). These examples provide a visual proof of the mapping capability achieved by ULISSE compared to the one exhibited by the full fine-tuning variant. Future research should focus on scaling up this approach to enable large-area processing of entire regions or countries, facilitating wall-to-wall mapping of bark beetle damage. Such work deserves a dedicated study to operationalize and fully automate the entire workflow, addressing challenges inherent to large-area processing and mapping, including the harmonization of multiple Sentinel-2 granules [10],

[10] `https://sentiwiki.copernicus.eu/web/s2-products`

(a) RGB (Czech Republic) (b) ULISSE, IoU=0.47 (c) FT, IoU=0.45



(d) RGB (Romania) (e) ULISSE, IoU=0.55 (f) FT, IoU=0.00

Figure 15: RGB images, and binary prediction maps produced with ULISSE and FT for two scenes in Czech Republic (Figs. 15(a)-15(c)) and Romania (Figs. 15(d)-15(f)), respectively. The red polygons delimit the ground-truth areas damaged by bark beetle infestations.

the mosaicking of semantic segmentation outputs, and conducting a comprehensive analysis and discussion of the resulting large-area thematic maps (Foody, 2002) to verify the forest disturbances caused by bark beetle outbreaks in the target case study.

Connection between model performance and disease severity: As introduced in Section 4.1, the documentation of DEFID2 reports that the Czech Republic dataset exhibits higher disease severity than the Romania one, making the latter a more challenging test bed for bark beetle damage detection. The experimental evaluation confirms this, as lower performance was generally achieved, with all competing methods, on the Romanian dataset in terms of absolute values. Nevertheless, general conclusions about the relationship between model performance and disease severity cannot be drawn from only two case studies. For this reason, a potential follow-up study could be devoted to

a comprehensive and rigorous analysis of model performance across different levels of disease severity. Such an analysis would aim to quantify the threshold severity level beyond which the proposed model becomes unreliable, as well as to characterize the correlation between model performance and disease severity.

Extension to other forest disturbances: Although ULISSE was developed for mapping bark beetle damage, the framework can be adopted to map other forest disturbances such as fires, windstorms, or droughts. While the remote sensing data pre-processing stage would remain largely unchanged, the main adaptations would involve selecting appropriate temporal intervals for the targeted disturbance and accessing relevant ground truth layers for the training phase. Like many mapping strategies, ULISSE is based on a supervised learning paradigm, where a classification model is trained using annotated data, more precisely, densely annotated data patches for semantic segmentation. If such ground truth data is available for the underlying mapping task, ULISSE can be adopted to map any forest disturbance effect that can be characterized and detected using Sentinel-2 multi-temporal data.

Finally, while this study has primarily focused on the design of the ULISSE methodological framework and its direct assessment, the perspectives outlined in this section highlight its potential to address pressing ecological and forest management challenges. By bridging the gap between methodological innovation and operational deployment, the discussed follow-ups can enable ULISSE to evolve from a proof-of-concept framework into a practical tool for large-area forest health monitoring in support of sustainable forest management.

6. Conclusion

In this paper, we have presented ULISSE: a deep learning methodology especially designed for the semantic segmentation of Sentinel-2 image time series. The proposed methodology adopts a convolutional encoder-decoder architecture with a multi-temporal encoder to handle time series of Sentinel-2 images. Each encoder branch is based on a multispectral pretrained ResNet block, which is

fine-tuned here for the considered downstream task using PEFT techniques. A comprehensive experimental assessment evaluates the performance of our proposed methodology in two case studies regarding the semantic segmentation of forest disturbances caused by bark beetle outbreaks in the Czech Republic and Romania. The evaluation results show that the use of a PEFT strategy outperforms the baseline approaches that use either training from scratch setting or the standard full fine-tuning setup. In particular, ULISSE with LoRA achieves the highest gain in accuracy with the lowest number of trainable parameters to update. Furthermore, the evaluation results show that ULISSE is able to gain accuracy accounting for multi-temporal Sentinel-2 images outperforming both the baseline considering Sentinel-2 images acquired at a single timestamp, as well as several related semantic segmentation methods formulated under both the spatial and spatio-temporal settings. Finally, explanatory insights obtained through occlusion-based analysis contribute to explaining differences in the decision processes of the semantic segmentation models under the two case studies.

Several promising directions exist for future work. First, the current approach requires fixed-length time series input. Extending ULISSE to handle variable-length time series data would be a valuable research direction, increasing the flexibility of our framework and accommodating operational scenarios in which no optical satellite images for a whole month are accessible. Second, a multimodal extension of the proposed methodology can be proposed to handle satellite images time series data acquired from different sensors (e.g., SAR Sentinel-1 and multi-spectral optical Sentinel-2). Third, a two-step approach, such as the one described by Illarionova et al. (2024) may also be explored in the future, to guide the analysis of medium resolution Sentinel-2 image time series as presented in this work, with information extracted from high resolution images such as the Google Earth-based images. In addition, following several remote sensing studies, e.g., Andresini et al. (2023b); Illarionova et al. (2024), as future work, the analysis can be enriched considering a selection of Spectral Vegetation Indexes (e.g., NDVI, GNDVI), which are sensitive to the presence of green vegetation, and commonly used to assess vegetation health and monitor

environmental changes. Fourth, human-in-the-loop approaches, such as active learning techniques, could be investigated to collaboratively identify the most informative labelled patches to be acquired, thereby improving the supervision and accuracy of the semantic segmentation model. Finally, while this study employs a supervised pretrained vision model to design the multi-temporal encoder, recent advances in Computer Vision have demonstrated the potential of self-supervised approaches as promising strategies for overcoming the need for manual labels in developing effective large-scale pretrained deep learning backbones (Caron et al., 2021). Since self-supervised pretraining for Earth Observation remains a rapidly evolving field (Xiao et al., 2025), future research can explore the challenges and opportunities of integrating a new self-supervised framework, especially tailored for remote sensing data, into the methodology proposed in this study.

Data availability statement

Data and code are available at `https://github.com/s4rgax/ULISSE`

Credits

Vito Recchia: Conceptualization, Methodology, Software, Data curation, Validation, Visualization, Writing - review & editing. **Giuseppina Andresini:** Conceptualization, Methodology, Investigation, Validation, Visualization, Supervision, Writing - original draft, Writing - review & editing. **Annalisa Appice:** Conceptualization, Methodology, Investigation, Validation, Writing - original draft, Project administration, Funding, Writing - review & editing. **Dino Ienco:** Conceptualization, Methodology, Investigation, Supervision, Writing - original draft, Writing - review & editing. **Giuseppe Fiameni:** Writing – review & editing. **Donato Malerba:** Conceptualization, Methodology, Writing – review & editing.

Acknowledgements

Annalisa Appice acknowledges support from the SWIFTT project, funded by the European Union under Grant Agreement 101082732. Dino Ienco acknowledges support from the Eco2Adapt project, funded by the European Union under Grant Agreement 101059498. Giuseppina Andresini, Donato Malerba and Vito Recchia are supported by the project FAIR - Future AI Research (PE00000013), Spoke 6 - Symbiotic AI, under the NRRP MUR program funded by the NextGenerationEU. We also acknowledge the use of Leonardo supercomputer for performing experiments with the support of CINECA-Italian Super Computing Resource Allocation, class C project satellIte iMage seMantic segmentatiOn foR foresT heALth monitoring – IMMORTAL (HP10CFR5D3).

References

Abdullah, H., Skidmore, A. K., Darvishzadeh, R., & Heurich, M. (2019). Sentinel-2 accurately maps green-attack stage of european spruce bark beetle (ips typographus, l.) compared with landsat-8. *Remote Sensing in Ecology and Conservation*, *5*, 87–106. doi:10.1002/rse2.93.

Abidi, A., Ienco, D., Abbes, A. B., & Farah, I. R. (2024). Multi-scale classification of sentinel-2 images for land cover mapping using two-branch convolutional neural network. In *IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2024* (pp. 4109–4113). IEEE. doi:10.1109/IGARSS53475.2024.10640796.

Aleem, S., Dietlmeier, J., Arazo, E., & Little, S. (2024). Convlora and adabn based domain adaptation via self-training. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)* (pp. 1–5). IEEE. doi:10.1109/ISBI56570.2024.10635661.

Andresini, G., Appice, A., Ienco, D., & Malerba, D. (2023a). Seneca: Change detection in optical imagery using siamese networks with active-transfer

learning. *Expert Systems with Applications*, *214*, 119123. doi:10.1016/j.eswa.2022.119123.

Andresini, G., Appice, A., Ienco, D., & Recchia, V. (2024a). DIAMANTE: A data-centric semantic segmentation approach to map tree dieback induced by bark beetle infestations via satellite images. *J. Intell. Inf. Syst.*, *62*, 1531–1558. doi:10.1007/S10844-024-00877-6.

Andresini, G., Appice, A., & Malerba, D. (2023b). SILVIA: An explainable framework to map bark beetle infestation in Sentinel-2 images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *16*, 10050–10066. doi:10.1109/JSTARS.2023.3312521.

Andresini, G., Appice, A., & Malerba, D. (2024b). A deep semantic segmentation approach to map forest tree dieback in sentinel-2 data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.*, *17*, 17075–17086. doi:10.1109/JSTARS.2024.3460981.

Andresini, G., Appice, A., & Malerba, D. (2024c). Leveraging sentinel-2 time series for bark beetle-induced forest dieback inventory. In *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing, SAC 2024,* (pp. 875–882). ACM. doi:10.1145/3605098.3635908.

Bhardwaj, D., Nagabhooshanam, N., Singh, A., Selvalakshmi, B., Angadi, S., Shargunam, S., Guha, T., Singh, G., & Rajaram, A. (2025). Enhanced satellite imagery analysis for post-disaster building damage assessment using integrated resnet-u-net model. *Multim. Tools Appl.*, *84*, 2689–2714. doi:10.1007/S11042-024-20300-0.

Bohlol, P., Hosseinpour, S., & Soltani Firouz, M. (2025). Improved food recognition using a refined resnet50 architecture with improved fully connected layers. *Current Research in Food Science*, *10*, 101005. doi:10.1016/j.crfs.2025.101005.

Brar, K. K., Goyal, B., Dogra, A., Mustafa, M. A., Majumdar, R., Alkhayyat, A., & Kukreja, V. (2025). Image segmentation review: Theoretical background and recent advances. *Information Fusion*, *114*, 102608. doi:10.1016/j.inffus.2024.102608.

Bárta, V., Hanuš, J., Dobrovolný, L., & Homolová, L. (2022). Comparison of field survey and remote sensing techniques for detection of bark beetle-infested trees. *Forest Ecology and Management*, *506*, 119984. doi:10.1016/j.foreco.2021.119984.

Bárta, V., Lukeš, P., & Homolová, L. (2021). Early detection of bark beetle infestation in norway spruce forests of central europe using sentinel-2. *International Journal of Applied Earth Observation and Geoinformation*, *100*, 102335. doi:10.1016/j.jag.2021.102335.

Cai, X., Bi, Y., Nicholl, P. N., & Sterritt, R. (2023). Revisiting the encoding of satellite image time series. In *34th British Machine Vision Conference 2023, BMVC 2023* (pp. 402–404). BMVA Press.

Candotti, A., De Giglio, M., Dubbini, M., & Tomelleri, E. (2022). A sentinel-2 based multi-temporal monitoring framework for wind and bark beetle detection and damage mapping. *Remote Sensing*, *14*. doi:10.3390/rs14236105.

Carletti, H., Gégout, J.-C., Dutrieux, R., Féret, J.-B., Vega, C., Belouard, T., Jolly, A., Cansell, J., & Piedallu, C. (2025). Sentinel-2 time series reveal species-specific responses in temperate conifer dieback. *European Journal of Remote Sensing*, *58*, 2547386. doi:10.1080/22797254.2025.2547386.

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

Cheng, B., Misra, I., Schwing, A. G., Kirillov, A., & Girdhar, R. (2022). Masked-attention mask transformer for universal image segmentation. In

IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022 (pp. 1280–1289). IEEE. doi:10.1109/CVPR52688.2022.00135.

Clasen, K. N., Hackel, L., Burgert, T., Sumbul, G., Demir, B., & Markl, V. (2025). reben: Refined bigearthnet dataset for remote sensing image analysis. In *IGARSS 2025 - 2025 IEEE International Geoscience and Remote Sensing Symposium* (pp. 1–5). doi:10.48550/arXiv.2407.03653.

Cong, Y., Khanna, S., Meng, C., Liu, P., Rozi, E., He, Y., Burke, M., Lobell, D. B., & Ermon, S. (2022). Satmae: pre-training transformers for temporal and multi-spectral satellite imagery. In *Proceedings of the 36th International Conference on Neural Information Processing Systems* NIPS '22. Curran Associates Inc.

Covert, I. C., Lundberg, S., & Lee, S.-I. (2021). Explaining by removing: a unified framework for model explanation. *J. Mach. Learn. Res.*, *22*.

Dalponte, M., Cetto, R., Marinelli, D., Andreatta, D., Salvadori, C., Pirotti, F., Frizzera, L., & Gianelle, D. (2023). Spectral separability of bark beetle infestation stages: A single-tree time-series analysis using planet imagery. *Ecological Indicators*, *153*, 110349. doi:https://doi.org/10.1016/j.ecolind.2023.110349.

Danish, S., Sadeghi-Niaraki, A., Khan, S. U., Dang, L. M., Tightiz, L., & Moon, H. (2026). A comprehensive survey of vision–language models: Pretrained models, fine-tuning, prompt engineering, adapters, and benchmark datasets. *Information Fusion*, *126*, 103623. doi:10.1016/j.inffus.2025.103623.

Darwis, H., Puspitasari, R., Purnawansyah, Astuti, W., Atmajaya, D., & Hasnawi, M. (2025). A deep learning approach for improving waste classification accuracy with resnet50 feature extraction. In *2025 19th International Conference on Ubiquitous Information Management and Communication (IMCOM)* (pp. 1–8). doi:10.1109/IMCOM64595.2025.10857536.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248–255). Ieee. doi:10.1109/CVPR.2009.5206848.

Dimitrovski, I., Spasev, V., Loshkovska, S., & Kitanovski, I. (2024). U-net ensemble for enhanced semantic segmentation in remote sensing imagery. *Remote Sensing*, *16*. doi:10.3390/rs16122077.

Fan, X., Yan, C., Fan, J., & Wang, N. (2022). Improved u-net remote sensing classification algorithm fusing attention and multiscale features. *Remote Sensing*, *14*. doi:10.3390/rs14153591.

Foody, G. M. (2002). Status of land cover classification accuracy assessment. *Remote sensing of environment*, *80*, 185–201. doi:10.1016/S0034-4257(01)00295-4.

Forzieri, G., Dutrieux, L., & et al. (2023). The database of European forest insect and disease disturbances: DEFID2. *Global Change Biology*, *29*, 6040–6065. doi:10.1111/gcb.16912.

Garnot, V. S. F., & Landrieu, L. (2020). Lightweight temporal self-attention for classifying satellite images time series. In *International Workshop on Advanced Analytics and Learning on Temporal Data* (pp. 171–181). Springer. doi:10.48550/arXiv.2007.00586.

Han, Z., Gao, C., Liu, J., Zhang, J., & Zhang, S. Q. (2024). Parameter-efficient fine-tuning for large models: A comprehensive survey. *Trans. Mach. Learn. Res.*, *2024*.

Hao, S., Zhou, Y., & Guo, Y. (2020). A brief survey on semantic segmentation with deep learning. *Neurocomputing*, *406*, 302–321. doi:10.1016/j.neucom.2019.11.118.

Hofmann, S., Kautz, M., & Schebeck, M. (2025). High plasticity in diapause responses benefits bark beetles in a changing climate. *Ecological Entomology*, *50*, 62–73. doi:https://doi.org/10.1111/een.13378.

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., & Gelly, S. (2019). Parameter-efficient transfer learning for nlp. In *International conference on machine learning* (pp. 2790–2799). PMLR.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W. et al. (2022). Lora: Low-rank adaptation of large language models. *ICLR*, *1*, 3. doi:10.48550/arXiv.2106.09685.

Huang, Z., Chen, Z., & Liu, Y. (2025). Fbinet: Few-shot semantic segmentation with foreground and background iteration. *IEEE Transactions on Instrumentation and Measurement*, *74*, 1–10. doi:10.1109/TIM.2025.3550211.

Illarionova, S., Tregubova, P., Shukhratov, I., Shadrin, D. G., Kedrov, A., & Burnaev, E. (2024). Remote sensing data fusion approach for estimating forest degradation: a case study of boreal forests damaged by polygraphus proximus. *Frontiers in Environmental Science*, (pp. 1–14). doi:https://doi.org/10.3389/fenvs.2024.1412870.

Inglada, J., Vincent, A., Arias, M., Tardy, B., Morin, D., & Rodes, I. (2017). Operational high resolution land cover map production at the country scale using satellite image time series. doi:10.3390/rs9010095.

Jamali, S., Olsson, P.-O., Ghorbanian, A., & Müller, M. (2023). Examining the potential for early detection of spruce bark beetle attacks using multitemporal sentinel-2 and harvester data. *ISPRS Journal of Photogrammetry and Remote Sensing*, *205*, 352–366. doi:10.1016/j.isprsjprs.2023.10.013.

Kerssies, T., Cavagnero, N., Hermans, A., Norouzi, N., Averta, G., Leibe, B., Dubbelman, G., & De Geus, D. (2025). Your vit is secretly an image seg-

mentation model. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 25303–25313). doi:10.1109/CVPR52734.2025.02356.

Kluczek, M., & Zagajewski, B. (2025). Mapping spatiotemporal mortality patterns in spruce mountain forests using sentinel-2 data and environmental factors. *Ecological Informatics*, *86*, 103074. doi:https://doi.org/10.1016/j.ecoinf.2025.103074.

Li, J., Cai, Y., Li, Q., Kou, M., & Zhang, T. (2024). A review of remote sensing image segmentation by deep learning methods. *International Journal of Digital Earth*, *17*, 2328827. doi:10.1080/17538947.2024.2328827.

Lillesand, T. M., Kiefer, R. W., & Chipman, J. W. (2015). *Remote Sensing and Image Interpretation*. (7th ed.). Hoboken, NJ: John Wiley & Sons.

Liu, S., Wang, C., Yin, H., Molchanov, P., Wang, Y. F., Cheng, K., & Chen, M. (2024a). Dora: Weight-decomposed low-rank adaptation. In *Forty-first International Conference on Machine Learning, ICML 2024*. OpenReview.net. doi:10.48550/arXiv.2402.09353.

Liu, S.-Y., Wang, C.-Y., Yin, H., Molchanov, P., Wang, Y.-C. F., Cheng, K.-T., & Chen, M.-H. (2024b). Dora: Weight-decomposed low-rank adaptation. In *Forty-first International Conference on Machine Learning*.

Louis, J., Debaecker, V., Pflug, B., Main-Knorn, M., Bieniarz, J., Mueller-Wilm, U., Cadau, E., & Gascon, F. (2016). Sentinel-2 sen2cor: L2a processor for users. In *Proceedings of the Living Planet Symposium 2016* (pp. 1–8). Spacebooks Onlin.

Lu, S., Guo, J., Zimmer-Dauphinee, J. R., Nieusma, J. M., Wang, X., VanValkenburgh, P., Wernke, S. A., & Huo, Y. (2025). Vision foundation models in remote sensing: A survey. *IEEE Geoscience and Remote Sensing Magazine*, *13*, 190–215. doi:10.1109/MGRS.2025.3541952.

Mai, Z., Zhang, P., Tu, C.-H., Chen, H.-Y., Nguyen, Q.-H., Zhang, L., & Chao, W.-L. (2025). Lessons and insights from a unifying study of parameter-efficient fine-tuning (peft) in visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 14845–14857). doi:10.1109/CVPR52734.2025.01383.

Mansourifar, H., Moskovitz, A., Klingensmith, B., Mintas, D., & Simske, S. J. (2022). Gan-based satellite imaging: A survey on techniques and applications. *IEEE Access*, *10*, 118123–118140. doi:10.1109/ACCESS.2022.3221123.

Marti Escofet, F., Blumenstiel, B., Scheibenreif, L., Fraccaro, P., & Schindler, K. (2026). Fine-tune smarter, not harder: Parameter-efficient fine-tuning for geospatial foundation models. In *Machine Learning and Knowledge Discovery in Databases. Research Track* (pp. 516–532). Cham: Springer Nature Switzerland.

Nwaiwu, S. (2025). Parameter-efficient fine-tuning for low-resource text classification: a comparative study of lora, ia3, and reft. *Frontiers in Big Data*, *Volume 8 - 2025*. doi:10.3389/fdata.2025.1677331.

Pasquadibisceglie, V., Recchia, V., Appice, A., Malerba, D., & Fiameni, G. (2025). GANDALF: A llm-based approach to map bark beetle outbreaks in semantic stories of sentinel-2 images. In *Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing, SAC 2025* (pp. 1074–1081). ACM. doi:10.1145/3672608.3707751.

Pelletier, C., Webb, G. I., & Petitjean, F. (2019). Temporal convolutional neural network for the classification of satellite image time series. *Remote Sensing*, *11*. doi:10.3390/rs11050523.

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

Ramos, L., & Sappa, A. (2025). Leveraging u-net and selective feature extraction for land cover classification using remote sensing imagery. *Scientific Reports*, *15*, 1–17. doi:10.1038/s41598-024-84795-1.

Recchia, V., Andresini, G., Appice, A., Fontana, G., & Malerba, D. (2024). An attention-based CNN approach to detect forest tree dieback caused by insect outbreak in sentinel-2 images. In *Discovery Science - 27th International Conference, DS 2024, Proceedings, Part II* (pp. 183–199). Springer volume 15244 of *Lecture Notes in Computer Science*. doi:10.1007/978-3-031-78980-9_12.

Reinosch, E., Backa, J., Adler, P., Deutscher, J., Eisnecker, P., Hoffmann, K., Langner, N., Puhm, M., Rüetschi, M., Straub, C., Waser, L. T., Wiesehahn, J., & Oehmichen, K. (2024). Detailed validation of large-scale sentinel-2-based forest disturbance maps across germany. *Forestry: An International Journal of Forest Research*, *98*, 437–453. doi:10.1093/forestry/cpae038.

Roscher, R., Russwurm, M., Gevaert, C., Kampffmeyer, M., Dos Santos, J. A., Vakalopoulou, M., Hänsch, R., Hansen, S., Nogueira, K., Prexl, J. et al. (2025). Better, not just more: Data-centric machine learning for earth observation. *IEEE Geoscience and Remote Sensing Magazine*, *13*, 512–532. doi:10.1109/MGRS.2024.3470986.

Rußwurm, M., & Körner, M. (2018). Multi-temporal land cover classification with sequential recurrent encoders. *ISPRS International Journal of Geo-Information*, *7*, 129.

Salehi, S. S. M., Erdogmus, D., & Gholipour, A. (2017). Tversky loss function for image segmentation using 3D fully convolutional deep networks. In *Machine Learning in Medical Imaging* (pp. 379–387). Springer. doi:10.1007/978-3-319-67389-9_44.

Sang, H., Zhou, Q., & Zhao, Y. (2020). Pcanet: Pyramid convolutional attention network for semantic segmentation. *Image and Vision Computing*, *103*, 103997. doi:10.1016/j.imavis.2020.103997.

Shelhamer, E., Long, J., & Darrell, T. (2017). Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*, 640–651. doi:10.1109/TPAMI.2016.2572683.

Sumbul, G., De Wall, A., Kreuziger, T., Marcelino, F., Costa, H., Benevides, P., Caetano, M., Demir, B., & Markl, V. (2021). Bigearthnet-mm: A large-scale, multimodal, multilabel benchmark archive for remote sensing image classification and retrieval [software and data sets]. *IEEE Geoscience and Remote Sensing Magazine*, *9*, 174–180.

Szwarcman, D., Roy, S., Fraccaro, P., Þorsteinn Elí Gíslason, Blumenstiel, B., Ghosal, R., de Oliveira, P. H., de Sousa Almeida, J. L., Sedona, R., Kang, Y., Chakraborty, S., Wang, S., Gomes, C., Kumar, A., Truong, M., Godwin, D., Lee, H., Hsu, C.-Y., Asanjan, A. A., Mujeci, B., Shidham, D., Keenan, T., Arevalo, P., Li, W., Alemohammad, H., Olofsson, P., Hain, C., Kennedy, R., Zadrozny, B., Bell, D., Cavallaro, G., Watson, C., Maskey, M., Ramachandran, R., & Moreno, J. B. (2024). Prithvi-eo-2.0: A versatile multi-temporal foundation model for earth observation applications. doi:10.48550/arXiv.2412.02732.

Taghanaki, S. A., Bentaieb, A., Sharma, A., Zhou, S. K., Zheng, Y., Georgescu, B., Sharma, P., Xu, Z., Comaniciu, D., & Hamarneh, G. (2019). Select, attend, and transfer: Light, learnable skip connections. In *Machine Learning in Medical Imaging: 10th International Workshop, MLMI 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings* (p. 417–425). Berlin, Heidelberg: Springer-Verlag. doi:10.1007/978-3-030-32692-0_48.

Tarasiou, M., Chavez, E., & Zafeiriou, S. (2023). Vits for SITS: vision transformers for satellite image time series. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023* (pp. 10418–10428). IEEE. doi:10.1109/CVPR52729.2023.01004.

Thomas Ramos, L., & Sappa, A. D. (2024). Multispectral semantic segmentation for land cover classification: An overview. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *17*, 14295–14336. doi:10.1109/JSTARS.2024.3438620.

Xiao, A., Xuan, W., Wang, J., Huang, J., Tao, D., Lu, S., & Yokoya, N. (2025). Foundation models for remote sensing and earth observation: A survey. *IEEE Geoscience and Remote Sensing Magazine*, . doi:10.48550/arXiv.2410.16602.

Xiao, Z., Chai, T., Li, N., Shen, X., Guan, T., Tian, J., & Li, X. (2024). Research advances in deep learning for image semantic segmentation techniques. *IEEE Access*, *12*, 175715–175741. doi:10.1109/ACCESS.2024.3496723.

Xin, Y., Luo, S., Zhou, H., Du, J., Liu, X., Fan, Y., Li, Q., & Du, Y. (2024). Parameter-efficient fine-tuning for pre-trained vision models: A survey. *arXiv e-prints*, (pp. arXiv–2402).

Xin, Y., Yang, J., Luo, S., Du, Y., Qin, Q., Cen, K., He, Y., Fu, B., Yang, X., Zhai, G., Yang, M.-H., & Liu, X. (2025). Parameter-efficient fine-tuning for pre-trained vision models: A survey and benchmark. URL: https://arxiv.org/abs/2402.02242. arXiv:2402.02242.

Xu, C., Förster, M., Gränzig, T., May, J., & Kleinschmit, B. (2024). Relating soil moisture and sentinel-2 vegetation index patterns to spruce bark beetle infestations prior to outbreak. *Forestry: An International Journal of Forest Research*, *97*, 728–738. doi:10.1093/forestry/cpae007.

Yang, M., Chen, J., Zhang, Y., Liu, J., Zhang, J., Ma, Q., Verma, H., Zhang, Q., Zhou, M., King, I. et al. (2024). Low-rank adaptation for foundation models: A comprehensive review. *arXiv preprint arXiv:2501.00365*, .

Yu, Y., & Geng, G. (2025). Deep learning methods for phase segmentation in backscattered electron images of cement paste and scm-blended sys-

tems. *Cement and Concrete Composites*, *155*, 105810. doi:10.1016/j.cemconcomp.2024.105810.

Yu, Y., Yang, C.-H. H., Kolehmainen, J., Shivakumar, P. G., Gu, Y., Ren, S. R. R., Luo, Q., Gourav, A., Chen, I.-F., Liu, Y.-C., Dinh, T., Filimonov, A. G. D., Ghosh, S., Stolcke, A., Rastow, A., & Bulyko, I. (2023). Low-rank adaptation of large language model rescoring for parameter-efficient speech recognition. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2023* (pp. 1–8). doi:10.1109/ASRU57964.2023.10389632.

Yuan, S., Liu, H., & Xu, H. (2024a). Bridging the gap between low-rank and orthogonal adaptation via householder reflection adaptation. *Advances in Neural Information Processing Systems*, *37*, 113484–113518. doi:10.52202/079017-3606.

Yuan, S., Liu, H., & Xu, H. (2024b). Bridging the gap between low-rank and orthogonal adaptation via householder reflection adaptation. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024*.

Zaken, E. B., Goldberg, Y., & Ravfogel, S. (2022). Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022* (pp. 1–9). Association for Computational Linguistics.

Zhang, J., Cong, S., Zhang, G., Ma, Y., Zhang, Y., & Huang, J. (2022). Detecting pest-infested forest damage through multispectral satellite imagery and improved unet++. *Sensors*, *22*. doi:10.3390/s22197440.

Zhou, W., Yue, Y., Fang, M., Qian, X., Yang, R., & Yu, L. (2023). Bcinet: Bilateral cross-modal interaction network for indoor scene understanding

in rgb-d images. *Information Fusion*, *94*, 32–42. doi:10.1016/j.inffus.2023.01.016.

Zhu, H., Zhu, Y., Xiao, J., Xiao, T., Ma, Y., Zhang, Y., & Dai, F. (2025). Exact: Exploring space-time perceptive clues for weakly supervised satellite image time series semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025* (pp. 14036–14045). Computer Vision Foundation / IEEE. doi:10.1109/CVPR52734.2025.01310.

Östersund, M., Honkavaara, E., Oliveira, R. A., Näsi, R., Hakala, T., Koivumäki, N., Pelto-Arvo, M., Tuviala, J., Nevalainen, O., & Lyytikäinen-Saarenmaa, P. (2024). Exploring forest changes in an ips typographus l. outbreak area: insights from multi-temporal multispectral uas remote sensing. *European Journal of Forest Research*, *143*, 871–18925. doi:10.1007/s10342-024-01734-5.