Ecological Informatics 89 (2025) 103129

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

Adapting a global plant identification model to detect invasive alien plant species in high-resolution road side images

Vincent Espitalier [a], Jean-Christophe Lombardo [b], Hervé Goëau [a],*, Christophe Botella [b], Toke Thomas Høye [c], Mads Dyrmann [d], Pierre Bonnet [a], Alexis Joly [b]

[a] CIRAD, UMR AMAP, Montpellier, Occitanie, France
[b] Inria, LIRMM, Univ Montpellier, CNRS, Montpellier, France
[c] Department of Ecoscience and Arctic Research Centre, Aarhus University, C. F. Møllers Allé 8, DK 8000, Aarhus C, Denmark
[d] AI Lab ApS, Aarhus, 8210, Denmark

ARTICLE INFO

Dataset link: https://zenodo.org/record/14013930, https://zenodo.org/record/13891416, https://github.com/plantnet/roadside-invasive-plant-identification

Keywords:
Invasive species
Roadside biodiversity monitoring
Plant recognition
Deep learning
Large-scale species identification
Tiling
Pl@ntNet

ABSTRACT

Early detection of invasive alien plant species is crucial for addressing their environmental impact. Recent advancements in vehicle-mounted equipment enable automatic analysis of high-resolution images to detect invasive plants along roadsides, a primary vector for their spread. Deep learning technologies show promise for processing this data efficiently, but the choice of approach significantly affects both computational and human resource costs. Object detection and segmentation methods require costly annotations, making them impractical for scaling to the thousands of invasive species worldwide. In contrast, multi-label classification, i.e. to predict all species present in the image, is less demanding but still challenging to implement without many annotated images for numerous species. However, large datasets from citizen science platforms such as Pl@ntNet or iNaturalist offer rich visual data for classifying individual plant species. In this article, we assess whether large plant identification models trained on such data can be leveraged for species detection in high-resolution images. Specifically, we explore two approaches: a multi-label classification model and a tiling-based model, using a vision transformer from the Pl@ntNet platform. We evaluate these models on high-resolution roadside images, both using a pre-trained model without fine-tuning and after applying fine-tuning. Our findings indicate that the tiling approach significantly outperforms other methods without fine-tuning and shows a slight advantage when fine-tuning is applied, demonstrating significant potential for detecting thousands of species without task-specific adaptation.

1. Introduction

Invasive species have been identified as major drivers of biodiversity loss and ecosystem disruption (Roy et al., 2023; Díaz et al., 2019), with their economic impact measured in billions globally (Haubrock et al., 2020). While early detection is crucial to managing their spread, current monitoring methods often struggle to cover broad areas effectively. Remote sensing and citizen science provide valuable monitoring data, but both approaches come with limitations. Remote sensing requires expensive, high-resolution imagery and may not allow for accurate identification of individual species, while citizen science data can suffer from geographic and taxonomic biases (Isaac and Pocock, 2015; Johnson et al., 2020).

Vehicle-mounted cameras offer a promising alternative for monitoring invasive alien plant species along transport networks, which are key pathways for species dispersal (Dyrmann et al., 2021; Kotowska et al., 2021). Such cameras, combined with deep learning-based image analysis, have great potential for large-scale plant monitoring. Deep learning methods like object detection (Hussain, 2023; Zong et al., 2022) and instance segmentation (Wang et al., 2022) can localize and identify plants in complex images, although they require considerable effort for manual annotation and updates to handle new species. However, these methods often struggle to generalize across diverse environments and species, particularly when faced with variation in plant growth stages and plant appearances.

Powerful species identification tools, such as those available through citizen science platforms like Pl@ntNet and iNaturalist, leverage extensive visual datasets. Using these tools, rather than traditional

* Corresponding author.
E-mail addresses: vincent.espitalier@cirad.fr (V. Espitalier), jean-christophe.lombardo@inria.fr (J.-C. Lombardo), herve.goeau@cirad.fr (H. Goëau), christophe.botella@inria.fr (C. Botella), tth@ecos.au.dk (T.T. Høye), mads@theailab.dk (M. Dyrmann), pierre.bonnet@cirad.fr (P. Bonnet), alexis.joly@inria.fr (A. Joly).

https://doi.org/10.1016/j.ecoinf.2025.103129
Received 23 August 2024; Received in revised form 31 March 2025; Accepted 1 April 2025
Available online 22 April 2025

object detection and instance segmentation methods, could enable broader deployment with the potential to detect a larger diversity of species. In this article, we explore an alternative approach that builds on these identification frameworks, aiming to improve scalability and efficiency in analyzing complex landscapes. Specially, we examine the Pl@ntNet model, trained on over six million images to identify more than forty thousand species, and evaluate its capacity to detect invasive plants from roadside images. However, it should be noted that there can be a significant gap, known as domain shift, between the training data used by Pl@ntNet model and the high-resolution images that need to be processed. Pl@ntNet primarily relies on zoomed-in, detailed images of individual plants, often focused on specific organs, such as on leave or a flower. In contrast, the target images are much larger, containing numerous species captured from a different angle and using different equipment, such as high-throughput professional cameras rather than the smartphones typically used by Pl@ntNet contributors. Our study assesses the model's performance without fine-tuning alongside a fine-tuned version with multi-species annotations to evaluate its suitability for large-scale, real-world applications.

2. Related work

A wide range of data sources and methods can be used to monitor invasive plants, including aerial imagery from airplanes or drones, satellite data (Müllerová et al., 2013; James and Bradshaw, 2020), roadside images captured with professional cameras (Dyrmann et al., 2021), smartphones photos (Pinzani and Ceschin, 2023), Google Street Views images (Kotowska et al., 2021, 2024), and even data derived from social media (Daume, 2016). In this paper, we focus on detecting invasive species in high-resolution plant survey images, with an initial evaluation based on roadside imagery from the dataset used in Dyrmann et al. (2021). Recording images from cars, trains and boats represents a cost-effective alternative to aerial or satellite campaigns, while still offering high spatial resolution necessary for fine-scale species identification. A key challenge with this type of data lies in the discrepancy between the large size of the images to be processed and the input size required by state-of-the-art deep learning models for image analysis. The images are often high-resolution and depict complex scenes containing numerous objects of interest, including potentially a large variety of plants and species in our case. Many deep learning architectures for image classification were originally developed and trained using datasets such as ImageNet, where input images are typically resized to resolutions typically between 224 × 224 and 518 × 518 pixels during pre-processing. Therefore, it is essential to develop technical solutions to process high-resolution images effectively while optimizing identification performance using these advanced models. For more than a decade, Convolutional Neural Networks (CNNs) (Krizhevsky et al., 2012; He et al., 2015; Szegedy et al., 2015; Woo et al., 2023) have been widely used for computer vision tasks, consistently achieving groundbreaking results across various image analysis problems. Since then, Vision Transformers (ViTs) have now become the standard in computer vision, especially with the advent of Self Supervised Learning (SSL) pretraining techniques (Bao et al., 2021; He et al., 2021; Oquab et al., 2023). In computer vision, SSL techniques train a network to predict missing or transformed parts of an image, allowing it to learn meaningful patterns and distinctions between visual elements without explicit labels. These techniques have demonstrated strong compatibility with the vision transformer architecture, enabling efficient pre-training on large datasets before fine-tuning for specific downstream tasks. The hidden part can depend on zooming and cropping like in DINO (Caron et al., 2021), or depend on partial masking of image patches like in MAE (He et al., 2021) or the Bidirectional Encoder representation from Image Transformer (BEiT) self-supervised learning method (Bao et al., 2021). By initially training on numerous unlabeled images through a self-supervised approach and subsequently fine-tuning the model with labeled data via supervised training, a more robust model can be generated compared to training it directly from scratch with supervision. However, while CNN models can be scaled up with image resolution by using relatively simple adaptive pooling operations, at the cost of increased memory usage, ViTs are more constrained by a fixed input resolution (Bao et al., 2021) requires 384 × 384 pixels images. Resizing high-resolution images to smaller resolutions creates an informational bottleneck that significantly compromises classification performance, particularly when the target plant species is small, as meaningful details may be lost, making detection increasingly challenging.

Moreover, for most invasive species, there is a lack of training data consisting of high-resolution (HR) images labeled with presence and absence information. The available datasets generally focus on a limited number of species and specific contexts. While training a deep learning model on these datasets can yield relatively good performance for the targeted species, it often comes at the cost of poor generalization to other contexts. In this study, one of our focal points is the scenario where the model is used without fine-tuning, meaning that it has not been specifically trained on any images of the final downstream task (i.e the HR images acquired by the vehicle-mounted camera). This scenario preserves the model's generalization capabilities across various species. For this study, we utilize the Pl@ntNet model that has been trained to identify more than 43,000 species using a dataset comprising over six million individual plant images from various countries and continents, independent to any specific downstream task. Leveraging a ViT architecture, we anticipate enhanced performance in detecting the presence of invasive species compared to previous CNNs.

3. Materials and methods

3.1. Dataset

In this paper, the methods were tested on a dataset used in a previous study (Dyrmann et al., 2021) to evaluate and compare high-throughput detection and classification of invasive plants. The dataset contains 15,529 high-resolution images showing roadside views taken in Denmark from a high-speed camera mounted on the roof of a car and oriented perpendicular to the direction of travel.

3.1.1. Description and cleaning

The dataset focuses on 11 invasive species that can be observed along a significant portion of Denmark's roads. Due to the striking visual similarity between sibling invasive species belonging to the same genus, Dyrmann et al. (2021) consolidated some species into a single class or meta-species (referred to as "spp". in the rest of this article), resulting in a total of seven distinct taxa: *Cytisus scoparius* (L.) Link, *Heracleum* spp. (*Heracleum mantegazzianum* Sommier & Levier, *Heracleum persicum* Desf. ex Fisch., C.A.Mey. & Avé-Lall., *Heracleum sosnowskyi* Manden.), *Lupinus polyphyllus* Lindl., *Pastinaca sativa* L., *Reynoutria spp.* (*Reynoutria japonica* Houtt., *Reynoutria sachalinensis* (F.Schmidt) Nakai), *Rosa rugosa* Thunb. and *Solidago* spp. (*Solidago canadensis* L., *Solidago gigantea* Aiton). The species related to the genus *Heracleum* and the associated images were removed from the detection analysis, due to too few images to obtain statistically significant results for this taxa (see Table 1). For the sake of clarity, we will colloquially term these selected categories as "invasive species" for the remainder of this paper. Dyrmann et al. (2021) insured that all the images underwent meticulous expert review to determine the presence/absence, as well as the identification, of the six considered invasive species.

The images show invasive plants at different stages of development (e.g., young plants, flowering, fruiting, or dying with a characteristic brownish-dry appearance), at different sizes, and at different distances from the camera. The plant may exhibit different shapes depending on the camera angle, appearing bushy (such as *Reynoutria* spp. and *R. rugosa*), occupying a significant portion of the image, or like *P. sativa*, with only the inflorescence emerging from the herbaceous cover.

Table 1
Number of pictures per species: this table provides the number of positively annotated pictures, for each species.

Invasive species	# images
Cytisus scoparius (L.) Link	1228
Heracleum spp.	25
Lupinus polyphyllus Lindl.	669
Pastinaca sativa L.	1031
Reynoutria spp.	716
Rosa rugosa Thunb.	1725
Solidago spp.	3201

Table 2
Statistics of the Danish road dataset: number of images containing zero, one, two or three invasive species. The 206 multi-species images, containing two or three species, are used to evaluate the multi-label detection capability of the models.

# invasive species	# images
0	6467
1	8135
2	205
3	1

From the initial dataset of Dyrmann et al. (2021), we used 14,808 high-resolution images for our study that maintained a consistent resolution of 4024 × 3036 pixels. 43.67% of the images do not feature any invasive species, whereas only a small percentage of images exhibit the presence of multiple invasive species (see Table 2). Note that while Dyrmann et al. (2021) focused on images with zero or only one species present, we incorporated 206 multi-species images (205 containing two species and one containing three species) for a specific evaluation experiment (see below). We will call hereafter this dataset the Danish road dataset.

Fig. 1 illustrates, for each species in the dataset, an example that highlights the significant gap between the training images of the Pl@ntNet model and the ones of the Danish road dataset. Pl@ntNet relies mainly on detailed, close-up images of individual plants, often focused on specific organs such as leaves, flowers, fruits, or bark. In contrast, the target images in the Danish road dataset are not only much larger but also feature wide views with numerous plants and are captured using high-throughput professional cameras. This setup often results in "tilted" image distortion, further reinforcing the domain shift.

Some of the image distortion originates from the sensor's rolling shutter effect and warrants attention. For instance, plants and trees often appear to lean to the right, despite being perfectly vertical in reality.

3.1.2. Dataset split

For the main experiments, we adopted the evaluation methodology applied by Dyrmann et al. (2021) to the Danish road dataset, enabling a direct comparison with their results. Specifically, we split the data into three subsets – hereafter 'train' (70%), 'val' (15%), and 'test'(15%) – as commonly done in machine learning, respectively for model training, validation and selection of the best model version, and testing the final method generalization on unseen data. The exact same images were used in each subset for this study and Dyrmann et al. (2021) to allow a direct comparison of the results. Note that, despite the benefits of a K-fold or bootstrap procedure for a more robust model validation, we avoided such strategies given the computational burden of training or fine-tuning the models (see below). In the three data subsets, the images contain either zero or only one invasive species. To prevent data leakage between train and test due to spatial proximity of the views, Dyrmann et al. (2021) clustered the images by acquisition location when less than 40 m apart, so that all images from the same cluster were assigned to the same set. They iteratively allocated the clusters based on a χ^2 goodness-of-fit test to maintain class distribution consistency. This location-aware clustering approach provides an unbiased evaluation while preserving a representative species distribution across all subsets. Further details regarding the data splitting process can be found in Dyrmann et al. (2021).

We set aside 206 multi-label images, each containing at least two invasive species (205 containing 2 species and 1 containing 3 species), from the total dataset of 14.8K images to serve as an additional test set for evaluating the efficiency of multi-species detection.

3.2. Proposed approaches

We investigate whether a high performance plant identification model pre-trained to predict a single label (species) per image, i.e. a classification model, can be efficiently adapted to a multi-species presence/absence detection problem without any additional fine-tuning on the final task, therefore without further data annotation, and what performance gain brings the fine-tuning of such a model.

We first consider to use a general plant classification model pre-trained on images of individual plants (a Pl@ntNet model) without fine-tuning it on the Danish road images. However, state-of-the-art image classification models, i.e. vision transformers, typically require input images of relatively small size (e.g. 384 × 384 pixels for BEiT, used for Pl@ntNet, see below) due to memory and computational power limitations. The usual pre-processing step for image classification involves resizing the image by reducing its resolution and possibly cropping it to the model's square input size. In our context, this approach would significantly reduce the performance when dealing with high-resolution images containing many plants, because the relevant visual information is most often too small to be retained after resizing. Furthermore, even if we adapt the Pl@ntNet model to the high-resolution input images, preserving all their visual information, we expect that the predicted probability per species for a given image will decrease with a larger the number of species detectable in this image, due to the model's monolabel training. This makes it difficult to interpret the predictions in terms of presence and absence.

Therefore, we propose two approaches to adapt the Pl@ntNet model to high-resolution image analysis: (i) adapting the model to higher-resolution input images (Variable Model Input Size or "VaMIS" approach), and (ii) decomposing the input high-resolution images into smaller sub-images of suited size for the model (tiles), predicting on each tile and aggregating the tile-level predictions (Tiling approach).

Overall, the main challenges are to compare and select the best approach for handling high-resolution images and multi-species presence detection, and to evaluate out-of-the-box pre-trained procedures that have the best generalization ability.

3.2.1. Pl@ntNet pre-trained model

Pl@ntNet is a large-scale collaborative plant species recognition application, available on mobile devices and online via a web interface, used by over 20 million people annually, with the goal of progressively covering all the world flora of vascular plants (Affouard et al., 2017). It relies on a regularly updated deep learning model trained on a large volume of data continuously curated by the user community (nearly 30 million images shared and identified by end of 2024). The architecture of the model, the learning methods, and the methodological best practices are regularly re-evaluated and integrated in response to the PlantCLEF challenges organized each year (Joly et al., 2019). The results of the PlantCLEF2022 challenge (Goëau et al., 2022) led to a major update of the Pl@ntNet model. This challenge advanced the state-of-the-art available datasets in terms of data volume and species diversity by introducing a plant identification task based on four million images covering 80,000 species. The results showed that approaches based on Vision Transformers (ViTs) architectures (Dosovitskiy et al., 2020) have supplanted CNN-based approaches, particularly when pre-trained using a Self-Supervised Learning (SSL) method (Xu et al., 2022). The Bidirectional Encoder representation from image Transformer (BEiT) self-supervised learning method (Bao et al., 2021) demonstrated an excellent balance between performance and memory

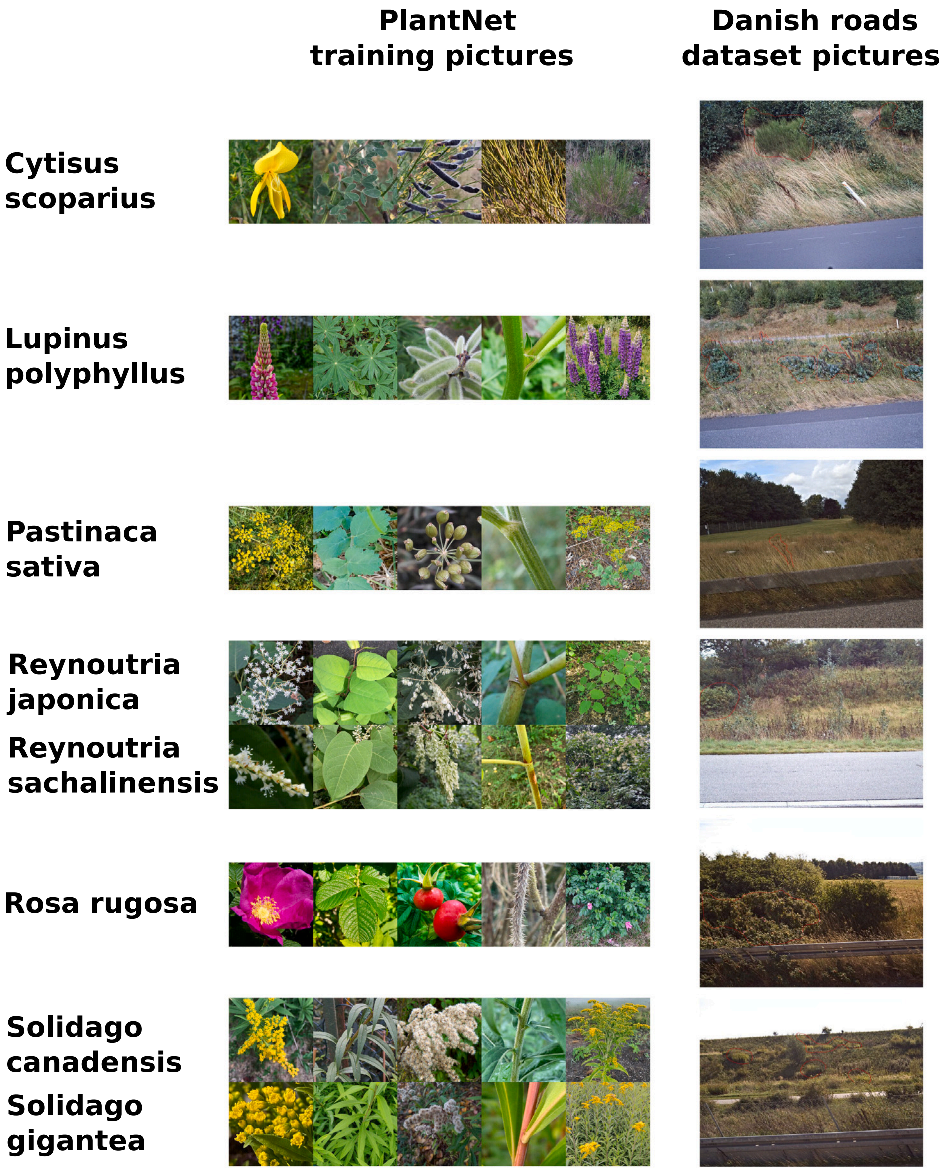

Fig. 1. Domain shift between the Pl@ntNet training dataset (left) and the Danish roads dataset (right). The Pl@ntNet model was trained to identify a single species per image, with multiple view types (flower, fruit, leaf, stem, whole plant), whereas the Danish roads dataset requires the recognition of multiple species within each image, using high-resolution analysis at varying camera-to-plant distances. Each image in the right-hand column contains at least one specimen of the species shown in the left-hand column.

usage. Following PlantCLEF2022, we trained a Pl@ntNet model based on the BEiT architecture. We used this specific pre-trained Pl@ntNet model in this study and detail below its training process.

BEiT architecture processes images with an input resolution of 384 × 384 pixels and an internal patch size of 16 × 16 pixels. We used a particular implementation of BEiT which underwent multi-stage trainings and that is publicly available.[1] To briefly describe this process, BEiT was first pre-trained from scratch without labels on ImageNet-22k, then fine-tuned on ImageNet-22k with supervision this time by adding a classification layer related to 22k classes. Finally, it was fine-tuned again on ImageNet-1k through transfer learning on 1k classes. For the remainder of the paper, we will refer to it as the off-the-shelf model.

This off-the-shelf BEiT model was downloaded and further fine-tuned on the Pl@ntNet training dataset. The Pl@ntNet training and validation data contained 6,585,369 images related to a total of 43,683 species. These images comprised the most validated Pl@ntNet observations, supplemented with public images from the GBIF (Wheeler, 2004) and private images of rare species or common species rarely imaged shared by collaborators and Pl@ntNet users. A relatively small subset of 45,000 images, covering 15,000 species, was reserved for validation, with the remaining images used for training. To prevent bias in species representation, no species-level balancing was enforced, and training was conducted with the natural class distribution observed in the dataset. However, to limit over-representation of common species, the number of images per species was capped at 800, while some rare species may be represented by a single image.

The training was carried out as distributed training on a high-performance computing cluster. The cluster comprised eight nodes, each equipped with four NVIDIA V100 GPUs (32 GB). Training leveraged the following hyperparameters: a batch size of 52 per GPU, an initial learning rate of 0.1625, and the SGD optimizer with a weight

[1] https://huggingface.co/timm/beit_base_patch16_384.in22k_ft_in22k_in1k.

decay of 10^{-4}. The learning rate was empirically chosen based on preliminary experiments on a validation subset, using a grid search. The selected value provided the best trade-off between convergence speed and generalization performance.

To make the model predictions robust to various transformations of the images, we employed several data augmentation techniques, including RandAugment (Cubuk et al., 2019), CutMix (Yun et al., 2019), MixUp (Zhang et al., 2018) and label smoothing. See[2] for the parameters values. These data augmentation techniques notably include shear transformations which address the rolling-shutter distortion present in the Danish road dataset. The training schedule utilized a learning rate plateau scheduler with a decay rate of 0.9 and a patience of one epoch, as well as 100 total epochs of training.

Training took 96 h. The final model was chosen based on the highest top-1 accuracy on the validation set, combined with a manual verification of species misclassifications for rare taxa. This training procedure, conducted prior to this paper, results in a generic model for plant identification, which we refer to as the Pl@ntNet model. More precisely, this is a candidate model trained in mid-2023 as part of the Pl@ntNet platform's ongoing update cycle. Although it was briefly evaluated, it was not deployed in production. For reproducibility purposes, we provide access to this model along with a classification head specifically trained on the invasive species considered in this study, and refer to it as the Pl@ntNet model throughout this paper for clarity.

3.2.2. *VaMIS approach: Adapting Pl@ntNet pre-trained model to larger image sizes*

Our pre-trained Pl@ntNet model is designed to take input images of 384 × 384 pixels. To leverage both the high-resolution of Danish road images and the pre-training of Pl@ntNet, a first option considered here is to adapt the Pl@ntNet pre-trained model to larger input images without further training. Through interpolation, it is possible to expand the receptive field of a pre-trained transformer and utilize rectangular input images. Upon loading the model, the BEiT Relative Positional Encoding (RPE) parameters undergo spatial interpolation to adjust for the larger image resolution. RPE facilitates the model's understanding of spatial relationships between different regions of the input image, thereby enhancing its ability to capture contextual information. Instead of using a uniform interpolation grid, we apply a non-uniform one with finer steps for nearby patches, which improves the encoding of short-range spatial relationships. This strategy provides more accurate interpolation for local details, important for visual tasks that rely on high-resolution, while remaining efficient for broader spatial relationships. We call this interpolation method "VaMIS", for Variable Model Input Size.

The RPE corresponds to 318k parameters compared to the 87 million of the entire BEiT transformer. During the VaMIS interpolation, the other parameters of the model remain the same. This is because, aside from the positional encoding mechanism, the computational procedure of a transformer model naturally scales up with the number of tokens. Specifically, the feed-forward block processes one token at a time, and the attention module operates on pairs of tokens, regardless of the total number of tokens. During inference, the RPE is added to the result of the attention matrix product to account for the spatial relative positions between tokens.

Even with the VaMIS approach, we did not use the full resolution of the original Danish road images of 4024 × 3036 pixels and resized them to 1024 × 768 pixels before feeding them to the Pl@ntNet model. This resolution was chosen for several reasons: it is one of the resolutions tested in Dyrmann et al. (2021), as well as one of the scales of the tiling method (scale two) that we experimented with and it was the highest resolution compatible with our GPU memory constraints. Further details can be found in the Appendix section titled 'VaMIS Interpolation'

The final output of the Pl@ntNet model, originally trained to identify single species, is a probability distribution across species. We removed the SoftMax layer, which transforms individual species scores ($\in \mathbb{R}$, called logits hereafter) into probabilities, and directly use the logits of the invasive species as the basis to predict their presence or absence.

The advantage of this training-free model is that other potential invasive species that have not been annotated could potentially be detected, even though we cannot test them here. In addition, this model can be used almost "out-of-the-box", as it does not require the technical AI skills or computational resources needed to customize or fine-tune a deep learning model to a multi-label classification dataset, such as the Danish road dataset used here.

3.2.3. *Tiling approach: Multi-scaled sampling*

While the VaMIS approach allows the Pl@ntNet model to be applied to larger images and reduces the need to downsize the original image (4024 × 3036 pixels), VaMIS still needed a drastic downsampling of the images to 1024 × 768 pixels due to a GPU RAM memory limit of 48 Gb. Another important limit of the VaMIS approach is that the image often contains many other plant species, which should increase the prediction ambiguity of Pl@ntNet, and it does not address the domain-shift problem (the Pl@ntNet pre-trained model was not exposed to comparable images in training). An alternative approach is to analyze the image "piece-by-piece", with a closer look, as if we were using a magnifying glass (Akyon et al., 2022). This is the principle of the tiling approach described below.

The tiling approach involves extracting square tiles at various scales from original high-resolution image. The tiles are resized to the Pl@ntNet model input size (384 × 384 pixels for BEiT) and a prediction is computed for each tile. We have defined eight tile scales, where the finest scale (scale eight) just extracts tiles from the original image resolution without any downsampling, and the coarsest scale resolution (scale one) extracts only two "big" tiles jointly covering the whole rectangular image, i.e. one square on the left and one on the right with an overlap area in the center of the image (see Fig. 2).

More precisely, scale one means to resize both tiles of size 3036 × 3036 to a 384 × 384 image. The last scale, number eight, uses the original resolution of the image as 8 * 384 = 3072, above 3036, which is the pixel-wise size of the smaller side of the images. For scales two to eight, we decided to keep an overlap of about 50 percent between two consecutive tiles, both on the x-axis and on the y-axis. This avoids that a highly informative area (e.g. a flowering individual) is split between two tiles during the process for a given scale. Thus, most pixels appear in four tiles for a given scale (two to eight). This results in 926 tiles for an initial resolution of 4024 × 3036 pixels, and a standard model input size of 384 × 384 pixels (Fig. 3).

The different scales allow the model to see different parts of the original image with a more or less focused view, representing different distances from the camera to the plant. This is particularly important because the photos were taken at different distances from the roadside, with different camera fields of view, and with plants at different stages of growth. Also, the Pl@ntNet training data contains images of plants at different distances. Thus, we expect some scales to naturally fit better for each image and each plant, allowing for some flexibility and robustness in the final predictor.

The tile-level predictions of the Pl@ntNet model notably output one logit per species, seen as a vector of presence scores. We sum the logits associated to the Latin species associated to each of our six invasive species to form their tile-level score. For each invasive species, we then took the maximum of the tile-level scores across all tiles of all scales to obtain the image-level score of that invasive species. This can be

[2] TIMM Parameters values: Randaugment: `--aa rand-m9-mstd0.5-inc1`, MixUp: `--mixup 0.8`, CutMix: `--cutmix 1.0`, Label smoothing: `--smoothing 0.1`.

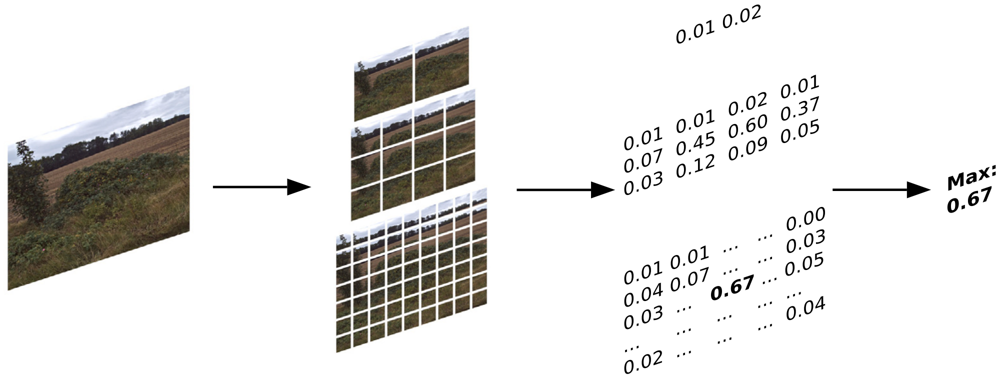

Fig. 2. Our tiling architecture. Example of *Rosa rugosa* Thunb. presence detection in a roadside image. Tiles are extracted from the high-resolution image, at eight different scales, and with 50% overlap. The BEiT Pl@ntNet model is used for inference and computes logits for each species. A max-pooling layer is then applied over all tiles to obtain the image-level logit. This method is a generic extension of a mono-label Standard Definition (SD) resolution input model to multi-label presence detection, High Resolution (HR) resolution image analysis.

interpreted as a common max pooling layer, as used in CNNs (see, for example, Simonyan and Zisserman, 2014).

We tested internally several other ways of aggregating the tile-level predictions to obtain image-level presence scores, for a given high-resolution image. The 926 tile sorted scores of a given meta-species can be viewed as an empirical distribution, and we can use a statistical parameter as an image-level presence score, such as the average, the maximum, or a quantile of this distribution. To avoid one high tile score to be mixed with the random noise of the other 925 tiles, we used the maximum tile score for each species at the image level. The quantile statistic is an alternative aggregation score, which could add robustness to the prediction relative to the maximum, but according to our internal tests, it did not perform better (50 percent, 90 percent and 99 percent quantile levels were tested). Another solution would be to use an additional linear layer covering all six species altogether, but this solution requires fine-tuning, and our experiments led to overfitting during the training, indicating too many aggregation parameters.

In summary, the tiling pipeline (see Fig. 4) consists of the following steps: (1) extract 926 tiles from the high-resolution image and resize them to the model input resolution, (2) compute deep features for each tile via inference with the BEiT Pl@ntNet model, (3) compute the six invasive species scores from the linear classifier of the Pl@ntNet model, and (4) aggregate tile-level predictions up to the image level by taking the maximum score over the tile per invasive species.

This tiling approach has several interesting properties, such as invariance through spatial translation in the image plane (a *R. rugosa* appearing in a tile in the upper left or the center of the image will be detected with identical statistics). There is also an approximate invariance by translation in the third dimension (plant-to-camera distance), because the multiscale tiling system ensures that at least one tile captures an entire individual at its best resolution. Finally, this approach is multi-species by design: after the shared model deep features inference, each species has a unique and dedicated presence/absence detection pipeline.

3.2.4. *Fine-tuning the models on the Danish road dataset (optional)*

Although the main focus of this paper is to capitalize on a pre-trained plant identification model, we also evaluate the scenario of fine-tuning our models on the train set of the Danish road images, for the purpose of comparison and discussion.

We aim to address a presence/absence detection problem, which is very similar to a multi-label classification problem. Therefore, we train our models using the recommended methodology in such case, i.e. a combination of a sigmoid layer and the binary cross-entropy loss.

The procedure is the following for the two approaches:

Table 3
Tiling of a 4024 × 3036 high-resolution image: distribution of the tiles across the eight scales. For scale one, two large 3036 × 3036 square tiles are extracted (left and right) and resized to 384 × 384 pixels. Scale eight uses 300 tiles (with layout 20 × 15), retrieved with native resolution of 384 × 384 pixels (no resizing) and about 50% overlap on both axes.

Scale	Resized image	n tiles X	n tiles Y	n tiles scale
1	384	2	1	2
2	768	4	3	12
3	1152	7	5	35
4	1536	10	7	70
5	1920	12	9	108
6	2304	15	11	165
7	2688	18	13	234
8	3036	20	15	300
All	–	–	–	926

Fine-tuning using the VaMIS approach. We aim to evaluate the performance improvement of the additional training on the Danish road by fine-tuning the previously introduced VaMIS model. The Pl@ntNet model, with VaMIS adaptation, uses higher resolution input images (1024 × 768 pixels input resolution). The performance of this out-of-the-box pre-trained model can be improved by fine-tuning on the Danish road dataset.

About the related training parameters, see Table 4.

Fine-tuning using the tiling approach. To evaluate the possible performance gain by fine-tuning, we optimize the presence/absence linear classification for each species, which takes 768 inputs (the deep features) and outputs a single logit per species, i.e. we optimize 4614 parameters in total. This layer is applied uniformly to all of the 926 tiles, thereby preserving the spatial translation invariance property.

The computation of the transformer deep features represents a bottleneck for both inference and training due to the large number of tiles for each image. Therefore, this computation is done once, as a pre-processing of the dataset, to efficiently fine-tune the model head.

The detection pipeline is unique for each species. Thus, during the training, these species-specific pipelines were optimized separately and then concatenated to obtain a model adapted to the Danish road dataset presented above, confirming the multi-species detection capabilities of the model by design.

3.3. *Experimental plan*

3.3.1. *VaMIS approach*

The VaMIS approach was tested both with and without fine-tuning on the Danish road dataset. These two experiments enabled us to measure the performance increase resulting from fine-tuning, compared

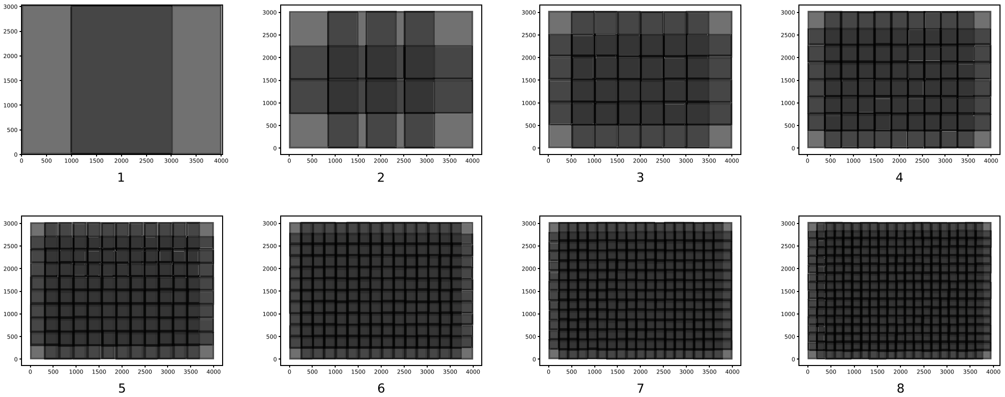

Fig. 3. Layout per scale: the receptive fields are plotted for each of the eight scales. Darker gray indicates overlapping tiles. This is directly related to Table 3, with the number of tiles per scale. Scale one uses two square tiles (left and right). For scale two, 12 tiles are extracted, with layout 4 × 3. Scale eight uses 300 tiles with layout 20 × 15.

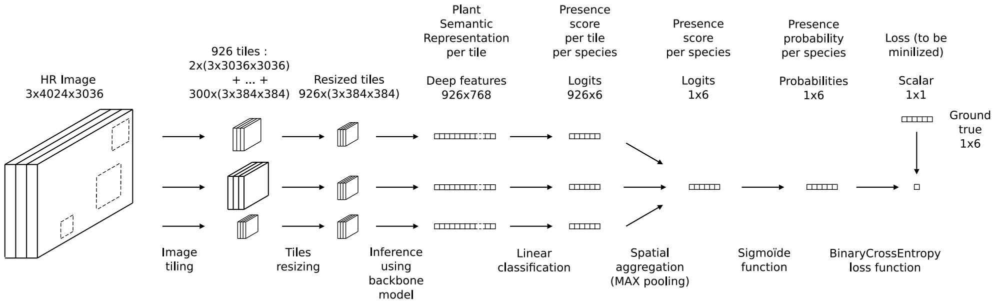

Fig. 4. Data flow processing for the tiling approach for an image: first, we extract 926 tiles from the original high-resolution image, at eight different scales (2 to 300 tiles per scale). On each resized tile, we apply Pl@ntNet's backbone and linear classifier. We get a presence score for each species in each tile. Spatial aggregation is done with a max-pooling layer that takes the best score over the tiles up to the image level. During fine-tuning, we add a sigmoid layer and use Binary Cross Entropy loss, and only optimize the linear classifier weights of the six invasive species. Thus, the computationally intensive deep feature extraction is performed only once, prior to training.

Table 4
Summary of the optimization parameters for the fine-tuning of the two models (VaMIS and tiling) on the Danish road dataset. The fine-tuning of VaMIS optimizes a full vision transformer with 87 million parameters and requires a server dedicated to deep learning. However, the fine-tuning of the tiling approach only requires optimizing the linear classifier of the head and could be performed on a standard laptop. Both models optimize the binary cross-entropy loss, which is added on top of a sigmoid layer, as commonly recommended for multi-label classification. Standard data augmentations were used (including rotate, sheer, and colorjitter). We did not apply any data augmentation to the tiling approach, since the model takes the deep features directly as input. ReduceLROnPlateau refers to the usual Reduce Learning Rate On Plateau strategy. We provide the epoch duration as additional information.

	VaMIS	Tiling
Optimized parameters	87 M	4614
# train image	10,302	10,302
# validation image	2160	2160
Computer GPU RAM	48 Gb	8 Gb
Batch size	8	256
Data augmentations	Standard	None
Learning rate	0.00001	0.02
LR decay	Cosine	ReduceLROnPlateau
Optimizer	ADAM	ADAM
Loss	Binary cross entropy	Binary cross entropy
Epoch duration	48 min	11 s

to the Task-independent VaMIS method. These models are labeled XP2 and XP3 in Table 5, respectively.

Furthermore, to specifically assess the effect of the pre-training of BEiT on Pl@ntNet images (yielding the Pl@ntNet pre-trained model), we tested VaMIS directly on the off-the-shelf BEiT transformer model, that had only been pre-trained on ImageNet but not on Pl@ntNet images. The linear classification head was initialized with random weights and outputs six logits, corresponding to the six invasive species. This model was fine-tuned using the same set of parameters. This model is labeled XP1 in Table 5.

3.3.2. Tiling approach

We also evaluated the tiling approach with and without fine-tuning on the Danish road dataset. In the experiment without fine-tuning on the training data, we performed inference on each tile using the pre-trained BEiT Pl@ntNet model, followed by a simple max function over the tiles. This out-of-the-box pre-trained method does not need any fine-tuning on the final dataset.It retains its ability to generalize, making it well-suited for detecting a wide range of invasive species. In the experiment with fine-tuning, we use the same BEiT Pl@ntNet tiling model presented above, and fine-tuned the model head on the Danish road dataset. The tiling models are labeled XP4 and XP5 in Table 5, corresponding to the models without and with fine-tuning, respectively.

3.3.3. Former ResNet and Yolo approaches

Dyrmann et al. (2021) investigated two approaches for detecting invasive species: image classification and object detection. Since our study builds upon the same dataset, we replicated these experiments to ensure a direct and meaningful comparison with our proposed methods. This replication allows us to assess how well our approaches – VaMIS and tiling – perform relative to existing CNN-based classification and object detection techniques.

In the image classification approach, the objective is to assign a category (i.e., select one species among the six or none) to each image in the test set. This was achieved using two well-established Convolutional Neural Network (CNN) architectures: ResNet50-v2 (He et al., 2016), known for its deep residual learning framework, and MobileNet-v2 (Sandler et al., 2019), optimized for efficiency on mobile and low-power devices. These models were trained and evaluated with

Table 5
Summary of the main experiments. The VaMIS model is an adaptation of BEiT to take 1024 × 768 images as input. The tiling approach consists of 926 square tiles on eight scales, extracted from the high-resolution input image, thus exploiting its true resolution. The models were pre-trained on either ImageNet only (IN), or ImageNet and then Pl@ntNet (PN). We tested both with and without fine-tuning. In the former one, 87 million of parameters are optimized in the VaMIS case, and 4614 for tiling (corresponding to the classification head). The multi-label experiment is not reported here.

	XP1	XP2	XP3	XP4	XP5
Type	VaMIS	VaMIS	VaMIS	Tiling	Tiling
Input (px)	1024 × 768	1024 × 768	1024 × 768	4024 × 3036	4024 × 3036
# Tiles	1	1	1	926	926
# Tiling scales	1	1	1	8	8
Pre-training	IN	IN + PN	IN + PN	IN + PN	IN + PN
Fine-tuning	Yes	Yes	No	No	Yes
Optimized parameters	87 M	87 M	–	–	4614

varying input resolutions, ranging from 128×96 to 2048×1536 pixels, to assess their performance across different scales.

To facilitate a comparison with Dyrmann's results, we reproduced the experiment using the ResNet50-v2 model at a resolution of 1024×768. This resolution was chosen as it offers a balance between computational efficiency and accuracy while aligning with the VaMIS configuration used in our study. Additionally, ResNet50-v2 was preferred over MobileNet-v2 due to its higher performance, making it a more relevant benchmark for evaluating Vision Transformers.

In the object detection approach, the goal is to localize and classify individual instances of species within images. To accomplish this, Dyrmann et al. (2021) employed YOLOv3 (You Only Look Once, version 3), a widely used object detection framework that balances speed and accuracy (Redmon and Farhadi, 2018). The model was trained on images with a resolution of 832 × 832 pixels and tested at 608 × 608 pixels, reflecting a trade-off between computational efficiency and spatial detail.

Since our objective is to compare the effectiveness of different presence/absence detection methods, we focused on reproducing the classification-based approach rather than the object detection approach. Object detection methods typically require extensive annotation efforts and are not directly comparable to our proposed tiling and VaMIS approaches, which operate at the image level rather than at the object instance level.

3.3.4. Summary of the experiments

We conducted a total of five experiments, summarized in Table 5, to compare the VaMIS and the tiling approaches. In addition, both approaches are tested with and without additional fine-tuning and we evaluated the multi-label approach. To facilitate comparison, we also reported the results from Dyrmann's image classification experiment, which employed ResNet50-v2 at a resolution of 768×1024. This allows us to assess how the VaMIS and tiling strategies compare to standard CNN-based classification and object detection approaches. Finally, we assess the effectiveness of presence/absence species detection when multiple species are simultaneously present for the five experimental models on the 206 multi-label images.

3.3.5. Evaluation metrics

Threshold selection. For all experiments, we used the logits as individual species presence scores (i.e., the output values of the neurons before applying SoftMax or Sigmoid layers). The threshold is chosen by optimization over the validation set and aims to maximize the balanced accuracy. The balanced accuracy is the average between the presence class and the absence class recalls, i.e. $\frac{1}{2}(\frac{TP}{P}+\frac{TN}{N})$. This statistic is more relevant than the standard accuracy statistic for unbalanced datasets. This can be interpreted graphically as selecting the threshold of the point on the ROC curve that is closest to the upper left corner of the graph, with coordinates (0,1), using Manhattan distance. This is the same as optimizing the Youden statistic (Youden, 1950), which is also calculated from the True Positive Rate and the False Positive Rate. Finally, once the threshold is chosen using the validation set, the model can decide, for each image, whether the species is present or not. We then evaluate the respective models on the test dataset.

Table 6
Result statistics on the test set (for 2140 images) for the five experiments. Fine-tuned models outperform the others (AUC). The non-fine-tuned methods, XP3 and XP4, benefit from Pl@ntNet pre-training. Tiling without fine-tuning has remarkably high performance (and generalization capabilities). Tiling and VaMIS approaches perform equally well.

	XP1	XP2	XP3	XP4	XP5
Type	VaMIS	VaMIS	VaMIS	Tiling	Tiling
Pretraining	IN	IN + PN	IN + PN	IN + PN	IN + PN
Fine-tuning	Yes	Yes	No	No	Yes
Bal. acc. (%)	89.91	91.82	66.76	84.87	**92.13**
AUC (%)	96.07	96.38	75.52	91.58	**97.29**

Metrics. In this article, we provide two metrics evaluated on the test set to assess the classification performance of the main experiments: the Area Under Curve (AUC) which is a global quality evaluation of a classifier and which does not rely on a threshold, and the balanced accuracy, seen above.

For the multi-label efficiency evaluation, we computed two statistics: the F1 score (calculated over the species for each image and then averaged over the dataset), and the Jaccard index, i.e. the so-called "intersection over union" (also calculated per plot and then averaged over the images): for each image, we divide the number of correctly detected species (presence) by the total number of species either detected or in the ground truth. We opted for multilabel metrics, such as F1 and Jaccard, over monolabel ones like AUC and balanced accuracy, to evaluate the models specifically for multilabel tasks using standard approaches tailored to such scenarios.

4. Results

Fig. 5 and Table 6 sum up the statistics of the experiments we previously introduced. The statistics per species for the tiling without fine-tuning are provided in Table 7. A closer look at the detection of *R. rugosa* are presented in Table 8. All but the VaMIS experiment without fine-tuning show exceptional performance, nearly or above 99%. We provide the ROC curve for *R. rugosa* detection, using the tiling without fine-tuning method (XP4) (see Fig. 6), confirming the efficient detection of this species. We evaluated the statistics over the multi-label dataset, composed of 206 pictures (each image contains at least two species within an image) in Table 9. The fine-tuned tiling method provides the best statistics.

The two approaches differ in both the inference and fine-tuning performance. On a standard laptop with 8 Gb GPU RAM and Nvidia RTX A2000, a single inference takes 0.36 s with the VaMIS approach and 12.4 s with the tiling method, for a time ratio of 34.5 in favor of the VaMIS method (See Table 10).

Fine-tuning a vision transformer, especially with VaMIS interpolation, requires a computer with enough GPU RAM to store the model, tokens, and gradients during optimization. On the other hand, the tiling model leverages Pl@ntNet pretraining and its resulting efficient deep feature extractor. Only the linear classifier head is optimized, which explains why it can be trained much faster. (See Table 11).

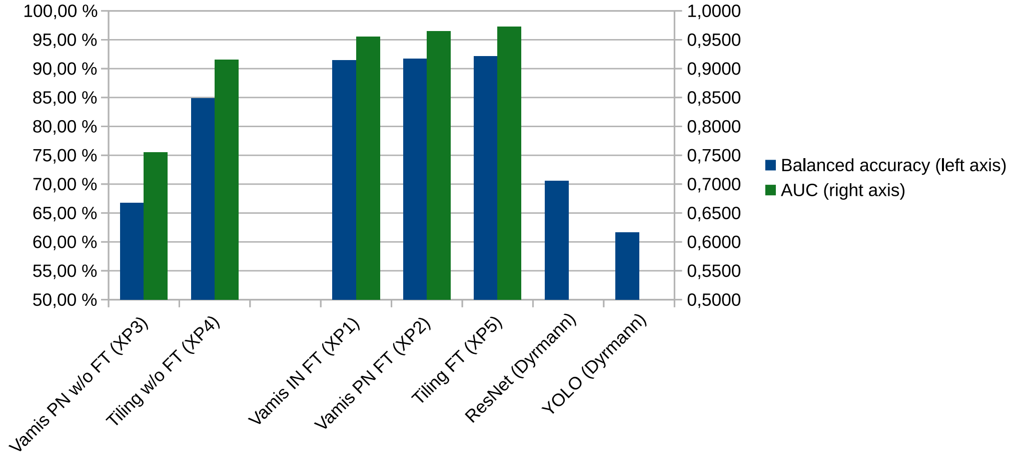

Fig. 5. Balanced accuracy and Area Under Curve (AUC) metrics, on the test set (for 2140 images). IN refers to ImageNet pre-training, and PN refers to Pl@ntNet pre-training (on top of ImageNet pre-training). On the left, we find the task-independent pre-trained models, that were not fine-tuned on the Danish road dataset. On the right, we compare the fine-tuned models. VaMIS stands for Variable Model Input Size (1024 × 768). Tiling takes the original 4024 × 3036 pixels image as input. ResNet and YOLO refer to the performance of models in Dyrmann et al. (2021), calculated from the confusion tables (which do not contain AUC information). ResNet is the standard ResNet50-v2 convolutional neural network model, with extended input size (1024 × 768), and YOLO refers to YOLOv3, the object detection model, with 608 × 608 pixels input size. The three fine-tuned vision transformers show comparatively strong performance, with an advantage for the tiling method.

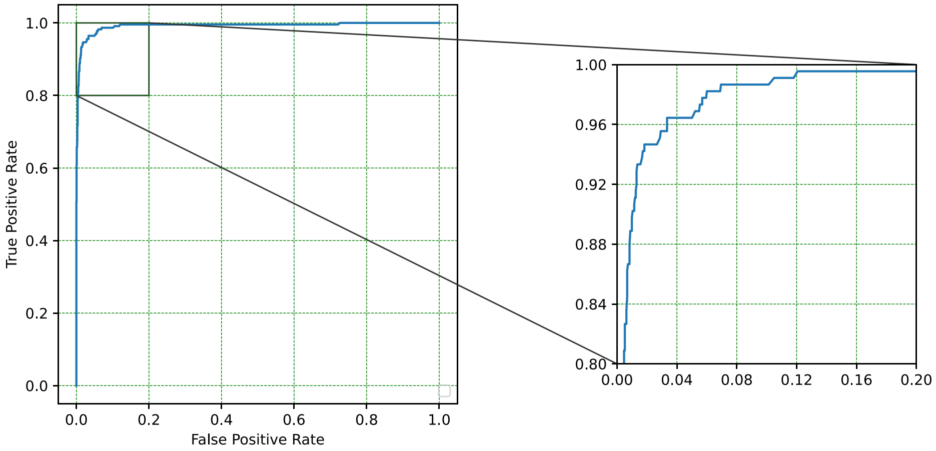

Fig. 6. XP4 — Tiling without fine-tuning, on the test set (for 2140 images): ROC curve for *Rosa rugosa* Thunb.. With 99.20% AUC (Area Under Curve), the non fine-tuned tiling is almost a perfect detector.

Table 7
Statistics for XP4 — Tiling without fine-tuning, on the test set (for 2140 images): P: Positive, N: Negative, TP: True Positive, FP: False Positive, TN: True Negative, FN: False Negative, bal.acc.: balanced accuracy, AUC : Area Under Curve. Weighted statistics use the number of positives P. Full invasive species names are *Solidago spp.*, *Cytisus scoparius* (L.) Link, *Rosa rugosa* Thunb., *Lupinus polyphyllus* Lindl., *Pastinaca sativa* L., *Reynoutria spp.*. The average AUC is above 91%, and the average balanced accuracy is nearly 85%. *Rosa rugosa* Thunb.'s AUC is 99.20%, the best among all the experiments.

	Soli.	Cyti.	Rosa	Lupi.	Past.	Reyn.	Weighted
P	451	161	225	66	124	118	
N	1689	1979	1915	2074	2016	2022	
TP	357	145	211	44	66	103	
FP	262	239	32	25	358	196	
TN	1427	1740	1883	2049	1658	1826	
FN	94	16	14	22	58	15	
Bal.acc. (%)	81.82	88.99	**96.05**	82.73	67.73	88.80	84.87
AUC (%)	89.84	95.33	**99.20**	95.99	75.31	93.24	91.58

Table 8
Rosa rugosa Thunb. presence detection statistics, on the test set (for 2140 images): Balanced Accuracy and Area Under Curve; All experiments (except for the VaMIS without fine-tuning) show exceptional detection for the *Rosa rugosa* Thunb. close to or above 99%.

	XP1	XP2	XP3	XP4	XP5
Type	VaMIS	VaMIS	VaMIS	Tiling	Tiling
Pretraining	IN	IN + PN	IN + PN	IN + PN	IN + PN
Fine-tuning	Yes	Yes	No	No	Yes
Bal. Acc. (%)	95.70	95.35	84.04	96.05	**96.07**
AUC (%)	99.16	99.09	91.26	**99.20**	98.99

5. Discussion

This study demonstrates the effectiveness of the two proposed approaches, tiling-based and VAMIS, in adapting a global plant image classification model, like Pl@ntNet's, for detecting species in high-resolution images. However, both approaches have their advantages and disadvantages in terms of data requirements and computational costs. Consequently, the choice of the best method depends on the context of use.

Table 9
Multi-label results (for 206 images): F1 is the F1 Score per image, averaged over the dataset. Jaccard index is the usual intersection over union statistic between predicted and ground truth present species. FT stands for fine-tuning. Scores are calculated on a separate dataset consisting of 206 images, each containing at least two species. The fine-tuned tiling provides the best statistics.

Experiment	XP	F1	Jaccard index
VaMIS ImageNet with FT	XP1	0.8414	0.7638
VaMIS Pl@ntNet with FT	XP2	0.7966	0.6958
VaMIS Pl@ntNet w/o FT	XP3	0.4815	0.3725
Tiling w/o FT	XP4	0.7301	0.6283
Tiling with FT	XP5	**0.8464**	**0.7642**

Table 10
Summary of the inference performance parameters for the two approaches (VaMIS and tiling) on the Danish road test set (2140 images). The VaMIS approach is, as expected, much faster (ratio × 34.5), since it requires only one “big” inference using a model input size of 1024 × 768 pixels with an internal representation of 3077 tokens. On the other hand, for a given high-resolution image, tiling relies on inference over 926 individual tiles, with a standard BEiT model input size of 384 × 384 pixels.

	VaMIS	Tiling
Test set inference time	12 m 50 s	7 h 22 m 20 s
# ViT tokens	3073	577
Batch size	4	128
Single image inference time	0.36 s	12.40 s
inference time ratio		× 34.5

Table 11
Summary of the fine-tuning performance parameters for the two approaches (VaMIS and tiling) on the Danish road dataset (10,302 images). As expected, fine-tuning the tiling model is much faster than fine-tuning the VaMIS model. For the VaMIS approach, a computer with 48 GB GPU RAM was utilized, with 44 GB being used continuously during training.

	VaMIS	Tiling
Optimized parameters	87 M	4614
Epoch duration	48 min	11 s
# epochs	31	112
Fine-tuning time	24 h 50 min	20 min 32 s
Batch size	8	256
Training time ratio	× 62	

5.1. Context 1: Known target species and availability of labeled images

When the set of species to be detected is known in advance and not too big, it may be possible to produce manually annotated data in the target domain at a reasonable cost. We then find ourselves in a classic context of supervised domain adaptation (Farahani et al., 2021), with annotated data in both the source and target domains. In this context, transfer learning is known to be an effective solution, especially when using large, pre-trained transformer models (Khan et al., 2022). Indeed, this makes it possible to fine-tune the model weights in a supervised way in order to improve performance. In our case, the balanced accuracy of the tiling approach can be increased from 84.87% to 92.13% thanks to fine-tuning. The performance gain is even stronger for the VAMIS approach, whose accuracy rises from 66.76% to 91.82% when fine-tuning the whole model.

Overall, the three fine-tuned models show comparatively strong performance, with an advantage for the tiling method on average (+0.31% on the accuracy and +0.91% on the AUC, compared to the Pl@ntNet pre-trained VaMIS). The tiling approach analyzes the image at its true resolution (4024 × 3036), while the VaMIS method requires a resizing preprocessing to 1024 × 768 pixels, thus losing some information. This may explain the difference in performance. Looking more closely at the VaMIS models, pre-training on a large plant dataset such as Pl@ntNet increases the accuracy by 1.91% and the AUC by 0.31%. In the particular case of multi-species detection (i.e. multi-label images), the fine-tuned methods provided the best detection performance. The ImageNet pre-trained VaMIS seems to be better than the Pl@ntNet pre-trained one. This may be attributed to the relatively small size of this dataset (206 images), which provides insights and general trends but does not allow for precise comparisons, as the limited number of images inherently impacts the statistical precision of the results calculated from it. As in the mono-label experiments, VaMIS Pl@ntNet without fine-tuning does not provide any good performance. Tiling without fine-tuning provides reasonably good performance (see the Jaccard index). The tiling model is a multi-label classifier by design, which may explain this result.

Regarding computational costs, it is worth noting that fine-tuning the tiling model involves optimizing only 4614 parameters (the linear classifier), which does not require powerful GPUs. In contrast, VaMIS has 87 million parameters to optimize, resulting in a learning time of nearly 20 min for the tiling model compared to 25 h for VaMIS (see Table 11). On the other side, the inference time and associated computational cost is much higher for the tiling model since it requires processing 926 tiles for a single high-resolution image. In a context of large-scale model deployment over a long period, it is therefore probably preferable to use the VAMIS approach, which offers a better tradeoff between quality and inference efficiency.

5.2. Context 2: Unknown target species and/or unavailability of labeled images

When the set of species to be detected is unknown in advance or when it is not possible to produce manually annotated data in the target domain at a reasonable cost, we find ourselves in a context of unsupervised domain adaptation (Liu et al., 2022), with annotated data in the source domain but not in the target domain. In this context, we note that the tiling approach provides the best performance, while maintaining very high generalization capabilities, with 84.87% balanced accuracy and 91.58% AUC. VaMIS without fine-tuning provides reduced performance, indicating that increasing the model input size of a vision transformer by relying solely on the interpolation of the spatial parameters (relative positional encoding) may not be sufficient.

The good performance of the out-of-the-box pre-trained tiling is due to the Pl@ntNet pre-training of the BEiT backbone model. In fact, tiling without fine-tuning can be seen as a direct extension of the Pl@ntNet model to high-resolution input images and multi-label classification. This adapted modeling approach, based on tiling and max-pooling aggregation, is quite generic and could be used in different contexts. It also requires only limited deep learning expertise to execute. Still, its drawback remains the computational inference time, since 926 inferences are required to analyze a single high-resolution image. In future work, we plan to experiment the design of a hybrid model between the tiling approach and the VAMIS approach. Such a hybrid model could provide a backbone that processes the image entirely like VAMIS, but with a classification head that enables tiling, similarly to object detection models (Shetty et al., 2021).

The growing development of automated species identification (Truong and Van der Wal, 2024), combined with citizen science platforms such as Pl@ntNet (Bonnet et al., 2020), iNaturalist (Di Cecco et al., 2021), Flora Incognita (Mäder et al., 2021) will undoubtedly make it possible to expand in the future the volume and diversity of visual data available for training deep learning models. This additional data could thus enable better coverage of the different growing and phenological stages of species, facilitating their detection by embedded sensors over longer periods (plant species are generally in flower or fruit only for short periods). In addition, the performances obtained within this work make it possible to consider the transfer of this methodology to analyze data produced from other types of device already used for monitoring invasive species, such as drones (Singh et al., 2024; Dash et al., 2019). This represents a tremendous opportunity for monitoring large areas and the early detection of invasive species in areas that are difficult to access by land.

6. Conclusion

In this work, we have presented a new methodology for analyzing high-resolution images with deep learning models. We have shown that tiling the image is a simple yet effective approach to take advantage of the entire area of the image. Moreover, this method is highly generic and enables transforming a medium resolution single-label classification model (one class per image) into a high-resolution multi-label classification model (multiple classes per image) without any additional training. The tiling model inherits the efficiency of the initial model and may even outperform other models that have been fine-tuned (e.g., Pl@ntNet tiling without fine-tuning on *R. rugosa*).

The VaMIS (Variable Model Input Size) model provides a natural extension of a transformer's receptive field, allowing the entire image to be analyzed at a higher resolution in a single inference. However, VaMIS requires a large amount of GPU RAM due to the quadratic complexity of the attention layer in the number of patches, which limits its effective resolution. Moreover, this method requires fine-tuning to achieve optimal classification performance. Since this is not possible in most contexts where additional labeled high-resolution images are unavailable, we recommend the tiling modeling approach based on the results presented in this study. This type of method could encourage the emergence of new protocols for monitoring species (Porcher et al., 2024).

Future work should focus on optimizing the tiling approach to minimize the computational resources required while maintaining good classification performance and generalization ability. In addition, we recommend to further analyze the presence of the detected species on the image surface. For instance, it may be possible to normalize the final score distribution using, for example, a layer norm. Due to its versatility, this approach could be easily integrated into a functional product without prior knowledge of the different usage contexts (e.g. different countries, acquisition protocols, target species, etc.). The approaches presented and evaluated in this paper contribute to scalable monitoring of plant populations using cost-effective recording and analysis methods.

Beyond performance, incorporating the results of this study with traditional observations of invasive species, such as those collected through citizen science, could enhance overall data quality and accuracy. The value of other AI methods for improving data quality has indeed, already been demonstrated in other contexts (Fraisl et al., 2024). Our model's predictions could eventually complete citizen science reports, contributing to future feedback loops aimed at refining both automated detection methods and human observations, while carefully avoiding the risk of reinforcing existing model biases. For instance, we could develop a collaborative platform where data from both sources are aggregated, enabling users to cross-reference automated predictions with verified citizen science data. Following recommendations of Ceccaroni et al. (2019) to ensure clarity and consistency in reporting, adopting standard metrics for accuracy that align with established protocols in the field is crucial. Additionally, engaging with citizen scientists through training workshops can help them leverage these findings, thereby enriching their observations and contributing to new and more robust ecological data.

CRediT authorship contribution statement

Vincent Espitalier: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Jean-Christophe Lombardo:** Software, Resources, Methodology, Investigation, Data curation, Conceptualization. **Hervé Goëau:** Writing – review & editing, Writing – original draft, Supervision, Resources, Methodology, Conceptualization. **Christophe Botella:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Conceptualization. **Toke Thomas Høye:** Writing – review & editing. **Mads Dyrmann:** Writing – review & editing, Data curation. **Pierre Bonnet:** Writing – review & editing, Supervision, Resources, Methodology, Investigation, Funding acquisition, Conceptualization. **Alexis Joly:** Writing – review & editing, Supervision, Resources, Methodology, Investigation, Funding acquisition, Conceptualization.

Statement

During the preparation of this work, the authors used large language model tools, such as ChatGPT from OpenAI, LeChat from Mistral AI and DeepL to improve language and readability.

After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The research described in this paper was funded by the European Commission through the MAMBO (Modern Approaches to the Monitoring of BiOdiversity) project and the GUARDEN (safeGUARDing biodivErsity aNd critical ecosystem services across sectors and scales) project. These projects received funding from the European Union's Horizon Europe research and innovation program under grant agreement 101060639 (start date: 01/09/2022; end date: 31/08/2026), and 101060693 (start date: 01/11/2022; end date: 31/10/2025) respectively. The authors declare no financial or personal conflicts of interest related to this work. Any ethical considerations were taken into account during the research process. The views and opinions expressed in this work are those of the authors only and do not necessarily reflect those of the GUARDEN or MAMBO partners, the European Commission. This project was provided with computer and storage resources by GENCI at IDRIS thanks to the grant 2024-A0171011389 on the supercomputer Jean Zay's the A100 partition.

Appendix A. VaMIS interpolation

A.1. Short introduction to transformers and vision transformers

Transformers (Vaswani et al., 2017) were introduced as an improvement over Convolutional Neural Network models (Hochreiter and Schmidhuber, 1997; Bahdanau et al., 2016) for language processing. The input text is projected into a sequence of tokens (for simplicity, assume one word corresponds to one token). A token represents the fundamental unit of data in transformers and is stored as a vector. A transformer relies on an attention mechanism, allowing the model to focus on specific tokens within the sequence. The attention layer processes intermediate representations of the token sequence (Queries, Keys, and Values, which are built with projections) and computes the so-called attention matrix, which is obtained by multiplying the queries and the keys. To enable the model to attend to multiple concepts simultaneously within a single attention layer, tokens are divided into equally sized vectors beforehand and processed by independent attention heads in parallel.

An attention layer is followed by a feed-forward layer (consisting of two successive, fully connected layers applied to each token), and together they constitute a transformer block. Repeating such blocks allows the model to build an abstract representation of the input text, which can then be utilized for various tasks, including text classification and translation.

Vision Transformers (Dosovitskiy et al., 2020) are a straightforward adaptation of transformers for image processing. An image is divided into small patches (e.g., 16 × 16 pixels), with each patch projected into a token. The resulting sequence of tokens is then processed using the transformer mechanism. Consequently, apart from the initial projection, the data processing for image analysis is fairly similar to that for text.

In addition to the attention mechanism described above, transformers usually rely on positional encoding to establish spatial relationships between tokens, thereby enabling an accurate understanding of their positions. In the Vision Transformer BEiT model, Relative Positional Encoding (RPE) provides this spatial awareness. This parameter constitutes the sole architectural limitation that prevents the model from handling arbitrary input sizes and is adapted through the VaMIS procedure.

Additionally, a special token, known as the CLS token, is inserted at the beginning of the token sequence and is independent of the input data. This token aggregates global information, serving as a cache memory that stores data at the image level. Unlike typical usage, the BEiT model does not use this token for classification. Instead, it computes the average of all other tokens (referred to as local tokens) and uses the CLS token to store intermediate global data, functioning as a register (Darcet et al., 2024). Therefore, the processing of the CLS token and its specific positional encoding is set aside, as it is not directly affected by the VaMIS procedure. To simplify the notation, further details on this aspect are omitted. Readers seeking additional information are encouraged to consult BEiT model (Bao et al., 2021) and its implementation (Wightman, 2019).

A.2. Towards BEiT's relative position encoding

In the original transformer architecture for language modeling (Vaswani et al., 2017), the authors introduced an absolute positional encoding derived from a deterministic scheme based on trigonometric functions. A fixed value, determined solely by the token's absolute position within the input sequence of words, is added to each token before applying the core mechanisms of the transformer blocks. Subsequently, a relative positional encoding was proposed by Shaw et al. (2018) to generalize transformers to variable input lengths. In this approach, a learned term is added to the projected key and value of the attention mechanism, with this additive bias depending only on the difference between the indices of the query and key tokens—that is, their relative position. As relative positional encoding requires explicit computation of the attention matrix, efforts have been made to simplify its implementation. For example, Stochastic Positional Encoding (Liutkus et al., 2021) was introduced to address this challenge. Additionally, a deterministic scheme called RoPE (Rotary Position Embedding) was proposed (Su et al., 2021) for relative positional encoding in the context of language modeling, incorporating the additional feature of decaying inter-token dependency with increasing relative distances. This approach was subsequently adapted for use in the EVA02 Vision Transformer (Fang et al., 2023).

Another version of relative positional encoding, referred to as T5 (Text-to-Text Transfer Transformer encoding), was introduced by Raffel et al. (2019). In this approach, a bias term is added to the attention matrix. Specifically, let this bias be denoted as $B = (b(i,j))_{i,j \leq N}$, where (i,j) represent the indices of the two local tokens under consideration, with N representing the total number of tokens. Let X denote the tokens after the preprocessing attention split, W_Q and W_K the respective weight matrices for the Query and Key projections, and d the dimensionality of the token embeddings. The formula for computing the attention matrix can be expressed as:

$$Att(X) = \text{Softmax}(\frac{XW_Q(XW_K)^T}{\sqrt{d}} + B)$$

with $b(i,j) = r_{min(i-j,K)}$, where r_i are learnable scalars and K is a hyperparameter that defines the positional shift range beyond which the positional encoding remains constant. The SoftMax function accepts a vector as input and is therefore applied to each row of the intermediate matrix. It is conventionally defined as follows:

$$\text{Softmax}(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}}$$

The BEiT model (Bao et al., 2021) employs a Relative Positional Encoding (RPE) adapted from T5's RPE for images. Consequently, it accounts for the relative positions of image patches along both the X and Y axes. The formula for BEiT's RPE is as follows:

$$B = b(i,j) = r_{x_j - x_i, y_j - y_i}$$

where (x_i, y_i) and (x_j, y_j) represent the respective spatial coordinates of the local tokens i and j under consideration. The $R = (r_{a,b})_{-N \leq a,b \leq N}$ matrix contains the positional encoding parameters (N represents the maximum positional difference between two tokens and depends on the image size.). In the context of image processing with the BEiT model, this matrix is referred to as the *Relative Position Bias Table*.

There are several differences compared to T5's RPE (Raffel et al., 2019). First, there is no K limit: all relative positions between tokens at the image level have a specific learned bias, without any restriction on maximum relative positions. Second, the bias table varies not only among the attention heads within a block of the Vision Transformer but also across blocks. In the original T5 version, all blocks share the same bias table, although it varies among the attention heads within each block. Finally, the positional encoding bias is represented as a matrix rather than a vector, as the data are structured in two dimensions (images) rather than as a linear sequence.

Dimension considerations: In this article, we use a BEiT model with a default input size of 384 × 384, and a patch size of 16 × 16, which matches a 'ViT-Base' size transformer (see Table 1 in Dosovitskiy et al. (2020)). An input image dimension of 384 × 384 and patch size of 16 × 16 result in a state space of $24 \times 24 = 576$ local tokens for the BEiT model.

The coordinates of a token (x, y) are bounded by $0 \leq x, y \leq 23$, and the relative positions of a pair of tokens, $(\delta_x = x_i - x_j, \delta_y = y_i - y_j)$, have the following boundaries: $-23 \leq \delta_x, \delta_y \leq 23$. Therefore, the initial Relative Position Bias Table R for BEiT has dimensions 47 × 47.

In this article, we analyze the performance of the BEiT model with a new resolution of 1024×768. With the same patch size of 16×16, this equates to $64 \times 48 = 3072$ local tokens. The relative position boundaries are $-63 \leq \delta_x \leq 63$ along the first axis and $-47 \leq \delta_y \leq 47$ along the second axis. Consequently, the new Relative Position Bias Table R has dimensions 127 × 95.

A.3. VaMIS interpolation

The VaMIS interpolation aims to increase the dimensions of the Relative Position Bias Table, thereby enabling an increase in input image resolution. This process can be viewed as a regridding of the spatial-wise bias.

The matrix R consists of parameters learned during the initial training phase using a gradient backpropagation algorithm (LeCun et al., 1989). As a result, it does not necessarily exhibit any inherent regularity or smoothness, making direct interpolation seemingly unsuitable. Nonetheless, the VaMIS procedure employs an interpolation method to increase the dimensions of R, allowing the Vision Transformer model to handle higher-resolution input images.

Central-wise Regridding: the goal is to select a regridding approach that prioritizes interpolation accuracy for nearby tokens, i.e., those with small relative positions, rather than for tokens with larger shifts. Nearby tokens are more likely to belong to the same object and characterize its structure, making it essential to preserve accurate relationships for short-range relative positions compared to long-range ones. This focus corresponds to the central terms of the bias table

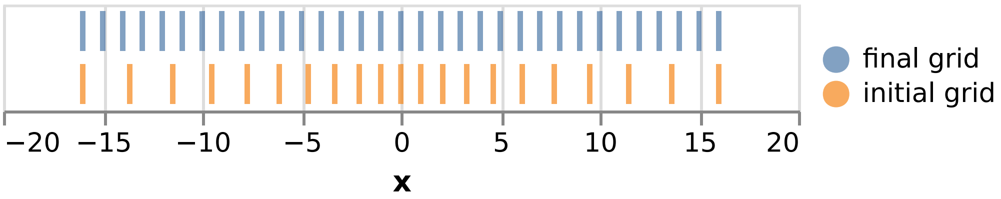

Fig. A.7. Example of Central-wise Regridding: increasing the number of points while ensuring that the central points remain aligned. In this example, the number of points increases from 21 to 33, with an initial stretched grid. For the purpose of clarity, we did not use the actual numbers.

R. Moreover, during inference, these central terms are used far more frequently than the side-positioned ones, as there are significantly more token pairs with small relative positions than with large relative positions. For example, 552 pairs of tokens share the relative position ($\delta_x = 1, \delta_y = 0$), whereas only a single pair of tokens shares the relative position ($\delta_x = 23, \delta_y = 23$). Therefore, maintaining the accuracy of these central terms is critical for the model's positional awareness and overall efficiency.

The BEiT implementation was made publicly available in 2021 (Bao et al., 2021), providing a solution to the previously mentioned issue. Subsequently, support for adaptation to arbitrary image sizes and aspect ratios was integrated into TIMM's BEiT implementation.[3] To increase the number of points in the R matrix while maintaining accuracy for short-range shifts, an unevenly spaced rectilinear grid is employed. This approach ensures that the points of the initial and final grids are aligned for $-1 \leq \delta_x, \delta_y \leq 1$ and approximately aligned near their center, thereby preserving an almost identical positional bias structure for short-range token pairs.

The VaMIS interpolation is based on a 2D regridding of the Relative Position Bias Table R and employs standard bilinear interpolation (Getreuer, 2011) to adapt the R matrix from the initial grid to the final one (see Fig. A.7).

The regridding of each dimension of the R matrix involves increasing the number of points from $2C + 1$ to $2D + 1$, with $C = 23$ and $D = 63$ or $D = 47$, depending on the dimension (Each dimension of the R matrix is an odd integer, reflecting the cardinality of relative positions along a single dimension: a symmetrical set that includes the zero value).

The initial grid is defined as rectilinear and unevenly spaced $(x_n^i)_{-C \leq n \leq C}$:

$$x_n^i = \begin{cases} \sum_{0 \leq j < -n} -r^j & \text{if } -C \leq n \leq -1 \\ 0 & \text{if } n = 0, \\ \sum_{0 \leq j < n} r^j & \text{if } 1 \leq n \leq C, \end{cases}$$

The final grid is rectilinear with fixed intervals of 1:

$$x_n^f = n \text{ for } -D \leq n \leq D \tag{1}$$

The central terms of both grids are $\{-1, 0, 1\}$, ensuring identical relative positional bias for neighboring tokens. The common ratio r of the geometric progression used to construct the initial grid is selected such that the extreme points of the two grids are aligned:

$$D = x_D^f = x_C^i = \sum_{0 \leq j < C} r^j = \frac{1 - r^C}{1 - r}, \text{ i.e. } \frac{1 - r^C}{1 - r} = D,$$

This formula is inverted through a dichotomy procedure in the BEiT implementation.

Appendix B. Tiling: Additional results

Table B.12
Result statistics on the test set (for 2140 images) for the five experiments. fine-tuned models outperform the others (AUC). The non-fine-tuned methods, XP3 and XP4, benefit from Pl@ntNet pre-training. Tiling without fine-tuning has remarkably high performance (and generalization capabilities). Tiling and VaMIS approaches perform equally well.

	XP1	XP2	XP3	XP4	XP5
Name	IN-VaMIS-FT	PN-VaMIS-FT	PN-VaMIS-noFT	PN-tiling-noFT	PN-tiling-FT
Fine-tuning	Yes	Yes	No	No	Yes
# params	87 M	87 M	–	–	4614
# epochs	36	31	–	–	137
Epoch	50 min	50 min	–	–	11 s
Bal. acc.	89.91	91.82	66.76	84.87	**92.13**
AUC	96.07	96.38	75.52	91.58	**97.29**

Table B.13
Statistics for XP5 - Fine-tuned tiling method, on the test set (for 2140 images): P: Positive, N: Negative, TP: True Positive, FP: False Positive, TN: True Negative, FN: False Negative, AUC: Area Under Curve, Bal. Acc.: Balanced Accuracy. Weighted statistics use the number of positives. The average AUC is above 97%, and the average accuracy is above 92%. *Rosa rugosa* Thunb.'s AUC is almost 99%.

	Soli.	Cyti.	Rosa	Lupi.	Past.	Reyn.	Weighted
P	451	161	225	66	124	118	
N	1689	1979	1915	2074	2016	2022	
TP	417	149	216	54	107	106	
FP	161	97	74	33	216	176	
TN	1528	1882	1841	2041	1800	1846	
FN	34	12	9	12	17	12	
Bal. acc. (%)	91.46	93.82	**96.07**	90.11	87.79	90.56	92.13
AUC (%)	97.14	98.17	**98.99**	98.34	93.60	96.67	97.29

Table B.14
Statistics for XP2 - fine-tuned VaMIS method, on the test set (for 2140 images): P: Positive, N: Negative, TP: True Positive, FP: False Positive, TN: True Negative, FN: False Negative, AUC: Area Under Curve, Bal. Acc.: Balanced Accuracy. Weighted statistics use the number of positives. The average AUC is above 96%, and the average accuracy is near 92%. *Rosa rugosa* Thunb.'s AUC is above 99%.

	Soli.	Cyti.	Rosa	Lupi.	Past.	Reyn.	Weighted
P	451	161	225	66	124	118	
N	1689	1979	1915	2074	2016	2022	
TP	390	148	213	58	92	118	
FP	87	29	76	73	210	198	
TN	1602	1950	1839	2001	1806	1824	
FN	61	13	12	8	32	0	
Bal. acc. (%)	90.66	95.23	**95.35**	92.18	81.89	95.10	91.82
AUC (%)	97.29	98.48	**99.09**	95.70	83.56	98.73	96.38

For a better understanding of the tiling approach without fine-tuning, we provide additional experimental results here. Table B.12 provides the main statistics for the five experiments, including additional quantitative information related to the fine-tuning process. Tables B.13 and B.14 show the statistics per species for the two best methods (see Fig. B.8).

[3] https://github.com/huggingface/pytorch-image-models/blob/47811bc05a2fdff2dedbbb8b8b3a4b9e8dba4bb3/timm/layers/pos_embed_rel.py#L193-L265.

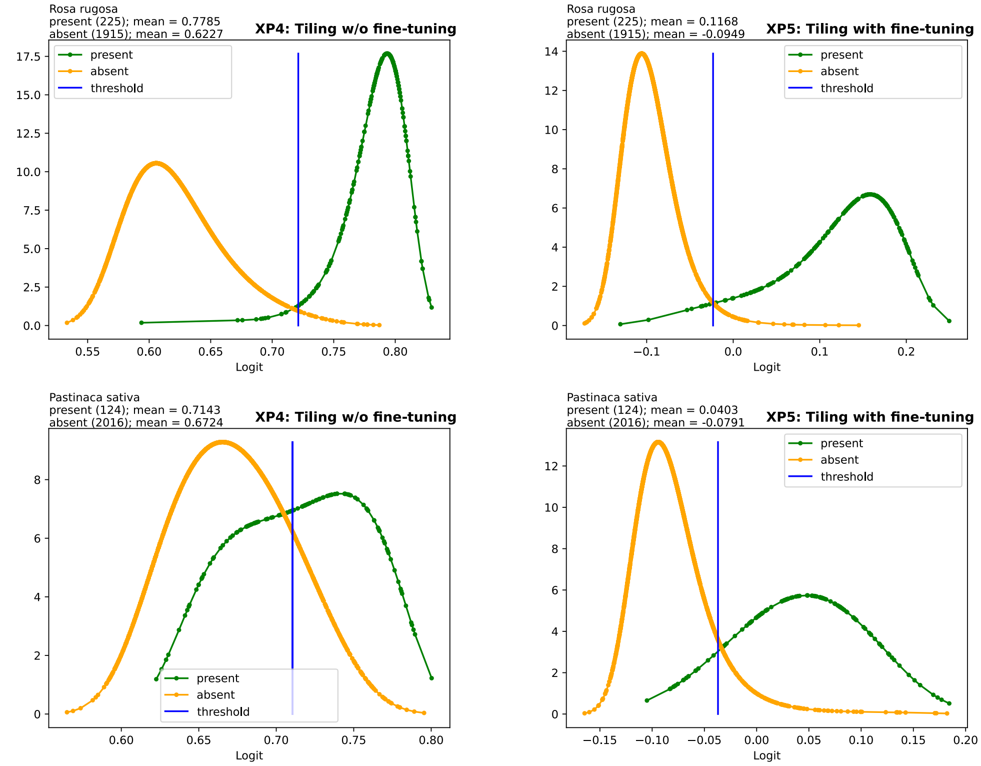

Fig. B.8. Presence and absence logit distributions, for *Rosa rugosa* Thunb. and *Pastinaca sativa* L., for the two tiling experiments (fine-tuned and not fine-tuned): Presence scores (also known as logits) were computed for all images in the test set, for these two species. Positive and negative image scores were separated (according to ground truth). The curves were smoothed by interpolation. The detection threshold, calculated by maximizing balanced accuracy over the validation set, is shown. We see the effect of the fine-tuning for the *Pastinaca sativa* L.: as expected, the fine-tuned tiling discriminates this species more efficiently than the non fine-tuned version. It is less clear for the *Rosa rugosa* Thunb., as the tiling model without fine-tuning is already an effective presence detector for this species.

Data availability

The images and the metadata used for the experiments are available in a package on Zenodo at https://zenodo.org/record/14013930 (DOI 10.5281/ zenodo.14013930 accessed in October 2024).

Pre-trained models used for the experiments are available in a packageon Zenodo at https://zenodo.org/record/13891416 (DOI 10.5281/ zen-odo.13891416 accessed in October 2024).

The code used for the experiments are available in a repository on github at https://github.com/plantnet/roadside-invasive-plant-identification.

References

Affouard, A., Goëau, H., Bonnet, P., Lombardo, J.-C., Joly, A., 2017. Pl@ntnet app in the era of deep learning. In: ICLR: International Conference on Learning Representations.

Akyon, F.C., Altinuc, S.O., Temizel, A., 2022. Slicing aided hyper inference and fine-tuning for small object detection. 2022 IEEE Int. Conf. Image Process. (ICIP) 966–970. http://dx.doi.org/10.1109/ICIP46576.2022.9897990.

Bahdanau, D., Cho, K., Bengio, Y., 2016. Neural machine translation by jointly learning to align and translate. URL https://arxiv.org/abs/1409.0473.

Bao, H., Dong, L., Wei, F., 2021. Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254 https://github.com/microsoft/unilm/tree/master/beit.

Bonnet, P., Joly, A., Faton, J.-M., Brown, S., Kimiti, D., Deneu, B., Servajean, M., Affouard, A., Lombardo, J.-C., Mary, L., et al., 2020. How citizen scientists contribute to monitor protected areas thanks to automatic plant identification tools. Ecol. Solut. Evid. 1 (2), e12023.

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.

Ceccaroni, L., Bibby, J., Roger, E., Flemons, P., Michael, K., Fagan, L., Oliver, J.L., 2019. Opportunities and risks for citizen science in the age of artificial intelligence. Citiz. Sci.: Theory Pr. 4 (1), Article–number.

Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V., 2019. RandAugment: Practical automated data augmentation with a reduced search space. URL https://arxiv.org/abs/1909.13719.

Darcet, T., Oquab, M., Mairal, J., Bojanowski, P., 2024. Vision transformers need registers.

Dash, J.P., Watt, M.S., Paul, T.S., Morgenroth, J., Hartley, R., 2019. Taking a closer look at invasive alien plant research: A review of the current state, opportunities, and future directions for UAVs. Methods Ecol. Evol. 10 (12), 2020–2033.

Daume, S., 2016. Mining Twitter to monitor invasive alien species — An analytical framework and sample information topologies. Ecol. Inform. 31, 70–82. http://dx.doi.org/10.1016/j.ecoinf.2015.11.014, URL https://www.sciencedirect.com/science/article/pii/S157495411500196X, 95 citations (Google Scholar) [2024-04-03].

Di Cecco, G.J., Barve, V., Belitz, M.W., Stucky, B.J., Guralnick, R.P., Hurlbert, A.H., 2021. Observing the observers: How participants contribute data to inaturalist and implications for biodiversity science. BioScience 71 (11), 1179–1188.

Díaz, S., Settele, J., Brondízio, E.S., Ngo, H.T., Agard, J., Arneth, A., Balvanera, P., Brauman, K.A., Butchart, S.H.M., Chan, K.M.A., Garibaldi, L.A., Ichii, K., Liu, J., Subramanian, S.M., Midgley, G.F., Miloslavich, P., Molnár, Z., Obura, D., Pfaff, A., Polasky, S., Purvis, A., Razzaque, J., Reyers, B., Chowdhury, R.R., Shin, Y.-J., Visseren-Hamakers, I., Willis, K.J., Zayas, C.N., 2019. Pervasive human-driven decline of life on earth points to the need for transformative change. Science 366 (6471), eaax3100. http://dx.doi.org/10.1126/science.aax3100, URL https://www.science.org/doi/abs/10.1126/science.aax3100.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Dyrmann, M., Mortensen, A.K., Linneberg, L., Høye, T.T., Bjerge, K., 2021. Camera assisted roadside monitoring for invasive alien plant species using deep learning. Sensors 21 (18), 6126.

Fang, Y., Sun, Q., Wang, X., Huang, T., Wang, X., Cao, Y., 2023. EVA-02: A visual representation for neon genesis. Image Vis. Comput. 149, 105171. http://dx.doi.org/10.1016/j.imavis.2024.105171.

Farahani, A., Voghoei, S., Rasheed, K., Arabnia, H.R., 2021. A brief review of domain adaptation. In: Advances in Data Science and Information Engineering: Proceedings from ICDATA 2020 and IKE 2020. Springer, pp. 877–894.

Fraisl, D., See, L., Fritz, S., Haklay, M., McCallum, I., 2024. Leveraging the collaborative power of AI and citizen science for sustainable development. Nat. Sustain. 1–8.

Getreuer, P., 2011. Linear Methods for Image Interpolation. Image Process. Line 1, 238–259. http://dx.doi.org/10.5201/ipol.2011.g_lmii.

Goëau, H., Bonnet, P., Joly, A., 2022. Overview of plantclef 2022: Image-based plant identification at global scale. In: CLEF 2022-Conference and Labs of the Evaluation Forum. Vol. 3180, pp. 1916–1928.

Haubrock, P., Cuthbert, R., Yeo, D., Banerjee, A., Liu, C., Diagne, C., Courchamp, F., 2020. Economic costs of invasive alien species across Europe. NeoBiota 67, 153–190. http://dx.doi.org/10.3897/neobiota.67.58196.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R., 2021. Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16000–16009.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep residual learning for image recognition.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Identity mappings in deep residual networks.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780. http://dx.doi.org/10.1162/neco.1997.9.8.1735.

Hussain, M., 2023. YOLO-v1 to YOLO-v8, the rise of YOLO and its complementary nature toward digital manufacturing and industrial defect detection. Machines 11 (7), URL https://www.mdpi.com/2075-1702/11/7/677.

Isaac, N.J., Pocock, M.J., 2015. Bias and information in biological records. Biol. J. Linnean Soc. 115 (3), 522–531.

James, K., Bradshaw, K., 2020. Detecting plant species in the field with deep learning and drone technology. Methods Ecol. Evol. 11 (11), 1509–1519.

Johnson, B.A., Mader, A.D., Dasgupta, R., Kumar, P., 2020. Citizen science and invasive alien species: An analysis of citizen science initiatives using information and communications technology (ICT) to collect invasive alien species observations. Glob. Ecol. Conserv. 21, e00812.

Joly, A., Goëau, H., Glotin, H., Spampinato, C., Bonnet, P., Vellinga, W.-P., Lombardo, J.-C., Planqué, R., Palazzo, S., Müller, H., 2019. Biodiversity information retrieval through large scale content-based identification: a long-term evaluation. In: Information Retrieval Evaluation in a Changing World: lessons learned from 20 years of CLEF. pp. 389–413.

Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M., 2022. Transformers in vision: A survey. ACM Comput. Surv. 54 (10s), 1–41.

Kotowska, D., Pärt, T., Żmihorski, M., 2021. Evaluating google street view for tracking invasive alien plants along roads. Ecol. Indic. 121, 107020.

Kotowska, D., Skórka, P., Pärt, T., Auffret, A.G., Żmihorski, M., 2024. Spatial scale matters for predicting plant invasions along roads. J. Ecol. 112 (2), 305–318. http://dx.doi.org/10.1111/1365-2745.14234, URL https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/1365-2745.14234.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet classification with deep convolutional neural networks. In: Pereira, F., Burges, C., Bottou, L., Weinberger, K. (Eds.), Advances in Neural Information Processing Systems. Vol. 25, Curran Associates, Inc..

LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D., 1989. Backpropagation Applied to Handwritten Zip Code Recognition. Neural Comput. 1 (4), 541–551. http://dx.doi.org/10.1162/neco.1989.1.4.541.

Liu, X., Yoo, C., Xing, F., Oh, H., El Fakhri, G., Kang, J.-W., Woo, J., et al., 2022. Deep unsupervised domain adaptation: A review of recent advances and perspectives. APSIPA Trans. Signal Inf. Process. 11 (1).

Liutkus, A., Cıfka, O., Wu, S.-L., Simsekli, U., Yang, Y.-H., Richard, G., 2021. Relative positional encoding for transformers with linear complexity. In: International Conference on Machine Learning. PMLR, pp. 7067–7079.

Mäder, P., Boho, D., Rzanny, M., Seeland, M., Wittich, H.C., Deggelmann, A., Wäldchen, J., 2021. The flora incognita app–interactive plant species identification. Methods Ecol. Evol. 12 (7), 1335–1342.

Müllerová, J., Pergl, J., Pyšek, P., 2013. Remote sensing as a tool for monitoring plant invasions: Testing the effects of data resolution and image classification approach on the detection of a model plant species Heracleum mantegazzianum (giant hogweed). Int. J. Appl. Earth Obs. Geoinf. 25, 55–65.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P., 2023. DINOv2: Learning robust visual features without supervision.

Pinzani, L., Ceschin, S., 2023. Smart (phone)-Monitoring (SPM): An efficient and accessible method for tracking alien plant species. Sustainability 15 (12), 9814.

Porcher, E., Bonnet, P., Damgaard, C., De Frenne, P., Deguines, N., Ehlers, B.K., Frei, J., García, M.B., Gros, C., Jandt, U., et al., 2024. Can we harmonize the monitoring of plants and pollinators? New Phytol. 244 (1), 39–42.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J., 2019. Exploring the limits of transfer learning with a unified Text-to-Text transformer.

Redmon, J., Farhadi, A., 2018. YOLOv3: An incremental improvement.

Roy, H.E., Pauchard, A., Stoett, P., Truong, T.R., Bacher, S., Galil, B.S., Hulme, P.E., Ikeda, T., Sankaran, K., McGeoch, M.A., et al., 2023. IPBES invasive alien species assessment: summary for policymakers. IPBES.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C., 2019. MobileNetV2: Inverted residuals and linear bottlenecks.

Shaw, P., Uszkoreit, J., Vaswani, A., 2018. Self-attention with relative position representations.

Shetty, A.K., Saha, I., Sanghvi, R.M., Save, S.A., Patel, Y.J., 2021. A review: Object detection models. In: 2021 6th International Conference for Convergence in Technology (I2CT). IEEE, pp. 1–8.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arxiv preprint arXiv:1409.1556.

Singh, K.K., Surasinghe, T.D., Frazier, A.E., 2024. Systematic review and best practices for drone remote sensing of invasive plants. Methods Ecol. Evol..

Su, J., Lu, Y., Pan, S., Murtadha, A., Wen, B., Liu, Y., 2021. RoFormer: Enhanced transformer with rotary position embedding.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2015. Rethinking the inception architecture for computer vision.

Truong, M.-X.A., Van der Wal, R., 2024. Exploring the landscape of automated species identification apps: Development, promise, and user appraisal. BioScience 74 (9), 601–613.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.

Wang, W., Dai, J., Chen, Z., Huang, Z., Li, Z., Zhu, X., Hu, X., Lu, T., Lu, L., Li, H., Wang, X., Qiao, Y., 2022. InternImage: Exploring large-scale vision foundation models with deformable convolutions.

Wheeler, Q.D., 2004. What if GBIF? BioScience 54 (8), 717.

Wightman, R., 2019. Pytorch image models. https://github.com/rwightman/pytorch-image-models.

Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I.S., Xie, S., 2023. ConvNeXt V2: Co-designing and scaling ConvNets with masked autoencoders. URL https://arxiv.org/abs/2301.00808.

Xu, M., Yoon, S., Jeong, Y., Lee, J., Park, D.S., 2022. Transfer learning with self-supervised vision transformer for large-scale plant identification. In: Faggioli, G., Ferro, N., Hanbury, A., Potthast, M. (Eds.), Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - To - 8th, 2022. In: CEUR Workshop Proceedings, Vol. 3180, CEUR-WS.org, pp. 2238–2252.

Youden, W.J., 1950. Index for rating diagnostic tests. Cancer 3 (1), 32–35.

Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y., 2019. CutMix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. ICCV.

Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D., 2018. Mixup: Beyond empirical risk minimization. URL https://arxiv.org/abs/1710.09412.

Zong, Z., Song, G., Liu, Y., 2022. DETRs with collaborative hybrid assignments training.