





Article

A Deep Learning-Based Solution for Large-Scale Extraction of the Secondary Road Network from High-Resolution Aerial Orthoimagery

Calimanut-Ionut Cira ^{1,*} , Ramón Alcarria ¹ , Miguel-Ángel Manso-Callejo ¹ 
and Francisco Serradilla ² 

¹ Departamento de Ingeniería Topográfica y Cartografía, E.T.S.I. en Geodesia, Topografía, Geodesia y Cartografía, Universidad Politécnica de Madrid, 28031 Madrid, Spain; ramon.alcarria@upm.es (R.A.); m.manso@upm.es (M.-Á.M.-C.)

² Departamento de Inteligencia Artificial, E.T.S.I. de Sistemas Informáticos, Universidad Politécnica de Madrid, 28031 Madrid, Spain; fserra@eui.upm.es

* Correspondence: ionut.cira@upm.es

Received: 21 September 2020; Accepted: 14 October 2020; Published: 17 October 2020



Abstract: Secondary roads represent the largest part of the road network. However, due to the absence of clearly defined edges, presence of occlusions, and differences in widths, monitoring and mapping them represents a great effort for public administration. We believe that recent advancements in machine vision allow the extraction of these types of roads from high-resolution remotely sensed imagery and can enable the automation of the mapping operation. In this work, we leverage these advances and propose a deep learning-based solution capable of efficiently extracting the surface area of secondary roads at a large scale. The solution is based on hybrid segmentation models trained with high-resolution remote sensing imagery divided in tiles of 256×256 pixels and their correspondent segmentation masks, resulting in increases in performance metrics of 2.7–3.5% when compared to the original architectures. The best performing model achieved Intersection over Union and F1 scores of maximum 0.5790 and 0.7120, respectively, with a minimum loss of 0.4985 and was integrated on a web platform which handles the evaluation of large areas, the association of the semantic predictions with geographical coordinates, the conversion of the tiles' format and the generation of geotiff results compatible with geospatial databases.

Keywords: aerial orthoimagery; deep learning; remote sensing; road extraction; semantic segmentation; web-based segmentation solution

1. Introduction

According to statistics from the Ministry of Development (Spanish: “Ministerio de Fomento”) [1], in 2019 the total mileage of roads in Spain was 655,322 km, of which 165,624 km (25%) are paved roads (including 17,021 km of routes of great capacity like highways). The remaining 499,698 km (75%) constitute the secondary road network (Spanish: “Red secundaria de carreteras”). Being able to obtain an accurate representation and distribution of these roads would be useful for government agencies, helping in the monitoring and detection of changes in order to provide up-to-date cartography. At the moment, at the national level, the road mapping operation is a challenging and time-consuming manual process, even with open access to high-resolution remote sensing imagery.

The advances in machine vision started with the development of Convolutional Neural Network (CNN) architectures and incentivized the emergence of modern semantic segmentation techniques [2–4]. We believe that these advancements can help in achieving an efficient large-scale extraction of the surface area of secondary roads. However, we have to take into account the great complexity of this

task. These roads have different spectral characteristics due to the various kinds of material used (asphalt, cement, gravel, etc.) and are often covered by obstructions present in the scenes (e.g., dense vegetation). Furthermore, these structures are complex in nature, due to the absence of clearly defined edges, the differences in widths, and the large curvature changes which can complicate the extraction of detailed information.

In this paper, we tackle the automation of the road mapping task by using openly available high-resolution aerial orthoimagery and propose an extraction solution based on semantic segmentation. For this, we explore a wide range of hybrid segmentation architectures to identify the most suitable one for this application and our dataset, and propose a novel approach by implementing the best performing model on a web application capable of analysing extended areas, and joining adjacent tiles before converting the output into road vectors compatible with geospatial databases.

To the best of our knowledge, this is the first intent of a large-scale road extraction system integrated on a web platform designed to eliminate the need for computationally expensive attention procedures. This deep learning-based solution can successfully analyse large areas, predict whether each pixel of the input aerial imagery belongs to a secondary road and overlap the predictions, and represents a complete unitary workflow for road surface extraction that could help the state administration reduce costs in the cartography updating task.

The main contributions of this work can be summarized as follows:

1. We contrast the performance of state-of-the-art segmentation architectures and proposed hybrid segmentation models trained for the road extraction task and analyse the effect of the interaction between the architectures and the base networks (backbones).
2. Different from previous works, we focus on challenging scenes where shadows and occlusions are present, and take into account roads present on a variety of soil types, that are easily confused with the surroundings (secondary routes), further complicating the object's extraction.
3. We carry out the experiments on a new dataset composed of tiles with a spatial resolution of 0.50 m and their corresponding annotated reference masks covering representative areas of the Spanish national territory.
4. To visually check the validity of our contributions, we built a system that combines the model's predictions with existing data. The web platform allows the users to assess the quality of the segmentation results and naturally detect the limitations of the study, in order to propose future directions.

The remainder of this article is organized as follows. Section 2 presents works related to road extraction using remote sensing data and machine learning techniques. Section 3 describes the dataset used in this paper. Section 4 explains the experiments carried out to obtain the best hybrid segmentation configuration. In Section 5, we conduct a statistical analysis of the performance metrics, study the effect of the interaction between architectures and backbones and discuss the results. Section 6 approaches the limitations encountered in the study. Lastly, Section 7 draws the conclusions of our work, while Section VIII presents the references.

2. Related Work

The emergence of modern semantic segmentation techniques was fuelled by the success of Convolutional Neural Network (CNN) architectures. Segmentation models like the Fully Convolutional Network (FCN) [5] are trained using a pixel-wise loss and introduce upconvolutions to resize the image to the input dimensions by making use of the feature maps produced by the CNNs (which act as backbone networks). FCN-8 variant reached a 0.62 mean Intersection over Union (IoU) score (or Jaccard Index) on the 2012 PASCAL VOC (Visual Object Classes) Segmentation Challenge [6] using pre-trained models on the 2012 ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [7]. The IoU score measures the similarity between sample sets by dividing the number of pixels in the intersection between the ground-truth mask and the predicted segmentation map to the size of the

union of the two. The IoU score can be calculated by means of the True Positives (*TP*), False Positives (*FP*) and False Negatives (*FN*) obtained from the confusion matrix of the segmentation classifier, $IoU = TP / (TP + FP + FN)$, and is widely used in image segmentation tasks. A value closer to 1 is desired; a score higher than 0.5 is considered a good prediction [8].

State-of-the-art segmentation models are focusing on a better understanding of the local context by linking different parts of the input image to improve the predictions and learn the relations between the objects. Recent architectures are generally composed of two linked parts: a convolutional part (or encoder), which downsamples the input image using convolutional and Rectified Linear Units (ReLU) layers [9] to generate a vector of features, and an upconvolutional part (or decoder), which uses transpose convolutions to expand the feature maps to match the input size while keeping the information location in the resulting maps. The upsampling part takes the vector of features as input and generates a map with the probabilities of belonging to a class for each pixel. One example of such architecture is U-Net [10], which received widespread interest in the scientific community and was continuously extended in recent works (LinkNet, Feature Pyramid Networks (FPN), Pyramid Scene Parsing Networks (PSPNet), etc.). LinkNet [11] uses the same U-shape structure, but replaces the convolutions in each of the encoder and decoder steps with residual blocks [12], while FPN [13] feed the outputs of each downsampling step into a pyramid level and use lateral connections to join low-resolution and high-resolution features, allowing the detection of objects at multiple scales.

The backbone networks also evolved to improve the efficiency of semantic segmentation. Moving away from using standard CNN architectures like VGGNet [14] or Inception-V3 [15] as backbone networks, architectures specialized in image segmentation emerged. One such example is MobileNetV2 [16], which obtains high computational efficiency by inserting linear bottleneck layers encoded by bottleneck inputs into the convolutional blocks. EfficientNet [17] focused on proposing a model scaling method and shows that balancing a network's depth, width, and resolution can lead to great improvements in performance. SEResNeXt uses "Squeeze-and-Excitation" (SE) blocks [18] that model the interdependencies between channels at the end of each non-identity branch of ResNeXt's [19] residual blocks (a ResNet-based architecture that introduced a hyper-parameter called number of independent paths, or cardinality, to adjust the model's capacity).

These computer vision advances enabled the automatic extraction of information from satellite imagery. The majority of works involving remotely sensed imagery tackle land cover classification tasks, where different land use classes are extracted from aerial images to produce maps (e.g., plantations, buildings, forests, etc.). However, the geospatial elements taken into consideration tend to have homogenous hyperspectral signatures and to be grouped into clearly defined regions covering smaller areas, unlike secondary road networks, which are far more complex in nature.

Most of the methods proposed to extract road structures from remote sensing imagery can be fitted into two categories: extraction of the road surface or extraction of road edges or centrelines, and use of supervised learning to extract the geometries using their radiometric, spatial and photometric features. These techniques have also been successfully applied to tasks related to road detection such as infrastructure monitoring or vehicle navigation [20].

Many of the existing road extraction studies use traditional machine learning and image segmentation methods. Liu et al. [21] tackle the secondary road extraction task by applying pre-processing with Gaussian filters to smooth and remove noise from the input and to enhance features specific to roads. As post-processing, they apply Hough transform to detect discontinuous boundaries of regions based on the global features of a single orthoimage of 7000×6000 pixels in size. In [22], the task is formulated as a Markov random field inference parameterized in terms of the location of the road edges and centrelines and makes use of OpenStreetMap to build an urban scene dataset covering 1.5 km^2 ; the model is trained using structured Support Vector Machines (SVMs). One obvious disadvantage of applying traditional machine learning techniques is represented by the need to analytically and mathematically model and understand the relations between the input and the expected predictions. The representations selected for training can result in models performing well

on small datasets, but may not be suitable for solving complex tasks that involve a large amount of data. To address the geospatial complexity of the road network, we chose the deep learning approach, which enables the building of robust segmentation models with a higher generalization capability that are more suitable for modelling complex functions as those needed to describe features specific to roads.

Apart from this, the main drawback found in relevant works is related to the reduced study areas taken into account (as also pointed out in [23]). Most existing works focus on small selected study areas, with favourable scenarios [24], which may not be representative for a large-scale extraction of these geospatial elements. Alshehhi et al. [25] propose a model consisting of five convolution layers, replacing the Fully Connected layers with Global Average Pooling (GAP) layers to extract road and building geometries from 50 aerial images with resolutions of 1500×1500 pixels. They extract spatial features from adjacent pixels to enhance the predictions and reach performance metrics as high as 81.6%. Kestur et al. [26] propose a FCN-based architecture called UFCN (U-shaped FCN), based on stacking convolutions and their mirrored “deconvolutions”, and train it on a limited dataset containing 76 images of roads subjected to real-time data augmentation. Henry et al. [27] recognize the complexity of the road extraction task when state-of-the-art networks obtained levels of precision of a maximum 72% on a single uncropped aerial orthophoto, even by applying filtering, data augmentation and post-processing operations. We believe that training the networks on small datasets can lead to inconsistent predictions, due to the lack of representative samples and lack of variety in data. The solution described in this paper addresses this drawback by building a new dataset composed of aerial orthoimages divided into tiles of 256×256 pixels covering extended areas of the Spanish territory.

We have also found that existing works generally apply concepts from popular architectures, while adding small tweaks (e.g., adding batch normalization after ReLU activations). For example, inspired by deep residual learning and U-Net, Zhang et al. [28] purpose ResUnet for road extraction by semantic segmentation on the Massachusetts roads dataset, replacing the convolutional blocks with residual modules. They obtained a decrease in parameters of $\frac{3}{4}$, while increasing the performance metrics by 1% when compared to the original U-Net. Liu et al. [29] propose the RoadNet framework, composed of a modified version of VGGNet architecture, to analyse and predict the surfaces, edges, and centrelines of roads in urban scenes. Similar to other FCN architectures, the CNNs are concatenated and use receptive fields of 1×1 pixels in size, obtaining F1 Scores of maximum 94%. The F1 score indicates the balance between correctness (Precision, P) and completeness (Recall, R) using the confusion matrix, $F1 = 2 \times [(P \times R)/(P + R)]$. Correctness (P) is the fraction of the pixels correctly labelled as roads and the total number of road pixels, $P = TP/(TP + FP)$, while completeness (R) is the fraction of all correctly labelled road pixels and the total number of correctly labelled pixels, $R = TP/(TP + FN)$ [30]. Similarly to the IoU score, the F1 score ranges from 0 (worst value) to 1 (best score). Wang et al. [31] built a semantic segmentation model for autonomous driving applications with VGGNet as the backbone and trained it to extract road boundaries using street scenes tagged with their semantic contour. The authors of [32] modify SegNet [33] by incorporating an Exponential Linear Unit (ELU) activation function and add conditional random fields (CRFs) for sharpening the road predictions. In this paper, we propose fifteen hybrid segmentation models based on state-of-the-art networks and apply transfer learning techniques to increase their generalization capacity.

Researchers also incorporated global and local attention procedures into neural network architectures as an attempt to consider positional information when working with larger remotely sensed areas. One such example is GL-Dense-U-Net, proposed by Xu et al. [34], which uses DenseNet [35] for the contracting part and U-Net’s symmetry to extract the local and global information through Attention Units, inspired by pyramid feature maps. The model obtained an F1 Score of 0.9516 on a dataset containing 224 Google Earth images. Similarly, Hu et al. [36] take a FCN-like approach to extract features from the last convolutional layer at multiple scales and encode them into global image features through feature coding techniques. In [37], this task is approached from a temporal perspective by modifying SegNet to extract roads from low resolution videos captured by Unmanned Aerial

Vehicles (UAVs). The model uses feature maps from the corresponding encoder stage to upsample and adds an iterative algorithm to process the temporal dimension. In our view, adding attention procedures to deep learning models can be considered a disadvantage when working at larger scales due to the additional computational effort required. We overcame this scenario by building a web platform architecture capable of managing the information related to the absolute position of the tiles.

In this work, we address the above-mentioned challenges found in the existing literature by proposing a unitary extraction solution based on hybrid semantic segmentation models trained with aerial orthoimagery covering extended representative areas of the Spanish territory. The solution can be considered a complete workflow for road surface extraction from aerial orthoimagery and employs a web platform capable of dealing with the evaluation of larger areas and the joining of segmentation predictions in an efficient manner.

3. Dataset

One of the major challenges specific to semantic segmentation is that it requires images annotated at the pixel level; we need aerial orthoimages containing secondary roads and their corresponding segmentation maps. Similar to other existing works [22,38], we used existent cartographic support to avoid a considerable tagging effort. The data were issued by public agencies and georeferenced in European Terrestrial Reference System 1989 (ETRS89) (compatible with World Geodetic System 1984, WGS84) and Universal Transversal Mercator (UTM) projection in the correspondent zone.

- Firstly, the needed imagery was obtained from PNOA [39] (National Plan of Aerial Orthophotography, Spanish: “Plan Nacional de Ortofotografía Aérea”) and has a spatial resolution of 0.50 m and a planimetric RMSE (root-mean-square error) ≤ 1 m. The photograms were acquired in optimal meteorological conditions at a low flight altitude using calibrated photogrammetric cameras equipped with 3-band RGB sensors (8 bits per band). The imagery was orthorectified, radiometrically corrected and has topographic corrections applied using ground points measured with accurate GPS systems.
- Secondly, the segmentation masks containing the correspondent geographic information were obtained from the National Topographic Map [40] (Spanish: “Mapa Topográfico Nacional”) in vectorial format at a scale of 1:25,000. These road vectors were rasterized to create segmentation masks in image format, the resulting masks representing our training labels. One thing worth noting is that the rasterized vectors do not always align completely with the imagery due to the distinct criteria used for tagging (variations in width and contour, depending on the road’s importance), due to the error-prone conversion of a vector to a raster or due to human errors. However, as stated in the introduction, we wanted to focus on very challenging instances, with irregular geometries, different spectral signatures of the materials used for pavement, and where obstructions from objects are present. The intuition is that by training the models with real data, they will be better prepared to handle difficult scenarios and will offer an intuition about their expected behaviour in large-scale applications.

To obtain the training set, an operator performed a visual comparison between the most recent orthoimages available and the existing cartographic support in vector format in different representative areas of the Spanish territory (Andalucía, Castilla y León, Castilla-La Mancha, Galicia, Islas Baleares, Murcia and Navarra) and manually tagged the tiles using a WMS (Web Map Service) viewer built for this purpose (described in [41]). We collected our training data as sets of aerial orthoimagery clipped into fixed tiles and their correspondent segmentation mask measuring 256 by 256 pixels (0.07 Megapixels) stored in the lossless png format (examples can be seen in Figure 1).

The dataset contains 7750 pairs of tiles and their rasterized masks of the ground-truth at pixel level, covers a land area of approximately 175 km², and has a size on disk of approximately 1.21 gigabytes. We are working towards increasing its size using crowdsourcing techniques and plan to make it openly available in the near future.

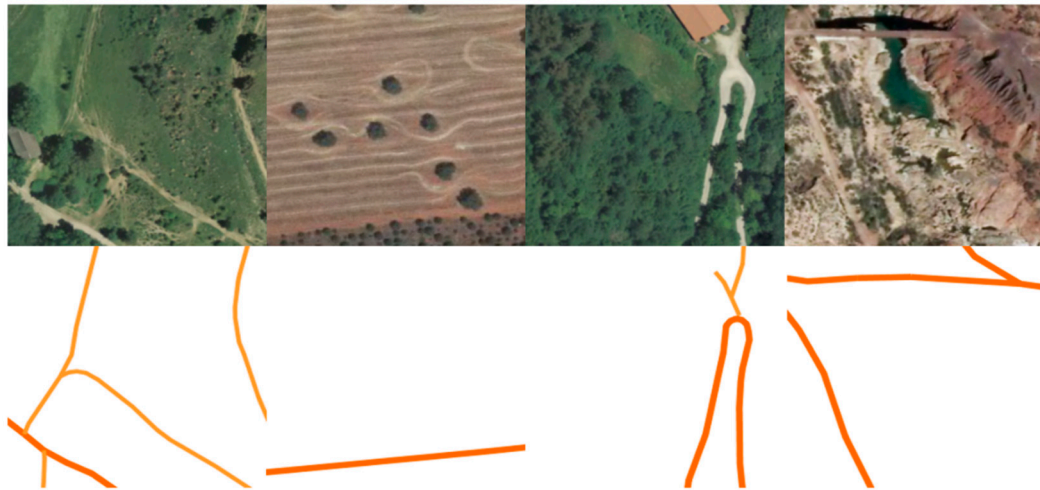


Figure 1. Example of tiles and their rasterized mask with labelled pixels used for training.

4. Methodology and Experiments

Semantic segmentation is about automatically extracting information from images. The notion is that our model will learn a correlation between the input image and the input segmentation maps and be capable of successfully predicting the class label of each pixel for new tiles. This task can be viewed as a series of binary image labelling problems, where “1” refers to a pixel belonging to the road surface and “0” refers to opposite case, with every pixel in the image belonging to one of these two classes (“road”/“no road”). The model has to find a binary segmentation function to correctly assign a state from the label space $Y = \{y_0, y_1\}$ to the element of a set of variables $X = \{x_1, x_2, \dots, x_n\}$ and predict the meaning of each pixel.

We need to consider that neighbouring pixels tend to correlate with each other (we cannot assume the independency and identical distribution of class labels y). As a result, the deep learning probabilistic approach where a model finds the parameters via maximum likelihood estimation are not applicable. The inference task is to predict a joint assignment of the class labels, instead of scalar ones (infer a joint label y from a given image x). The naïve approach would imply learning the corresponding segmentation mapping through successive transformation of features from the input tile; however, this would be computationally inefficient [42]. One approach that allows image segmentation models to make use of probability theory is to follow the encoder/decoder structure where we downsample the input tile to feature mappings of lower resolution (learning this way to discriminate between classes), and then upsample the feature mappings into a segmentation map with the same resolution as the input tile [43]. In Figure 2, we can see how a random tile is processed using this segmentation structure.

For training, we employed the open-source deep learning “Segmentation Models” library [44] (based on Keras [45] with TensorFlow 1.14 [46] as backend) using an NVIDIA 2060 GPU (Graphics Processing Unit). We split the dataset presented in Section 3 by randomly assigning the tiles and their ground-truth segmentation maps into the following sets:

- A full training set of 6200 tiles and their corresponding segmentation maps (80% of the data) and five training subsets containing 90% of the training set (5580 tiles and segmentation maps) to perform the weights initialization. The five additional subsets represent variations of the training population and are used to statistically analyse the performance of the proposed networks and study the effect of the interaction effect between architectures and backbones.
- A test set (20% of the data) formed by 1550 sets of tiles and their segmentation maps for tuning the model’s hyperparameters and evaluating the model’s predictive performance.

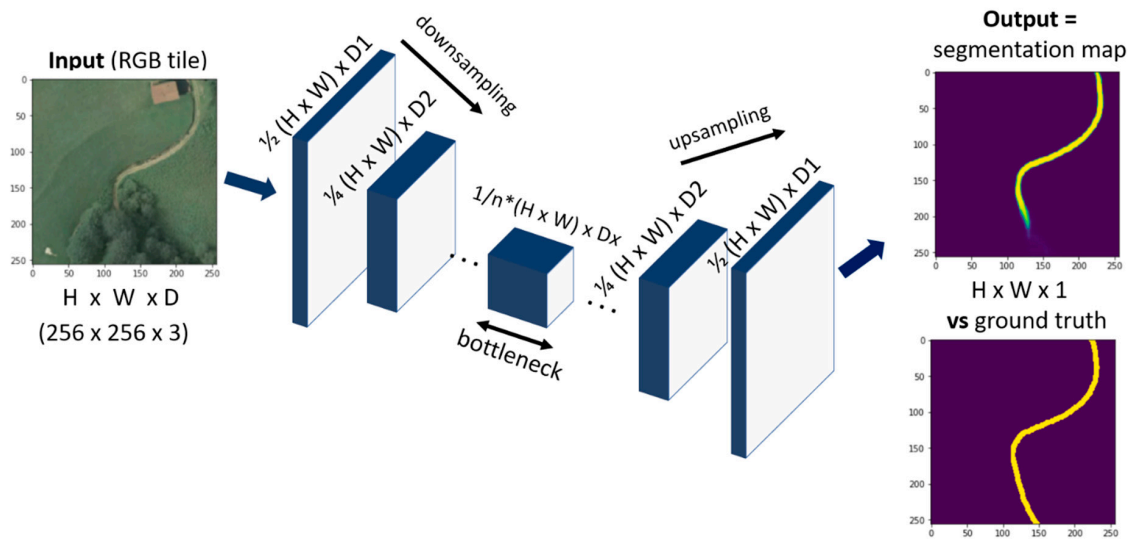


Figure 2. Example of processing a random tile through the encoder–decoder structure.

Given that we are tackling a complex segmentation task with a limited dataset, we carried out the experiments on the data split describe above, allowing a higher ratio of tiles to be used for training [47–49].

Segmentation models are usually built using a segmentation architecture coupled with a base network (or backbone). In the paper, we propose the following fifteen hybrid networks: configurations 1 to 5 with U-Net [10] as segmentation architecture and MobileNetV2 [16], EfficientNet [17], InceptionV3 [15], VGG16 [14] and SEResNeXt50 [18] respectively, as backbones, configurations 6 to 10 using the LinkNet architecture [11] with the same backbones, and configurations 11 to 15 with FPN [13] as segmentation architecture coupled with the same above-mentioned base networks. We chose these combinations to leverage their particularities mentioned in Section 2. For comparison reasons, we also took into account the original U-Net, LinkNet and FPN architectures (as proposed by the authors). In broad terms, all these models follow the encoder–decoder pattern presented in Figure 2.

The fifteen proposed configurations and the original architectures were trained with sets of input tiles and their corresponding rasterized segmentation masks via stochastic gradient descent to optimize the model’s weights using estimations of the gradient and to approximately learn the unknown data distribution given the labelled training data.

The models take an RGB image of size $256 \times 256 \times 3$ and output a segmentation map with the number of channels equal to the number of filters in the last transposed convolution layer. This third dimension is squeezed down using a 1-by-1 convolution layer to the number of classes and subsequently collapsed into a single segmentation map (of size $256 \times 256 \times 1$) by keeping the depth-wise argmax of each pixel.

We begin the pre-processing by normalizing the intensity values of the pixels between 0 and 1 (dividing them by 255). As a means of overcoming the limited training dataset, we controlled the overfitting behaviour by applying aggressive data augmentation to the training set described above [50]. These transformations included random crops, horizontal and vertical flips, random rotation and brightness, contrast and gamma shifts, random scale rotation, elastic transformations, and grid and optical distortions. The aggressive data augmentation also helps with the radiometric unbalance present among the Spanish cartography at the national level, by exposing the model to more aspects of the data.

As for the hyperparameters, we used Adam [51] (considered to be the fastest to converge [52]) to optimize the binary crossentropy Jaccard loss function. This loss function measures the similarity between the prediction (pr) and the ground-truth (gt) when working with a network ending in a sigmoid function, $L(gt, pr) = -gt \times \log(pr) - (1 - gt) \times \log(1 - pr)$, the Jaccard criterion being $1 - (A \cap B / A \cup B)$

for any two sets, A and B . The starting learning rate of 0.01 was reduced with a factor of 0.1 when performance metrics plateaued for more than 10 epochs up to a minimum of 1×10^{-5} in all training scenarios. In addition, we set the IoU score as the metric to monitor, the learning process stopping when the performance stalled for more than 10 epochs or when the network started to display overfitting behaviour.

For weights initialization of the proposed configurations, we applied transfer learning (called Feature Mapping in segmentation tasks) from the ILSVRC dataset [7] to start with pre-trained weights instead of random initialized ones, this way ensuring a good and faster convergence even when the amount of available training data is insufficient to train a model from scratch [47,53–55]. In [55], it was proven that a good convergence can be achieved by fine-tuning the convolutional layers in the encoder path, even if the feature types of the source and the target task differ. For comparison, the original architectures were trained from scratch (using random weight initialization).

5. Results and Discussions

We evaluated the prediction capability of the networks on the testing set (1550 tiles and ground-truth masks) using common performance metrics for binary operations: the IoU score, $IoU = TP/(TP + FP + FN)$, and the F1 score, $F1 = 2 \times [(P \times R)/(P + R)]$. For analysis purposes, we also accounted the loss on the testing set as a performance indicator (a lower loss value being desired).

We compared the performances of the proposed hybrid configurations and the standard architectures trained from scratch using one-way analysis of variance (ANOVA) with the metrics IoU score, F1 score and loss as dependent variables and the configurations as fixed factors. To test the null hypothesis that the performances of all models are equal, we computed the F -statistic from the ANOVA table (results are reported in Table 1 for each of the metrics). The alternative hypothesis is that the performance of at least one configuration is different than the others.

Table 1. Mean (M) and standard deviation (SD) of the performance metrics (IoU score, F1 score and Loss) obtained by the networks.

Configuration	IoU Score		F1 Score		Loss	
	M	SD	M	SD	M	SD
1	0.5335	0.0064	0.6694	0.0072	0.5529	0.0077
2	0.5593	0.0075	0.6958	0.0060	0.5231	0.0090
3	0.5028	0.0059	0.6641	0.0069	0.5591	0.0080
4	0.5466	0.0130	0.6819	0.0123	0.5372	0.0158
5	0.5644	0.0086	0.6988	0.0078	0.5178	0.0106
Standard U-Net	0.5299	0.0087	0.6638	0.0044	0.5588	0.0055
6	0.5302	0.0047	0.6683	0.0026	0.5559	0.0044
7	0.5626	0.0159	0.6979	0.0140	0.5187	0.0175
8	0.5293	0.0095	0.6663	0.0099	0.5576	0.0103
9	0.5372	0.0098	0.6744	0.0082	0.5450	0.0100
10	0.5515	0.0062	0.6877	0.0049	0.5294	0.0065
Standard LinkNet	0.5184	0.0080	0.6552	0.0094	0.5687	0.0088
11	0.5298	0.0039	0.6650	0.0036	0.5570	0.0039
12	0.5437	0.0088	0.6827	0.0046	0.5361	0.0033
13	0.5276	0.0095	0.6656	0.0081	0.5579	0.0090
14	0.5421	0.0069	0.6781	0.0071	0.5420	0.0074
15	0.5509	0.0033	0.6871	0.0018	0.5307	0.0034
Standard FPN	0.5233	0.0093	0.6621	0.0075	0.5641	0.0105
F -Statistic	11.344		13.119		12.886	
p -value	0.000		0.000		0.000	

In Table 1, we can see that all p -values are smaller than 0.001 (a p -value < 0.05 implies that the null hypothesis is to be rejected at a level of confidence of 95%); therefore, configurations are significantly different in terms of the IoU score as well as the F1 score and loss. The results obtained on our dataset are homogenous and show a significant improvement over the standard state-of-the-art models (2.76–3.45%), proving that more complex hybrid semantic segmentation models have a stronger ability to extract complex geospatial elements such as secondary roads. However, the analysis of variance of F -statistics and their p -values does not reveal which configurations are different from the others when there is a significant difference. To have a detailed comparison of the hybrid configurations in terms of performance metrics, in Figure 3 we can find the boxplots for the fifteen proposed configurations.

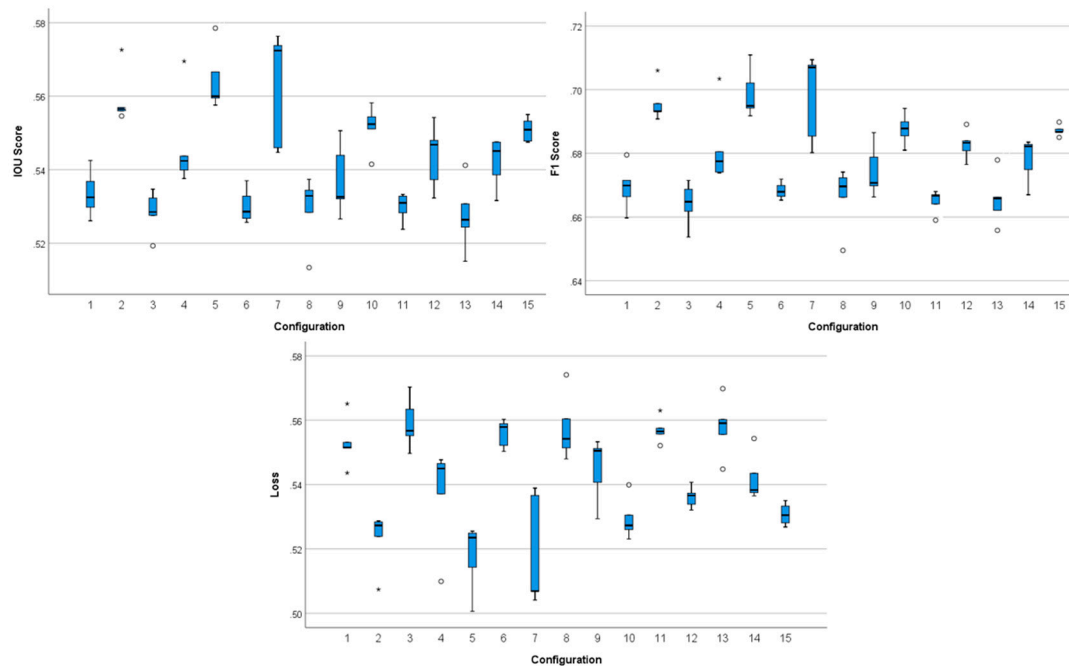


Figure 3. Boxplots of performance metrics obtained by the proposed hybrid configurations.

The performance of the configurations seems to follow a similar pattern. Figure 3 shows that Configuration 7 (with LinkNet as base architecture and EfficientNet as backbone) has the highest median IoU score and median F1 score, and is closely followed by configurations 5 and 2. However, configuration 7 also has the highest variability in the performance metrics.

Next, we applied Tukey's HSD (Honestly Significant Difference) test to compare the performances in terms of the IoU Score and identify the best performing ones. The test conducts a comparison between two configurations at a time using a t -test adjusted for overall variability of the data, while maintaining the level of significance (or the probability of type I error) at 5%. In Table 2, the post-hoc test results are presented in terms of homogenous subsets of configurations for the IoU Score using Tukey's HSD test described above.

The homogenous subsets reported in Table 2 contain proposed configurations whose performances are not significantly different from each other at a level of significance of 5%. For example, configurations 12, 4, 15, 10, 2 and 7 do not have a significantly different IoU Score. On the other hand, configurations that are not common in two homogenous subsets have significantly different performance. These post-hoc test results support our observations in Figure 3's boxplots.

When grouping the proposed networks by their backbones, we observe significant differences in the performance scores. The highest average IoU score and the lowest average loss is attained by SEResNeXt50, closely followed by EfficientNet, while the highest average F1 score is attained by EfficientNet, closely followed by SEResNeXt50. These two networks seem to have a similar

performance, being the best performing backbones of homogenous subsets based on post-hoc tests. The worst performer is VGG16, one of the most used CNN for classification tasks (as seen in Figure 4, it obtained the lowest average IoU and F1 scores and the highest average loss).

Table 2. Homogeneous subsets of configurations with post-hoc tests for the IoU score.

Config. No.	Homogeneous Subsets				
	1	2	3	4	5
13	0.5276				
3	0.5285				
8	0.5293				
11	0.5298				
6	0.5302				
1	0.5335	0.5335			
9	0.5372	0.5372			
14	0.5421	0.5421	0.5421		
12	0.5437	0.5437	0.5437	0.5437	
4	0.5466	0.5466	0.5466	0.5466	0.5466
15		0.5509	0.5509	0.5509	0.5509
10		0.5515	0.5515	0.5515	0.5515
2			0.5593	0.5593	0.5593
7				0.5626	0.5626
5					0.5644
Sig.	0.055	0.092	0.127	0.059	0.098

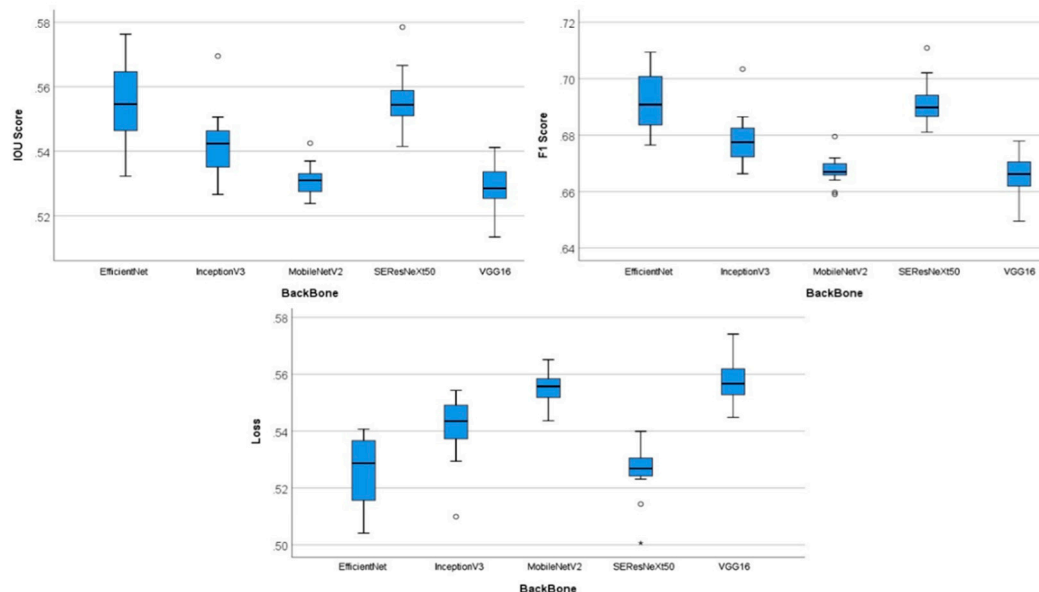


Figure 4. Boxplots of performance metrics for different backbones.

When applying Tukey's HSD test to the performance metrics grouped by the base architectures, FPN is the worst performer (with average IoU and F1 scores of 0.5388 and 0.6757, respectively, and an average loss of 0.5447). This may be due to the FPN architecture, where a tile is scaled at different sizes, each scaled result being processed individually to provide the predictions. U-Net is the best performer (average IoU and F1 scores of 0.5465 and 0.6820, respectively, and an average loss of 0.5380), the results being significantly different when compared to FPN. However, LinkNet (with an average IoU score of 0.5422, an average F1 score of 0.6789 and an average loss of 0.5413) is not significantly different from FPN or U-Net and we do not see any significant difference in their performance when

base architectures are considered alone. Figure 5 presents the average IoU scores, F1 scores and the losses obtained by the base architectures.

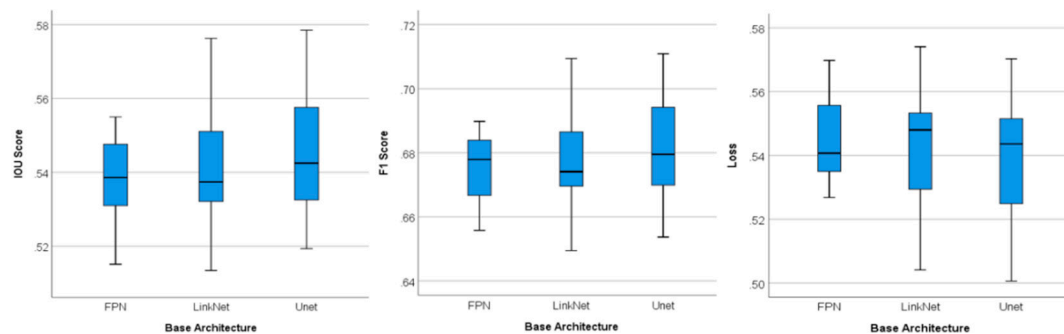


Figure 5. Boxplots of performance metrics grouped by the base architectures.

To study the effect of backbones and the base architectures, we use two-factor analysis of variance models, and the results are presented in Table 3.

Table 3. ANOVA results for the effects of backbones and base architectures on performance.

Dependent Variable	Source	Sum of Squares	Degrees of Freedom	Mean Square	F	Sig.
IoU score	Backbone	0.0099	4	0.0025	33.367	0.000
	Base Architecture	0.0007	2	0.0004	4.978	0.010
	Backbone and Base Architecture	0.0011	8	0.0001	1.924	0.073
	Error	0.0044	60	7.41×10^{-5}		
	Total	0.0162	74			
F1 score	Backbone	0.0096	4	0.0024	40.382	0.000
	Base Architecture	0.0005	2	0.0002	4.166	0.020
	Backbone and Base Architecture	0.0008	8	0.0001	1.726	0.111
	Error	0.0036	60	5.94×10^{-5}		
	Total	0.0140	74			

The main effects of backbones and base architectures on performance are significant at a level of significance of 5%. The main effect null hypothesis studies the marginal effect of a factor when all other factors are kept at a fixed level and states that the effect is not significant. In Table 3, we can see that the effect of backbones on performance metrics when all other factors are kept at a fixed level is significant on the performance metrics and we can reject the null hypothesis and conclude that the effect is significant (p -value of 0.000; smaller than 0.05). Thus, there is a difference in performance due to different backbones. Similarly, there is a significant difference in performances due to different base architectures (p -values of 0.01 and 0.02, respectively).

For the interaction effect, the null hypothesis is that the effect of backbones on the performance does not vary with the base architecture. The interaction effects of backbone and base architectures on IoU and F1 scores are not significant with p -values 0.073 and 0.111, respectively. So, we can conclude that there is not enough evidence to suggest that the effects of backbones are different for different base architectures. The null hypothesis is not rejected, the p -values being greater than 0.05.

When trained on the full dataset, the best performing network was the hybrid configuration No. 5 (with U-Net [10] as segmentation architecture and SEResNeXt50 [18] as backbone), obtaining an IoU score of 0.5788, a F1 score of 0.7120, and a loss of 0.4985. These performance scores were computed from the confusion matrix obtained by evaluating the test set (presented in Figure 6 with a support of 1550 tiles, or 101,580,800 pixels).

Studying the relationships between the values of the confusion matrix, we can see that the best performing hybrid model correctly classified 97.87% of the samples, while incorrectly classifying 2.13% of the samples. Configuration 5 correctly detected 94.95% of the pixels as “No Road” instances (True Negatives) and 2.92% of the “Road” instances (True Positives or TP), while incorrectly labelling 0.94% instances of “Road” (False Positives-FP) and missing 1.19% instances of “No road” labels (False Negatives-FN). To gain a better interpretation of these results, in Figure 7, we can find examples of predictions (pixel labels) compared to the ground-truth labels.

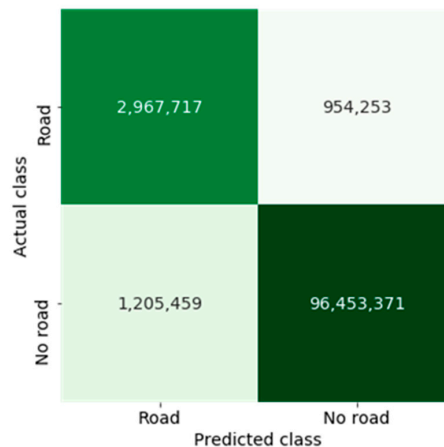


Figure 6. Confusion matrix obtained by configuration 5 trained on the full dataset. Note: The IoU score metric monitored during training does not consider True Negatives (TN); therefore, TN had no impact on the network’s training evolution.

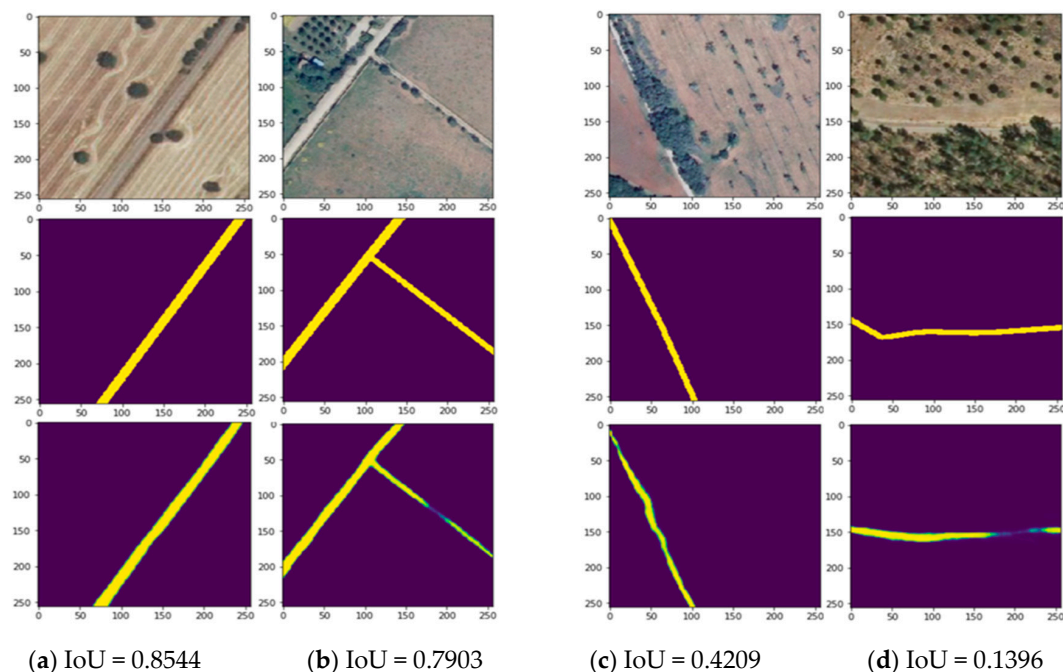


Figure 7. Predictions (third row) and IoU scores obtained by configuration 5 trained on 7750 tiles together with their correspondent ground-truth masks (second row). Note: The probability scores are plotted using a colormap that allows a better identification of ambiguous areas (where probability scores are close to 0.5). This way, purple is used to represent pixels labelled with “No road”, yellow is used to represent pixels belonging to the class “Road Exists”, and scales of blue were chosen to represent pixels with probability scores in proximity to 0.5.

The hybrid configuration is capable of correctly recognizing secondary roads in aerial images and has learned that the expected representation of road surfaces is a continuous line of fixed width (as seen in the examples from Figure 7a,b). Analysing the predictions' errors, we found that the proposed model generally predicted more False Positive labels in areas where the materials used in the road pavement have a similar spectral signature with their surroundings (ambiguous scenarios where the roads are not clearly delimited, as seen in Figure 7d), or in areas where geospatial objects with similar features are present (such as dry riverbeds, railroads or irrigation canals). On the other hand, a higher rate of False Negatives was encountered in sections where other objects cover large portions of the roads (trees, dense vegetation, etc). However, in Figure 7c, we can see that by training with real-life data, the model was capable of inferring small road segments in scenes where obstructions are present.

In general, the segmentation classifier seems to favour the detection of road labels, at the cost of a higher rate of False Positives. However, this metric trade-off is preferable given that, in our case, false negatives yield a lower importance (we want to avoid losing misclassified "Road" labels at the cost of adding additional checks to verify the legitimacy of the FN labels). The rates might also be affected by the fact that a considerable number of secondary roads have not been mapped yet and some of them might be present in our dataset. The qualitative analysis supports the statistical analysis, and taking into account the complexity of the task and the nature of the dataset, these results can be considered satisfactory.

6. Visualization, Limitations and Future Directions

Our efforts were focused next on building a web platform that combines the model's predictions with existing cartographic support. This platform allows an operator to select a geographic extent and feed that area to the prediction system, obtaining a segmentation result and allowing the user to view it overlaid onto the orthoimages, with different presentation styles. The architecture of the web platform is presented in Figure 8 and represents a novel approach to semantic segmentation of very large areas.

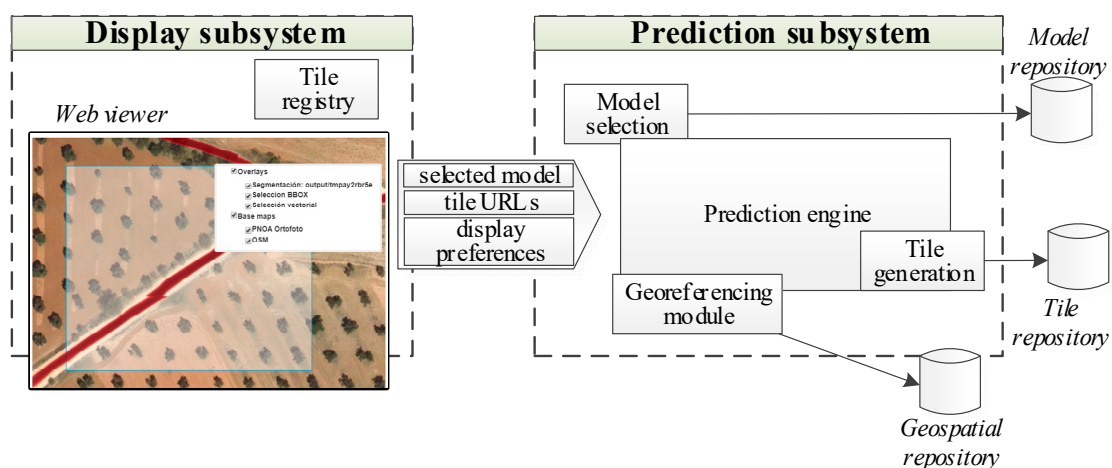


Figure 8. Architecture of the web platform.

The *display subsystem* consists of a *web viewer* where the work area is displayed. The information represented is a layer of aerial ortho-photography from a WMS-Tiled service, offered by the PNOA. This layer is integrated into the map through the Openlayers libraries, which also allow controls for the selection and deletion of areas on the map. Once the workspaces are selected, the *tile registry* component generates a list of URLs addressing tiles of zoom level 18 (scale 1:2257), and supplies this information to the prediction engine, along with the operator's preferences for displaying results.

The *model selection* component maintains a record of the prediction models delivered and available for use on the web platform in h5 format. The operator can select which model of those available in the *model repository* he wants to use to make the prediction.

The *prediction subsystem* receives the information of the selected model, the display preferences, and the list of tile URLs to process. With this information, it performs tile downloads for the work area, if they have not been previously downloaded, and, in the *prediction engine*, applies the semantic segmentation function to each tile, according to the selected model.

The result of this process is a set of segmentation matrices to which the tile generation module applies a colour ramp and a transparent background if the operator has selected it (an example can be seen in Figure 9). These generated tiles are stored in a temporary directory, inside the *tile repository*, which is returned to the *display subsystem*, which adds this set of tiles as a new layer in the *web viewer*. The layer selection and deselection controls allow the operator to visually compare this result with the application of other models, and with the information on roads present in the orthophoto, in order to adjust post-processing techniques.



Figure 9. Per-pixel predictions of the best performing hybrid segmentation model covering an area of 5.24 km² (160 tiles). In the map scheme’s representation, bright red means a higher probability of the pixel belonging to a road while transparent is used for pixels labelled with “No road”.

For post-processing based on Geographic Information Systems (GIS), the platform offers the possibility of downloading a geotiff, georeferenced file with the segmentation results that can be directly imported into the GIS processing program [56]. This file is generated through the *georeferencing module*, using the OSGeo/gdal [57] Python libraries.

The main limitations of this study are related to the imperfections present in the reference maps containing the roads. However, backed up by the statistical analysis carried out in Section 5, it was proved that complex segmentation models can obtain high quality results even without having segmentation maps that are accurate at the pixel level and require a considerable tagging effort.

Another downside is that even the best segmentation classifier would sometimes omit potential roads that might be identifiable to human experts or would misclassify similar background geospatial objects (dry riverbeds, railroads, or irrigation canals) as roads. We consider that the predictions can be improved by adding more training data from these challenging areas. The intuition is that by exposing a network to more real-life data, we can boost its performance and reduce the FP and FN rates. However, if the problem persists, a multiclass classification could be a preferable approach.

Another problem found was related to the overlook of connection points, resulting in unconnected road segments. We have also identified the challenge of inaccurate extraction near the borders due to the lack of context. However, the platform is capable of following a partial large-scale processing strategy by evaluating tiles with a lower stride (e.g., 128 × 128 pixels), and subsequently overlap the central parts of adjacent images, resulting in predictions with no obvious division boundaries.

Another solution to this challenge would be working with larger tiles to provide more semantic context (e.g., training with tiles of 512×512 or 1024×1024 pixels).

The goal of this study was not to create of the best possible road maps, but to demonstrate that hybrid segmentation models implemented on a flexible web platform component can improve the road surface area's extraction from remotely sensed data. We have seen that a solution based on a combination of frameworks can be more efficient for large-scale extraction operations when compared to single end-to-end deep learning models.

7. Conclusions

We made steps towards reducing the tedious and time-consuming task of obtaining road geometries from aerial imagery data by proposing a large-scale road extraction solution based on semantic segmentation methods. We trained the proposed hybrid configurations on a dataset containing aerial information and ground-truth road masks from representative areas of the Spanish territory and obtained maximum IoU and F1 scores of 0.5790 and 0.7120, respectively, and a minimum loss of 0.4985 by applying feature mapping techniques. The best performing hybrid model (with U-Net as segmentation architecture with SEResNeXt50 as backbone network) achieved gains in performance metrics in the order of 3.5% when compared to the original architecture trained from scratch.

The results obtained are promising, but we identified various factors that influenced the model's predictive behaviour, such as obstructions and occlusions in the scenes or other geospatial elements with similar hyperspectral properties (e.g., dry riverbeds, railroads or irrigation canals). We have found that sometimes the model failed to segment all roads within an area, or would omit connection segments or would classify pixels to the wrong class. These limitations can be addressed with post-processing operations, as the web platform allows easy integration of additional processing algorithms. We believe that the proposed solution will improve the utility of the aerial imagery projects for automatic mapping of secondary roads, and can provide a technical base for large-scale road mapping by assigning pixel information with geographical information.

The next step is to further increase the size of the dataset with samples from challenging areas and apply post-processing techniques based on generative learning, this way obtaining more robust models trained with varied data, capable of a more efficient extraction of the road surface areas. We also believe that an ideal solution requires a good balance between machine-generated features and manual mapping, and we will continue to improve the platform in order to make the task process smoother by designing an extension to solve incidents by an operator, this way introducing humans into the decision-making process.

Author Contributions: In this study, Conceptualization, C.-I.C., R.A. and M.-Á.M.-C.; Data curation, C.-I.C. and R.A. and M.-Á.M.-C.; Formal analysis, C.-I.C., R.A. and M.-Á.M.-C.; Funding acquisition, M.-Á.M.-C. and F.S.; Investigation, C.-I.C., R.A. and M.-Á.M.-C.; Methodology, C.-I.C., R.A. and M.-Á.M.-C.; Project administration, R.A., M.-Á.M.-C. and F.S.; Resources, R.A., M.-Á.M.-C. and F.S.; Supervision, R.A., M.-Á.M.-C. and F.S.; Validation, C.-I.C., R.A., M.-Á.M.-C. and F.S.; Visualization C.-I.C., R.A., and M.-Á.M.-C.; Writing—original draft, C.-I.C., R.A. and M.-Á.M.-C.; Writing—review & editing, C.-I.C., R.A., M.-Á.M.-C. and F.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received funding from the Cartobot project, in collaboration with Instituto Geográfico Nacional (IGN), Spain.

Acknowledgments: We thank Mathias Gatti and all other Cartobot participants for their help in the initial phases of the research and in generating the dataset.

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

References

1. Catálogo y Evolución de la Red de Carreteras|Ministerio de Transportes, Movilidad y Agenda Urbana. Available online: <https://www.mitma.gob.es/carreteras/catalogo-y-evolucion-de-la-red-de-carreteras> (accessed on 27 January 2020).

2. Pritt, M.; Chern, G. Satellite Image Classification with Deep Learning. In Proceedings of the 2017 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), Washington, DC, USA, 10–12 October 2017; IEEE: Washington, DC, USA, 2017; pp. 1–7.
3. Tuia, D.; Persello, C.; Bruzzone, L. Domain Adaptation for the Classification of Remote Sensing Data: An Overview of Recent Advances. *IEEE Geosci. Remote Sens. Mag.* **2016**, *4*, 41–57. [[CrossRef](#)]
4. Camps-Valls, G.; Tuia, D.; Bruzzone, L.; Benediktsson, J.A. Advances in Hyperspectral Image Classification: Earth Monitoring with Statistical Learning Methods. *IEEE Signal Process. Mag.* **2014**, *31*, 45–54. [[CrossRef](#)]
5. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; IEEE: Boston, MA, USA, 2015; pp. 3431–3440.
6. Everingham, M.; Eslami, S.M.A.; Van Gool, L.; Williams, C.K.I.; Winn, J.; Zisserman, A. The Pascal Visual Object Classes Challenge: A Retrospective. *Int. J. Comput. Vis.* **2015**, *111*, 98–136. [[CrossRef](#)]
7. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [[CrossRef](#)]
8. Forczmański, P. Performance Evaluation of Selected Thermal Imaging-Based Human Face Detectors. In Proceedings of the 10th International Conference on Computer Recognition Systems CORES 2017, Polanica Zdroj, Poland, 22–24 May 2017; Kurzynski, M., Wozniak, M., Burduk, R., Eds.; Springer International Publishing: Cham, Switzerland, 2018; Volume 578, pp. 170–181, ISBN 978-3-319-59161-2.
9. Nair, V.; Hinton, G.E. Rectified Linear Units Improve Restricted Boltzmann Machines. In Proceedings of the 27th International Conference on Machine Learning (ICML-10), Haifa, Israel, 21–24 June 2010; Fürnkranz, J., Joachims, T., Eds.; Omnipress: Madison, WI, USA, 2010; pp. 807–814.
10. 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.
11. Chaurasia, A.; Culurciello, E. LinkNet: Exploiting Encoder Representations for Efficient Semantic Segmentation. In Proceedings of the 2017 IEEE Visual Communications Image Processing (VCIP), St. Petersburg, FL, USA, 10–13 December 2017; pp. 1–4. [[CrossRef](#)]
12. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; IEEE: Las Vegas, NV, USA, 2016; pp. 770–778.
13. Lin, T.-Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; IEEE: Honolulu, HI, USA, 2017; pp. 936–944.
14. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015; 1409. Available online: <https://arxiv.org/abs/1409.1556> (accessed on 16 October 2020).
15. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; IEEE: Las Vegas, NV, USA, 2016; pp. 2818–2826.
16. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; IEEE: Salt Lake City, UT, USA, 2018; pp. 4510–4520.
17. Tan, M.; Le, Q.V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the 36th International Conference on Machine Learning, ICML 2019, Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114. Available online: <http://proceedings.mlr.press/v97/tan19a.html> (accessed on 16 October 2020).
18. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; IEEE: Salt Lake City, UT, USA, 2018; pp. 7132–7141.
19. Xie, S.; Girshick, R.; Dollar, P.; Tu, Z.; He, K. Aggregated Residual Transformations for Deep Neural Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; IEEE: Honolulu, HI, USA, 2017; pp. 5987–5995.

20. Woźniak, M.; Damaševičius, R.; Maskeliūnas, R.; Malūkas, U. Real Time Path Finding for Assisted Living Using Deep Learning. *JUCS J. Univ. Comput. Sci.* **2018**. [CrossRef]
21. Liu, J.; Qin, Q.; Li, J.; Li, Y. Rural Road Extraction from High-Resolution Remote Sensing Images Based on Geometric Feature Inference. *ISPRS Int. J. Geo-Inf.* **2017**, *6*, 314. [CrossRef]
22. Mattyus, G.; Wang, S.; Fidler, S.; Urtasun, R. Enhancing Road Maps by Parsing Aerial Images Around the World. In Proceedings of the Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; IEEE Computer Society: Columbia, DC, USA, 2015; pp. 1689–1697.
23. Dong, R.; Li, W.; Fu, H.; Gan, L.; Yu, L.; Zheng, J.; Xia, M. Oil palm plantation mapping from high-resolution remote sensing images using deep learning. *Int. J. Remote Sens.* **2020**, *41*, 2022–2046. [CrossRef]
24. Alshaikhli, T.; Liu, W.; Maruyama, Y. Automated Method of Road Extraction from Aerial Images Using a Deep Convolutional Neural Network. *Appl. Sci.* **2019**, *9*, 4825. [CrossRef]
25. Alshehhi, R.; Marpu, P.R.; Woon, W.L.; Mura, M.D. Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks. *ISPRS J. Photogramm. Remote Sens.* **2017**, *130*, 139–149. [CrossRef]
26. Kestur, R.; Farooq, S.; Abdal, R.; Mehraj, E.; Narasipura, O.; Mudigere, M. UFCN: A fully convolutional neural network for road extraction in RGB imagery acquired by remote sensing from an unmanned aerial vehicle. *J. Appl. Remote Sens.* **2018**, *12*, 1. [CrossRef]
27. Henry, C.; Azimi, S.M.; Merkle, N. Road Segmentation in SAR Satellite Images with Deep Fully-Convolutional Neural Networks. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 1867–1871. [CrossRef]
28. Zhang, Z.; Liu, Q.; Wang, Y. Road Extraction by Deep Residual U-Net. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 749–753. [CrossRef]
29. Liu, Y.; Yao, J.; Lu, X.; Xia, M.; Wang, X.; Liu, Y. RoadNet: Learning to Comprehensively Analyze Road Networks in Complex Urban Scenes from High-Resolution Remotely Sensed Images. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 2043–2056. [CrossRef]
30. Sujatha, C.; Selvathi, D. Connected component-based technique for automatic extraction of road centerline in high resolution satellite images. *EURASIP J. Image Video Process.* **2015**, *2015*. [CrossRef]
31. Wang, Q.; Gao, J.; Yuan, Y. Embedding Structured Contour and Location Prior in Siamesed Fully Convolutional Networks for Road Detection. *IEEE Trans. Intell. Transp. Syst.* **2018**, *19*, 230–241. [CrossRef]
32. Panboonyuen, T.; Jitkajornwanich, K.; Lawawirojwong, S.; Srestasathiern, P.; Vateekul, P. Road Segmentation of Remotely-Sensed Images Using Deep Convolutional Neural Networks with Landscape Metrics and Conditional Random Fields. *Remote Sens.* **2017**, *9*, 680. [CrossRef]
33. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
34. Xu, Y.; Xie, Z.; Feng, Y.; Chen, Z. Road Extraction from High-Resolution Remote Sensing Imagery Using Deep Learning. *Remote Sens.* **2018**, *10*, 1461. [CrossRef]
35. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017; IEEE Computer Society: Columbia, DC, USA, 2017; pp. 2261–2269.
36. Hu, F.; Xia, G.-S.; Hu, J.; Zhang, L. Transferring Deep Convolutional Neural Networks for the Scene Classification of High-Resolution Remote Sensing Imagery. *Remote Sens.* **2015**, *7*, 14680–14707. [CrossRef]
37. Luque, B.; Morros, J.R.; Ruiz-Hidalgo, J. Spatio-temporal Road Detection from Aerial Imagery using CNNs. In Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, Porto, Portugal, 27 February–1 March 2017; SCITEPRESS—Science and Technology Publications: Porto, Portugal, 2017; pp. 493–500.
38. Bonafilia, D.; Gill, J.; Basu, S.; Yang, D. Building High Resolution Maps for Humanitarian Aid and Development with Weakly- and Semi-Supervised Learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, 16–20 June 2019; pp. 1–9. Available online: https://openaccess.thecvf.com/content_CVPRW_2019/html/cv4gc/Bonafilia_Building_High_Resolution_Maps_for_Humanitarian_Aid_and_Development_with_CVPRW_2019_paper.html (accessed on 16 October 2020).
39. Instituto Geográfico Nacional Plan Nacional de Ortofotografía Aérea. Available online: <https://pnoa.ign.es/caracteristicas-tecnicas> (accessed on 25 November 2019).

40. Instituto Geográfico Nacional Centro de Descargas del CNIG (IGN). Available online: <http://centrodedescargas.cnig.es> (accessed on 3 February 2020).
41. Gómez-Barrón, J.P.; Alcarria, R.; Manso-Callejo, M.-Á. Designing a Volunteered Geographic Information System for Road Data Validation. *Proceedings* **2019**, *19*, 7. [CrossRef]
42. Li, F.-F.; Johnson, J.; Yeung, S. Lecture 11: Detection and Segmentation. 95. Available online: http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture11.pdf (accessed on 16 October 2020).
43. Jordan, J. An Overview of Semantic Image Segmentation. Available online: <https://www.jeremyjordan.me/semantic-segmentation/> (accessed on 4 February 2020).
44. Yakubovskiy, P. *Segmentation Models*; GitHub: California, CA, USA, 2019; Available online: https://github.com/qubvel/segmentation_models (accessed on 16 October 2020).
45. Chollet, F. *Keras*. 2015. Available online: <https://github.com/fchollet/keras> (accessed on 16 October 2020).
46. Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; et al. TensorFlow: A system for large-scale machine learning. In Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation, Savannah, GA, USA, 2–4 November 2016; p. 21.
47. Cira, C.-I.; Alcarria, R.; Manso-Callejo, M.-Á.; Serradilla, F. Evaluation of Transfer Learning Techniques with Convolutional Neural Networks (CNNs) to Detect the Existence of Roads in High-Resolution Aerial Imagery. In *Applied Informatics*; Florez, H., Leon, M., Diaz-Nafria, J.M., Belli, S., Eds.; Springer International Publishing: Cham, Switzerland, 2019; Volume 1051, pp. 185–198. ISBN 978-3-030-32474-2.
48. Zhu, X.; Vondrick, C.; Fowlkes, C.C.; Ramanan, D. Do We Need More Training Data? *Int. J. Comput. Vis.* **2016**, *119*, 76–92. [CrossRef]
49. Cira, C.-I.; Alcarria, R.; Manso-Callejo, M.-Á.; Serradilla, F. A Framework Based on Nesting of Convolutional Neural Networks to Classify Secondary Roads in High Resolution Aerial Orthoimages. *Remote Sens.* **2020**, *12*, 765. [CrossRef]
50. Sharma, S.; Ball, J.E.; Tang, B.; Carruth, D.W.; Doude, M.; Islam, M.A. Semantic Segmentation with Transfer Learning for Off-Road Autonomous Driving. *Sensors* **2019**, *19*, 2577. [CrossRef] [PubMed]
51. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015; Available online: <http://arxiv.org/abs/1412.6980> (accessed on 16 October 2020).
52. Chen, X.; Liu, S.; Sun, R.; Hong, M. On the Convergence of a Class of Adam-Type Algorithms for Non-Convex Optimization. In Proceedings of the 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, 6–9 May 2019; Available online: <https://openreview.net/pdf?id=H1x-x309tm> (accessed on 16 October 2020).
53. Garcia-Garcia, A.; Orts-Escolano, S.; Oprea, S.; Villena-Martinez, V.; Martinez-Gonzalez, P.; Garcia-Rodriguez, J. A survey on deep learning techniques for image and video semantic segmentation. *Appl. Soft Comput.* **2018**, *70*, 41–65. [CrossRef]
54. Zhang, S.; Zhang, Z.; Sun, L.; Qin, W. One for All: A Mutual Enhancement Method for Object Detection and Semantic Segmentation. *Appl. Sci.* **2019**, *10*, 13. [CrossRef]
55. Gupta, S.; Girshick, R.B.; Arbeláez, P.A.; Malik, J. Learning Rich Features from RGB-D Images for Object Detection and Segmentation. In Proceedings of the Computer Vision—ECCV 2014—13th European Conference, Zurich, Switzerland, 6–12 September 2014; pp. 345–360. [CrossRef]
56. Ye, Y.; Yilmaz, A. An automatic pipeline for mapping roads from aerial images. In Proceedings of the 1st ACM SIGSPATIAL Workshop on High-Precision Maps and Intelligent Applications for Autonomous Vehicles—AutonomousGIS'17, Redondo Beach, CA, USA, 7 November 2017; pp. 1–4. [CrossRef]
57. Open Source Geospatial Foundation. GDAL/OGR contributors GDAL/OGR Geospatial Data Abstraction software Library. 2020. Available online: <https://gdal.org/index.html> (accessed on 30 March 2020).

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



© 2020 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 (<http://creativecommons.org/licenses/by/4.0/>).