



Article

A Weakly Supervised and Self-Supervised Learning Approach for Semantic Segmentation of Land Cover in Satellite Images with National Forest Inventory Data

Daniel Moraes, Manuel L. Campagnolo and Mário Caetano



Article

A Weakly Supervised and Self-Supervised Learning Approach for Semantic Segmentation of Land Cover in Satellite Images with National Forest Inventory Data

Daniel Moraes ^{1,2,*} , Manuel L. Campagnolo ³  and Mário Caetano ^{1,2} 

¹ Nova Information Management School (NOVA IMS), Universidade Nova de Lisboa, Campus de Campolide, 1070-312 Lisbon, Portugal; mario.caetano@dgterritorio.pt

² Direção-Geral do Território, Rua da Artilharia Um 107, 1099-052 Lisbon, Portugal

³ Forest Research Centre, Associate Laboratory TERRA, School of Agriculture, University of Lisbon, Tapada da Ajuda, 1349-017 Lisbon, Portugal; mlc@isa.ulisboa.pt

* Correspondence: g20180450@novaims.unl.pt

Abstract: National Forest Inventories (NFIs) provide valuable land cover (LC) information but often lack spatial continuity and an adequate update frequency. Satellite-based remote sensing offers a viable alternative, employing machine learning to extract thematic data. State-of-the-art methods such as convolutional neural networks rely on fully pixel-level annotated images, which are difficult to obtain. Although reference LC datasets have been widely used to derive annotations, NFIs consist of point-based data, providing only sparse annotations. Weakly supervised and self-supervised learning approaches help address this issue by reducing dependence on fully annotated images and leveraging unlabeled data. However, their potential for large-scale LC mapping needs further investigation. This study explored the use of NFI data with deep learning and weakly supervised and self-supervised methods. Using Sentinel-2 images and the Portuguese NFI, which covers other LC types beyond forest, as sparse labels, we performed weakly supervised semantic segmentation with a convolutional neural network to create an updated and spatially continuous national LC map. Additionally, we investigated the potential of self-supervised learning by pretraining a masked autoencoder on 65,000 Sentinel-2 image chips and then fine-tuning the model with NFI-derived sparse labels. The weakly supervised baseline achieved a validation accuracy of 69.60%, surpassing Random Forest (67.90%). The self-supervised model achieved 71.29%, performing on par with the baseline using half the training data. The results demonstrated that integrating both learning approaches enabled successful countrywide LC mapping with limited training data.

Keywords: land cover; deep learning; convolutional neural networks; masked autoencoder; national forest inventory; sentinel-2



Academic Editor: Zhenwei Shi

Received: 30 December 2024

Revised: 13 February 2025

Accepted: 18 February 2025

Published: 19 February 2025

Citation: Moraes, D.; Campagnolo, M.L.; Caetano, M. A Weakly Supervised and Self-Supervised Learning Approach for Semantic Segmentation of Land Cover in Satellite Images with National Forest Inventory Data.

Remote Sens. **2025**, *17*, 711.

<https://doi.org/10.3390/rs17040711>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Land cover is a descriptor of the Earth's surface, an important climate variable [1,2] and crucial for the study of the environment [3]. Land cover information is essential for environmental modeling, resource management [4,5] and understanding territorial dynamics [6]. Further societal benefits include areas such as agriculture and disaster management [7].

National Forest Inventories (NFIs) provide valuable land cover information focusing on forest resources. NFIs collect data to produce reliable statistics and cartography at

the national level. However, forest inventories only collect data sparsely in the form of points [8] or small plots [9,10] and normally with long time intervals, e.g., every 10 years. Therefore, NFI land cover information lacks the temporal frequency and continuous spatial coverage needed for many applications.

Remote sensing has emerged as an effective method for deriving timely and spatially continuous land cover information, as it provides adequate spatial coverage and systematic observations [11,12]. The extraction of thematic land cover information from satellite data has relied on advanced machine learning techniques, with supervised methods being the most widely used [13,14]. These methods require training samples to perform semantic segmentation, which consists of assigning a class to each pixel in an image. NFIs have been used to provide training data for pixel-level land cover classification [9,15,16]. Given the sparse nature of NFI data, most studies have adopted classifiers that treat each pixel independently, such as Random Forest [9,15,17] and Support Vector Machines [18]. Unlike pure pixel-level classifiers, deep learning convolution-based methods leverage the spatial context between neighboring pixels to learn spatial features, leading to improved classification accuracy [19–22]. Notable network architectures include DeepLab [23], PSPNet [24], Feature Pyramid Networks (FPN) [25] and U-Net [26], which can be coupled with distinct backbones, such as ResNet, VGG [27] or ConvNext [28]. These methods, however, need fully annotated images, which can be costly and hard to obtain, especially for complex tasks such as land cover classification.

To address the challenges associated with acquiring training data, alternative approaches such as weakly supervised and self-supervised learning have been proposed. Weakly supervised approaches reduce the reliance on fully annotated training images by using incomplete or inexact annotations, such as points or image-level labels [29]. Self-supervised learning also tackles this problem by distilling representative features from unlabeled data, which are largely available in the Earth Observation (EO) domain [30]. Both methods have shown encouraging results in settings with limited training data [31,32]. However, their potential for large-scale land cover mapping is yet to be investigated.

In this study, we explore how dated, point-based NFI data can be used in conjunction with deep learning and weakly supervised and self-supervised learning techniques to produce an updated and spatially continuous national land cover map. For model training and validation, we select four Sentinel-2 tiles representative of the Portuguese landscape. We use land cover data from a regular grid of points of the Portuguese NFI, which includes not only forest but other non-forest land cover types, to serve as sparse labels for a weakly supervised semantic segmentation with a convolutional neural network and multi-temporal Sentinel-2 data. We use a U-Net architecture based on the ConvNext-V2 model, a modernized version of a standard ResNet. We develop an extensive preprocessing protocol to ensure the NFI provides clean and reliable labels. In addition, this study evaluates the potential of self-supervised learning, represented by a masked autoencoder (MAE), to improve the weakly supervised model's performance. We pretrain a ConvNext-V2-based MAE on over 65,000 Sentinel-2 image chips systematically sampled within our representative areas. Then, the pretrained model is fine-tuned using the sparse labels provided by the NFI. By integrating NFI data—widely available in many countries—with weakly supervised and self-supervised learning, our study offers a novel approach to overcome the practical limitations in training data availability for land cover mapping with convolution-based models. We expect to demonstrate how such integration is possible and to what extent it can produce more accurate national-scale land cover maps.

The subsequent sections of this paper are structured as follows: Section 2—Related Works, Section 3—Data and Study Area, Section 4—Methods, Section 5—Results and Discussion and Section 6—Conclusion.

2. Related Works

2.1. Training Data for Land Cover Classification

Different data sources and strategies have been used to collect training data for land cover classification [33]. Studies have used distinct pre-existing land cover products to collect training data autonomously, including land cover maps [11,34,35] and point-based datasets [36,37]. To ensure the acquisition of accurate samples, various cleaning protocols have been proposed, such as intersecting different reference datasets [35,38], creating spectral threshold filters [11] and removing areas where changes may have occurred [39].

NFIs have also been used as a source of training data [9,15]. These datasets normally provide information in the form of a regular grid of circular plots. In most approaches, individual pixels inside the plots are extracted to train pixel-based classifiers, with Random Forest being the predominant choice [17,40]. In terms of scope, studies using NFI data were often limited to country-scale mapping of forest species, not including non-forest land cover classes [41,42].

2.2. Weak Supervision for Semantic Segmentation of Land Cover

Weakly supervised learning encompasses three approaches: incomplete, inexact and inaccurate supervision [29]. Incomplete supervision refers to learning with a small, insufficient number of labels, such as a few points or pixels per image [43]. Inexact supervision relates to using labels from a higher aggregation level, such as image-level labels or pixels with a lower resolution [44,45]. Inaccurate supervision means learning from noisy, erroneous labels [24,46]. In the EO domain, most studies have focused on inexact supervision, using image-level [32,47,48] or lower-resolution annotations [49]. Incomplete supervision using point-based [50] and sparse, partial pixel-level annotation [51] has shown encouraging results. Further research on this approach is needed, as it can benefit from sparse labels derived from reference datasets, such as NFIs.

2.3. Masked Autoencoders

Masked autoencoders (MAEs) are self-supervised pretraining techniques based on masked image modeling, where models are taught to reconstruct masked image patches, thereby learning useful image representations for downstream tasks such as classification and semantic segmentation [52]. Initially designed for transformer architectures [53], recent approaches have managed to successfully integrate MAEs with modern convolution-based networks [54]. In the EO domain, research on MAEs has gained momentum but remains very incipient. Adaptations to transformer-based MAEs were proposed, considering the multispectral and multi-temporal characteristics of EO data [55,56]. Multi-modal EO data including optical, SAR, temperature, precipitation and geolocation were used to pretrain convolution-based multi-pretext task MAEs [57]. Experiments with benchmark datasets have shown encouraging results [58]. These studies have consistently shown that fine-tuning pretrained MAEs leads to higher accuracy compared with training from scratch. However, the potential of MAE pretraining for typical large-scale EO downstream tasks, such as semantic segmentation for national land cover mapping, is yet to be explored.

3. Data and Study Area

3.1. National Forest Inventory

The National Forest Inventory (NFI) is a database of statistics and cartography that characterizes the forest resources in Portugal. In this study, we used data from the sixth and most recent NFI (NFI6), which corresponds to the reference year 2015 [8]. Data collection for the NFI is based on a two-stage sampling process. The first stage aims to characterize the territory's land use and land cover, using a sample that consists of a regular grid of

points spaced 500 m apart, covering the whole territory of continental Portugal. These points, referred to as photo-points, total approximately 360,000. The second stage aims to evaluate vegetation parameters at the forest stand level, based on fieldwork conducted in a sample of approximately 12,000 locations. Since our focus was on land cover classification, we used data from the first stage only.

The photo-points were characterized according to predefined land use and land cover nomenclature, covering both forest and non-forest classes, and four attributes: stand type, percentage of tree cover, understory and patch dimension. Stand type, percentage of cover and understory are attributes unique to photo-points whose land use is forest. This process was based on the interpretation of 30 cm aerial images from 2015. Each photo-point was classified based on the interpretation of the patch of land where the photo-point occurs. The NFI defines patch of land as a portion of land with an area equal to or greater than 5000 m² and average width equal to or greater than 20 m, encompassing a homogeneous area in terms of land use and land cover [8].

3.2. Satellite Data

Image acquisition, preprocessing and composite generation followed the methodology described in [59]. Sentinel-2 images with less than 50% cloud cover were acquired from the Theia Land Data Centre for the agricultural year of 2023 (October 2022 to September 2023). These images were processed with the MAJA algorithm, a method known for producing accurate cloud and cloud shadow masks [60,61]. We used MAJA's mask to convert masked pixels to missing data.

We used all the Sentinel-2 bands, except bands 1, 9 and 10, which are primarily useful for atmospheric studies. Bands acquired at a 20 m and 60 m spatial resolution were disaggregated to 10 m, creating images with a uniform pixel size across all bands. Additionally, we computed five spectral indices: the Normalized Difference Vegetation Index (NDVI), Normalized Burn Ratio (NBR), Normalized Difference Water Index (NDWI), Normalized Difference Built-up Index (NDBI) and Normalized Difference Middle Infrared Index (NDMIR).

Images and spectral indices were used to create 12 monthly composites for the agricultural year of 2023, based on the median value of the monthly observations. A linear interpolation was conducted to fill the gaps caused by the absence of month-long cloud-free observations. By summing the 12 monthly composites for 10 bands and 5 spectral indices, we obtained 180 features for the pixel-level classification.

A time series of images from Landsat 5, 7 and 8 available on the Google Earth Engine catalog was used to compute a change mask with a change detection algorithm available on the same platform (Section 4.2.4). We selected surface reflectance images (Level 2 Collection 2 Tier 1) with less than 30% of cloud cover from January 2005 to September 2023. Cloud-contaminated and defective pixels were masked using the quality assessment bands. The selected Landsat bands were blue, green, red, near infrared and shortwave infrared 1 and 2. We also computed the spectral indices NDVI and NBR.

3.3. Auxiliary Data

A digital terrain model (DTM) of the Portuguese territory at a 25 m spatial resolution was also used. We included the elevation of the DTM as a classification feature, hence reaching a final number of 181 classification features for the pixel-level classification.

As part of our NFI filtering strategy, we relied on a sample of 3×3 pixel windows with the NFI photo-point representing the central pixel. A stratified sample of 1700 windows was created and manually annotated based on the visual interpretation of high-resolution orthophotos and Sentinel-2 images. We labeled each window as homogeneous or non-

homogeneous. This first sample encompassed all the primary land cover classes except for urban. A second sample of 165 windows corresponding to urban land cover was created. Such a separation is justified as urban produces windows of lower uniformity, therefore demanding special treatment. An example of the annotated windows is exhibited in Appendix A.

Additionally, the Copernicus High-Resolution Layer (HRL) Dominant Leaf Type (DLT) for the 2018 reference year was used to filter the NFI photo-points. The HRL DLT product provides information at a 10 m spatial resolution derived from Sentinel-2 time series regarding the dominant leaf type in a pixel, discriminating between broadleaved and coniferous [62].

For the accuracy assessment, we used a stratified random sample with 1009 sampling units designed to validate past official land cover maps, which was updated to serve as the reference for the year of 2023. We refer to this dataset as an independent validation dataset.

3.4. Study Area

Our study area encompasses the territory of continental Portugal, as our ultimate goal is to create a national land cover map. For training and validating our machine learning models, we selected specific regions referred to as sample areas (Figure 1).

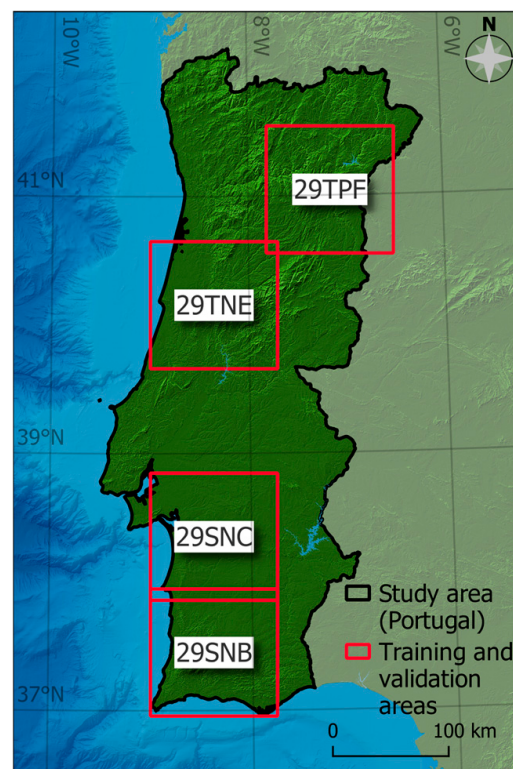


Figure 1. Study area and location of sample areas used for model training and validation.

The sample areas are aligned with Sentinel-2 tiles, namely, tiles 29SNB, 29SNC, 29TNE and 29TPF, and represent the diversity of the Portuguese landscape. They comprise forest-rich regions with an abundance of maritime pine and eucalyptus, cork and holm oak agroforestry systems, intensive agriculture, orchards and shrubs.

4. Methods

The objective of our study was to create a land cover map of continental Portugal at a 10 m spatial resolution for the agricultural year of 2023. To accomplish this goal, we

used NFI data to provide labels for a training dataset used for semantic segmentation, i.e., pixel-level classification, of land cover.

4.1. Definition of Land Cover Nomenclature

The land cover nomenclature was defined based on prior national land cover products, the availability of sufficient NFI photo-points and compatibility with satellite-based classification. We used two nomenclatures with different aggregation levels [63], one for the training and validation of the machine learning models and the other for the final map (Table 1).

Table 1. Map and training nomenclature. All the training classes correspond to original NFI classes, except for shrubland and non-vegetated surfaces, which derive from the NFI classes in parentheses.

Map Class	Training Class
Urban	Urban
Winter crops	Winter crops
Summer crops	Summer crops Rice fields
Other agriculture	Orchards Vineyards Olive trees Irrigated pastures
Cork and holm oak	Cork oak Holm oak
Eucalyptus	Eucalyptus
Other broadleaf	Oaks Other broadleaf
Maritime pine	Maritime pine
Stone pine	Stone pine
Other coniferous	Other coniferous
Shrubland	Shrubland (shrubs, tall shrubs)
Pastures and natural grasslands	Pastures
Non-vegetated surfaces	Non-vegetated surfaces (unproductive)
Water and wetlands	Water and wetlands

4.2. Preprocessing NFI Data

The class label and geographic location of the NFI photo-points were associated with the Sentinel-2 and DTM data in order to create training examples automatically. Since photo-points characterize a patch of land instead of the point's exact location, class labels may not reflect the actual land cover on the terrain (Figure 2). Additionally, NFI6 provides data for the reference year 2015, while our goal is to create a land cover map for the agricultural year of 2023. Therefore, we designed and applied an extensive filtering protocol to address the inherent mislabeling caused either by the patch/point disagreement or by the data being outdated. This protocol is presented in the following subsections.

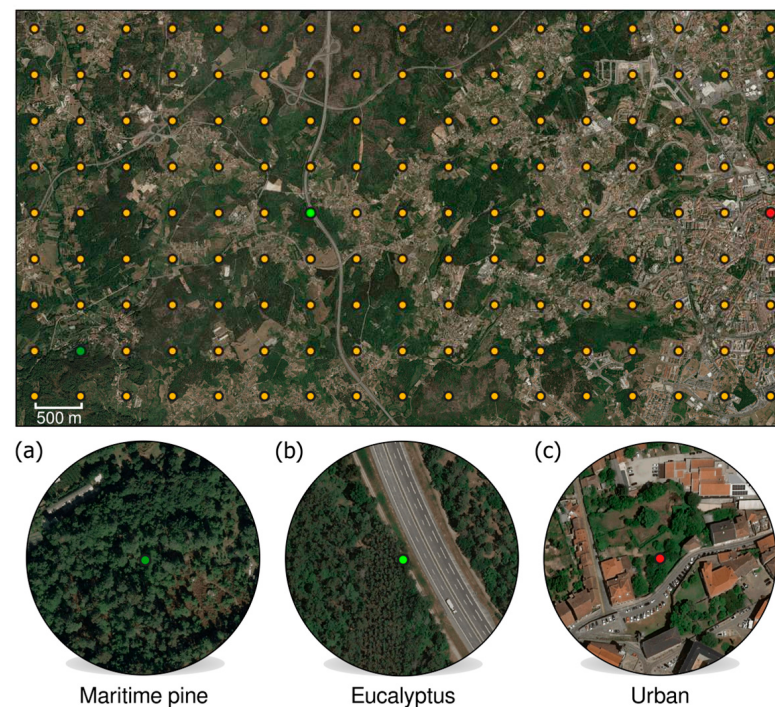


Figure 2. Example of NFI photo-points: (a) with matching point-patch labels; (b) located at the interface between distinct land covers; and (c) with mismatching point-patch labels.

4.2.1. Filter Based on Attributes

We used NFI photo-point attributes to clean our dataset. First, we removed points whose primary land cover was burnt shrubs or a mix of permanent crops, as these classes did not fit into our intended map nomenclature. Next, we applied filters based on the queries exhibited in Table 2. Our goal was to filter out points corresponding to potentially heterogeneous areas, which are more prone to mislabeling. We ensured that our sample only contained points from patches with homogeneous land cover (filter 1) and a relatively large size (filter 2), the latter to prevent acquiring points from areas with multiple small parcels with distinct land cover types. For the forest classes, we used filters 3 and 4 to remove clear cuts and stands with sparse trees, respectively.

Table 2. Attribute filters and respective queries. Only points that satisfied the queries were included.

Filter Number	Query	Scope
1	Primary land cover = secondary land cover	All classes
2	Patch dimension > 2 ha	All classes
3	Stand type = 'Standing stand'	Forest only
4	Percentage of tree cover $\geq 30\%$	Forest only

4.2.2. Homogeneity Filter

An examination of the NFI photo-points overlaid on high-resolution orthophotos revealed that many points were located on the interface between different landscape elements or land cover classes (Figure 2b). In these cases, the vicinity of the photo-point was highly heterogeneous and more likely to result in the acquisition of points whose spectral pattern did not match their label. To address this issue, we proposed a homogeneity filter.

We used our sample of homogeneous and non-homogeneous windows, described in Section 3.3, with the goal of identifying NFI-centered 3×3 -pixel windows corresponding to homogeneous areas. For that purpose, we relied on a Random Forest classification. We computed the mean and standard deviation of the 9 pixels in each window for the monthly

composites of January, April and June of the visible and near-infrared bands, NDVI and NDMIR. These amounted to 36 classification features that were used to train a Random Forest binary classifier. The classifier was trained on 75% of the data with 200 trees and achieved a test set accuracy of approximately 89%.

The trained model was used to classify all the windows around our NFI points in our sample areas. Windows classified as non-homogeneous were excluded from our dataset. The homogeneity filter was not applied on the windows corresponding to cork and holm oak and stone pine forest, as these tend to occur with a predominantly sparse tree pattern. As a result, the windows are inherently more heterogeneous, making this filtering strategy ineffective for those particular species.

4.2.3. Spectral Filter

A common filtering strategy consists of using spectral thresholds to remove or include sampling units [11,64]. We used this method to remove photo-points from our sample that did not fit the typical spectral pattern of the land cover class indicated by their label. Our approach consisted of first conducting hierarchical clustering with points of the same land cover class. We created 6 to 10 clusters for each class. With these clusters, we expected to identify not only groups corresponding to the typical spectral pattern of a given class but also groups related to mislabeled examples. By doing so, we were able to determine the most significant groups of outliers and define spectral thresholds tailored to separate these from the true class. We set loose, conservative thresholds to prevent excessively pruning our sample. Thresholds were used primarily to separate classes with very distinct spectral patterns, e.g., vegetation and bare soil. For instance, we set that the maximum monthly NDVI value must be less than 0.4 in order to filter out vegetation pixels from photo-points labeled as non-vegetated surfaces. Our spectral filter was designed to exclude clear outliers, not to address more nuanced mislabeling, such as between shrubs and forest.

4.2.4. Land Cover Change Filter

Changes in land cover between 2015 (NFI6 reference year) and 2023 could lead to acquiring mislabeled training data. To address this issue, we created a change mask by applying the Continuous Change Detection and Classification (CCDC) algorithm [65] to a time series of Landsat data. CCDC uses a harmonic regression to model time series, decomposing it into seasonality, trends and breaks. Breaks refer to a disruption in a spectro-temporal pattern, being indicative of land cover change. The method has shown reliable change detection results [66–69].

We created a mask containing pixels where CCDC identified at least one break from January 2015 to September 2023. We considered 30×30 m windows and removed photo-points whose windows intersected the mask. Landsat images were used, as Sentinel-2 surface reflectance data were not available prior to 2017, starting in January 2005, as CCDC seems to benefit from a longer time series [69]. The method's parameters are exhibited in Table 3.

Table 3. CCDC parameterization. NIR and SWIR refer to near and shortwave infrared, respectively.

Parameters	Value/Settings
Breakpoint bands	Blue, green, red, NIR, SWIR1, SWIR2, NDVI, NBR
Tmask bands	Green, SWIR1
Min_obs	6
Chi-square	0.99
minYears	1
Lambda	0.02

4.2.5. Leaf Type Concordance

Despite the application of all the previous filters, we still encountered potentially mislabeled photo-points due to confusion between forest species with different leaf types. Most notably, mixing up eucalyptus for maritime pine, and vice versa. These may be due to the difficulty, in some cases, to distinguish both species based on the interpretation of aerial images. In order to reduce such confusion in our sample, we applied a rule to ensure that the photo-point label was in accordance with the information from the Copernicus HRL DLT product. We selected photo-points from all the forest classes, except for cork and holm oaks as these did not exhibit significant confusion with other species, and removed those whose leaf types (broadleaf or coniferous) did not match the HRL DLT.

4.2.6. Additional Samples for Non-Vegetated Surfaces

Our class nomenclature included non-vegetated surfaces, with the sample drawn from the NFI unproductive class. However, such a sample lacked examples from a spectral class related to bare soil. To bridge this gap, we collected additional samples corresponding to bare soil by selecting photo-points in croplands whose maximum NDVI throughout the year was less than 0.3, indicating no agricultural activity that year.

4.3. Weakly Supervised Learning

We used the NFI photo-points along with their neighboring pixels within the 3×3 windows as sparse, partial labels to train a semantic segmentation model. Hence, only a small number of pixels around the photo-points were annotated, while the rest of the image remained unlabeled. This contrasts with fully supervised approaches, where dense annotations are required, with labels for every pixel in the image (Figure 3). By not using dense, fully annotated images, we refer to our approach as weakly supervised.

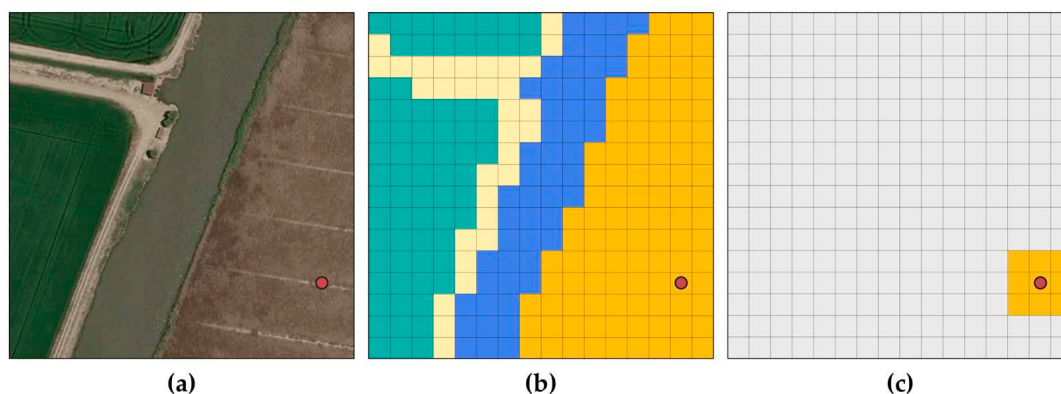


Figure 3. Illustration of distinctly labeled training data. High-resolution image (a), dense labels used in typical fully supervised methods (b) and sparse labels used in our weakly supervised approach (c). Colored and grey pixels correspond to labeled and unlabeled pixels, respectively. The labels in (c) are derived from the photo-point, seen in the center of the 3×3 window.

Our weakly supervised approach aimed to leverage NFI information in combination with a state-of-the-art convolutional neural network-based semantic segmentation model. While labeled pixels represent only a small fraction of the image, the model sees all the pixels during training, thus being able to compute global features such as texture and shape. By leveraging these features, along with spectral information, we expect the model to produce accurate classifications and spatially coherent maps.

4.3.1. Training Data Preparation

The semantic segmentation model was trained on 64×64 -pixel image chips, generated by cropping the Sentinel-2 composites using the NFI photo-points as anchors. Rather than being fixed at the center, points were randomly located within chips (Figure 3c). Hence, the position of the labeled pixels varied across the chips, introducing a spatial variability that could improve the model's spatial learning [50].

We partitioned the data into training (80%) and validation (20%) splits based on stratified random sampling. To capture landscape diversity in both sets, we used a map of landscape units of Portugal [59]. By concatenating the landscape unit code with the class label, we created a unique stratum identifier, ensuring that the sampling process preserved the distribution of both landscape types and class labels.

Since the NFI sample was imbalanced, we conducted an oversampling of the minority classes, creating additional image chips for these classes by moving the box around the labeled pixel. While the spectral information of labeled pixels was the same, their location within the chips and the chips themselves varied, which contrasts with simply duplicating the data. The additional number of image chips increased the occurrence of the minority classes in the mini batches, hence reinforcing the contribution of these classes in the gradient computation. Additionally, we also undersampled the majority classes. In this case, we randomly removed the labels of some pixels in the image chips, leaving each majority class image with only two labeled pixels rather than the original nine. This allowed us to maintain the sample diversity as we did not discard any image chips, thereby losing likely redundant information only. Overall, these strategies managed to deliver a fairly balanced training dataset.

We computed the mean and standard deviation of our 181 bands using all the pixels from the four Sentinel-2 tiles in order to standardize our data to have a mean of 0 and standard deviation of 1. This approach exhibited superior training stability and higher validation accuracy when compared to simply normalizing the data.

4.3.2. Model Architecture and Training

A U-Net [70] architecture was employed for semantic segmentation, with ConvNext-V2 [54] serving as the encoder. The original ConvNext model [28] is a modern convolutional neural network architecture, which incorporates design improvements inspired by the success of vision transformers while maintaining the simplicity and efficiency of convolution-based networks. Key improvements include modifications to the macro design, use of an inverted bottleneck and larger kernel sizes as well as fewer activation functions and normalization layers. Our chosen model, ConvNext-V2, improves upon its predecessor by introducing a Global Response Normalization layer that enhances feature diversity. Most importantly, the ConvNext-V2 model was co-designed with a self-supervised learning framework, which is discussed in Section 4.4.

We used the ConvNext-V2 Atto model, a lightweight model with 3.7M parameters, and made adaptations to optimize it for semantic segmentation of medium-resolution satellite images. In the original model, the first convolutional layer downsamples the input to feature maps that are one-fourth of the input size. This may result in losing valuable information obtained from high-resolution feature maps, which could be useful for semantic segmentation tasks. To address this limitation, we followed the modifications suggested by [57]. Essentially, we replaced the first layer with an initial convolutional layer using a kernel size of 3 and a stride of 1 to capture feature maps at the input resolution, followed by a depthwise convolutional downsampling layer. Our U-Net design followed the ConvNext-V2 U-Net implementation also used in [57], using skip connections to concatenate feature

maps from the encoder with the corresponding upsampling blocks. Figure 4 exhibits the architecture of our network.

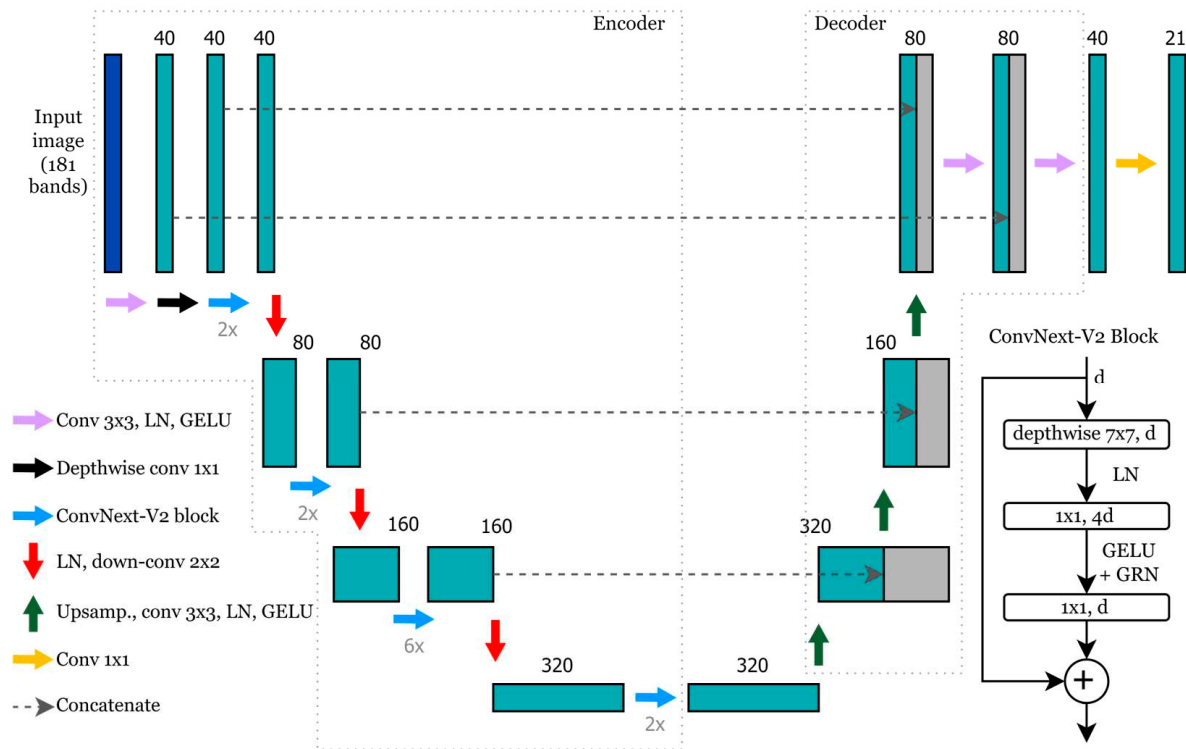


Figure 4. Network architecture of our ConvNext-V2 Atto U-Net. The figure also exhibits the ConvNext-V2 block. LN, GRN and GELU stand for Layer Normalization, Global Response Normalization and Gaussian Error Linear Unit, respectively. Conv $K \times K$ refers to a convolutional layer with a kernel size of $K \times K$.

The model was trained for 50 epochs with a batch size of 32, using the AdamW optimizer and a fixed learning rate of 0.0001. The batch size and learning rate were determined by performing a grid search with batch sizes of 32, 64 and 128 and learning rates of 0.00001, 0.0001 and 0.001. The training was limited to 50 epochs since we observed no significant changes in model performance past this point across a number of tests. Random horizontal and vertical flips were used for data augmentation. We computed a weighted cross-entropy loss, in which we assigned the unlabeled pixels to an additional class and set its weight to zero. Hence, only labeled pixels effectively contributed to the loss computation.

4.4. Self-Supervised Learning

Training with sparse labels under a weakly supervised approach may lead to limited feature learning, particularly in regions lacking labeled data. To bridge this gap, we propose to train a self-supervised model to learn general useful representations from large amounts of unlabeled data. We refer to this process as model pretraining. A pretrained model can be subsequently fine-tuned for specific tasks, improving the performance of models with limited supervision [31,71,72] such as ours.

We trained an MAE, a special self-supervised technique. In computer vision, MAEs learn by solving a pretext task, which consists of reconstructing masked parts of an image, referred to as masked image patches (Figure 5). Learned feature representations can be transferred and applied to downstream tasks, such as image classification and semantic segmentation, and may outperform fully supervised models [52,53].

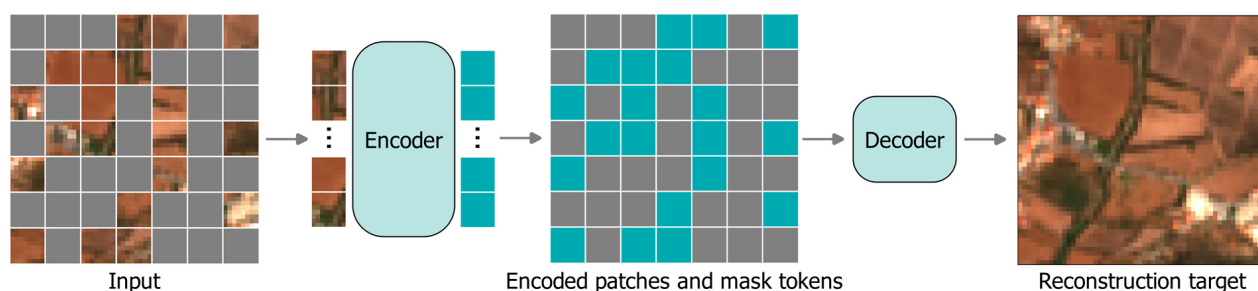


Figure 5. MAE architecture, illustrating the reconstruction of masked patches. Image representations learned at the encoder can be transferred and applied to different downstream tasks. Each patch corresponds to 8×8 pixels.

4.4.1. Training Data for MAE

Each of the four Sentinel-2 tiles comprising our sample areas were divided into a uniform grid of 560×560 m cells. Then, for each grid, we selected alternating grid cells, resulting in a systematic sampling pattern across the sample area. Our goal was to create a sufficiently large sample for the self-supervised pretraining with even and comprehensive coverage of our sample area, minding the limits of our compute resources during the model training. Selected cells were used to crop the composites into 56×56 -pixel image chips, containing 181 bands each. We opted for the 56-pixel image chip size as it integrates more effectively with our MAE model, as explained in the following subsection. In total, we collected over 65,000 image chips for training.

4.4.2. MAE Model and Training

Original MAEs were developed using the transformer architecture [53]. Here, we built our MAE model upon the ConvNext-V2 architecture, which was co-designed with a fully convolutional MAE framework in mind [54]. We used a ConvNext-V2 Atto model with the same modifications explained in Section 4.3.2. The encoder was the same as the one used in the U-Net model exhibited in Figure 4. For decoding, instead of using multiple upsampling blocks, we used a shallow decoder with just one ConvNext-V2 block as it has shown good performance but with a faster runtime [54,57]. In addition, no skip connections were used. This results in an asymmetric architecture, such as the transformer-based MAEs.

When adapting the ConvNext-V2 model to Sentinel-2 images, preserving the model's original patch layout is crucial for the performance of the pretrained model in downstream tasks [57]. Therefore, we chose to work with 56-pixel images and a patch size of 8, resulting in a configuration of 7×7 patches, consistent with the original design. During training, patches are randomly masked, and the masked image can be represented as a two-dimensional sparse array of pixels. A mask ratio of 0.6 was used. The encoder processes only the unmasked, visible pixels, while the decoder reconstructs the image using the encoded pixels and mask tokens. The model loss consists of the mean squared error between the reconstructed and original image, calculated only on the masked patches. Unlike most existing approaches, our MAE was trained to reconstruct all 181 bands simultaneously, thereby preserving the temporal information.

Our training used a batch size of 128 and the AdamW optimizer. The model was initially trained with a learning rate of 0.0015 for 700 epochs, followed by an additional 300 epochs at a reduced learning rate of 0.00015. We defined the batch size and learning rate by conducting a grid search with batch sizes of 64, 128 and 256 and initial learning rates of 0.0001, 0.001 and 0.0015. Random horizontal and vertical flips were used for data augmentation.

4.4.3. Fine-Tuning

We transferred the encoder of the MAE-pretrained model to our ConvNext-V2 U-Net and then fine-tuned this resulting model with the train and validation sets described in Section 4.3.1, following the same weakly supervised approach. The hyperparameters were the same as those used for the weakly supervised training, namely, 50 epochs, a batch size of 32 and a learning rate of 0.0001.

4.5. Baseline Comparison

We compared the results of the weakly supervised model, which we refer to as the baseline model, against the self-supervised pretrained and fine-tuned model. Pixel-level classification was also conducted using the Random Forest algorithm, which is widely used in land cover classification due to its simplicity and effectiveness [11,73,74]. Besides being well suited to our point-based training data, the algorithm is also robust to noise [75], which is advantageous for our application given the potential presence of noisy labels despite preprocessing. We used Random Forest with 300 trees and the number of features at each split as the square root of the total number of features.

4.6. Accuracy Assessment

Model performance was evaluated based on the pixel-level agreement using the overall accuracy and F1-score metrics. We ran each model 10 times to ensure consistency in our results. For the ConvNext-V2 models, we report the accuracies of the last epoch. Accuracy was evaluated on two datasets, namely, the validation split and the independent validation dataset described in Section 3.3. Since the validation split may contain a more homogeneous sample as a result of the preprocessing, using an independent dataset unconnected to the NFI data helps ensure the validity of our results.

5. Results and Discussion

The self-supervised learning (SSL) pretrained ConvNext-V2 model achieved the best overall accuracy compared to the baseline model and the Random Forest classifier. Improvements over the baseline model were consistent across the training epochs (Figure 6), culminating in a ~2% increase in accuracy by the end of the training. The result demonstrates the advantages of self-supervised pretraining in the context of weakly supervised semantic segmentation with sparse labels. While the weakly supervised model may learn features with limited discriminative power, the pretrained model can offer more useful and discriminative features for fine-tuning.

The self-supervised pretrained model also consistently outperformed its counterparts in tests with different training set sizes (Figure 7). The pretrained model exhibited comparable results to the baseline when trained with only 50% of the data. These results highlight the benefits of self-supervised pretraining in settings with limited training data. Both deep learning models were more effective at leveraging increasing data availability to improve performance compared to Random Forest, which only outperformed the baseline model in the setting with 10% of the training data.

The accuracy assessment on the independent validation dataset showed similar results (Table 4), with the self-supervised pretrained model achieving the highest overall accuracy, followed by the baseline model and Random Forest. Despite the lower accuracies, the model rankings remained identical regardless of the validation dataset, thereby sustaining our findings that the self-supervised pretrained model improves upon the baseline. Contrary to our expectations, the decision tree-based classifier did not perform on par with the deep learning-based models, despite its widespread use for pixel-level land cover classification of satellite images with point-based training data and its robustness to noisy labels.

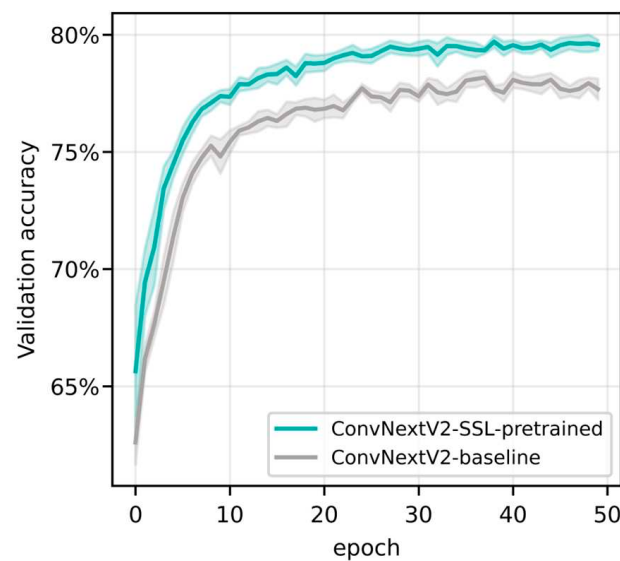


Figure 6. Overall accuracy of the baseline and self-supervised pretrained models. The values represent the average of 10 runs with a 95% confidence interval and were computed on the validation split.

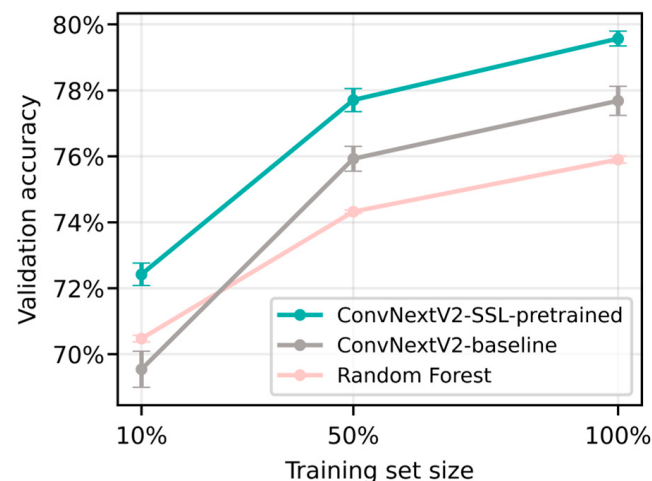


Figure 7. Validation split accuracy of the three tested models with distinct training set sizes. The reported values are the average of 10 runs with a 95% confidence interval.

Table 4. Accuracy assessment on the independent validation dataset. The reported values are the average of 10 runs and a 95% confidence interval.

Model	Overall Accuracy
Random Forest	67.90% $\pm$ 0.20%
ConvNext-V2-baseline	69.60% $\pm$ 0.25%
ConvNext-V2-SSL-pretrained	71.29% $\pm$ 0.69%

Overall, these results provide further evidence about the benefits of pretraining MAEs, which has consistently exhibited improvements over training from scratch in different studies [55,56]. However, since most MAE studies focus on transformer architectures, more research using distinct architectures such as ours is needed in order to establish a comparative basis. In addition, to the best of our knowledge, there is a lack of studies using MAEs for the semantic segmentation of land cover with medium-resolution imagery, such as Sentinel-2, which limits the scope of our comparisons.

An analysis of classification performance per class (Figure 8) shows the pretrained model was notably more accurate in classifying urban, a mix of permanent crops, maritime pine and non-vegetated surfaces. All three models failed to accurately classify samples of winter crops, summer crops and pastures and grasslands. This could be a result of imperfect training data, as NFI samples from these classes tended to be spatially adjacent and exhibit a similar spectral pattern, especially winter crops and pastures and grasslands, leading to the collection of mislabeled samples despite the NFI preprocessing efforts. The models also exhibited limited accuracy in the classification of other broadleaf and other coniferous, which was likely caused by the intricate nature of such classes and the lack of diversity in the training data, as these classes needed oversampling.

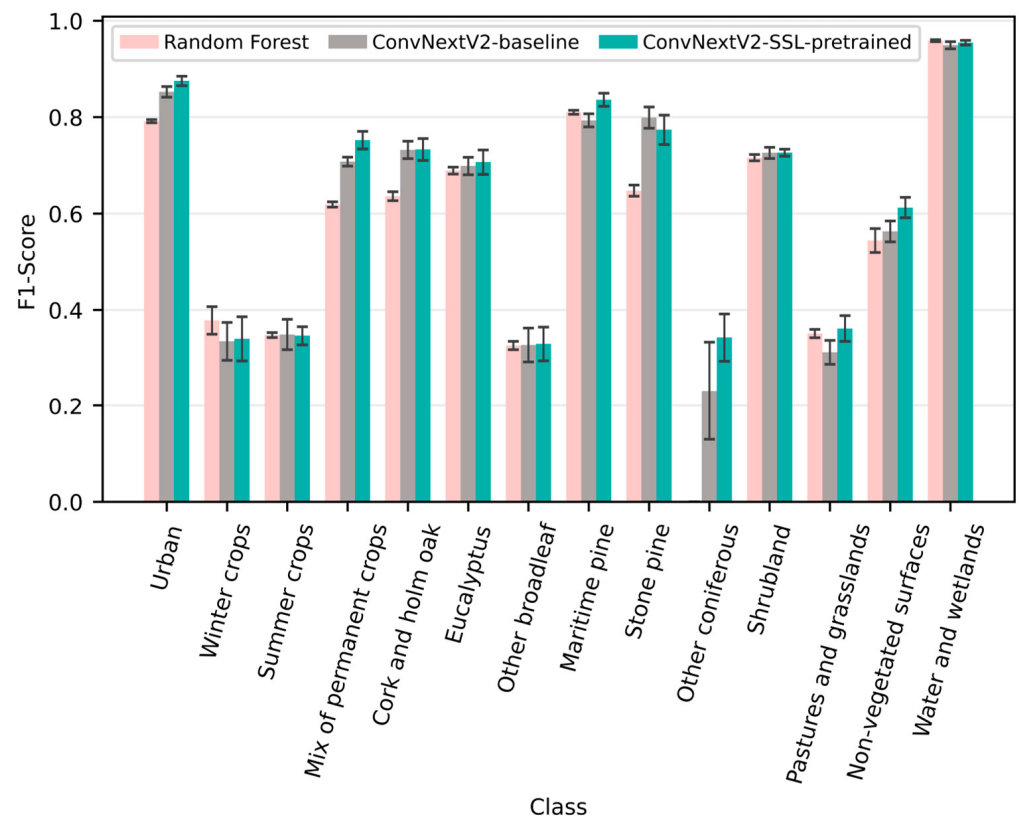


Figure 8. Model performance per land cover class measured by the F1-score. For other coniferous, no F1-score was reported for Random Forest, as the model did not predict any sampling units belonging to this class.

A qualitative assessment of the maps generated by the semantic segmentation methods (Figure 9) revealed similarities and key differences. On a higher level, the maps displayed a similar picture of the landscape, capturing most objects and shapes equivalently. However, a few differences were noticeable. In terms of classification, Random Forest and the ConvNext-V2 baseline model exhibited less accurate maps overall. Common confusions observed in these maps were between urban and non-vegetated surfaces and shrubland and forest. Those classification errors were also recurrent in other land cover mapping initiatives of the Portuguese landscape [59], demonstrating the complexity in correctly distinguishing these classes due to their similar spectral patterns.

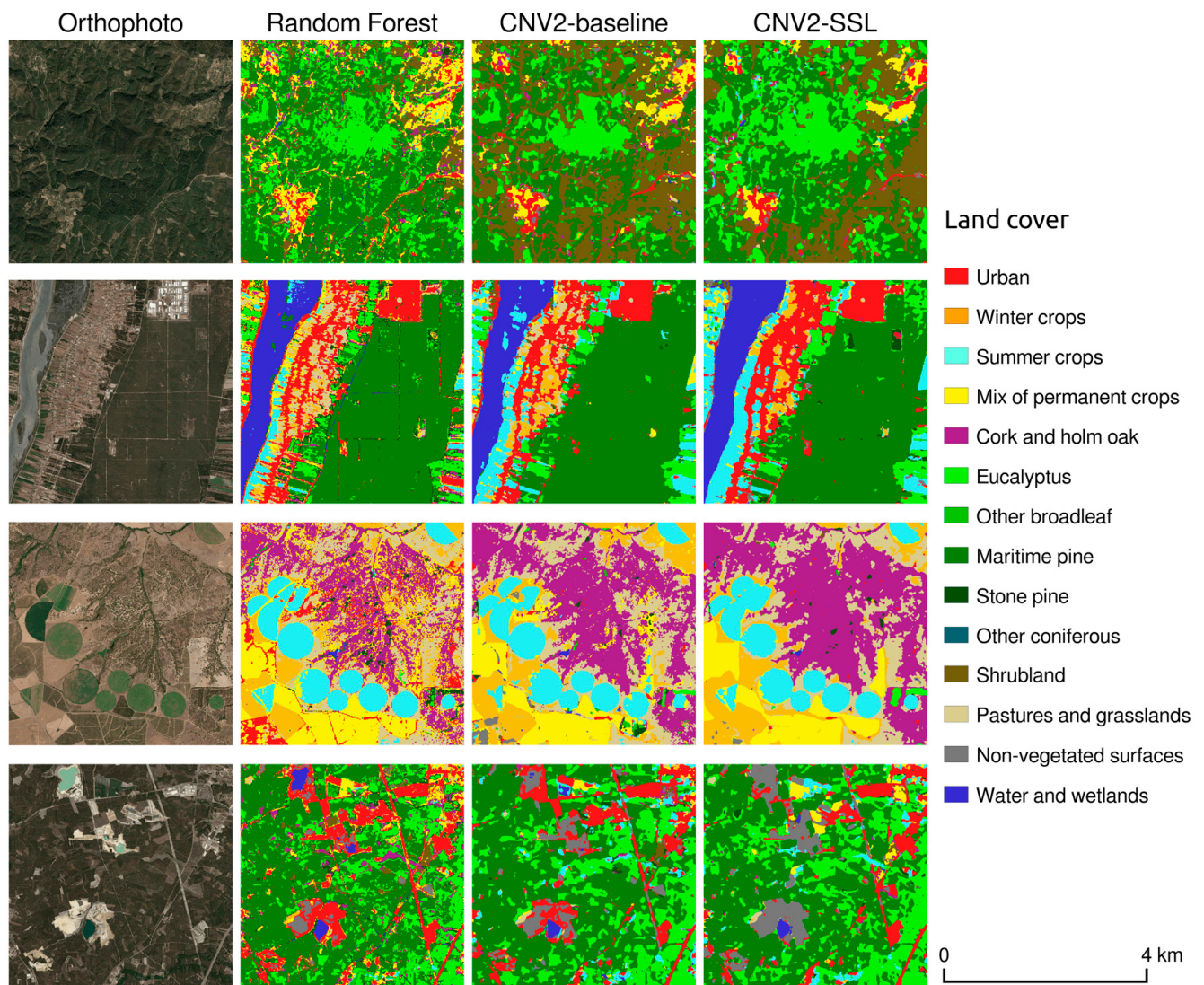


Figure 9. Example of land cover maps produced by Random Forest, ConvNext-V2 baseline and ConvNext-V2 self-supervised pretrained models.

Random Forest produced maps with finer spatial detail, which was particularly notable in areas where small features are abundant, namely, urban areas, small crop parcels and sparse forest. Despite the finer detail, Random Forest often produced results with a salt-and-pepper effect. In contrast, the ConvNext-V2 models produced smoother maps but with relatively coarser spatial detail. These differences were expected and previously documented [22,76]. While Random Forest is a purely pixel-based model, the ConvNext-V2 models use convolutions to compute spatial features, which aggregates information from neighboring pixels and involves downsampling operations that reduce spatial detail despite the use of skip connections. The implications of choosing between the decision tree-based and the neural network-based methods will depend on the intended application, region of interest and scale. For instance, if the precise delineation of small features, such as building and roads, is important, then Random Forest's finer spatial detail could be particularly beneficial. The method could also be preferable for a finer representation of a fragmented landscape, such as the montado ecosystem in Southern Portugal (Figure 9, third row), which is characterized by disperse cork and holm oak tree cover among agro-forestry systems [77]. On the other hand, the coarser spatial detail but higher thematic accuracy

of the ConvNext-V2 model may be more advantageous for large-scale environmental modeling, studying climate dynamics and hydrological modeling.

The national land cover map for the reference year of 2023, produced with the best performing model, is exhibited in Figure 10. The map successfully depicts the main characteristics of the Portuguese landscape, for instance, the abundant forest of eucalyptus and maritime pine in the center, large areas of pastures and grasslands mixed with cork and holm oaks in the South and the two main urban centers.

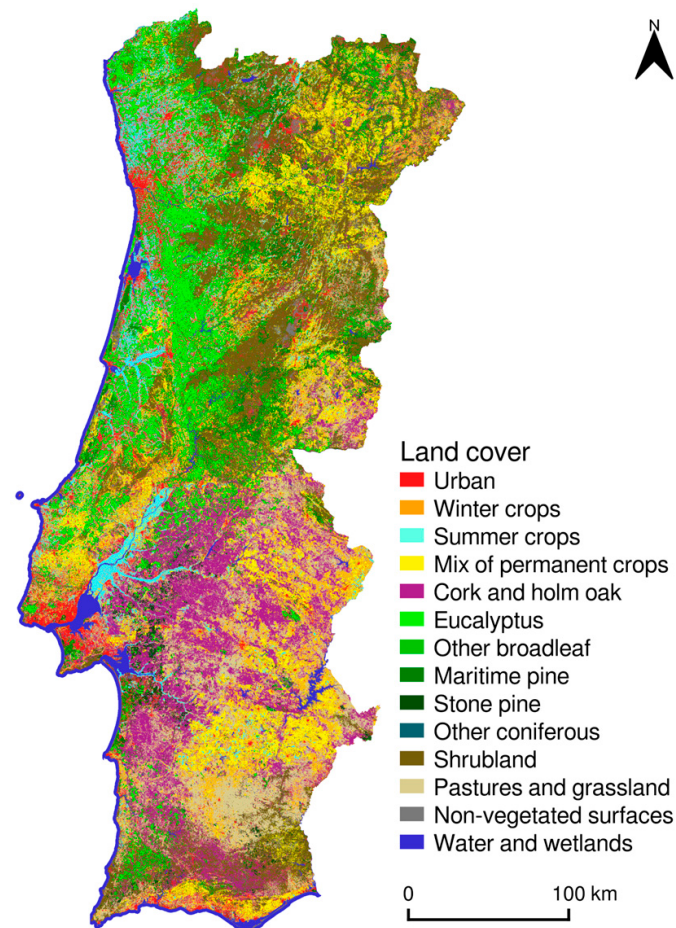


Figure 10. Land cover map of Portugal (2023).

While this study represents a step forward in the use of dated, point-based NFI data, deep learning and pretraining models to map land cover at a large scale, certain limitations should be outlined. Firstly, despite the efforts to reduce label noise in the NFI-derived training sample, the adopted filtering strategy could have been insufficient. For instance, the spectral threshold filter aimed to remove only the most obvious mislabeling rather than filtering sampling units with subtler differences (e.g., shrubland and forest). Secondly, the algorithm used in the land cover change filter can produce false negatives, meaning it may fail to filter out changed areas and hence lead to the acquisition of potentially mislabeled sampling units. In contrast, the filtering protocol led to the removal of a number of valid sampling units, for instance, linear features, such as small roads and riverside broadleaved trees. Since mislabeled data in the training sample degrade model performance, proper NFI data cleaning remains a challenge. Thirdly, since the goal was to produce a yearly land cover map, we had to predict a single label for an entire 1-year period. However, since our training data should be temporally stable, as a result of the land cover change filter, it should only be reliably used for classifying other temporally stable examples. This

means areas where disturbances occurred during the reference year were likely to be poorly classified. Future work could implement a protocol to process these areas in a separate classification. Finally, as pointed out earlier, the final map exhibited coarser spatial detail. In that regard, future research may test modifications to the network architecture to reduce downsampling and/or include additional skip connections in order to preserve the spatial detail of the native resolution. With respect to the self-supervised pretraining, subsequent research can explore the combination of spatial with spectral and temporal masking for learning improved representations when training masked autoencoders, thus exploiting the potential of multispectral Sentinel-2 time series.

6. Conclusions

This study aimed to create spatially continuous land cover maps using data from the Portuguese National Forest Inventory. NFI data in the form of photo-points were cleaned with an extensive filtering protocol and then used as training data for the semantic segmentation of land cover. We explored the use of photo-points as sparse labels to train a weakly supervised convolutional neural network. A self-supervised pretrained masked autoencoder was used to improve the predictive capabilities of our weakly supervised model. The results showed the potential of the weakly supervised approach, which managed to produce spatially coherent and accurate maps, outperforming a Random Forest classification. Furthermore, using the pretrained masked autoencoder showed clear advantages, producing more accurate results than Random Forest and the baseline weakly supervised method. The pretrained model performed on par with the baseline weakly supervised model using half of the training data, revealing the utility of the image representations learned through self-supervision. In summary, our work showcased the joint use of NFI data and advanced machine learning techniques to successfully map land cover at the national level. Future work can explore adjustments to the NFI preprocessing protocol to improve training data quality, modifications to the network architecture to preserve spatial detail and the use of spectral and temporal masks for training autoencoders.

Author Contributions: Conceptualization, D.M., M.L.C. and M.C.; methodology, D.M., M.L.C. and M.C.; software, D.M.; validation, D.M., M.L.C. and M.C.; formal analysis, D.M., M.L.C. and M.C.; investigation D.M.; resources, M.L.C. and M.C.; data curation, D.M.; writing—original draft, D.M.; writing—review and editing, M.L.C. and M.C.; supervision, M.L.C. and M.C.; project administration, M.L.C. and M.C.; funding acquisition, M.L.C. and M.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by national funds through the FCT (Fundação para a Ciência e a Tecnologia), under the project UIDB/04152/2020 (DOI: 10.54499/UIDB/04152/2020)—Centro de Investigação em Gestão de Informação (MagIC)/NOVA IMS and grant number PRT/BD/153517/2021.

Data Availability Statement: The data presented in this study are available in the Theia data centre at www.theia-land.fr and in the National Institute for Nature and Forest Conservation (ICNF) at www.icnf.pt.

Acknowledgments: The value-added data were processed by CNES for the Theia data centre, www.theia-land.fr, using Copernicus products. The processing uses algorithms developed by Theia's Scientific Expertise Centres. We thank Francisco D. Moreira for his contributions on the definition of the land cover nomenclature.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of this study; in the collection, analyses or interpretation of data; in the writing of this manuscript; or in the decision to publish the results.

Appendix A

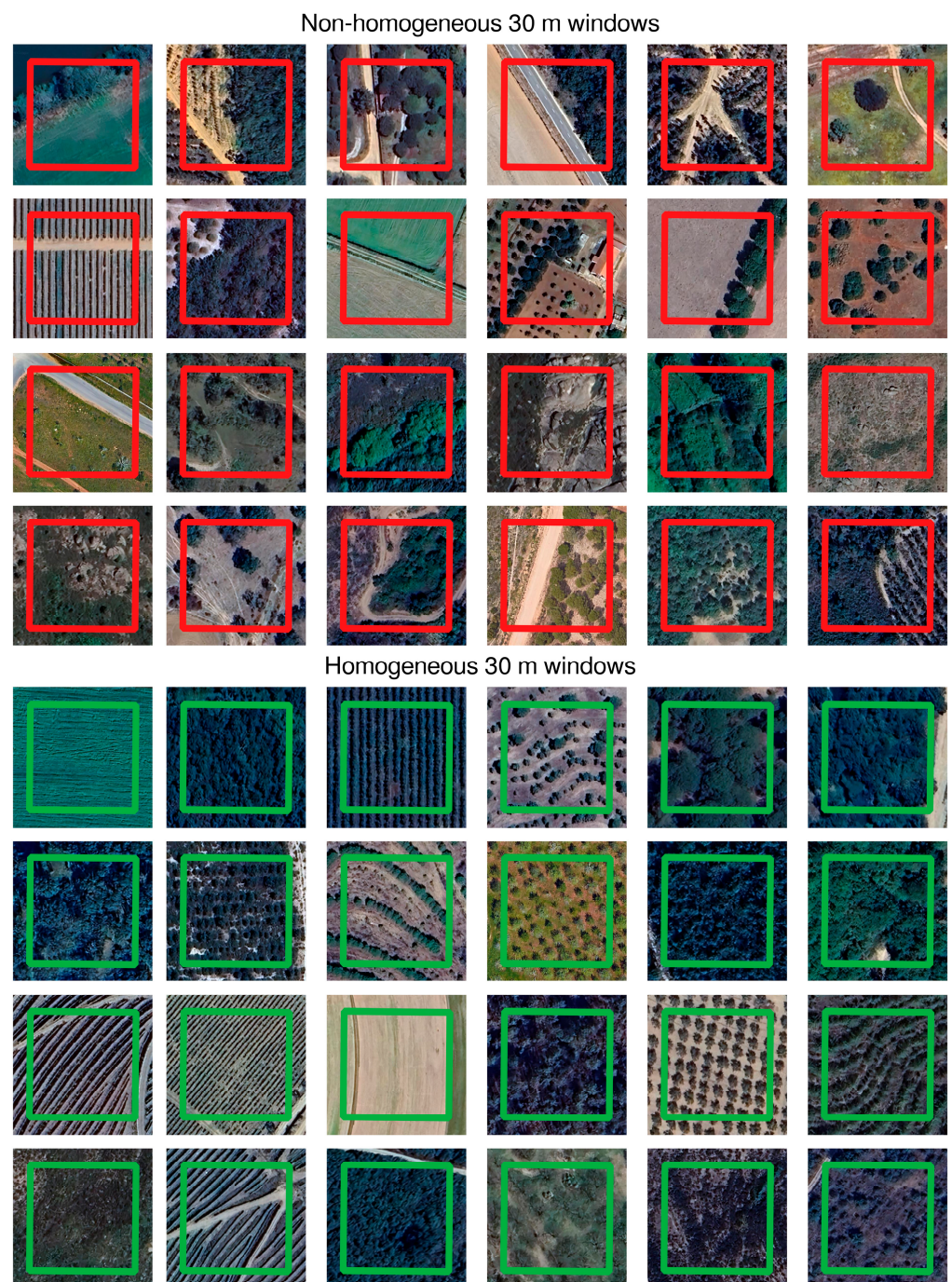


Figure A1. Example of 30×30 m windows used for training a Random Forest classifier for the homogeneity filter. Annotations as non-homogeneous or homogeneous considered not only the high-resolution images (seen in the figure) but also Sentinel-2 images.

References

1. Alkama, R.; Cescatti, A. Climate Change: Biophysical Climate Impacts of Recent Changes in Global Forest Cover. *Science* **2016**, *351*, 600–604. [[CrossRef](#)]
2. Feddema, J.J.; Oleson, K.W.; Bonan, G.B.; Mearns, L.O.; Buja, L.E.; Meehl, G.A.; Washington, W.M. Atmospheric Science: The Importance of Land-Cover Change in Simulating Future Climates. *Science* **2005**, *310*, 1674–1678. [[CrossRef](#)]
3. Herold, M.; Latham, J.S.; Di Gregorio, A.; Schmullius, C.C. Evolving Standards in Land Cover Characterization. *J. Land Use Sci.* **2006**, *1*, 157–168. [[CrossRef](#)]

4. Hermosilla, T.; Wulder, M.A.; White, J.C.; Coops, N.C.; Hobart, G.W. Regional Detection, Characterization, and Attribution of Annual Forest Change from 1984 to 2012 Using Landsat-Derived Time-Series Metrics. *Remote Sens. Environ.* **2015**, *170*, 121–132. [\[CrossRef\]](#)
5. Song, W.; Deng, X. Land-Use/Land-Cover Change and Ecosystem Service Provision in China. *Sci. Total Environ.* **2017**, *576*, 705–719. [\[CrossRef\]](#)
6. Alves, A.; Marcelino, F.; Gomes, E.; Rocha, J.; Caetano, M. Spatiotemporal Land-Use Dynamics in Continental Portugal 1995–2018. *Sustainability* **2022**, *14*, 15540. [\[CrossRef\]](#)
7. Wulder, M.A.; White, J.C.; Goward, S.N.; Masek, J.G.; Irons, J.R.; Herold, M.; Cohen, W.B.; Loveland, T.R.; Woodcock, C.E. Landsat Continuity: Issues and Opportunities for Land Cover Monitoring. *Remote Sens. Environ.* **2008**, *112*, 955–969. [\[CrossRef\]](#)
8. Instituto da Conservação da Natureza e das Florestas. 6° Inventário Florestal Nacional (IFN6)—2015 Relatório Final; Instituto da Conservação da Natureza e das Florestas: Alges, Portugal, 2019.
9. Blickensdörfer, L.; Oehmichen, K.; Pflugmacher, D.; Kleinschmit, B.; Hostert, P. National Tree Species Mapping Using Sentinel-1/2 Time Series and German National Forest Inventory Data. *Remote Sens. Environ.* **2024**, *304*, 114069. [\[CrossRef\]](#)
10. Schelhaas, M.-J.; Clerckx, S.; Lerink, B. 7th Dutch National Forest Inventory: 2017–2021; Wettelijke Onderzoekstaken Natuur & Milieu: Wageningen, The Netherlands, 2022.
11. Rybicki, M.; Gromny, E.; Malinowski, R.; Lewiński, S.; Jenerowicz, M.; Michał, K.; Nowakowski, A.; Wojtkowski, C.; Krupiński, M.; Kraetzschmar, E.; et al. Automated Production of a Land Cover/Use Map of Europe Based on Sentinel-2 Imagery. *Remote Sens.* **2020**, *12*, 3523. [\[CrossRef\]](#)
12. Yifang, B.; Gong, P.; Gini, C. Global Land Cover Mapping Using Earth Observation Satellite Data: Recent Progresses and Challenges. *ISPRS J. Photogramm. Remote Sens.* **2015**, *103*, 1–6.
13. Talukdar, S.; Singha, P.; Mahato, S.; Shahfahad; Pal, S.; Liou, Y.A.; Rahman, A. Land-Use Land-Cover Classification by Machine Learning Classifiers for Satellite Observations—A Review. *Remote Sens.* **2020**, *12*, 1135. [\[CrossRef\]](#)
14. Wulder, M.A.; Coops, N.C.; Roy, D.P.; White, J.C.; Hermosilla, T. Land Cover 2.0. *Int. J. Remote Sens.* **2018**, *39*, 4254–4284. [\[CrossRef\]](#)
15. Francini, S.; Schelhaas, M.J.; Vangi, E.; Lerink, B.J.; Nabuurs, G.J.; McRoberts, R.E.; Chirici, G. Forest Species Mapping and Area Proportion Estimation Combining Sentinel-2 Harmonic Predictors and National Forest Inventory Data. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *131*, 103935. [\[CrossRef\]](#)
16. Gilichinsky, M.; Sandströma, P.; Reesea, H.; Kivinenb, S.; Moenb, J.; Nilsona, M. Application of National Forest Inventory for Remote Sensing Classification of Ground Lichen in Northern Sweden; Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences-ISPRS Archives: Haifa, Israel, 2010; pp. 146–152.
17. Guindon, L.; Manka, F.; Correia, D.L.P.; Villemaire, P.; Smiley, B.; Bernier, P.; Gauthier, S.; Beaudoin, A.; Boucher, J.; Boulanger, Y. A New Approach for Spatializing the Canadian National Forest Inventory (SCANFI) Using Landsat Dense Time Series. *Can. J. For. Res.* **2024**, *54*, 793–815. [\[CrossRef\]](#)
18. Denisova, A.Y.; Kavelenova, L.M.; Korchikov, E.S.; Prokhorova, N.V.; Terentyeva, D.A.; Fedoseev, V.A. Tree Species Classification for Clarification of Forest Inventory Data Using Sentinel-2 Images. Proceedings of Seventh International Conference on Remote Sensing and Geoinformation of the Environment (RSCy2019), Paphos, Cyprus, 18–21 March 2019; Volume 11174.
19. Hu, Y.; Zhang, Q.; Zhang, Y.; Yan, H. A Deep Convolution Neural Network Method for Land Cover Mapping: A Case Study of Qinhuangdao, China. *Remote Sens.* **2018**, *10*, 2053. [\[CrossRef\]](#)
20. Rezaee, M.; Mahdianpari, M.; Zhang, Y.; Salehi, B. Deep Convolutional Neural Network for Complex Wetland Classification Using Optical Remote Sensing Imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 3030–3039. [\[CrossRef\]](#)
21. Tong, X.Y.; Xia, G.S.; Lu, Q.; Shen, H.; Li, S.; You, S.; Zhang, L. Land-Cover Classification with High-Resolution Remote Sensing Images Using Transferable Deep Models. *Remote Sens. Environ.* **2020**, *237*, 111322. [\[CrossRef\]](#)
22. Boston, T.; Van Dijk, A.; Larraondo, P.R.; Thackway, R. Comparing CNNs and Random Forests for Landsat Image Segmentation Trained on a Large Proxy Land Cover Dataset. *Remote Sens.* **2022**, *14*, 3396. [\[CrossRef\]](#)
23. Chen, T.H.K.; Qiu, C.; Schmitt, M.; Zhu, X.X.; Sabel, C.E.; Prishchepov, A.V. Mapping Horizontal and Vertical Urban Densification in Denmark with Landsat Time-Series from 1985 to 2018: A Semantic Segmentation Solution. *Remote Sens. Environ.* **2020**, *251*, 112096. [\[CrossRef\]](#)
24. Zhao, X.; Hong, D.; Gao, L.; Zhang, B.; Chanussot, J. Transferable Deep Learning from Time Series of Landsat Data for National Land-Cover Mapping with Noisy Labels: A Case Study of China. *Remote Sens.* **2021**, *13*, 4194. [\[CrossRef\]](#)
25. Yuan, P.; Zhao, Q.; Zheng, Y.; Wang, X.; Hu, B. Capturing Small Objects and Edges Information for Cross-Sensor and Cross-Region Land Cover Semantic Segmentation in Arid Areas. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 983–997. [\[CrossRef\]](#)
26. Zhang, H.; Liu, M.; Wang, Y.; Shang, J.; Liu, X.; Li, B.; Song, A.; Li, Q. Automated Delineation of Agricultural Field Boundaries from Sentinel-2 Images Using Recurrent Residual U-Net. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *105*, 102557. [\[CrossRef\]](#)
27. Li, J.; Cai, Y.; Li, Q.; Kou, M.; Zhang, T. A Review of Remote Sensing Image Segmentation by Deep Learning Methods. *Int. J. Digit. Earth* **2024**, *17*, 2328827. [\[CrossRef\]](#)

28. Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A Convnet for the 2020s. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 11976–11986.
29. Zhou, Z.-H. A Brief Introduction to Weakly Supervised Learning. *Natl. Sci. Rev.* **2018**, *5*, 44–53. [\[CrossRef\]](#)
30. Wang, Y.; Albrecht, C.M.; Braham, N.A.A.; Mou, L.; Zhu, X.X. Self-Supervised Learning in Remote Sensing: A Review. *IEEE Geosci. Remote Sens. Mag.* **2022**, *10*, 213–247. [\[CrossRef\]](#)
31. Felfeliyan, B.; Forkert, N.D.; Hareendranathan, A.; Cornel, D.; Zhou, Y.; Kuntze, G.; Jaremko, J.L.; Ronsky, J.L. Self-Supervised-RCNN for Medical Image Segmentation with Limited Data Annotation. *Comput. Med. Imaging Graph.* **2023**, *109*, 102297. [\[CrossRef\]](#) [\[PubMed\]](#)
32. Schmitt, M.; Prexl, J.; Ebel, P.; Liebel, L.; Zhu, X.X. Weakly Supervised Semantic Segmentation of Satellite Images for Land Cover Mapping—Challenges and Opportunities. In Proceedings of the ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Nice, France, 31 August–2 September 2020; Volume V-3–2020, pp. 795–802.
33. Moraes, D.; Campagnolo, M.L.; Caetano, M. Training Data in Satellite Image Classification for Land Cover Mapping: A Review. *Eur. J. Remote Sens.* **2024**, *57*, 2341414. [\[CrossRef\]](#)
34. Zhou, Q.; Tollerud, H.; Barber, C.; Smith, K.; Zelenak, D. Training Data Selection for Annual Land Cover Classification for the Land Change Monitoring, Assessment, and Projection (LCMAP) Initiative. *Remote Sens.* **2020**, *12*, 699. [\[CrossRef\]](#)
35. Hermosilla, T.; Wulder, M.A.; White, J.C.; Coops, N.C. Land Cover Classification in an Era of Big and Open Data: Optimizing Localized Implementation and Training Data Selection to Improve Mapping Outcomes. *Remote Sens. Environ.* **2022**, *268*, 112780. [\[CrossRef\]](#)
36. Venter, Z.S.; Sydenham, M.A.K. Continental-Scale Land Cover Mapping at 10 m Resolution over Europe (Elc10). *Remote Sens.* **2021**, *13*, 2301. [\[CrossRef\]](#)
37. Weigand, M.; Staab, J.; Wurm, M.; Taubenböck, H. Spatial and Semantic Effects of LUCAS Samples on Fully Automated Land Use/Land Cover Classification in High-Resolution Sentinel-2 Data. *Int. J. Appl. Earth Obs. Geoinf.* **2020**, *88*, 102065. [\[CrossRef\]](#)
38. Zhang, J.; Fu, Z.; Zhu, Y.; Wang, B.; Sun, K.; Zhang, F. A High-Performance Automated Large-Area Land Cover Mapping Framework. *Remote Sens.* **2023**, *15*, 3143. [\[CrossRef\]](#)
39. Xie, S.; Liu, L.; Yang, J. Time-Series Model-Adjusted Percentile Features: Improved Percentile Features for Land-Cover Classification Based on Landsat Data. *Remote Sens.* **2020**, *12*, 3091. [\[CrossRef\]](#)
40. Polyakova, A.; Mukharamova, S.; Yermolaev, O.; Shaykhutdinova, G. Automated Recognition of Tree Species Composition of Forest Communities Using Sentinel-2 Satellite Data. *Remote Sens.* **2023**, *15*, 329. [\[CrossRef\]](#)
41. Breidenbach, J.; Waser, L.T.; Debella-Gilo, M.; Schumacher, J.; Rahlf, J.; Hauglin, M.; Puliti, S.; Astrup, R. National Mapping and Estimation of Forest Area by Dominant Tree Species Using Sentinel-2 Data. *Can. J. For. Res.* **2021**, *51*, 365–379. [\[CrossRef\]](#)
42. Welle, T.; Aschenbrenner, L.; Kuonath, K.; Kirmaier, S.; Franke, J. Mapping Dominant Tree Species of German Forests. *Remote Sens.* **2022**, *14*, 3330. [\[CrossRef\]](#)
43. Bearman, A.; Russakovsky, O.; Ferrari, V.; Li, F.-F. *What's the Point: Semantic Segmentation with Point Supervision*; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 549–565.
44. Yao, X.; Han, J.; Cheng, G.; Qian, X.; Guo, L. Semantic Annotation of High-Resolution Satellite Images via Weakly Supervised Learning. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 3660–3671. [\[CrossRef\]](#)
45. Chan, L.; Hosseini, M.S.; Plataniotis, K.N. A Comprehensive Analysis of Weakly-Supervised Semantic Segmentation in Different Image Domains. *Int. J. Comput. Vis.* **2021**, *129*, 361–384. [\[CrossRef\]](#)
46. Kaiser, P.; Wegner, J.D.; Jaggi, M.; Hofmann, T.; Schindler, K. Learning Aerial Image Segmentation from Online Maps. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 6054–6068. [\[CrossRef\]](#)
47. Chen, J.; He, F.; Zhang, Y.; Sun, G.; Deng, M. SPMF-Net: Weakly Supervised Building Segmentation by Combining Superpixel Pooling and Multi-Scale Feature Fusion. *Remote Sens.* **2020**, *12*, 1049. [\[CrossRef\]](#)
48. Fu, K.; Lu, W.; Diao, W.; Yan, M.; Sun, H.; Zhang, Y.; Sun, X. WSF-NET: Weakly Supervised Feature-Fusion Network for Binary Segmentation in Remote Sensing Image. *Remote Sens.* **2018**, *10*, 1970. [\[CrossRef\]](#)
49. Chen, Y.; Zhang, G.; Cui, H.; Li, X.; Hou, S.; Ma, J.; Li, Z.; Li, H.; Wang, H. A Novel Weakly Supervised Semantic Segmentation Framework to Improve the Resolution of Land Cover Product. *ISPRS J. Photogramm. Remote Sens.* **2023**, *196*, 73–92. [\[CrossRef\]](#)
50. Wang, S.; Chen, W.; Xie, S.M.; Azzari, G.; Lobell, D.B. Weakly Supervised Deep Learning for Segmentation of Remote Sensing Imagery. *Remote Sens.* **2020**, *12*, 207. [\[CrossRef\]](#)
51. Zhang, W.; Tang, P.; Corpetti, T.; Zhao, L. WTS: A Weakly towards Strongly Supervised Learning Framework for Remote Sensing Land Cover Classification Using Segmentation Models. *Remote Sens.* **2021**, *13*, 394. [\[CrossRef\]](#)
52. Balestrieri, R.; Ibrahim, M.; Sobal, V.; Morcos, A.; Shekhar, S.; Goldstein, T.; Bordes, F.; Bardes, A.; Mialon, G.; Tian, Y.; et al. A Cookbook of Self-Supervised Learning. *arXiv* **2023**, arXiv:2304.12210. [\[CrossRef\]](#)
53. He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; Girshick, R. Masked Autoencoders Are Scalable Vision Learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 16000–16009.

54. Woo, S.; Debnath, S.; Hu, R.; Chen, X.; Liu, Z.; Kweon, I.S.; Xie, S. ConvNext-V2: Co-Designing and Scaling Convnets with Masked Autoencoders. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 16133–16142.
55. Lin, J.; Gao, F.; Shi, X.; Dong, J.; Du, Q. SS-MAE: Spatial-Spectral Masked Autoencoder for Multisource Remote Sensing Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5531614. [\[CrossRef\]](#)
56. Cong, Y.; Khanna, S.; Meng, C.; Liu, P.; Rozi, E.; He, Y.; Burke, M.; Lobell, D.; Ermon, S. SatMAE: Pre-Training Transformers for Temporal and Multi-Spectral Satellite Imagery. In Proceedings of the Advances in Neural Information Processing Systems, New Orleans, LA, USA, 28 November–9 December 2022; Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2022; Volume 35, pp. 197–211.
57. Nedungadi, V.; Kariryaa, A.; Oehmcke, S.; Belongie, S.; Igel, C.; Lang, N. MMEarth: Exploring Multi-Modal Pretext Tasks for Geospatial Representation Learning. *arXiv* **2024**, arXiv:2405.02771.
58. Sosa, J.; Aloulou, M.; Rukhovich, D.; Sleimi, R.; Changaival, B.; Kacem, A.; Aouada, D. How Effective Is Pre-Training of Large Masked Autoencoders for Downstream Earth Observation Tasks? *arXiv* **2024**, arXiv:2409.18536.
59. Costa, H.; Benevides, P.; Moreira, F.D.; Moraes, D.; Caetano, M. Spatially Stratified and Multi-Stage Approach for National Land Cover Mapping Based on Sentinel-2 Data and Expert Knowledge. *Remote Sens.* **2022**, *14*, 1865. [\[CrossRef\]](#)
60. Baetens, L.; Desjardins, C.; Hagolle, O. Validation of Copernicus Sentinel-2 Cloud Masks Obtained from MAJA, Sen2Cor, and FMask Processors Using Reference Cloud Masks Generated with a Supervised Active Learning Procedure. *Remote Sens.* **2019**, *11*, 433. [\[CrossRef\]](#)
61. Skakun, S.; Wevers, J.; Brockmann, C.; Doxani, G.; Aleksandrov, M.; Batič, M.; Frantz, D.; Gascon, F.; Gómez-Chova, L.; Hagolle, O.; et al. Cloud Mask Intercomparison EXercise (CMIX): An Evaluation of Cloud Masking Algorithms for Landsat 8 and Sentinel-2. *Remote Sens. Environ.* **2022**, *274*, 112990. [\[CrossRef\]](#)
62. European Environmental Agency (EEA). *High Resolution Layer Forest, Dominant Leaf Type 2018*; European Environmental Agency (EEA): Copenhagen, Denmark, 2021.
63. Moraes, D.; Benevides, P.; Costa, H.; Moreira, F.D.; Caetano, M. Exploring Different Levels of Class Nomenclature in Random Forest Classification of Sentinel-2 Data. In Proceedings of the IGARSS 2022—IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 17–22 July 2022; pp. 2279–2282.
64. Lin, C.; Du, P.; Samat, A.; Li, E.; Wang, X.; Xia, J. Automatic Updating of Land Cover Maps in Rapidly Urbanizing Regions by Relational Knowledge Transferring from Globeland30. *Remote Sens.* **2019**, *11*, 1397. [\[CrossRef\]](#)
65. Zhu, Z.; Woodcock, C.E. Continuous Change Detection and Classification of Land Cover Using All Available Landsat Data. *Remote Sens. Environ.* **2014**, *144*, 152–171. [\[CrossRef\]](#)
66. De Souza, A.A.; Galvão, L.S.; Korting, T.S.; Almeida, C.A. On a Data-Driven Approach for Detecting Disturbance in the Brazilian Savannas Using Time Series of Vegetation Indices. *Remote Sens.* **2021**, *13*, 4959. [\[CrossRef\]](#)
67. Zhu, Z.; Zhang, J.; Yang, Z.; Aljaddani, A.H.; Cohen, W.B.; Qiu, S.; Zhou, C. Continuous Monitoring of Land Disturbance Based on Landsat Time Series. *Remote Sens. Environ.* **2020**, *238*, 111116. [\[CrossRef\]](#)
68. Mulverhill, C.; Coops, N.C.; Achim, A. Continuous Monitoring and Sub-Annual Change Detection in High-Latitude Forests Using Harmonized Landsat Sentinel-2 Data. *ISPRS J. Photogramm. Remote Sens.* **2023**, *197*, 309–319. [\[CrossRef\]](#)
69. Moraes, D.; Barbosa, B.; Costa, H.; Moreira, F.D.; Benevides, P.; Caetano, M.; Campagnolo, M. Continuous Forest Loss Monitoring in a Dynamic Landscape of Central Portugal with Sentinel-2 Data. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *130*, 103913. [\[CrossRef\]](#)
70. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015, Munich, Germany, 5–9 October 2015; Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 234–241.
71. Van Gansbeke, W.; Vandenhende, S.; Georgoulis, S.; Proesmans, M.; Van Gool, L. SCAN: Learning to Classify Images Without Labels. In Proceedings of the European Conference on Computer Vision (ECCV); Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 268–285.
72. Zhai, X.; Oliver, A.; Kolesnikov, A.; Beyer, L. S4L: Self-Supervised Semi-Supervised Learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1476–1485.
73. Inglada, J.; Vincent, A.; Arias, M.; Tardy, B.; Morin, D.; Rodes, I. Operational High Resolution Land Cover Map Production at the Country Scale Using Satellite Image Time Series. *Remote Sens.* **2017**, *9*, 95. [\[CrossRef\]](#)
74. Zhu, Z.; Gallant, A.L.; Woodcock, C.E.; Pengra, B.; Olofsson, P.; Loveland, T.R.; Jin, S.; Dahal, D.; Yang, L.; Auch, R.F. Optimizing Selection of Training and Auxiliary Data for Operational Land Cover Classification for the LCMAP Initiative. *ISPRS J. Photogramm. Remote Sens.* **2016**, *122*, 206–221. [\[CrossRef\]](#)
75. Pelletier, C.; Valero, S.; Inglada, J.; Champion, N.; Sicre, C.M.; Dedieu, G. Effect of Training Class Label Noise on Classification Performances for Land Cover Mapping with Satellite Image Time Series. *Remote Sens.* **2017**, *9*, 173. [\[CrossRef\]](#)

76. Ma, L.; Liu, Y.; Zhang, X.; Ye, Y.; Yin, G.; Johnson, B.A. Deep Learning in Remote Sensing Applications: A Meta-Analysis and Review. *ISPRS J. Photogramm. Remote Sens.* **2019**, *152*, 166–177. [[CrossRef](#)]
77. Costa, H.; Machado, I.; Moreira, F.D.; Benevides, P.; Moraes, D.; Caetano, M. Exploring the Potential of Sentinel-2 Data for Tree Crown Mapping in Oak Agro-Forestry Systems. In Proceedings of the 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, Brussels, Belgium, 11–16 July 2021; pp. 5807–5810.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.