Article

Semantic Segmentation of Land Covers Using Deep Learning with a Pre-trained Backbone: A Case Study in in the Franciacorta Wine-growing Area

Girma Tariku * , Isabella Ghiglieno , Andres Sanchez Morchio , Luca Facciano , Celine Birolleau , Anna Simonetto , Ivan Serina , Gianni Gilioli

Posted Date: 27 June 2024

doi: 10.20944/preprints202406.1947.v1

Keywords: Land cover mapping; Semantic segmentation; Deep learning, Satellite imagery, pre trained backbone

Article

Semantic Segmentation of Land Covers Using Deep Learning with a Pre-Trained Backbone: A Case Study in in the Franciacorta Wine-growing Area

Girma Tariku [1,*], Isabella Ghiglieno [2], Andres Sanchez Morchio [2], Luca Facciano [2], Celine Birolleau [2], Anna Simonetto [2], Ivan Serina [1] and Gianni Gilioli [2]

[1] University of Brescia – Department of Information Engineering (DII), –via Branze 38, 25123, Brescia, Italy
[2] Agrofood Research Hub, University of Brescia – Department of Civil, Environmental, Architectural Engineering, and Mathematics – via Branze 43, 25123, Brescia, Italy
* Correspondence: g.tariku@unibs.it

Abstract: Land cover mapping, essential for understanding global land use patterns, relies on satellite imagery for monitoring changes, assessing ecosystem health, and supporting conservation ef- forts. However, significant challenges remain in managing large, complex satellite imagery datasets, acquiring specialized datasets due to high costs and labor intensity, and a lack of comparative studies for optimal deep learning model selection. Additionally, a scarcity of aerial datasets specifically tailored for agricultural areas exists. This study addresses these gaps by presenting a method for semantic segmentations of land covers in agricultural areas using satellite images and deep learning models with pre-trained backbones. We introduce an efficient methodology for preparing semantic segmentation datasets and contribute the "Land Cover Aerial Imagery" (LICAID) dataset for semantic segmentation. The study focuses on the Franciacorta area, Lombardy Region, leveraging the rich diversity of the dataset to effectively train and evaluate the models. We conduct a comparative study, employing cutting-edge deep learning-based segmentation models (U-Net, SegNet, DeepLabV3) with various pre-trained backbones (ResNet, Inception, DenseNet, EfficientNet) on our dataset acquired from Google Earth Pro. Through meticulous data acquisition, preprocessing, model selection, and evaluation, we demonstrate the effectiveness of these techniques in accurately delineating land cover classes. Integrating pre-trained feature extraction networks significantly improves performance across various metrics. Additionally, addressing challenges such as data availability, computational resources, and model interpretability is essential for advancing the field of remote sensing and supporting sustainable environmental stewardship worldwide.

Keywords: land cover mapping; semantic segmentation; deep learning; satellite imagery; pre trained backbone

1. Introduction

Land cover mapping serves as a foundational tool for gaining insights into the intricate patterns of land use across the globe. By accurately delineating various land cover types such as forests, agricultural lands, wetlands, and urban areas, it provides essential information for assessing ecosystem health, biodiversity conservation, and monitoring changes over time. Advancements in remote sensing technology have expanded the availability of imagery sources, with satellites offering diverse characteristics. Satellite imagery is essential for land cover mapping, providing detailed insights for monitoring changes, assessing ecosystem health, and supporting conservation efforts globally.

Classical machine learning techniques such as SVM [1], Decision Trees [2], and Random Forest [3] are foundational in land cover mapping. Maximum Likelihood Classification (MLC) categorizes land cover based on pixel probabilities but faces challenges like expert analysis and resolution limitations [4]. In seeking cost-effective solutions, Sentinel2 satellite imagery and machine learning techniques are utilized for land cover mapping in developing countries, despite computation time constraints and data processing bottlenecks [5]. High-resolution satellite imagery is emphasized for land cover mapping, and the limitations of using traditional classification methods are noted [6]. These authors showcase the effectiveness of Random Forests and Support Vector Machines with high-resolution satellite imagery data but also highlight sensitivity to classifier parameters and input features, as well as class imbalance issues, emphasizing the need for careful optimization. Urban planning and global challenges benefit from land cover data utilization, as shown by the study [7], despite persistent challenges in satellite imagery analysis, such as managing large volumes and complex classifications. To address the challenges associated with Very High-Resolution (VHR) imagery, an integrated OBIA and random forest approach has been proposed [8]. However, this approach also faces issues like spectral redundancy and the computational demands of incorporating temporal features.

Recent technological advances have driven the evolution of image analysis, notably with deep learning algorithms [9]. While traditional methods like object-based image analysis (OBIA) show effectiveness, they face challenges due to spectral signature similarities and class heterogeneity. Deep learning offers automatic feature extraction, exemplified by the TASSEL framework [10], which combines CNN-based feature extraction with object-based image analysis classification. Recent advancements in deep learning, particularly Convolutional Neural Networks (CNNs), offer significant potential for land cover classification and object detection using high-resolution imagery, as highlighted in a study [11]. However, challenges remain in enhancing the generalization and robustness of these models, necessitating the use of diverse training datasets. Additionally, a precise LULC mapping technique on Landsat 5/7 images using pre-trained neural networks is presented, potentially accelerating map production but limiting generalizability [12].

In land cover mapping with deep learning in computer vision, semantic segmentation excels by dividing images into meaningful segments, offering superior accuracy, especially in vegetation mapping [13–16]. To address the limitations of single-scale convolution kernels, a Multi-Scale Fully Convolutional Network (MSFCN) is introduced [17]. Similarly, the U-Net Temporal Attention Encoder (U-TAE) semantic segmentation approach is enhanced with temporal attention for improved classification [18]. Additionally, the Remote Sensing Segmentation Transformer (RSSFormer) is introduced, tailored for high spatial resolution remote sensing land cover segmentation, effectively tackling challenges like large-scale variation and imbalanced foreground-background distribution [19]. Meanwhile, CNNs are explored for per-pixel classification in high-resolution remote sensing, leveraging detailed data for precise object classification [20]. Despite advancements, challenges persist in acquiring specialized datasets due to high costs associated with data collection, demanding significant time, labor, and financial resources for activities like field surveys and manual annotation of vegetation classes.

Simultaneous segmentation for a comprehensive range of land cover classes in agricultural areas, including grasslands, arable land, herb-dominated habitats, hedgerows, vineyards, tree-dominated man-made habitats, and Olea europea groves. To address this gap, we introduce a unique land cover dataset with seven manually annotated classes suitable for semantic segmentation. We present a method for semantic segmentation of land covers in agricultural areas using satellite images and pre-trained deep learning models. The study focuses on the Franciacorta area located in Lombardy Region, Italy. Unlike existing works that lack rigorous comparisons to identify the optimal deep learning (DL)-based model for land cover semantic segmentation, we perform a comparative study of cutting-edge deep learning-based segmentation models. For this, we leverage a range of segmentation models: UNet, SegNet, and DeepLabV3, combined with four backbone convolutional neural networks (ResNet50, InceptionV2, DenseNet121, and EfficientNetB0).

Our method addresses research gaps by focusing on collecting sufficient image data from satellite imagery using Google Earth Pro and preparing mask images for a comprehensive land cover analysis in specific regions. In contrast to previous studies, our method provides a detailed approach on how to gather a sufficiently enough dataset for semantic segmentation. The main contributions can be summarized as follows:

- We develop a cost-effective and efficient methodology for preparing semantic segmentation datasets, mitigating the high costs associated with data collection.
- We contribute to the field by presenting a unique land cover dataset manually annotated for seven distinct classes, addressing the segmentation of grasslands, arable land, herb-dominated habitats, hedgerows, vineyards, tree-dominated man-made habitats, and Olea europea groves simultaneously.
- Through meticulous data acquisition, preprocessing, and model selection, we demonstrate the effectiveness of deep learning models such as UNet, SegNet, and DeepLabV3 in accurately delineating land cover classes.
- Conduct a comparative analysis of three semantic segmentation models, each with different backbones, to identify the most suitable model for aerial imagery mapping, thereby improving the accuracy and efficiency of land cover mapping efforts.
- The results of our experiments underscore the importance of incorporating deep learning techniques and a pre-trained backbone in land cover mapping applications, offering scalable and efficient solutions for environmental monitoring and conservation efforts.

2. Materials and Methods

A. Study Site

This study focuses on the Franciacorta area, a famous Italian wine-growing region located in Lombardy, Italy (as shown in Figure 1). Franciacorta, nestled in the picturesque province of Brescia, is renowned for its exquisite landscapes, rich history, and world-class wine production.

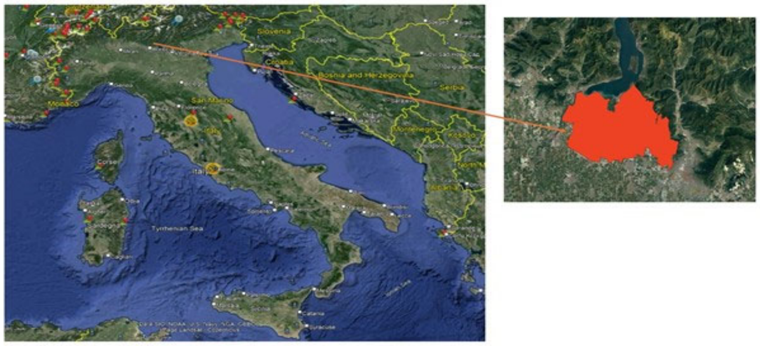

Figure 1. Geographical Context of the study area of Franciacorta, Lombardy Region, Italy. Left: Overview showing the position of the Lombardy Region map within Italy. Right: Close-up view displaying a detailed perspective of the study area of Franciacorta.

B. Data Acquisition

For the sake of maximum diversity of the dataset, we manually selected 18 orthophoto tiles from different places in Franciacorta. The following steps were taken to prepare the image and the corresponding mask image:

Step 1: The process initiates with the acquisition of satellite imagery from Google Earth Pro, capturing both the visual image and its accompanying shape file in .kmz format. This imagery encompasses a broad spatial scope and provides a comprehensive view of the target area. Subsequently, the georeferencing of the acquired imagery is executed using ArcGIS software, a crucial step in aligning the satellite image with geographic coordinates for accurate spatial analysis.

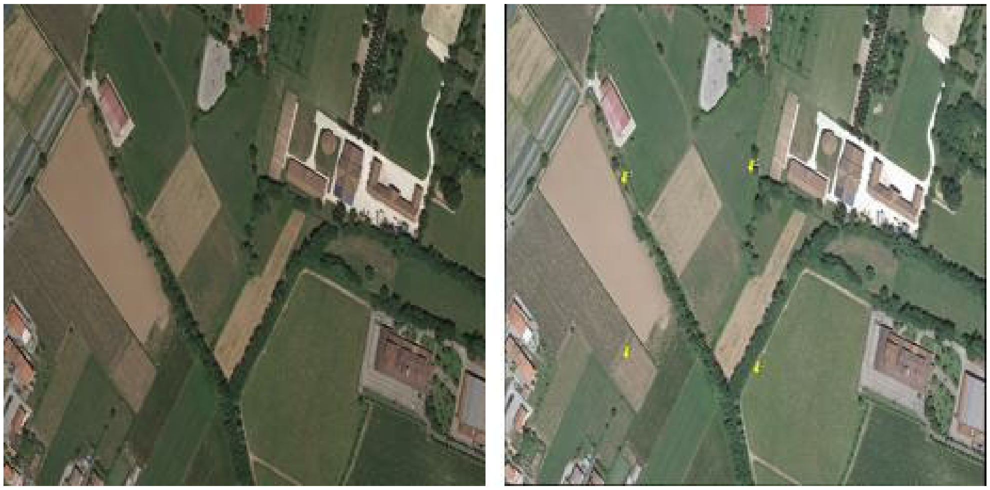

Figure 2. Satellite imagery acquisition and georeferencing process.

Step 2: Once georeferencing is completed, the imagery undergoes segmentation using the multiresolution segmentation tool within the eCognition software platform. This segmentation technique partitions the image into homogeneous regions based on similarities in spectral and spatial characteristics, facilitating the delineation of distinct features and objects within the landscape.

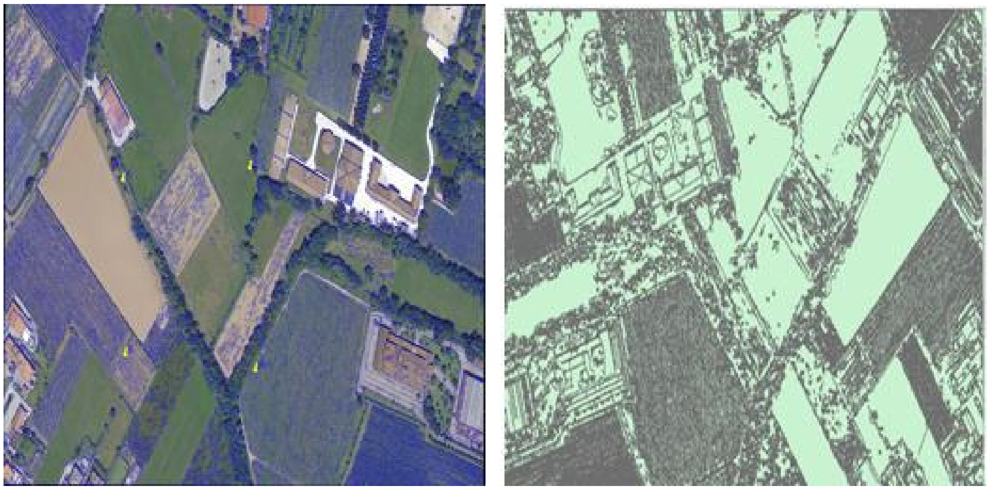

Figure 3. Segmentation Process Using Multiresolution Segmentation in eCognition Software.

Step 3: Following segmentation, the resulting segmented shape file and image are subject to meticulous examination and validation by a knowledgeable plant expert. Leveraging the capabilities of QGIS software, the expert meticulously inspects each segmented polygon, assigning appropriate plant names based on their botanical characteristics and distribution within the landscape.

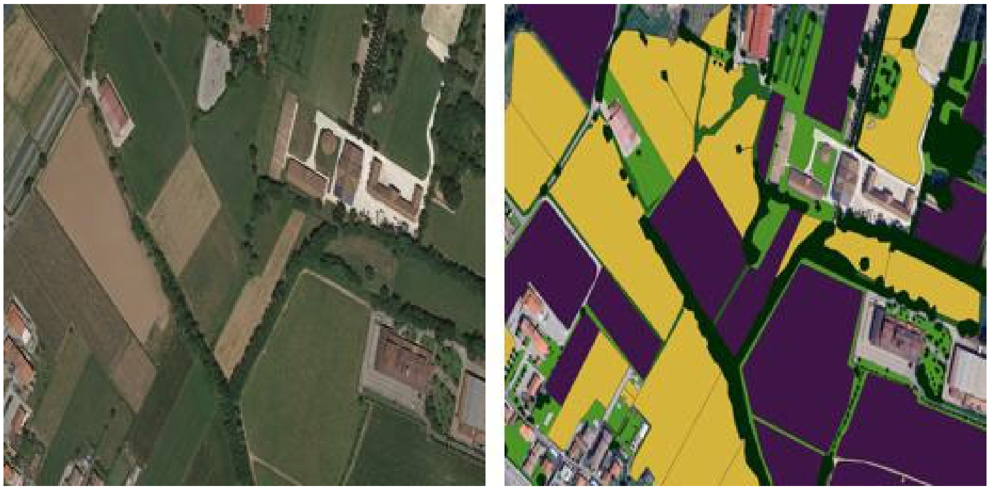

Figure 4. Validation Process of Segmented Shape File in QGIS: On the left, the original image; on the right, segmented classes are represented. Green denotes grasslands, yellow indicates arable land, medium-dark green suggests herb-dominated habitats, purple represents vineyards, and black signifies tree-dominated man-made habitats. Additionally, brown indicates Olea europaea groves.

Classes:

1. Grassland: A type of habitat dominated by grasses, with few or no trees. Grasslands can be found in various regions and climates, from tropical to temperate.
2. Arboreal land: Refers to land or habitats that are predominantly covered with trees. These areas may include forests, woodlands, and other tree-dominated ecosystems.
3. Herb-dominated habitats: Habitats where herbs, or nonwoody plants, are the dominant vegetation. These habitats can range from meadows and prairies to marshes and wetlands.
4. Hedgerows: Linear strips of vegetation, typically consisting of shrubs, small trees, and grasses, often used to mark boundaries or provide wildlife habitat in agricultural landscapes.
5. Vineyards: Agricultural landscapes specifically cultivated for growing grapevines, typically for wine production. Vineyards can vary in size and management practices.
6. Tree-dominated man-made habitats: Human-modified landscapes where trees are the predominant vegetation, such as urban parks, orchards, and landscaped gardens.
7. Olea europaea groves: Groves or orchards of olive trees, primarily cultivated for the production of olives and olive oil. These groves are commonly found in Mediterranean regions

3. Experiment

In this work, the original images were large in pixel size. This presented a challenge for deep learning models, as large image sizes require significant memory and can slow down the training process. To address this challenge, we prepared a Python code that efficiently processes the large images. The code starts by iterating through the directories and subdirectories within a specified root directory, which contains both image and mask files. For each image and mask file found, the code resizes them to the nearest size divisible by a predefined patch size. The resized images and masks are then divided into non-overlapping patches using the patchify library from the TensorFlow function, ensuring each patch fits the specified dimensions. These patches are saved as individual image files, preserving their association with their corresponding masks as shown in Figure 5. Additionally, the code utilizes the split folders library from Tensorflow to split the patched images and masks into training and validation datasets with an 80-20 split ratio, respectively. The resulting split datasets are saved in separate directories, with the training dataset comprising 612 images and the testing dataset comprising 153 images. This streamlined process facilitates efficient model training and evaluation.

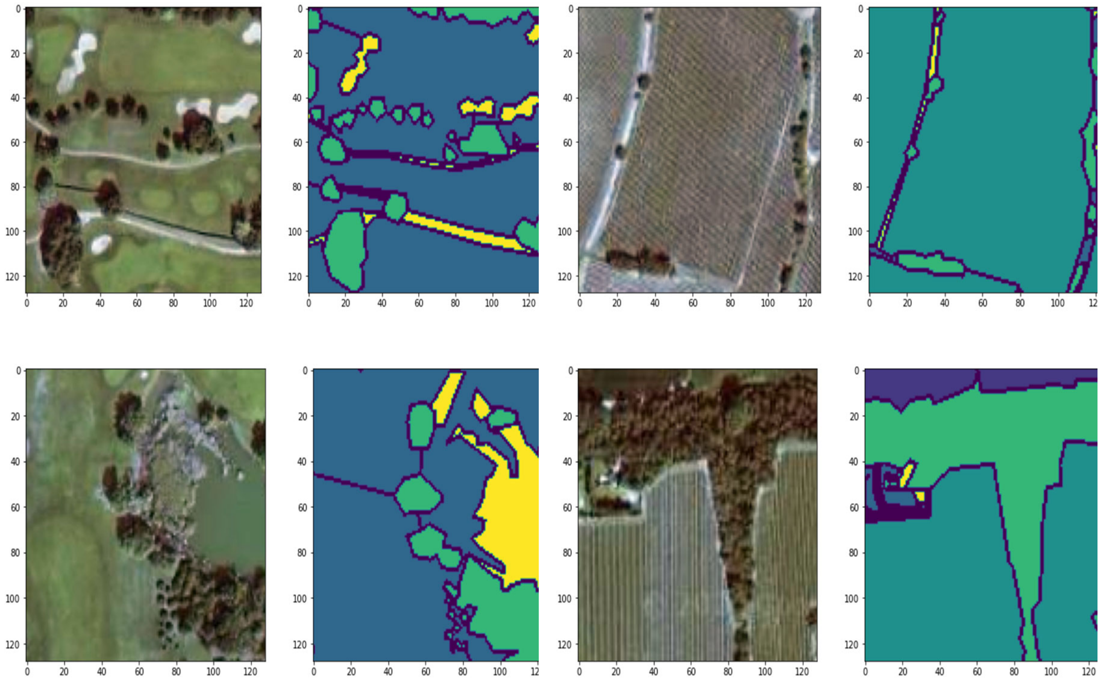

Figure 5. Illustration of image patches extracted from the original images along with their corresponding masks.

In general, the following steps were taken as shown in Figure 6 to perform semantic segmentation:

Step 1: Collect the image and the corresponding mask image from QGIS software.

Step 2: We partition the larger training and mask images into smaller segments, each with a size of 128x128 pixels. These smaller segments are used to train the U-Net segmentation model. The purpose of this model is to perform image segmentation and identify specific features within the images. Once the U-Net segmentation model has been trained successfully, we save the model for future use.

Step 3: Save the model to predict and map large images.

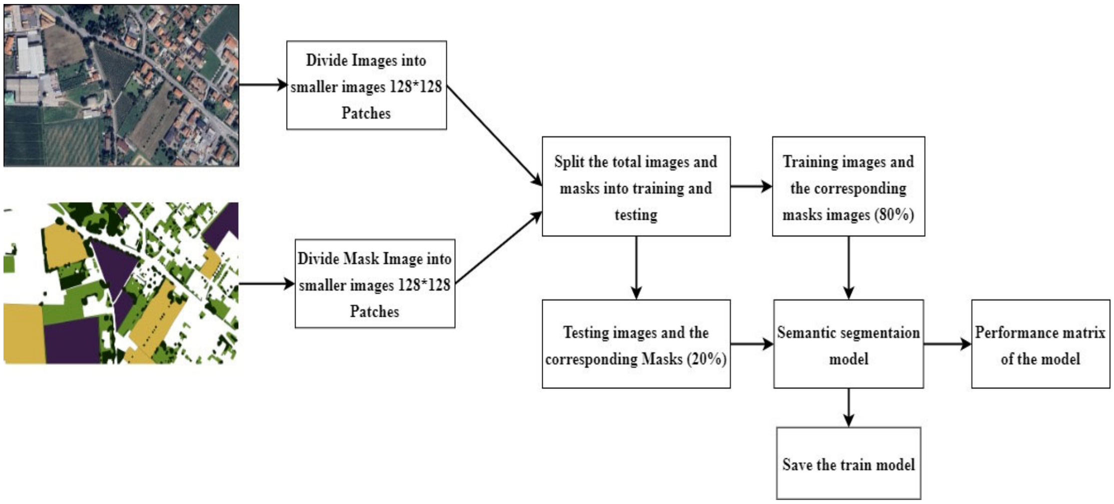
Divide Images into smaller images 128*128 Patches
Divide Mask Image into smaller images 128*128 Patches
Split the total images and masks into training and testing
Training images and the corresponding masks images (80%)
Testing images and the corresponding Masks (20%)
Semantic segmentaion model
Performance matrix of the model
Save the train model

Figure 6. Illustrates the General Workflow of Semantic Segmentation for Large Images.

doi:10.20944/preprints202406.1947.v1

3.1. Backbones

In our study, we performed two approaches to semantic segmentation: with and without leveraging a backbone. Semantic segmentation without a backbone involves training models from scratch, without the use of pre-trained networks. This method relies solely on the model's architecture to extract features directly from the input data. While this approach offers independence from pre-existing biases and data patterns embedded in pre-trained networks, it typically requires more extensive training and larger datasets to achieve comparable performance to models with pre-trained backbones.

Conversely, leveraging a pre-trained backbone is a widely adopted strategy aimed at enhancing model performance in semantic segmentation tasks. Backbones, such as ResNet, InceptionV2, DenseNet, and EfficientNet, are well-established architectures in deep-learning CNNs. These models are pre-trained on large-scale datasets (like ImageNet) and have learned to extract meaningful features from images, encompassing spatial relationships and contextual information crucial for accurate segmentation.

By fine-tuning these pre-trained backbones for specific segmentation tasks, models can efficiently adapt to new datasets and extract relevant features tailored to the nuances of land cover classification. This process leverages transfer learning, where knowledge gained from previous tasks (e.g., image classification) is transferred to improve performance on the current segmentation task. This approach not only accelerates model convergence during training but also enhances segmentation accuracy by leveraging the rich feature representations learned by the backbone networks.

A. ResNet

ResNet [21], or Residual Network, is a pioneering convolutional neural network architecture devised to tackle the challenge of vanishing gradients in deep networks. Developed by Microsoft Research, ResNet introduces residual connections, enabling the training of exceptionally deep networks by learning residual functions. These connections address the degradation problem associated with increasing network depth, resulting in state-of-the-art performance across various computer vision tasks.

B. Inception

InceptionV3 [22], an evolution of Google's Inception architecture, is renowned for its computational efficiency and high accuracy. It introduces several novel features, including the inception module, which enables the network to efficiently capture features at multiple scales through parallel convolutions. InceptionV3 employs factorized convolutions, reducing computational cost while maintaining expressive power. Additionally, it incorporates batch normalization and auxiliary classifiers, aiding in both convergence speed and regularization.

C. DensNet

DenseNet [23], short for "Densely Connected Convolutional Networks," is a convolutional neural network architecture known for its densely connected layers. Unlike traditional architectures where each layer is connected only to the subsequent layers, DenseNet introduces dense connections, where each layer receives direct inputs from all preceding layers. This design fosters feature reuse and facilitates the flow of gradients, leading to improved parameter efficiency and better gradient propagation.

D. EfficientNet

EfficientNet is a family of convolutional neural network architectures developed by [24]. It is specifically designed to achieve state-of-the-art accuracy while simultaneously being highly efficient in terms of computational resources. The key innovation behind EfficientNet is the compound scaling method, which uniformly scales network depth, width, and resolution in a systematic manner. This approach ensures that the network's parameters and computational cost are optimized for a given resource constraint, resulting in models that are both accurate and efficient across a wide range of tasks. In our study, we specifically utilized EfficientNetB0, which is the smallest variant in the EfficientNet family. EfficientNetB0 strikes a balance between model complexity and performance,

doi:10.20944/preprints202406.1947.v1

making it suitable for various computer vision tasks, including image classification, object detection, and semantic segmentation.

3.2. Semantic Segmentation Models

A. UNet

The UNet [25] semantic segmentation model is a deep learning architecture designed for pixel-wise classification tasks, particularly in image segmentation. In this sample, we utilize the UNet model to perform semantic segmentation on UAV RGB images. The model employs a symmetrical encoder-decoder structure, with skip connections between corresponding encoder and decoder layers to preserve spatial information. This aids in capturing both local and global features, making it particularly effective for tasks like object detection and boundary delineation.

Our UNet semantic segmentation model, illustrated in Figure 7, is designed to process 128x128 pixel RGB images with a structured architecture aimed at capturing intricate spatial details. The model architecture is structured into two main sections: a contraction path (c1 to c5) and an expansive path (u6 to u9).

In the contraction path, successive convolutional blocks progressively increase the number of filters to extract and encode features from the input image. Max-pooling layers (p1 to p4) are strategically placed to downsample the spatial dimensions, thereby reducing computational load while preserving essential features. At the bottleneck layer (c5), the model captures high-level context and semantic information crucial for accurate segmentation.

Conversely, the expansive path of the UNet model focuses on upsampling the feature maps using transpose convolutional layers. These layers help recover spatial resolution lost during the contraction path and enable the model to reconstruct detailed information. Each block in the expansive path (c6 to c9) sequentially reduces the number of filters, allowing the model to refine segmentation boundaries and enhance localization accuracy.

`The final layer of the UNet model (outputs) employs a 1x1 convolutional layer with a sigmoid activation function, facilitating pixel-wise predictions for each class in the segmentation task. This setup enables the model to output probability maps where each pixel indicates the likelihood of belonging to a specific land cover class.

During training, the model is compiled with the Adam optimizer, which efficiently adjusts learning rates for each parameter during training. The choice of binary cross-entropy loss function aligns with the pixel-wise nature of the segmentation task, optimizing the model to produce accurate binary predictions for each pixel. The accuracy metric provides insights into the model's performance by measuring the proportion of correctly classified pixels compared to the total number of pixels.

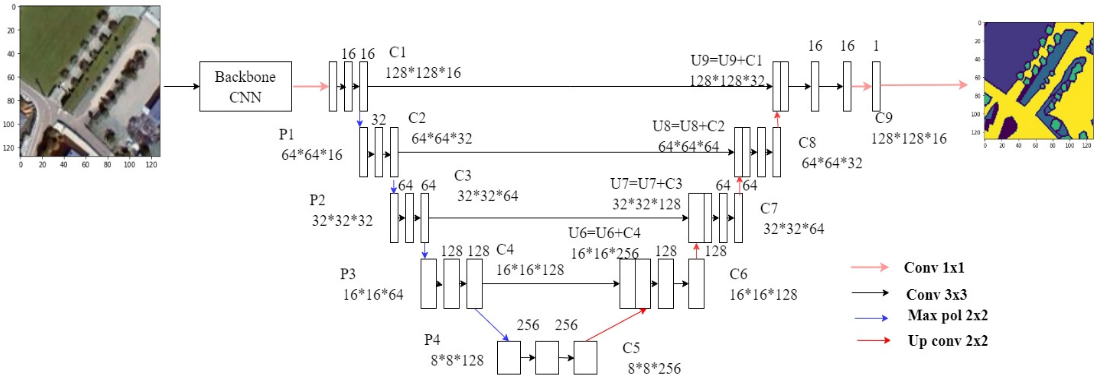

Figure 7. The architecture of the UNet Semantic Segmentation Model with backbone.

B. SegNet

SegNet [26] is a convolutional neural network (CNN) architecture specifically tailored for semantic segmentation tasks, where the goal is to assign a class label to each pixel in an image. It features an encoder-decoder structure, which comprises an encoding path to extract hierarchical features from the input image and a decoding path to generate a pixel-wise segmentation map. Our SegNet semantic segmentation model, depicted in Figure 8, is optimized with the EfficientNetB0 architecture serving as its foundational backbone, omitting fully connected layers to streamline feature extraction from input images. The model is structured around an encoder-decoder architecture, each fulfilling distinct roles in the segmentation process.

The encoder, or contracting path, initiates with convolutional layers followed by batch normalization and max-pooling operations. This sequence progressively condenses spatial dimensions while effectively capturing essential features from the input images. By utilizing max-pooling layers, the encoder enhances computational efficiency by reducing the complexity of feature maps without compromising significant visual information.

In contrast, the decoder, or expansive path, employs up-sampling layers to restore the spatial resolution of feature maps. These layers are crucial for recovering fine-grained details lost during the down-sampling process in the encoder. Furthermore, the decoder integrates features extracted from the EfficientNetB0 backbone with the up-sampled feature maps to refine segmentation accuracy. This fusion of hierarchical features ensures that the model can leverage both high-level semantic information and detailed spatial context for precise pixel-wise predictions. Subsequent convolutional layers further refine the features, and the output layer generates pixel-wise predictions through a softmax activation function. The model, constructed with the backbone input and output layer, is designed to classify pixels in images.

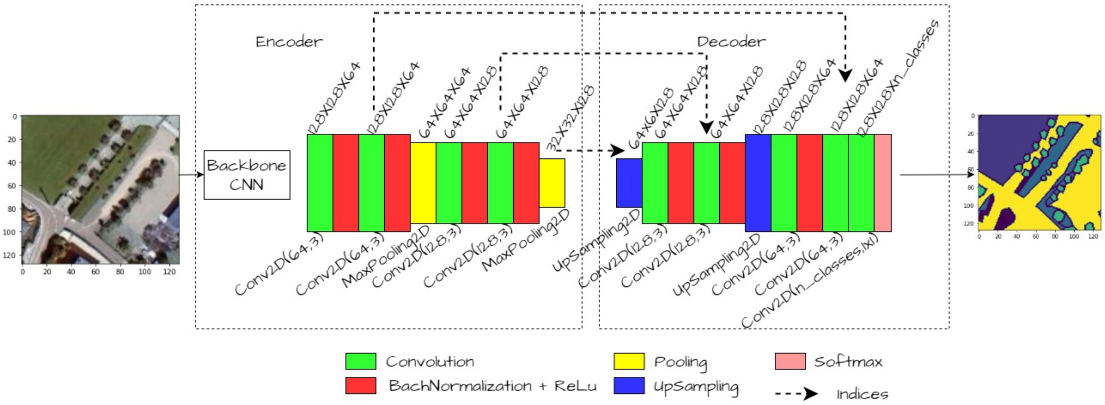

Figure 8. Block diagram of SegNet with backbone CNN as an encoder and decoder.

C. DeepLab

DeepLabV3[27] is a cutting-edge convolutional neural network architecture tailored for semantic image segmentation. It leverages atrous convolution to capture multi-scale contextual information efficiently. Key features include a feature pyramid network for hierarchical feature extraction, atrous spatial pyramid pooling for multi-scale feature aggregation, and efficient up-sampling methods for high-resolution segmentation maps. Our DeepLabV3+ model, depicted in Figure 9, utilizes an EfficientNetB0 backbone for feature extraction without fully connected layers, serving as the encoder. This encoder processes input images and extracts features, which are then passed to the decoder. The decoder consists of an Atrous Spatial Pyramid Pooling (ASPP) module followed by global average pooling and concatenation of low-level features. The ASPP module employs convolutional layers with varying dilation rates to capture multi-scale contextual information. Additionally, global average pooling is performed to capture global context information. The decoder then combines the ASPP outputs, global context, and low-level features using concatenation. Further convolutional layers refine the features, followed by up-sampling layers to restore spatial information. Finally, a 1x1 convolutional layer with SoftMax activation generates pixel-wise predictions for semantic segmentation. The model is compiled with the Adam optimizer, employing categorical cross-entropy loss and accuracy as evaluation metrics.

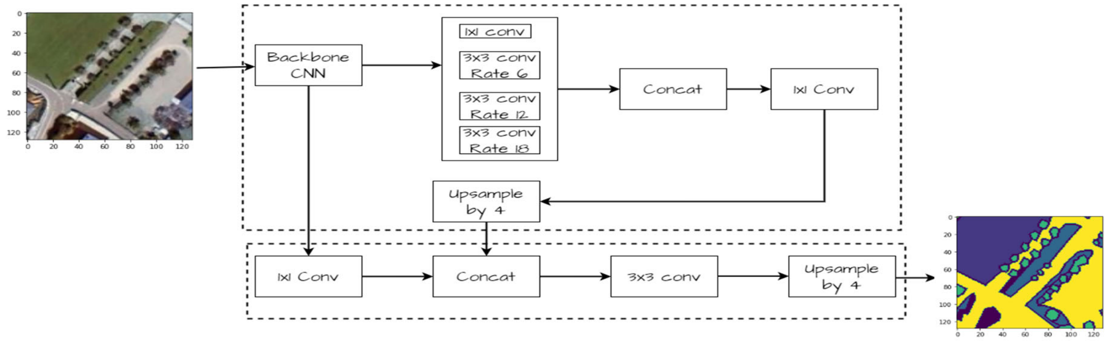

Figure 9. Block diagram of DeeplabV3 with backbone CNN as the encoder and decoder.

We implemented U-Net, SegNet, and DeepLabV3 segmentation models for the semantic segmentation of satellite images. Initially, the image and mask datasets undergo preprocessing by patchifying them into smaller patches to facilitate training. Each patch is then scaled using Min-Max scaling, and the RGB masks are converted into integer labels based on predefined RGB values. These models are constructed using the EfficientNetB0 backbone, with and without encoder and decoder paths. A SoftMax activation function is applied at the output layer for multi-class segmentation. Following compilation with the Adam optimizer and categorical cross-entropy loss function, training commences for 100 epochs with a batch size of 16. Evaluation is conducted using accuracy and IoU metrics on both training and validation datasets, and the trained model is saved for future use. Performance metrics, including accuracy, precision, recall, F1 score, and Jaccard coefficient, are computed, and the segmentation performance is visually demonstrated on randomly selected test images.

4. Result

Our semantic segmentation models (Unet, SegNet, and DeepLabV3) utilize TensorFlow and the segmentation model's library to perform image segmentation. The dataset consists of a total of 765 images, divided into training and testing sets with a ratio of 0.80 and 0.20, respectively. Each image is annotated with one of seven classes: grasslands, arable land, herb-dominated habitats, hedgerows, vineyards, tree-dominated man-made habitats, and Olea europea groves. Initially, the images and their corresponding masks are loaded from the specified directories and pre-processed accordingly. The masks are encoded using label encoding and then split into training and testing datasets. Next, the model is configured with specific parameters, including the EfficientNet backbone architecture, and a combination of Dice loss and Categorical Focal loss as the optimization objective. The model is trained over 100 epochs, and the training progress is visualized using plots for loss and IOU score. After training, the model is saved for future use. Performance metrics such as accuracy, precision, recall, F1 score, Jaccard score, and Mean IoU are computed using the trained model on the test dataset. Finally, the results are printed to evaluate the segmentation model's performance.

- Accuracy: This metric measures the overall correctness of the segmentation by calculating the ratio of correctly predicted pixels to the total number of pixels.
- Precision: Precision quantifies the model's ability to correctly identify positive predictions among all predicted positives. It's calculated as the ratio of true positives to the sum of true positives and false positives.
- Recall: Recall, also known as sensitivity, measures the ability of the model to detect all relevant instances of the class in the image. It's calculated as the ratio of true positives to the sum of true positives and false negatives.
- F1 Score: The F1 score is the harmonic mean of precision and recall. It provides a balanced measure between precision and recall and is calculated as 2 * (precision * recall) / (precision + recall).

- Mean IoU: Mean IoU calculates the average IoU across all classes. It's a popular metric for semantic segmentation tasks as it provides an overall measure of segmentation accuracy across different classes.
- Jaccard Score (IoU): The Jaccard score, or Intersection over Union (IoU), measures the ratio of the intersection of the predicted and ground truth segmentation masks to their union. It evaluates the overlap between the predicted and ground truth regions.

Table 1 presents a comparative analysis of performance metrics for three semantic segmentation models Unet, SegNet, and DeeplabV3 under two distinct conditions: with and without backbone integration. These metrics serve as quantitative indicators of model effectiveness in accurately segmenting images. Across all models, the incorporation of backbone architecture consistently leads to improvements in performance metrics. Specifically, models equipped with backbone demonstrate enhanced accuracy, precision, recall, and F1 scores compared to their counterparts lacking backbone integration. Accuracy, a measure of correct pixel classification, ranges from 0.638 to 0.763 with backbone and from 0.667 to 0.762 without. Similarly, precision, recall, and F1 scores exhibit higher values when backbone architecture is utilized. Additionally, metrics like the Jaccard Coefficient (IOU) and Mean IOU, which evaluate the overlap between predicted and ground truth masks, show substantial improvements with backbone integration. These findings underscore the importance of incorporating pre-trained feature extraction networks, or backbones, to enhance the performance of semantic segmentation models, thereby advancing their applicability in diverse image analysis tasks.

Table 1. Performance metrics comparison of UNet, SegNet, and DeeplabV3 with and without a backbone.

Performance metrics	UNet		SegNet		DeeplabV3	
	Without backbone	With backbone	Without backbone	With backbone	Without backbone	With backbone
Accuracy	0.574	0.653	0.564	0.673	0.681	0.763
Precision	0.590	0.657	0.563	0.678	0.6855	0.761
Recall	0.574	0.653	0.566	0.673	0.681	0.763
F1 score	0.573	0.646	0.559	0.672	0.674	0.756
Jaccard Coefficient (IOU)	0.411	0.500	0.399	0.528	0.5308	0.626
Mean IOU	0.306	0.323	0.290	0.407	0.410	0.520

Based on the provided performance metrics, it can be observed that the DeeplabV3 model consistently outperforms Unet and SegNet models across the metrics, especially when backbone integration is considered. With backbone integration, DeeplabV3 achieves the highest values for accuracy, precision, recall, F1 score, Jaccard Coefficient (IOU), and Mean IOU among the three models. This indicates that DeeplabV3, particularly when equipped with backbone architecture, offers superior image segmentation capabilities compared to Unet and SegNet. Therefore, in terms of overall performance and effectiveness in segmenting images, DeeplabV3 stands out as the preferred model choice.

Our investigation also evaluated the effectiveness of various pre-trained backbones for semantic segmentation of land cover using DeepLabV3. As shown in Table 2, Efficient NetB0 emerged as the leader, achieving the highest overall accuracy (76.35%) and F1-score (75.60%). This indicates its superior ability to accurately segment land cover classes in the imagery. Resnet-34 followed closely with an accuracy of 70.83% and F1-score of 70.77%. While InceptionV3 and DenseNet exhibited lower overall accuracy (around 71- 75%), they maintained good precision and recall values for land cover class segmentation (refer to Table 2 for detailed results). Mean Intersection-over-Union (mIOU) and Jaccard Score mirrored these trends, with EfficientNetB0 achieving the highest values (52.05% and 62.60%, respectively). These findings highlight the importance of selecting an appropriate pre-trained backbone for DeepLabV3 in land cover segmentation tasks. The optimal choice can significantly impact the model's ability to differentiate and delineate land cover classes.

Table 2. Comparison of Semantic Segmentation Performance using DeepLabV3 with Different Pre-trained Backbones.

Backbone	Accuracy	Precision	Recall	F1 score	Mean IOU	Jaccard Coefficient (IOU)
Resnet 34	70.83	73.29	70.84	70.77	46.32	56.90
EfficientNetB0	76.33	76.10	76.30	75.60	52.00	62.60
InceptionV3	71.46	74.00	71.46	72.01	47.50	61.40
Densest	75.07	76.18	75.05	74.66	50.70	61.75

After loading the trained model, we generate a batch of test images and their corresponding masks using the validation data generator. Next, we use the loaded model to compute predictions for the test images. We then convert the predicted masks from categorical format to integer format for visualization and IoU calculation. Finally, to qualitatively assess the model's performance, we visualize a randomly selected test image along with its ground truth mask and predicted mask, as shown in Figure 10.

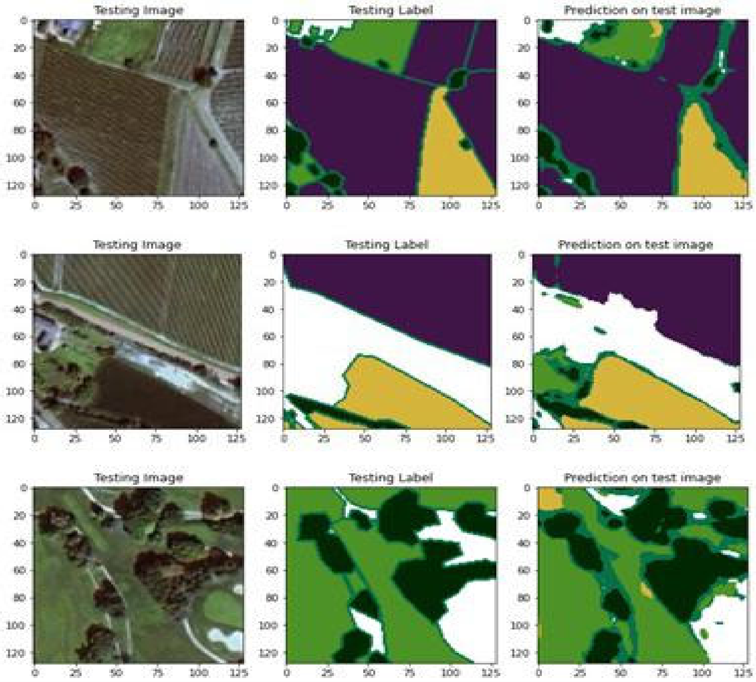

Figure 10. Visualization of a randomly selected test image, ground truth mask, and predicted mask. The image illustrates the qualitative assessment of the model's performance, showing the alignment between the original test image, its corresponding ground truth mask, and the predicted mask generated by the model.

Finally, we applied the trained model to a large image (as shown in Figure 11). To achieve this, we segmented the image into patches of an appropriate size for processing. The prediction process was then carried out on each patch. Finally, the resulting segmented patches were stitched together to reconstruct the predicted mask for the entire large image. This approach facilitates the application of the model to images beyond the validation dataset's size.

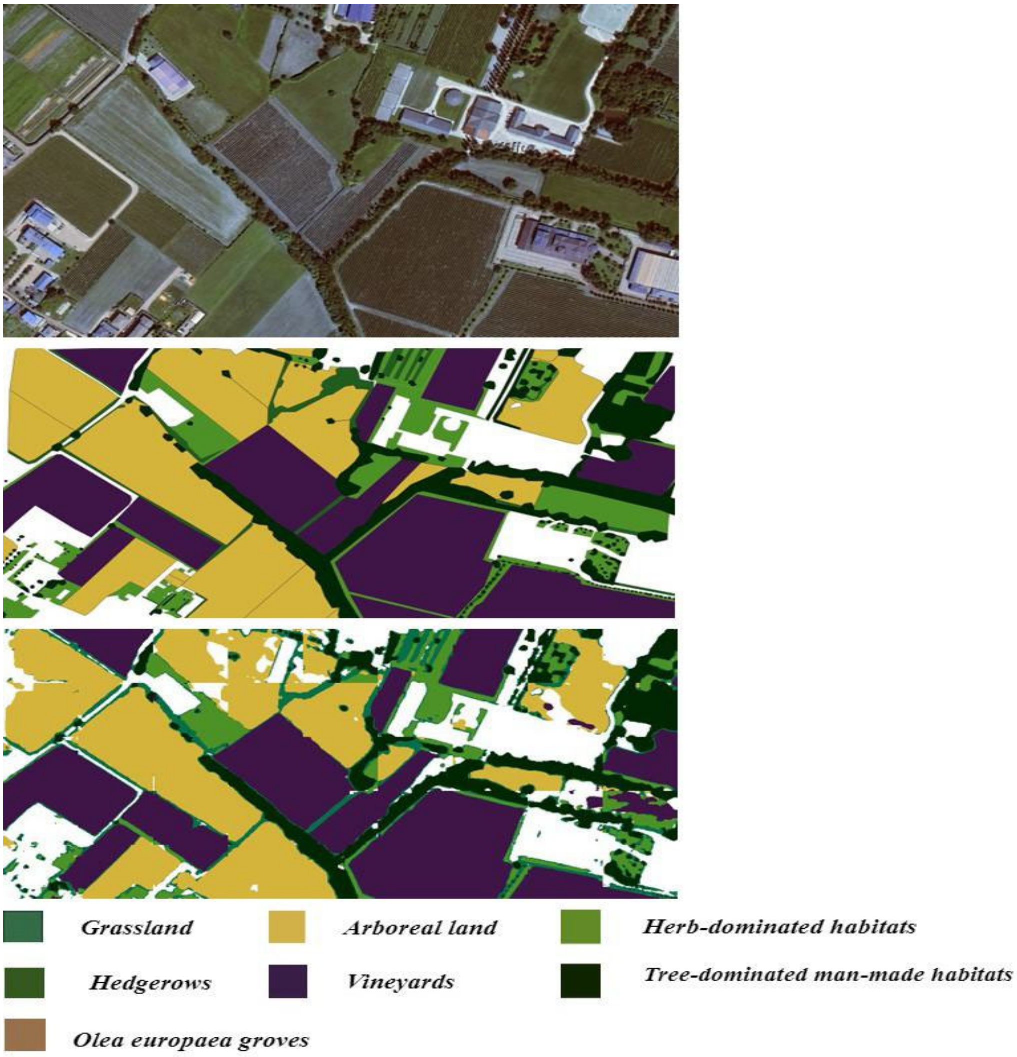

Figure 11. This figure illustrates the original image, ground truth mask, and the final segmented image mask obtained by applying a trained model. The model segments the large image into patches, generates predictions for each patch, and seamlessly stitches the segmented patches together to reconstruct the predicted mask for the entire large image. The colors represent different land cover classes: green denotes grasslands, yellow indicates arable land, medium-dark green suggests herb-dominated habitats, purple represents vineyards, black signifies tree-dominated man-made habitats, and brown indicates Olea europaea groves.

5. Discussion

In this study, we have addressed key challenges and contributed significantly to the field of land cover mapping in agricultural areas. By focusing on semantic segmentation, using satellite imagery and deep learning models with a transfer learning backbone, we have provided insights into the effectiveness of advanced techniques in accurately delineating land cover classes.

Firstly, we introduced a unique land cover dataset annotated for seven distinct classes, including grasslands, arable lands, herb-dominated habitats, hedgerows, vineyards, tree-dominated man-made habitats, and Olea europea groves. This dataset fills a crucial gap in the existing literature [19,20] by simultaneously addressing the segmentation of multiple land cover types essential for comprehensive agricultural analysis.

However, our approach also addresses a drawback highlighted in previous studies [5,7,8]. While traditional land cover mapping techniques have utilized classical machine learning algorithms and object-based image analysis (OBIA), they often struggle with spectral signature similarities, class heterogeneity, and limited generalizability. Additionally, acquiring specialized datasets for semantic segmentation has been challenging due to the high costs associated with data collection and manual annotation.

With careful data gathering, preparation, and model selection, we proved that deep learning models like UNet, SegNet, and DeepLabV3 excel at accurately mapping different land cover types from satellite imagery. By integrating pre-trained feature extraction networks, or backbones, into these models, we observed significant improvements in segmentation performance across various metrics. Our findings underscore the importance of leveraging advanced methodologies to achieve more accurate and reliable land cover mapping results. These advanced methodologies offer scalable and efficient solutions for environmental monitoring, conservation, and land management. By harnessing the power of satellite imagery and deep learning models, we can gain valuable insights into ecosystem health, biodiversity conservation, and land use dynamics.

Looking ahead, further research could focus on enhancing the robustness and generalizability of deep learning models for land cover mapping. This includes incorporating additional datasets, refining model architectures, and exploring novel techniques for feature extraction and classification. Additionally, efforts to address challenges such as data availability, computational resources, and model interpretability will be crucial for advancing the field of remote sensing and supporting sustainable environmental stewardship worldwide.

6. Conclusions

This study has addressed critical challenges in land cover mapping by leveraging satellite imagery and advanced deep learning techniques. We introduced a cost-effective methodology for preparing semantic segmentation datasets, which significantly reduces the high costs typically associated with data collection. Our contribution includes the development of the Land cover aerial imagery dataset specifically tailored for the Franciacorta, featuring seven manually annotated land cover classes crucial for agricultural areas.

Through meticulous data acquisition, preprocessing, and model selection, we demonstrated the efficacy of deep learning models UNet, SegNet, and DeepLabV3 enhanced by various pre-trained backbones like ResNet, Inception, DenseNet, and EfficientNet. Our comparative analysis highlighted that DeepLabV3 consistently outperforms other models across multiple performance metrics, underscoring its suitability for aerial imagery mapping and land cover classification tasks.

This research underscores the transformative potential of integrating deep learning techniques with satellite imagery for scalable and efficient environmental monitoring and conservation efforts. Moving forward, further enhancements could focus on refining model architectures, incorporating additional datasets to improve generalizability, and addressing computational challenges to facilitate broader adoption in remote sensing applications. By advancing these methodologies, we aim to support sustainable land management practices and contribute to global efforts in biodiversity conservation and ecosystem health assessment.

Author Contributions: Conceptualization, Girma Tariku, Gianni Gilioli, Ivan Serina and Isabella Ghiglieno; methodology, Girma Tariku; software, Girma Tariku, Andres Sanchez Morchio, Luca Facciano, and, Celine Birolleau; validation, Isabella Ghiglieno, Gianni Gilioli, Ivan Serina, and Anna Simonetto; formal analysis, Girma Tariku, Isabella Ghiglieno, Gianni Gilioli, and Ivan Serina; investigation, Girma Tariku; resources, Girma Tariku, Andres Sanchez Morchio, and Celine Birolleau; data curation, Girma Tariku, and Celine Birolleau; writing—original draft preparation, Girma Tariku; writing—review and editing, Girma Tariku, Isabella Ghiglieno, Gianni Gilioli, and Ivan Serina; visualization, Isabella Ghiglieno, and Gianni Gilioli; supervision, Gianni Gilioli; project administration, Gianni Gilioli. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been partially supported by "Fondazione Cariplo" (Italy) and "Regione Lombardia" (Italy) under the project: Progetto "Biodiversità, suolo e servizi ecosistemici: Metodi e tecniche per food system robusti, resilienti e sostenibili" – Bando Emblematici Maggiori 2020

Data Availability Statement: The dataset presented in this study is available at the link.

Acknowledgments: Not available.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Pal M, Mather PM. Support vector machines for classification in remote sensing. International Journal of Remote Sensing [Internet]. 2005 Mar 1 [cited 2024 Feb 20];26(5):1007–11. Available from: https://doi.org/10.1080/01431160512331314083
2. Pal M, Mather PM. An assessment of the effectiveness of decision tree methods for land cover classification. Remote Sensing of Environment [Internet]. 2003 Aug 30 [cited 2024 Feb 20];86(4):554–65. Available from: https://www.sciencedirect.com/science/article/pii/S0034425703001329
3. Cutler DR, Edwards Jr. TC, Beard KH, Cutler A, Hess KT, Gibson J, et al. Random Forests for Classification in Ecology. Ecology [Internet]. 2007 [cited 2024 Feb 20];88(11):2783–92. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1890/07-0539.1
4. Gauci A, Abela J, Austad M, Cassar LF, Zarb Adami K. A Machine Learning approach for automatic land cover mapping from DSLR images over the Maltese Islands. Environmental Modelling & Software [Internet]. 2018 Jan 1 [cited 2024 Apr 2];99:1–10. Available from: https://www.sciencedirect.com/science/article/pii/S136481521630487X
5. Mardani M, Mardani H, De Simone L, Varas S, Kita N, Saito T. Integration of Machine Learning and Open Access Geospatial Data for Land Cover Mapping. Remote Sensing [Internet]. 2019 Jan [cited 2024 Apr 2];11(16):1907. Available from: https://www.mdpi.com/2072-4292/11/16/1907
6. Pelletier C, Valero S, Inglada J, Champion N, Dedieu G. Assessing the robustness of Random Forests to map land cover with high-resolution satellite image time series over large areas. Remote Sensing of Environment [Internet]. 2016 Dec 15 [cited 2024 Apr 2];187:156–68. Available from: https://www.sciencedirect.com/science/article/pii/S0034425716303820
7. Ouchra, H.; Belangour, A.; Erraissi, A. Machine Learning Algorithms for Satellite Image Classification Using Google Earth Engine and Landsat Satellite Data: Morocco Case Study. *IEEE Access* **2023**, *11*, 71127–71142. Available online: https://ieeexplore.ieee.org/document/10177754 (accessed on 2 April 2024).
8. Han R, Liu P, Wang G, Zhang H, Wu X. Advantage of Combining OBIA and Classifier Ensemble Method for Very High-Resolution Satellite Imagery Classification. Journal of Sensors [Internet]. 2020 Nov 25 [cited 2024 Apr 2];2020: e8855509. Available from: https://www.hindawi.com/journals/js/2020/8855509/
9. Ma L, Liu Y, Zhang X, Ye Y, Yin G, Johnson BA. Deep learning in remote sensing applications: A meta-analysis and review. ISPRS Journal of Photogrammetry and Remote Sensing [Internet]. 2019 Jun 1 [cited 2024 Feb 20]; 152:166–77. Available from: https://www.sciencedirect.com/science/article/pii/S0924271619301108
10. Zaabar N, Niculescu S, Kamel MM. Application of Convolutional Neural Networks with Object-Based Image Analysis for Land Cover and Land Use Mapping in Coastal Areas: A Case Study in Ain Témouchent, Algeria. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing [Internet]. 2022 [cited 2024 Apr 2]; 15:5177–89. Available from: https://ieeexplore.ieee.org/document/9803238
11. Zhang X, Han L, Han L, Zhu L. How Well Do Deep Learning-Based Methods for Land Cover Classification and Object Detection Perform on High-Resolution Remote Sensing Imagery? Remote Sensing [Internet]. 2020 Jan [cited 2024 May 13];12(3):417. Available from: https://www.mdpi.com/2072-4292/12/3/417
12. Ienco D, Gbodjo YJE, Gaetano R, Interdonato R. Weakly Supervised Learning for Land Cover Mapping of Satellite Image Time Series via Attention-Based CNN. IEEE Access [Internet]. 2020 [cited 2024 Apr 2]; 8:179547–60. Available from: https://ieeexplore.ieee.org/document/9197599
13. Yang N, Tang H. Semantic Segmentation of Satellite Images: A Deep Learning Approach Integrated with Geospatial Hash Codes. Remote Sensing [Internet]. 2021 Jan [cited 2024 Feb 20];13(14):2723. Available from: https://www.mdpi.com/2072-4292/13/14/2723
14. Yuan X, Shi J, Gu L. A review of deep learning methods for semantic segmentation of remote sensing imagery. Expert Systems with Applications [Internet]. 2021 May 1 [cited 2024 Feb 20];169:114417. Available from: https://www.sciencedirect.com/science/article/pii/S0957417420310836
15. Garcia-Garcia A, Orts-Escolano S, Oprea S, Villena-Martinez V, Garcia-Rodriguez J. A Review on Deep Learning Techniques Applied to Semantic Segmentation [Internet]. arXiv; 2017 [cited 2024 Feb 20]. Available from: http://arxiv.org/abs/1704.06857
16. Marmanis D, Wegner JD, Galliani S, Schindler K, Datcu M, Stilla U. SEMANTIC SEGMENTATION OF AERIAL IMAGES WITH AN ENSEMBLE OF CNNS. ISPRS Ann Photogramm Remote Sens Spatial Inf Sci [Internet]. 2016 Jun 6 [cited 2024 Feb 20]; III–3:473–80. Available from: https://isprs-annals.copernicus.org/articles/III-3/473/2016/
17. Li R, Zheng S, Duan C, Wang L, Zhang C. Land cover classification from remote sensing images based on multi-scale fully convolutional network. Geo-spatial Information Science [Internet]. 2022 Apr 3 [cited 2024 Apr 2];25(2):278–94. Available from: https://doi.org/10.1080/10095020.2021.2017237

18. Tzepkenlis A, Marthoglou K, Grammalidis N. Efficient Deep Semantic Segmentation for Land Cover Classification Using Sentinel Imagery. Remote Sensing [Internet]. 2023 Jan [cited 2024 Apr 2];15(8):2027. Available from: https://www.mdpi.com/2072-4292/15/8/2027
19. Xu R, Wang C, Zhang J, Xu S, Meng W, Zhang X. RSSFormer: Foreground Saliency Enhancement for Remote Sensing Land-Cover Segmentation. IEEE Transactions on Image Processing [Internet]. 2023 [cited 2024 Apr 2];32:1052–64. Available from: https://ieeexplore.ieee.org/abstract/document/10026298
20. Längkvist M, Kiselev A, Alirezaie M, Loutfi A. Classification and Segmentation of Satellite Orthoimagery Using Convolutional Neural Networks. Remote Sensing [Internet]. 2016 Apr [cited 2024 Apr 2];8(4):329. Available from: https://www.mdpi.com/2072-4292/8/4/329
21. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. Available online: https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html (accessed on 22 April 2023).
22. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826. Available online: https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Szegedy_Rethinking_the_Inception_CVPR_2016_paper.html (accessed on 22 April 2023).
23. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. arXiv 2018, doi:10.48550/arXiv.1608.06993.
24. Tan M, Le QV. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks [Internet]. arXiv; 2020 [cited 2023 Sep 14]. Available from: http://arxiv.org/abs/1905.11946
25. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation [Internet]. arXiv; 2015 [cited 2024 Feb 13]. Available from: http://arxiv.org/abs/1505.04597
26. Badrinarayanan V, Kendall A, Cipolla R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence [Internet]. 2017 Dec [cited 2024 Apr 5];39(12):2481–95. Available from: https://ieeexplore.ieee.org/document/7803544
27. Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence [Internet]. 2018 Apr [cited 2024 Apr 5];40(4):834–48. Available from: https://ieeexplore.ieee.org/document/7913730

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.