remote sensing

MDPI

Article

State-Level Mapping of the Road Transport Network from Aerial Orthophotography: An End-to-End Road Extraction Solution Based on Deep Learning Models Trained for Recognition, Semantic Segmentation and Post-Processing with Conditional Generative Learning

Calimanut-Ionut Cira [1,*], Miguel-Ángel Manso-Callejo [1], Ramón Alcarria [1], Borja Bordel Sánchez [2] and Javier González Matesanz [3]

[1] Departamento de Ingeniería Topográfica y Cartografía, E.T.S.I. en Topografía, Geodesia y Cartografía, Universidad Politécnica de Madrid, C/Mercator 2, 28031 Madrid, Spain
[2] Departamento de Sistemas Informáticos, E.T.S.I. de Sistemas Informáticos, Universidad Politécnica de Madrid, C/Alan Turing, s/n, 28031 Madrid, Spain
[3] Subdirección General de Geodesia y Cartografía, Dirección General del Instituto Geográfico Nacional, C/Gral. Ibáñez de Ibero 3, 28003 Madrid, Spain
* Correspondence: ionut.cira@upm.es

Abstract: Most existing road extraction approaches apply learning models based on semantic segmentation networks and consider reduced study areas, featuring favorable scenarios. In this work, an end-to-end processing strategy to extract the road surface areas from aerial orthoimages at the scale of the national territory is proposed. The road mapping solution is based on the consecutive execution of deep learning (DL) models trained for ① road recognition, ② semantic segmentation of road surface areas, and ③ post-processing of the initial predictions with conditional generative learning, within the same processing environment. The workflow also involves steps such as checking if the aerial image is found within the country's borders, performing the three mentioned DL operations, applying a $p = 0.5$ decision limit to the class predictions, or considering only the central 75% of the image to reduce prediction errors near the image boundaries. Applying the proposed road mapping solution translates to operations aimed at checking if the latest existing cartographic support (aerial orthophotos divided into tiles of 256 × 256 pixels) contains the continuous geospatial element, to obtain a linear approximation of its geometry using supervised learning, and to improve the initial semantic segmentation results with post-processing based on image-to-image translation. The proposed approach was implemented and tested on the openly available benchmarking SROADEX dataset (containing more than 527,000 tiles covering approximately 8650 km^2 of the Spanish territory) and delivered a maximum increase in performance metrics of 10.6% on unseen, testing data. The predictions on new areas displayed clearly higher quality when compared to existing state-of-the-art implementations trained for the same task.

Keywords: road mapping solution; road recognition; road surface area extraction; road predictions post-processing

Citation: Cira, C.-I.; Manso-Callejo, M.-Á.; Alcarria, R.; Bordel Sánchez, B.; González Matesanz, J. State-Level Mapping of the Road Transport Network from Aerial Orthophotography: An End-to-End Road Extraction Solution Based on Deep Learning Models Trained for Recognition, Semantic Segmentation and Post-Processing with Conditional Generative Learning. *Remote Sens.* **2023**, *15*, 2099. https://doi.org/10.3390/rs15082099

Academic Editors: Massimo Losa and Nicholas Fiorentini

Received: 16 February 2023
Revised: 14 April 2023
Accepted: 15 April 2023
Published: 16 April 2023

1. Introduction

At the moment, at the national level, the road mapping task is a challenging, time-consuming, and manual process, even with open access to high-resolution, remotely sensed imagery [1]. It is also important to mention that secondary roads are mostly dismissed, due to the considerable efforts required to digitalize them, although they may become important in the short and medium term, considering the rise of autonomous cars. Thus, being able

to obtain an accurate representation and distribution of the road transport network is a priority for government agencies.

Recent advances in computer vision [2–4] can enable a higher level of automation of the traditional feature extraction approach from aerial imagery and can help achieve road mapping workflows that are applicable at a large scale. However, current road extraction approaches based on machine learning (generally applying modifications to popular algorithms and neural network architectures [5,6]) feature drawbacks such as favorable study areas, the implementation of methods trained on reduced datasets that may not contain the features required to model the road transport network or the implementation of models that are not capable enough to learn a correct road extraction function [7–9]. Existing implementations found in the specialized literature (presented and discussed in Section 2) indicate that the road mapping operation needs optimization.

Moreover, a process tackling the extraction of road representations from aerial images must consider the great complexity of the operation. First, the use of aerial imagery can be challenging due to the defects present in the images (e.g., noise, occlusions, etc.) and the complexity of the ground scenes. Second, roads can be structured (clearly marked highways and city roads, or primary road network) or unstructured (the secondary road network), further complicating their detection and extraction. In this respect, secondary roads often present borders that are not clearly delimited (no markings) and can be easily confused with their surroundings, due to the absence of clearly defined edges or centerlines. Third, roads have different spectral characteristics because of the various kinds of material used (asphalt, cement, gravel, etc.) and are often covered by obstructions present in the scenes (e.g., dense vegetation). Fourth, these structures are complex in nature, due to the differences in widths and the large curvature changes, which can complicate the extraction of detailed information. In addition, imperfections in the ground-truth segmentation masks and particularities of the applied segmentation algorithms also influence the process. For these reasons, the recognition and extraction of the road network can be considered complicated. Scenes encountered in existing official Spanish road cartography that feature some of these mentioned problems are illustrated in Figure 1.

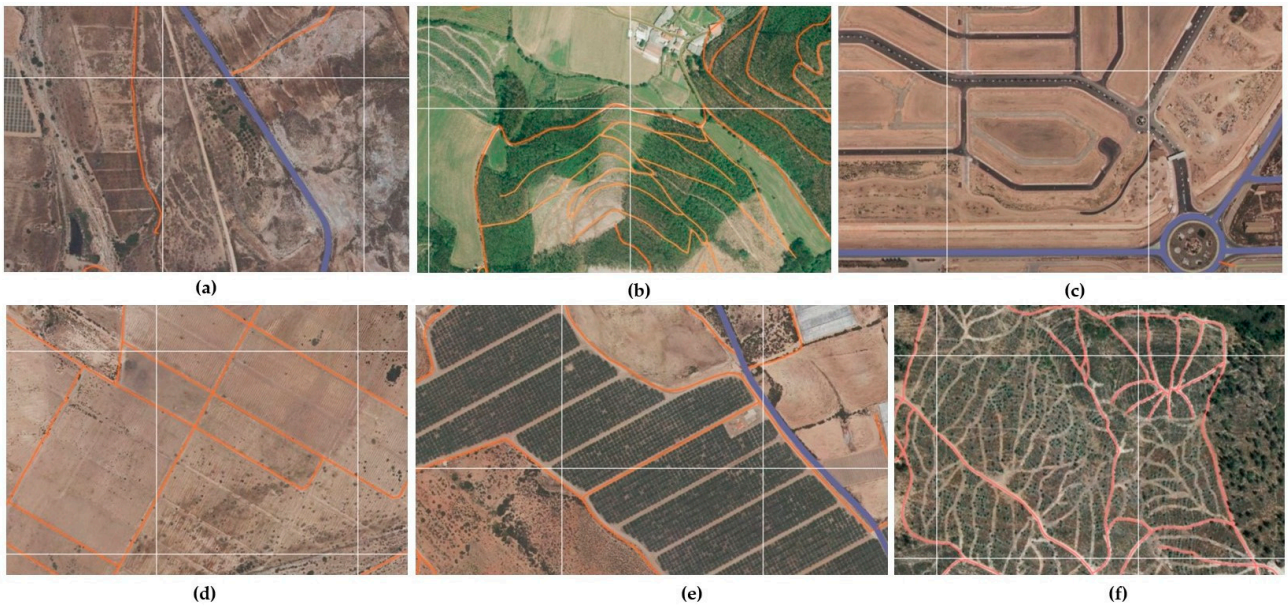

Figure 1. Examples of imperfections found in existent official road cartography: (**a**) impossibility to visually define roads and their category, (**b**) errors caused by large curvature changes, (**c**) incomplete digitalization of roads and subsequently, incorrect cartographic masks, (**d**) secondary roads easily confused with their surroundings, (**d**,**e**) incomplete road cartography and impossibility to decide which path is a road, and (**f**) systematic inconsistencies in differentiating the roads network as a continuous geospatial object.

In this work, the methodology proposed in [10] for extracting point features (where 19,617 wind turbines present in the Spanish peninsular territory were successfully extracted from aerial orthoimages, while obtaining false positive, or FP, rates of 6.4% and false negative, or FN, rates of 1.2%) is expanded, and a novel end-to-end road surface area mapping solution is proposed. "End-to-end" refers to the design framework that enables the full procedure to be executed within the same processing environment, from start to finish—the process developed can be implemented from beginning to end to deliver a fully functional solution and is applicable for road surface area mapping in new regions. The goal is to implement a strategy based on deep learning (DL) models to extract the road transport network (multiline feature) on a large scale (for example, for a whole country, or autonomous region).

The road extraction procedure is tackled as a chained processing workflow of aerial orthophotographs, based on three operations: a road recognition task, ①, to identify tiles where roads are present and avoid unnecessary processing at pixel level of tiles where roads are not present, a semantic segmentation extraction of road surface area, ②, evaluating only the tiles that contain road elements to obtain an initial approximation, and a third operation, ③, of post-processing the initial semantic segmentation results obtained in the second operation with the generator of a conditional GAN trained for image-to-image translation.

The proposed extraction solution was first validated metrically, on a new dataset containing road information from high-resolution aerial orthophotography, covering 8636.94 km^2 of the Spanish territory. The performance levels were evaluated through appropriate machine learning (ML) metrics, such as Intersection over Union (IoU) score ($IoUscore = TP/(TP + FP + FN)$), obtained on unseen data and analyzed with statistical methods to extrapolate the findings to the whole national territory—improvements of maximum 11% in the studied performance metrics were observed. Afterwards, a large-scale proof of concept was performed by qualitatively evaluating a single representative cartographic sheet covering a land area of 28 km × 19 km from the Spanish region of Murcia, unseen during training, to identify the strengths and disadvantages of the proposed solution.

The main contributions of this research are to be applied to road mapping from aerial image data with DL techniques and are summarized as follows.

1. A processing methodology that combines image classification, semantic segmentation, and post-processing operations with DL model components, which delivers considerably improved road surface area extraction results when compared to other state-of-the-art models trained for the same task, is proposed. To the best of our knowledge, this is the first time such an objective has been approached, and a large-scale production solution for road mapping has been applied in this manner.
2. The proposed end-to-end road mapping solution was implemented and evaluated at a very large scale on the SROADEX dataset (containing image tiles of 256 × 256 pixels). The application of the methodology resulted in a maximum increase in the IoU score of 10.6% over the state-of-the-art U-Net [3]—Inception-ResNet-v2 [11] semantic segmentation model (in a metrical comparison on a test set of more than 12,500 unseen images).
3. A large-scale, qualitative assessment of the capabilities was carried out on a single full, unseen orthoimage covering 532 km^2 to identify the advantages and challenges posed by the proposed processing workflow, and future lines of work that could improve the results are discussed.

The remainder of the manuscript is organized as follows. Section 2 describes works related to road mapping using remotely sensed data and ML techniques and focuses on discussing the main drawbacks found in existing literature and on explaining the design decisions taken. Section 3 presents the proposed end-to-end road extraction procedure. Section 4 describes the data used, the large-scale implementation of the extraction solution, together with a metrical comparison of the results obtained by other state-of-art approaches trained for the same task. In Section 5, the results delivered by the extraction solution on a novel, unseen area are analyzed from a non-numerical, qualitative perspective. An

extensive discussion regarding the advantages and disadvantages identified is carried out in Section 6. Lastly, the conclusions of the research conducted are drawn in Section 7.

2. Related Work

One of the most important computer vision applications is related to the analysis of remotely sensed images to detect and extract geospatial objects. However, most existing works focus on geospatial objects with defined boundaries that are independent of the background. What happens when we have to deal with more difficult, continuous geospatial objects such as the road transport network?

Many works tackling road extraction apply traditional ML and image segmentation techniques. For example, Liu et al. [12] applied pre-processing with Gaussian filters to remove input noise and enhance the secondary roads extracted with Hough Transform from an orthoimage of 7000 × 6000 pixels. Mattyus et al. [13] extracted road edges and centerlines from openly available OpenStreetMap data covering 1.5 km^2 of an urban scene using Support Vector Machine [14] combined with a Markov random field inference parameterized in terms of the location. Wang et al. [15] proposed an urban road extract approach based on spatial textures and, subsequently, selected feature like brightness, standard deviation to build a road knowledge model based on mathematical morphology. The traditional ML approach requires the analytical and mathematical modelling of the input-output relations, and the models can deliver a good performance on small datasets but are not suitable for tackling the geospatial complexity of the road transport network. For this reason, a deep learning approach was chosen in this work, as it can deliver models with a higher generalization capacity and can model more complex functions (more capable of describing road-specific, complex features).

Hutchison et al. [16] were among first to approach the road detection task using deep learning. The authors applied supervised methods to build a road extraction model and used unsupervised learning to obtain filters that improved the performance and initial predictions of the model. Post-processing the initial predictions has been traditionally applied by means of conditional random fields (CRFs) [17]. In [18], the authors further improved this proposed network by adding CRFs for sharpening the road predictions and obtaining higher quality extraction results. Dong et al. [19] applied shape filtering to improve the extraction the roads' centerlines by combining high-resolution imagery with LiDAR data and vectoral data from OpenStreetMap. Liu et al. [12] improved the extracted road segments by constructing a geometric knowledge base of rural roads to detect disconnected boundaries.

Various methods to extract roads from remote sensing images have been proposed since then, and most of them focus on extracting the road surface area or extracting the road edges and centerlines. The proposed studies generally apply supervised learning to extract the road representations using radiometric, geometric, or photometric features and transfer learning to improve the model's performance. Existing research related to road extraction with deep learning can be classified according to the type of neural network (NN) applied [20] as follows.

2.1. Approaches Based on Convolutional Neural Networks

In works where the road extraction approach is based on convolutional neural networks (CNNs), the images are evaluated at a patch level using CNNs, and the road predictions are delivered by combing the labelled patches. Zhong et al. [21] combined the outputs of different types of pooling layer with the score of the final layers to obtain extraction precision rates of maximum 78% on the Massachusetts road dataset [22], containing optical satellite images with a spatial resolution of 1 m. Wei et al. proposed RSRCNN [23], based on refined CNNs to road extraction in aerial images, and introduced fusion layers to improve the road extraction operation. The model also incorporates a cross-entropy loss based on minimum Euclidian distance between pixels belonging to the same road section. Wang et al. [24] developed an NN for road extraction that incorporates

a module that improves the classification operation by highlighting high-level information. Alshehhi et al. [25] proposed a patch-based CNN to extract road elements and introduced a post-processing operation based on spatial features to integrate low-level geometrics characteristics and connect ungrouped road sections, achieving performance metrics as high as 82%. The CNN-based technique implemented by Li et al. [26] is capable of anticipating the possibility of each pixel to belong to a road segment and features a centerline extraction module with morphological operators to achieve an IoU score of 0.78.

2.2. Approaches Based on Semantic Segmentation Models

However, road extraction approaches based on semantic segmentation networks are applied by most existing works. Here, the fully connected layers of a CNN are replaced by layers that upsample the last feature maps to the $height \times width$ size of the input image to predict the road labels. The approaches following the encoder–decoder learning structure, where the input image is downsampled to extract the representations up a bottleneck, where the process is reversed, and the feature maps are resized to the original $height \times width$ dimensions [27].

Li et al. proposed Y-Net [28], which includes feature extraction modules (composed of a downsampling–upsampling subnetwork combined with a convolutional subnetwork) for a more detailed feature extraction operation and a fusion extraction module to combine the segmented road classes and deliver improved road segment predictions. Xin et al. implemented DenseUnet [8], based on U-Net's encoder–decoder architecture to extract low-level features such as road edges and textures, obtaining IoU scores of maximum 0.70 on the Massachusetts dataset [22]. Buslaev et al. [29] developed a model based on U-Net [3] and ResNet [30] to extract roads from remotely sensed images and applied a loss function combining binary cross-entropy with Jaccard score to reduce the model's cost and achieved an IoU score of 0.64 on unseen data. Xu et al. [31] introduced M-Res-U-Net, a model based on U-Net and ResNet that features pre-processing with Gaussian filters to reduce the noise. The model was trained on vectorial road data that were rasterized but underperformed in areas where other geospatial objects had similar colors with the road distribution. Zhang et al. developed ResUnet [6], an U-Net-like residual network (decrease in the number of parameters of 75% when compared to U-Net) to semantically segment roads from the Massachusetts roads dataset [22], and obtained In increIse in the performance metrics of 1% when compared to the original U-Net.

Cheng et al. introduced CasNet [9], which includes two cascaded networks—one for detecting road regions and the other for extracting the road centerlines, while taking advantage of the feature maps learned by the first network. The model was trained and tested on a dataset composed of 224 Google Earth images [23] and achieved an IoU score of maximum 0.88. However, the authors recognized the unsuitability of the network for processing areas where trees occlusions are present. Zhou et al. [32] introduced the D-LinkNet model, also based on the encoder–decoder architecture, combined with dilated convolutions and a pre-trained encoder to extract roads from RS imagery. The authors raised concerns regarding the extraction of the road connection points.

Liu et al. built RoadNet [5], composed of a modified version of VGGNet [2] architecture concatenated three times, to analyze and predict the surfaces, edges, and centerlines of roads in urban scenes, and obtained an F1 score of maximum 94%. Wang et al. [33] used VGGNet as backbone of the semantic segmentation model built for autonomous driving and trained it to learn the features of road boundaries and extract them using RGB street scenes tagged at pixel level. Panboonyuen et al. [34] modified SegNet [35] to incorporate the Exponential Linear Unit (ELU) [36] activation function and trained the proposed network for the road extraction task on the Massachusetts dataset while applied data augmentation techniques (angle rotations).

Doshi [37] integrated a ResNet-based model with an Inception [38]-like encoder to extract roads from satellite imagery and obtained an IoU score of 0.61 on a dataset containing 6226 images with a spatial resolution of 0.5 m. He et al. [39] fine-tuned the first layers of an improved U-Net model with cross-modal data converted by an autoencoder. However, the IoU score obtained was only 0.42, and an efficient road extraction approach was not possible in areas where objects with a similar reflectance were present. Xia et al. [40] focused on building a dataset containing roads that have different spectral signatures using a semiautomatic approach and applied the DeepLab [41] segmentation architecture and post-processing by means of several morphological algorithms. The quality of the predictions obtained demonstrate the challenging nature of the road extraction operation.

Gao et al. [42] introduced RDRCNN, a the framework based on a Residual Connected Unit (consisting of a modified ResNet) and a Dilated Perception Unit (consisting of a dilated convolution layer and a convolutional layer), combined with a post-processing stage to extract roads from the Massachusetts dataset and achieve an IoU score of 0.66. Xie et al. proposed HgNet [43], based on LinkNet [44], which introduced a block between the encoder and decoder to "preserve global, long-distance context semantic information and the dependencies among feature channels" ([43], page 11). The model obtained a IoU score of 0.71 on the SpaceNet dataset [45] (a publicly available dataset containing 2213 training and 567 test images of 512 $\times$ 512 pixels) and a IoU score of 0.83 on the DeepGlobe dataset [46] (consisting of 4971, 622, and 622 training, validation, and test images, respectively, with a resolution of 1024 $\times$ 1024 pixels).

As a consideration, in most existing research, the road elements tend to have homogenous hyperspectral signatures and are grouped into clearly defined regions covering smaller areas. A large-scale road extraction operation must consider both structured roads (clearly marked highways, together with rural and urban roads) and unstructured roads (no obvious borderers), with distinct spectral characteristics caused by the various kinds of materials used in pavement (e.g., asphalt, cement, gravel, etc.). This is challenging due to the noise, obstructions, and complexity of the scenes, present in aerial imagery from extended areas, which further complicate the road detection and extraction and the effect that these scenarios have on the extraction operation needs to be reduced.

Furthermore, most relevant studies focus on reduced selected areas featuring favorable scenarios [7] that may not be representative for extracting road elements on a very large scale (as also pointed out in [17]). For example, Kestur et al. [47] proposed UFCN, based on the encoder–decoder structure but trained it on a dataset containing 76 images. Henry et al. [48] obtained IoU scores of a maximum of 45.5% when evaluating the performance of popular NNs in detecting and extracting roads from three test areas covered by satellite images. Alshehhi et al. [25] developed an NN consisting of five convolution layers to extract road geometries from 50 publicly available urban aerial images of 1500 $\times$ 1500 pixels in size, with a spatial resolution of 1 m.

In this work, the mentioned drawbacks are addressed by applying the proposed extraction methodology to the SROADEX dataset, introduced in [49]. SROADEX contains openly available road representations present in official cartography, namely aerial orthoimages from extended areas of the Spanish territory, divided in tiles of 256 $\times$ 256 pixels, and their correspondent segmentation masks. The dataset adds real-world complexity to the road extraction task, avoiding focusing on ideal study scenes to achieve a deep learning-based solution capable of delivering high quality results.

2.3. Approaches Based on Unsupervised Learning

Recently, the road extraction task started to be approached from an unsupervised learning perspective. The identified studies are generally based on conditional generative adversarial networks (GANs), where a generator, G, which is trained to learn the distribution of the data, is used to obtain the road representations. The training process is carried out in an unsupervised setting, where G does not have access to the training data and improves its predictions using the feedback provided by the discriminator D.

In this task, *G*'s training objective is to "deceive" the discriminator by generating new, artificial samples that are closely mimicking those coming from the real data distribution. In parallel, the discriminator's job is to better differentiate data coming from the real and the fake distributions.

Varia et al. [50] proposed a model based on FCN [51] and Pix2pix [4] to extract roads from a dataset containing 189 training and 23 test images, but observed high rates of FN predictions. Shi et al. [52] developed a cGAN architecture using SegNet [35] (based on the encoder–decoder architecture) as *G* to segment roads in aerial images and achieved a 3.5% increase in the F1 score compared the network trained in a supervised setting. Yang et al. [53] applied ensembling techniques to a standard GAN to extract road geometries from rural areas in China and obtained an IoU score of 0.73. Hartmann et al. [54] trained a GAN architecture to synthesize Ioad information and enrich the attributes in areas where the extraction is complicated (e.g., where discontinuities are present, or in complex areas, such as intersections or highway ramps). Costea et al. [55] proposed road extraction solution featuring a GAN stage to detect road edges and a post-processing operation of the initial results using smoothing-based methods. Zhang et al. implemented a Multi-conditional Generative Adversarial Network (McGAN) [56] composed of two discriminators (one to employ the original spectral information and the other to refine the road network topology) and a generator based on ResNet to refine the road topology and obtain more complete road networks graphs. Belli and Kipf [57] approach the road extraction task by using a generative model that recurrently generates road graph candidates based on the conditional information and a multilayer perceptron as attention mechanism. Liu et at. [58] focused on modifying the standard Pix2pix model [4] to apply image-to-image translation and obtain vectors of road markings using point clouds provided by mobile laser scanners.

As a consideration, post-processing extracted road geometries is an active area of research, and unsupervised conditional learning is directly applicable to improving the extraction of geospatial elements from remote sensing images. In this regard, conditional generative learning for image-to-image translation was applied in [59] to transfer the images from a source domain (initial segmentation results) to a target domain (road representation from official cartography) to obtain high-quality road representations. In [60], a conditional GAN architecture for deep inpainting operations, capable of filling the gaps present in the extracted segmentation masks and significantly improving the performance metrics, was implemented.

3. Methodology Proposal of an End-to-End Road Surface Area Mapping Solution Combining Classification, Semantic Segmentation, and Post-Processing with Conditional Generative Adversarial Learning

In this section, the novel processing design and approach for an end-to-end road surface area mapping is presented. As mentioned previously, the methodology presented in [10] is expanded to propose an end-to-end extraction procedure for multiline features that is adapted to avoid the issues related to the extraction of road elements from aerial orthoimages discussed in the "Introduction" section and in Section 2.

The unitary road extraction solution proposed in this work is based on a consecutive execution of a framework for recognizing roads, a semantic segmentation model for extracting their geometries, and a cGAN for post-processing the initial predictions. The proposal features a road recognition stage, introduced before semantic segmentation (known to be computationally expensive [61]), to determine whether the evaluated tile contains or not a road element and to avoid unnecessarily evaluating of tiles that do not contain roads. Second, a hybrid semantic segmentation model is implemented to semantically segment only tiles that contain the studied geospatial element, reducing, in this way, the false predictions. In this regard, modifications to avoid duplicated features caused by the spatial overlap of adjacent orthoimage files are applied. Third, a post-processing step by means of a conditional generative learning model is added to enhance the extracted road features and translate the semantic segmentation predictions to the domain of the roads present in manually tagged cartographic support. The end-to-end solution proposed can

be considered a complete workflow for the road surface extraction from remote sensing imagery and can evaluate large land areas and efficiently joining the predictions to deliver higher quality results.

The methodological solution applied to extract the roads at state-level consists of a script that processes each orthoimage file stored in ECW format, the national administrative boundaries, stored in a shapefile, and the polygons that define the theoretical extent of each orthoimage, also stored in a shapefile.

- The first step is the determination of the bounding box of the region to be processed as the most restrictive coordinates between the orthoimage and the theoretical coordinates for that orthoimage. This procedure is applied to avoid processing twice the additional geographic extent covered by the orthoimages in ECW format.
- From the obtained bounding box, the number of tiles (columns $\times$ rows) to be processed is determined, knowing that, for each 256 $\times$ 256-pixel tile, with resolution 0.5 m/pixel extracted, only the central 192 $\times$ 192 pixels will be considered in the final predictions. This reduces the problem of inaccurate extraction near the edges of the tiles displayed by segmentation networks.
- Before processing each tile, it is checked whether it is contained in, or overlaps with, national administrative boundaries, which allows to discard the processing tiles from sea, or ocean areas, and outside the country's borders.
- The tile satisfying the criteria is processed with the binary classification model, trained for road recognition, to determine whether the tile contains road elements, or not (in operation ①). The pixels of the tiles not fulfilling these criteria will have a probability $p = 0$ assigned. The training procedure of the road recognition DL component and the relation between its input and output are presented in Section 4.2.
- If the tile contains roads (prediction p of road recognition model is higher than 0.5, $p > 0.5$), it is processed with the semantic segmentation network (in operation ②) and stored as a georeferenced GeoTIFF image of a single band, containing the probability that the tile pixels belong to the "Road" class. The stored tile maintains the original size of 256 $\times$ 256 pixels. On the contrary, a $p = 0$ will assigned to every pixel belonging to tiles tagged with the "No_Road" label.
- To subsequently evaluate the results, the probability that the tile contains a road (after passing through the recognition network) is stored in a PostgreSQL [62] with PostGIS [63] database, together with the associated rectangle with its coordinates. The training procedure of the DL component trained for the semantic segmentation of the road surface areas and the relation between the encoding and decoding of the input and output are presented in Section 4.3.
- The processing continues by binarizing the initial segmentation predictions with a decision limit of 0.5.
- Next, the binarized semantic segmentation predictions are evaluated with the generator G of the cGAN network proposed in [59], trained on the SROADEX dataset (in operation ③), to obtain the post-processed, synthetic road predictions. The training procedure and the relation between the input and output of the DL component trained for the post-processing are presented in Section 4.4.
- The values obtained from the GAN network are binarized afterwards with the same threshold $p = 0.5$, with the final result being a binary image that contains the "Road"/"No_Road" labels, at pixel level. The file is stored as a single-band georeferenced image that contains only the central 192 $\times$ 192 pixels; adjacent tiles are processed with the corresponding overlap of 10% (to reduce the effect of prediction errors near image boundaries).

- Finally, after processing all the tiles organized in rows and columns of the orthoimage (approximately 59,000 tiles), a mosaic is constructed with the resulting post-processed predictions and stored in a GeoTIFF file containing the final road mapping predictions.

The proposed methodology is described in pseudocode form in Algorithm 1 and schematically represented in Figure 2. Regarding the notations used in Algorithm 1, "$tile_{ij}$" refers to any processed image tile, "$\rightarrow$" refers to the output of the mentioned step, "p_tile_{ij}" refers to the prediction delivered by the recognition model (probability of a tile to belong to the "Road" class), "p_pixel_{ij}" refers to the prediction delivered by the semantic segmentation model (probability of a pixel to belong to the "Road" class), while $p\_pixel\_cGAN_{ij}$ refers to the prediction delivered by the cGAN model (post-processed probability of belonging to the "Road" class, at pixel level).

Algorithm 1: Road Mapping Procedure

Input: ECW orthoimage file, national boundary and official grid division limits, road recognition model, semantic segmentation model, and cGAN model for postprocessing
Output: extracted road surface areas (image format)

Compute extension of image to process from ECW file and its grid division
Calculate #rows and #columns of 256×256 image to process
for i in #rows **do**
    for j in #columns **do**
        calculate $tile_{ij}$ boundary
        if $tile_{ij}$ inside country boundary **then**
            extract $tile_{ij}$ from ECW file and store in memory
            evaluate $tile_{ij}$ with road recognition model $\rightarrow p\_tile_{ij}$
            if $p\_tile_{ij} > 0.5$ **then**
                evaluate $tile_{ij}$ with semantic segmentation model $\rightarrow p\_pixel_{ij} \in [0,1]$
                store p_pixel_{ij} as GeoTIFF file using $tile_{ij}$ coordinates
            else
                store $p\_pixel_{ij} = 0$ as GeoTIFF file
            end if
        end if
    end for
end for
for each GeoTIFF image file stored **do**
    read p_pixel_{ij}
    binarize p_pixel_{ij} by applying threshold value $p = 0.5$
    evaluate p_pixel_{ij} with postprocessing cGAN model $\rightarrow p\_pixel\_cGAN_{ij} \in [0,1]$
    infer class by thresholding $p\_pixel\_cGAN_{ij}$ with decision limit $p = 0.5$
    store binarized $p\_pixel\_cGAN_{ij}$ overwriting GeoTIFF image file
end for each
mosaic all GeoTIFF image files $\rightarrow$ extracted road surface areas (image format)
clean intermediate files (GeoTIFF tiles)

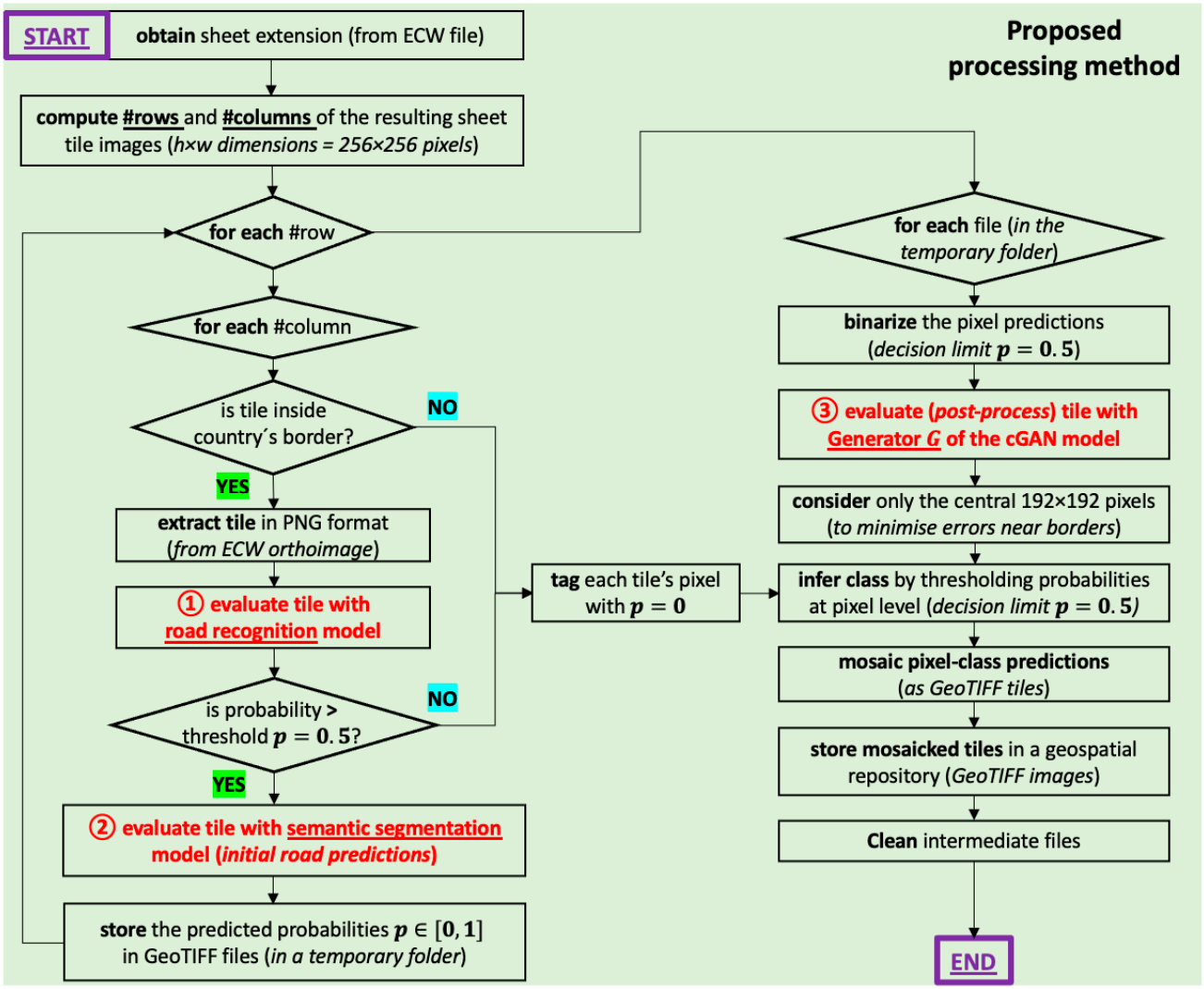

Figure 2. Flowchart describing the proposed end-to-end, large-scale road mapping procedure.

4. Implementation of the Proposed End-to-End Road Mapping Solution on the SROADEX Dataset

The road extraction solution proposed in Section 3 takes as input orthoimages with a spatial resolution of 0.5 m, divided in tiles of size 256 × 256 pixels—these tiles will be processed with specialized DL models to recognize, extract, and improve the surface area predictions of the road elements (an extended graphical description can be found in Figure 3). To ensure a higher generalization capacity of the resulting DL models, the three artificial neural network components were trained on the SROADEX large-scale dataset [49]—an Ubuntu server featuring four Nvidia Tesla V100 Graphical Processing Units with 16 GB of VRAM was employed to train the components and implement the solution.

Next, the three operations applied will be presented, and a discussion of the input-output modelling of the will be introduced in the sections corresponding to the implementation of each DL component. It is important to mention that the image classification and semantic segmentation are supervised operations (where the input data are remotely sensed images and their corresponding tags, and the goal is to learn how to predict if an image contains a road or not using the labels of the input data). On the contrary, conditional generative learning is an unsupervised learning task, where only input variables, X, (and no output variables, y), are given to the deep learning model.

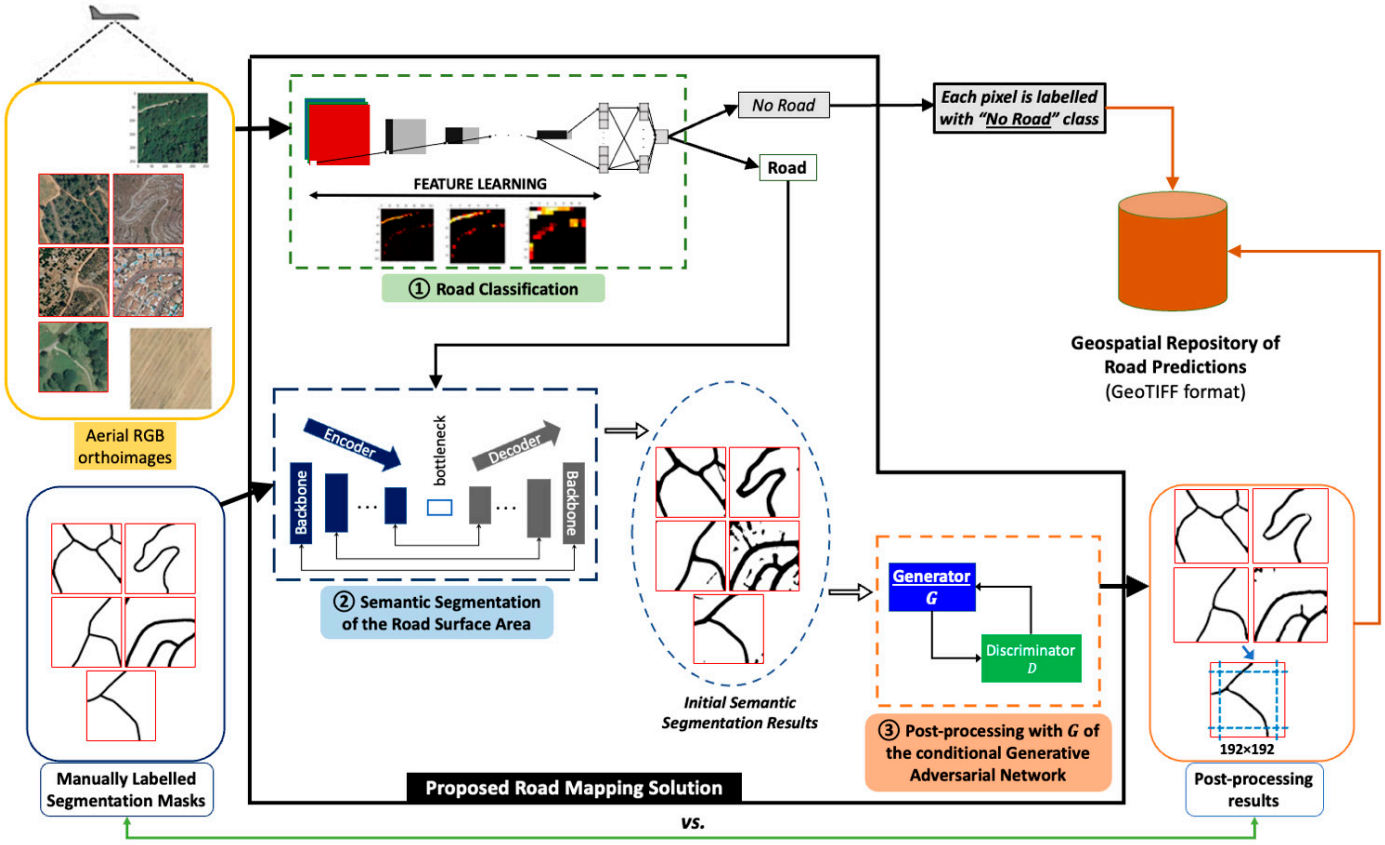

Figure 3. Graphical overview of the implementation of the road extraction solution proposed (from the perspective of the trained deep learning models).

4.1. Data

In previous studies [59,60,64,65], having used datasets containing around 10,000 tiles of 256 × 256 pixels from representative areas, it was observed that one of the main drawbacks of the training processes was the lack of sufficient data, as inconsistent predictions were frequently present. To implement and test the large-scale road extraction solution, the SROADEX dataset [49] was used (hosted in the Zenodo repository [66], openly available under a CC-BY 4.0 licence). The data were selected because the information and variables it contains are produced and distributed by public agencies (under the supervision of the National Geographical Institute of Spain). The strategy applied for creating the SROADEX dataset is detailed in Section 2 of [49].

The SROADEX dataset covers a land area of 8636.94 km^2 of the Spanish territory and contains n = 527,157 tiles of 256 × 256 pixels with a spatial resolution of 0.5 m with information related to main and secondary transport routes, urban, and rural roads (tile examples can be found in Figure 2 of [49]). The first component of the SROADEX dataset is represented by the digital RGB aerial orthophotographs from the entire Spanish territory (distributed under the "National Plan of Aerial Orthophotography" [67], or PNOA product). In this regard, the aerial imagery was acquired with calibrated instruments, corrected, and rectified, and its traceability can be ensured. The second component of the SROADEX dataset are the ground-truth predictions, coming from BTN information related to road elements such as highways, conventional roads, tracks, and paths that were manually checked and corrected by an operator (the ground truth road digitalization results have been generated by specialized operator and crosschecked before being published). Finally, the vectorial representations were rasterized to achieve the ground-truth masks of the road elements and provide their pixel-level semantic representation.

SROADEX dataset was randomly pre-split with a division ratio of 90:5:5% (splitting criterion chosen because it allows for more training data, as recommended in [64]) to

generate the training, validation, and test sets, respectively, and ensure that the same data are used for the training and testing of the models.

4.2. Road Classification Operation

In the proposed extraction workflow, the road recognition step is first applied to avoid the unnecessarily processing of the areas with semantic segmentation models (as many of the tiles do not contain road elements). For task ①, the training approach described in [64] was followed to obtain the classification models specialized in road recognition. The data used for the road recognition task contain the same type of representations as the tiles presented in Figures 1 and 2 of [64]. By applying the 90:5:5% data split criterion to SROADEX, positive samples (tagged with "Road" label) were divided as follows: 225,397 for the training set, 12,523 in the validation set, and 12,522 for the test set; the negative samples (tagged with "No_Road" label), were divided as follows: 249,043 in the training set, 13,836 in the validation set, and 13,836 in the test set.

To lower and optimize the computational requirements of an ensemble framework (stacking multiple weak learners together, although increases the performance, also increases the computational needs and, consequently, the evaluation time—an important aspect in the design of a production solution), two neural network architectures based on VGGNet, named VGG-v1 (featuring 15,240,513 parameters), and VGG-v2 (featuring 25,733,953 parameters), were trained. These convolutional neural network architectures networks were introduced in Table 1 of [10]. They both feature the default VGG16 configuration as convolutional base but display a different disposition of the classifier layers. VGG-v1 features three fully connected (FC) layers of [512, 512, 1] units, while VGG-v2 features three FC layers with [3072, 3072, 1] units. Both networks use MSRA initialization [68] for the first FC layer of the classifier and apply the feature extraction method of transfer learning to re-use the weights learned on the ImageNet Large Scale Visual Recognition Challenge [69]. For comparison reasons, the original VGGNet (featuring 65,058,625 parameters, for an input size of 256 × 256 pixels) from scratch using random initialization was also trained.

As for the input–output modelling, each aerial image of 256 × 256 × 3 was first downsampled (as shown in Figure 3) by being passed through the learning structure defined by the convolutional base (composed of a succession of convolutional and pooling layers). This enabled the recognition models to learn filters that detect different road-specific features. In the classifier part of the model, the modelled classification function's response to the loss computed between the expected label input and the predicted label (a single "Road"/"No_Road" tag, at tile level) controls the evolution of learned filter representations, favoring those that helped obtain correct predictions. Additional explanations are provided in Sections 4 and 5 of [64].

During training, a weight matrix, $w_j = n_{samples}/(n_{classes} \times n_{samples_j})$ was applied to increase the importance of the minority class (as the SROADEX data for road recognition dataset is slightly imbalanced). The weight, w, for each class, j, was represented by the division of the total number of tiles in the dataset and the product two classes in the dataset and the number of tiles in the class j—resulting a weight matrix $w_{i,j} = \{1.0525, 0.9525\}$. Similarly to the learning approach presented in [64], the model's loss was calculated with the formula $\mathcal{L}(\hat{y}_i, y_i) = \frac{1}{m}\sum_{i=1}^{m} y_i * \log \hat{y}_i + (1 - y_i) * \log(1 - \hat{y}_i)$, as the average negative log-likelihood for each event output value over a target (considering the actual and the predicted probabilities). To reduce the computed loss, backpropagation with the Adam [70] optimizer with a weight update rule of $1e^{-5}$ was applied. The models were trained until signs of overfitting were observed (lower cost on the training set when compared to the validation set).

The experiments were repeated three times to obtain the mean values and the standard deviation of the performance metrics. The performance reported by the trained architectures can be found in Table 1, in terms of loss value and accuracy $((TP + TN)/(TP + FP + TN + FN))$, precision $(TP/(TP + FP))$, recall $(TP/(TP + FN))$,

F1 score $\left(TP/\left(TP+\frac{1}{2}(FP+FN)\right)\right)$, and AUC-ROC score (shows how much a model is capable of distinguishing between classes) metrics.

Table 1. Mean (M) and standard deviation (SD) of the performance metrics obtained by the configurations trained for road recognition with n = 474,437 tiles on the validation (containing n = 26,359 tiles) and the test set (containing n = 26,358 tiles).

Performance Metric		VGG-v1		VGG-v2		VGGNet Trained from Scratch	
		Validation	Test	Validation	Test	Validation	Test
Loss	M	0.2536	0.2511	0.2550	0.2555	0.2622	0.2581
	SD	0.0039	0.0023	0.0021	0.0033	0.0040	0.0046
Accuracy	M	0.8956	0.8967	0.8946	0.8961	0.8928	0.8953
	SD	0.0008	0.0007	0.0014	0.0018	0.0020	0.0017
Precision (weighted)	M	0.8964	0.8974	0.8946	0.8961	0.8929	0.8953
	SD	0.0006	0.0005	0.0015	0.0018	0.0019	0.0016
Recall (weighted)	M	0.8956	0.8967	0.8946	0.8961	0.8928	0.8953
	SD	0.0008	0.0007	0.0014	0.0018	0.0020	0.0017
F1 score (weighted)	M	0.8957	0.8967	0.8946	0.8961	0.8928	0.8953
	SD	0.0009	0.0007	0.0015	0.0018	0.0020	0.0017
AUC-ROC score	M	0.9634	0.9640	0.9605	0.9608	0.9579	0.9590
	SD	0.0003	0.0012	0.0008	0.0016	0.0004	0.0007

The best performing road classification architecture was VGG-v1, as it achieved the lowest average loss and yielded higher average accuracy, recall, F1 and AUC-ROC scores on the test set. The best performing VGG-v1 model obtained a minimum loss value of 0.2506 on the validation set used during training, and maximum accuracy, weighted precision, weighted recall, weighted F1 score, and AUC-ROC score values of 0.8957, 0.8961, 0.8957, 0.8957, and 0.9635, respectively; and a loss value of 0.2485 on the test set containing unseen data, and accuracy, weighted precision, weighted recall, weighted F1 score, and AUC-ROC score values of 0.8974, 0.8978, 0.8974, 0.8975, and 0.9639, respectively. This network was trained for 30 epochs, with an average duration per epoch of 4985 s (approximately 1 h and 23 min) and delivered lower error ratios on the validation and test sets.

4.3. Semantic Segmentation of Road Surface Areas Operation

The semantic segmentation component was trained and tested on the aerial tiles labelled with the "Road" label in the SROADEX dataset, together with their corresponding ground-truth masks (tagged with road information at pixel level, from MTN50), following the learning techniques tested in [65]. The data covers a land area of 4103.24 km^2 and contains 250,442 pairs of tiles that were divided by applying the 90:5:5% data split criterion. Road representations used for semantic segmentation are similar to the road data presented in Figure 1 of [65]. Following the methodology described in Section 3, task ② will be applied only to tiles that contain road elements, due to the prior use of the road recognition operation (in which the 65,536 pixels of tiles that do not contain road elements are automatically tagged with the label corresponding to the "No_Road" category).

The models trained for Image segmentation are based on U-Net and were implemented using the "Segmentation Model" deep learning library [71]. Following [65], the default encoder and decoder backbones were replaced with NNs specialized in extracting more complex objects. The semantic segmentation models considered for training are following the architecture–backbone configurations described next: U-Net—SEResNeXt50 [72] (featuring 34,594,177 parameters), together with the U-Net as base architecture, coupled with Inception-ResNet-v2 (featuring 57,868,721 parameters), and the original U-Net architecture from scratch (as proposed by the authors). These implemented configurations

represent state-of-the-art semantic segmentation networks and follow the encoder–decoder learning structure.

In line with the road extraction approach applied in [65], Adam optimizer [70] was applied, with a variable learning rate and early stopping to optimize the binary cross entropy combined with Jaccard loss function. The loss function, $L(gt, pr) = -gt \times log(pr) - (1 - gt) \times log(1 - pr)$, measures the similarity between prediction (pr) and the ground-truth (gt) to reduce the cost via backpropagation. To control the evolution of the training, performance metrics specific to image segmentation that correctly handle the unbalanced nature of the two classes considered ("Road" and "No_Road", as road pixels generally occupy around 5–10% of the pixels in any given tile) were computed.

As for the input–output modeling, the training required pairs of aerial images of size $256 \times 256 \times 3$ and their ground truth masks of size $256 \times 256 \times 1$. The aerial images were downsampled in the encoder by convolutional structures to extract road-specific representations in feature maps, up to a bottleneck, when the process is reversed in the decoder by means of transposed convolutions that upsample the tensor to its original size. The encoder–decoder structure also uses skip connections to transfer information between processing levels containing feature maps of same index (as shown in Figure 3). The classifier part makes use of the filters learned (detecting road-specific features) to model the function capable of delivering predictions at pixel level (keeping the depth-wise argmax of each pixel) at pixel level, to output a single channel image of $256 \times 256 \times 1$. Additional explanations are provided in Section 4 of [65].

The training experiments were repeated three times to obtain their mean and standard deviation of the IoU score, F1 score, precision, and recall (for both the positive and negative class), and the pixel accuracy and the loss performance values were delivered by the implemented semantic segmentation models. The results computed on the validation and test sets are presented in Table 2.

Table 2. Mean (M) and standard deviation (SD) of the performance metrics obtained by the semantic segmentation networks trained with $n = 225,397$ tiles on the validation set (containing $n = 12,523$ tiles) and on the test set (containing $n = 12,522$ tiles).

Performance Metric		U-Net (Trained from Scratch)		U-Net—SEResNeXt50		U-Net—Inception-ResNet-v2	
		Validation	Test	Validation	Test	Validation	Test
Loss	M	0.5527	0.5473	0.5170	0.5212	0.5057	0.5108
	SD	0.0037	0.0212	0.0028	0.0027	0.0039	0.0040
IoU score (positive class)	M	0.2941	0.2903	0.3197	0.3160	0.3470	0.3435
	SD	0.0035	0.0034	0.0117	0.0115	0.0212	0.0214
IoU score (negative class)	M	0.8674	0.8664	0.8798	0.8772	0.8887	0.8876
	SD	0.0051	0.0055	0.0038	0.0042	0.0264	0.0274
IoU score	M	0.5808	0.5784	0.5998	0.5966	0.6179	0.6155
	SD	0.0027	0.0029	0.0043	0.0039	0.0210	0.0216
F1 score (positive class)	M	0.4271	0.4223	0.4555	0.4511	0.4848	0.4805
	SD	0.0042	0.0043	0.0128	0.0128	0.0222	0.0225
F1 score (negative class)	M	0.9148	0.9150	0.9230	0.9215	0.9277	0.9273
	SD	0.0057	0.0060	0.0045	0.0048	0.0222	0.0230
F1 score	M	0.6709	0.6687	0.6893	0.6863	0.7062	0.7039
	SD	0.0027	0.0029	0.0038	0.0042	0.0198	0.0203
Pixel accuracy	M	0.8710	0.8701	0.8831	0.8806	0.8916	0.8907
	SD	0.0051	0.0055	0.0038	0.0042	0.0262	0.0272
Precision (positive class)	M	0.2986	0.2947	0.3242	0.3204	0.3527	0.3492
	SD	0.0036	0.0035	0.0121	0.0120	0.0226	0.0229

Table 2. *Cont.*

Performance Metric		U-Net (Trained from Scratch)		U-Net—SEResNeXt50		U-Net—Inception-ResNet-v2	
		Validation	Test	Validation	Test	Validation	Test
Precision (negative class)	M	0.9977	0.9976	0.9979	0.9980	0.9978	0.9979
	SD	0.0001	0.0001	0.0001	0.0000	0.0003	0.0002
Precision	M	0.6481	0.6462	0.6611	0.6592	0.6753	0.6735
	SD	0.0018	0.0017	0.0061	0.0060	0.0112	0.0114
Recall (positive class)	M	0.8597	0.8554	0.8695	0.8626	0.8662	0.8592
	SD	0.0053	0.0013	0.0014	0.0008	0.0059	0.0064
Recall (negative class)	M	0.8692	0.8682	0.8814	0.8788	0.8904	0.8893
	SD	0.0051	0.0056	0.0038	0.0042	0.0265	0.0275
Recall	M	0.8644	0.8618	0.8755	0.8707	0.8783	0.8743
	SD	0.0044	0.0022	0.0024	0.0024	0.0123	0.0126

The best performing segmentation architecture for road extraction was U-Net—InceptionResNet-v2, because it achieved the lowest average loss and yielded the highest average accuracy, recall, F1 and AUC-ROC scores, both on the test and the validation sets. The results are aligned with the ones obtained in [65] but present an increase of 4–5% in the IoU score, due to the use of a significantly bigger dataset. These performance values are to be expected when performing a large-scale extraction of the road surface areas.

The best performing U-Net—Inception-ResNet-v2 model achieved a maximum IoU and F1 scores of 0.6264 (0.3409 for the positive class and 0.9118 for the negative class), and 0.7132 (0.4788 for the positive class and 0.9476 for the negative class), respectively, a pixel accuracy of 0.9149, a mean precision of 0.6722 (0.3463 positive class, 0.9980 for the negative class), and a mean recall of 0.8872 (0.8610 positive class, 0.9134 for the negative class), together with a minimum loss of 0.5113. This model was trained for 30 epochs, with an average duration per epoch of 6325 s (approximately 1 h and 45 min) and delivered the lowest error ratios on the validation and test sets, in both the positive and negative classes. The maximum values obtained by the best segmentation model represent the metrics that are to be improved with the post-processing suboperation.

4.4. Post-Processing of the Initial Segmentation Masks with Conditional Generative Adversarial Learning

Similar to [65] (where the results that state-of-the-art semantic segmentation architectures deliver when trained for the road surface area extraction task were analyzed), in this study, the predictions delivered by the best semantic segmentation model were not always satisfactory (discontinuities, overlooked connection points, or isolated road segments were present). For the post-processing operation ③, the SROADEX training and validation sets with the ground-truth masks from task ② were joined to obtain a bigger training set (resulting in 237,920 tiles, as unsupervised learning does not require a validation set). The same test data from task ② (containing 12,522 tiles) were used to evaluate the performance of the trained cGAN. The data used for post-processing with cGANs are similar to the road data representations from Figure 2 of [59], and Figure 2 of [60].

The post-processing operation, added to improve the initial segmentation results, follows the learning procedure of the conditional GAN architecture proposed in [59]. The initial segmentation mask (of size of 256 × 256 × 1) is added as a condition to the latent space z (containing Gaussian noise), and the training goal is to obtain improved road representations that are similar to the ground-truth domain (from an input of 256 × 256 × 1) via image-to-image translation operations. The conditional model for post-processing is trained through unsupervised learning from z.

The cGAN features a generator G, based on U-Net, which is trained to produce realistic synthetic data, $G(z|x)$, and a discriminator D based on PatchGAN, which is trained to distinguish between real images (from the ground-truth domain, Y) and the fake generated

data ($G(z|x)$). During training, the original ground-truth mask of 256 × 256 × 1 is only provided to the discriminator. However, with the feedback received from the discriminator, G will improve at generating realistic data, while D will become better at differentiating synthetic data to the real data distribution (target domain containing ground-truth representations of roads). The objective function of the model can be expressed with $\mathcal{L}(G, D) = E_{x,y}[\log D(x, y|x)] + E_{x,z}(1 - \log D(x, G(z|x)))$ [59], $E_{x,z}$ being the expected value over all generated fake instances $G(z|x)$, given the condition x, and $E_{x,y}$ being the expected value over all real data instances (belonging to the density distribution to be replicated), given x [4,59]. After training, the trained G will be used to improve the initial segmentation masks of size 256 × 256 × 1 and output an improved, post-processed version of the same size. An in-depth explanation of the training procedure can be found in Sections 3–6 of [59].

The cGAN model was trained three times to obtain the mean and standard deviations of the IoU score, F1 score, precision, and recall (for both the positive and negative classes), and the pixel accuracy performance values delivered by the implemented model. The results computed on the test set present significant increases in the considered performance metrics and can be found in Table 3.

Table 3. Comparison between the performance metrics obtained by applying the post-processing operation with a conditional GAN trained on n = 237,920 tiles and the best performing semantic segmentation model on the test set containing n = 12,522 unseen tiles (covering a land area of 205.16 km^2).

Performance Metric	Best Semantic Segmentation Model (*)	Post-Processing with the cGAN Proposed in [59]			
		Mean Value and Standard Deviation	Mean Percentage Difference (**)	Maximum Result	Maximum Improvement (***)
IoU score (positive class)	0.3409	0.4959 ± 0.0053	+15.50%	0.5012	+16.03%
IoU score (negative class)	0.9118	0.9629 ± 0.0005	+5.11%	0.9628	+5.10%
IoU score	0.6264	0.7294 ± 0.0026	+10.30%	0.7320	+10.57%
F1 score (positive class)	0.4788	0.6118 ± 0.0051	+13.30%	0.6175	+13.87%
F1 score (negative class)	0.9476	0.9764 ± 0.0003	+2.88%	0.9763	+2.87%
F1 score	0.7132	0.7941 ± 0.0024	+8.09%	0.7969	+8.37%
Pixel accuracy	0.9149	0.9640 ± 0.0005	+4.91%	0.9639	+4.90%
Precision (positive class)	0.3463	0.6170 ± 0.0116	+27.07%	0.6088	+26.25%
Precision (negative class)	0.9980	0.9899 ± 0.0006	−0.81%	0.9904	−0.76%
Precision	0.6722	0.8034 ± 0.0056	+13.13%	0.7996	+12.75%
Recall (positive class)	0.8610	0.6592 ± 0.0190	−20.18%	0.6762	−18.48%
Recall (negative class)	0.9134	0.9725 ± 0.0010	+5.91%	0.9719	+5.85%
Recall	0.8872	0.8158 ± 0.0091	−7.14%	0.8241	−6.32%

Notes: (*) refers to the performance values obtained by the best semantic segmentation model (i.e., metrics to be improved by the cGAN model); (**) refers to the difference with respect to the initial semantic segmentation results; (***) refers to the difference between the maximum result and the initial semantic segmentation results.

In Table 3, it can be observed that the post-processing operation delivered a mean IoU score of 0.7294 ± 0.003, an average increase of 10.3% over the maximum IoU score of 0.6264 obtained by U-Net—Inception-ResNet-v2. The results are aligned with those obtained and discussed in [59,60] and display the precision-recall trade-off scenario, where an average of 7.14% of the recall performance delivered by the semantic segmentation model was sacrificed to achieve average a mean increase of 13.1% in precision. This trade-off scenario, where the precision values are increased at the cost of the recall metric (a higher precision involves minimizing false positive rates; a higher recall involves minimizing false negative rates) is to be expected, considering the imbalanced classes present in the training data (featuring significantly fewer positive samples), due to the nature of the

studied geospatial object. Nonetheless, precision and recall scores should not be discussed in isolation, and for this reason, the F1 score, and the pixel accuracy performance values were also computed. In this regard, the cGAN model achieved a mean increase of 8.1% over the initial F1 score value of 0.7132. In addition, the correct predictions have a higher ratio compared to the initial segmentation masks, a mean increase of 4.9% in the pixel accuracy being observed when compared to the initial value of 0.9149 obtained by the best-performing segmentation model.

5. Evaluation of a Single, Unseen Orthoimage File Covering 532 km^2

Next, a large-scale evaluation of a new, unseen, and untagged area was performed with the processing procedure proposed in Figures 2 and 3. For this proof-of-concept pilot, a large-scale extraction of the road transport network from a selected territory in the Region of Murcia (the PNOA orthoimage corresponding to the area covered by the sheet 0997-Águilas of MTN50) was conducted to study, in a qualitative manner, the significance of the performance metrics was computed in Sections 4.2–4.4.

The selected PNOA orthophotograph covers a land area of approximately 28 km × 19 km, is georeferenced in ETRS89 (compatible with WGS84), UTM zone 31, and can be downloaded in the ECW (Enhanced Compression Wavelet) file format from *Centro de Descargas del CNIG* [73], maintained by the Spanish *Instituto Geográfico Nacional*. This land surface was chosen for its localization within the Mediterranean region and features coastlines, and hilly regions, dry and green vegetation—scenarios similar to those used for training the DL models. The area was evaluated to identify the strengths and disadvantages of the proposed method, and to compare the results with those obtained by the standard U-Net trained from scratch.

Example of predictions from rural scenes can be found in Figure 4—the first column presents the aerial orthoimages, the second column features the road predictions delivered by U-Net trained from scratch, while the third column features the predictions obtained with the proposed road mapping solution. Examples of results delivered in urban regions are presented in Figure 5—the first column features the aerial orthoimages, the second column represents the road predictions delivered by U-Net trained from scratch, while the third column features the predictions obtained with the proposed end-to-end road surface area extraction solution presented in this work.

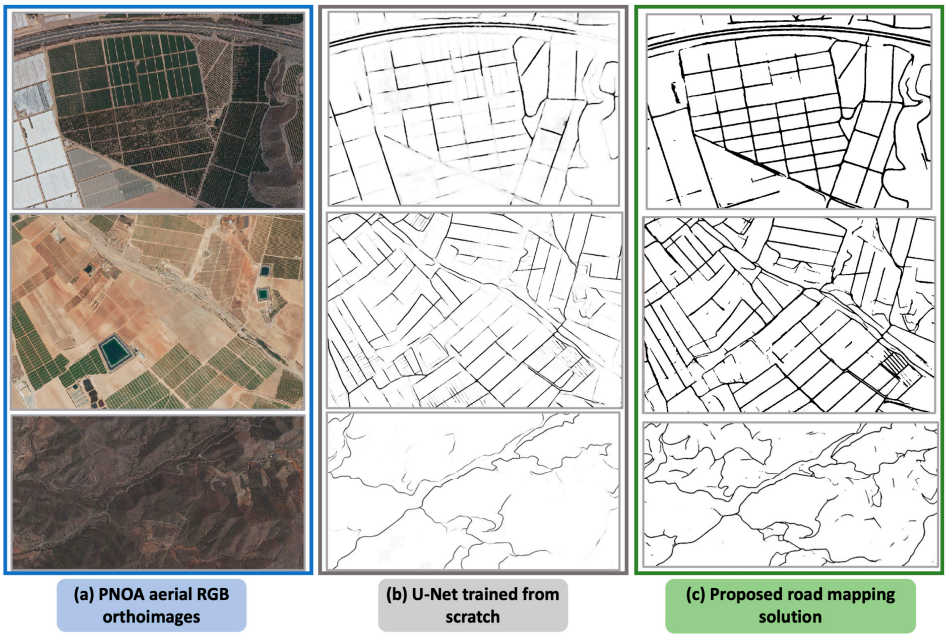

Figure 4. Examples of rural scenes considered in the qualitative evaluation of the road surface area extraction process from a PNOA orthoimage covering 532 km^2.

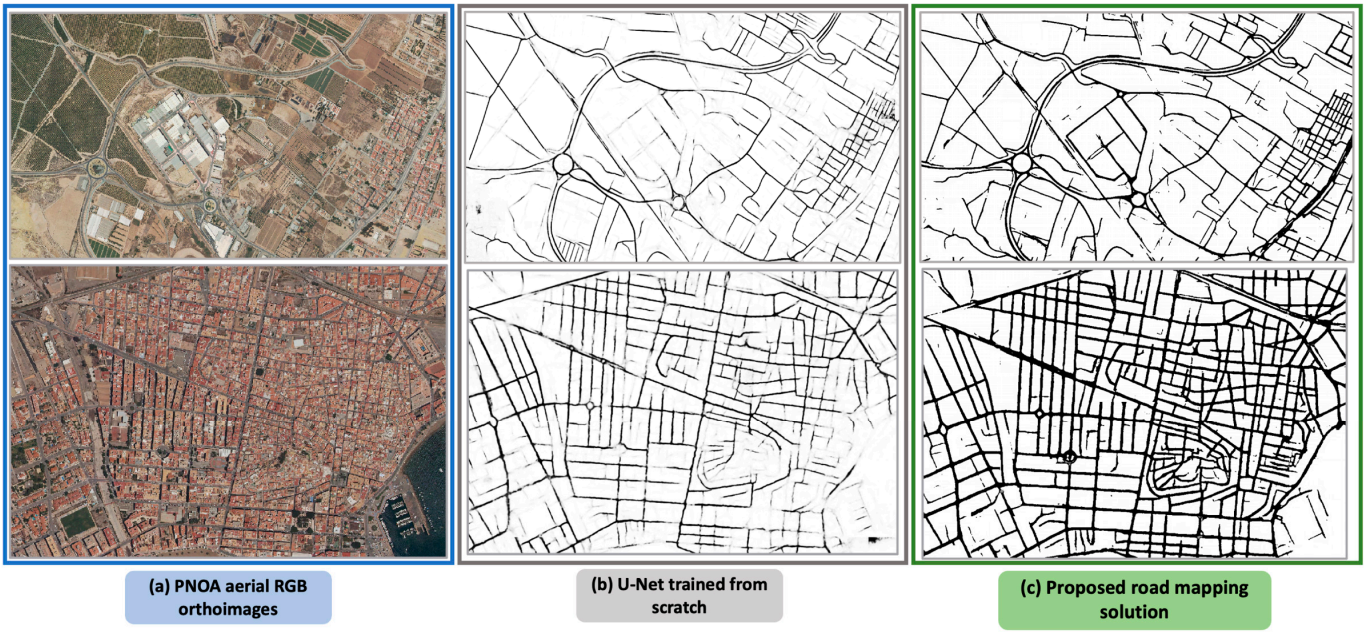

Figure 5. Examples of urban scenes considered in the qualitative evaluation of the road surface area extraction process from a PNOA orthoimage covering 532 km^2.

The semantic segmentation with U-Net trained from scratch delivered the worse road extraction predictions, as the evaluated areas present high rates of false predictions (as found in Figures 4b and 5b)—the road geometries would require a lot of debugging by a mapping operator to generate a final road extraction dataset. In contrast, the consecutive execution of road recognition, road extraction via semantic segmentation step, combined with the post-processing step resulted in significantly reduced processing artefacts and errors and improved road surface area predictions (as found in Figures 4c and 5c). The proof-of-concept experiment supported the performance metrics values achieved throughout the project; the workflow delivered significantly improved road mapping results, and, subsequently, reduced the rates of FP and FN predictions.

6. Discussion

The road extraction solution proposed presented as an end-to-end processing strategy based on deep learning models is capable of successfully analyzing extended, unseen areas, and evaluating each pixel of the input aerial imagery to deliver final "Road"/"No_Road" predictions. The objective of this study–automatically obtain road representation similar to those present in official cartography by applying the proposed extraction workflow–was successfully reached, as the generated results display high performance metrics, and the qualitative interpretations carried out proved the obtention of high-quality road representations.

6.1. Advantages of the Proposed Solution

The performance metrics obtained when training the component models show considerable increases (gains in the IoU score of +10.57% when compared to U-Net—Inception-ResNet-v2), a significant reduction in the FP errors, together with a significant increase in the TN ratios. In the qualitative evaluation of extended unseen areas, it was observed that the predictions have notably improved when compared to U-Net model, proving that the extraction of the road surface areas using the proposed workflow can deliver improved results when compared to existing implementations applied for the same large-scale extraction task.

Different from existing road extraction solutions, here, the semantic segmentation task is preceded by a road classification task (by means of a binary classification network) that helps reduce the unnecessary processing of areas that do not contain road elements

and is succeeded by a conditional GAN model (trained to improve the extracted road representations). This approach eliminated the unnecessary processing of all the tiles with semantic segmentation models (known to be computationally more expensive, and an important source of the FP and FN predictions) by including a filtering of the images to be semantically segmented and automatically tagging the pixels of tiles that do not contain roads, or those that fall outside the administrative boundaries, with the "No_Road" label.

From an economic viewpoint, the proposal enables a higher degree of automation of the road cartography generation and updating processes and has the potential to greatly reduce costs by minimizing the human factors. Moreover, the proposed solution can play an important role in road infrastructure monitoring and its digital data integration, as it enables a great degree of exploitation of existing, openly available geographic support for automatic extraction tasks. In this regard, the mapping procedure can be applied to enhance other similar extraction tasks of continuous geospatial elements (such as the mapping of riverbeds or railroads), by retraining the DL model components on datasets containing the specific object considered. The research could also help map other non-continuous geospatial objects (after large appropriate datasets were created for the supervised learning part) or serve as base for developing additional extraction workflows from remote sensing images.

As for the development and efficiency of the proposed solution, the increased processing complexity resulted from applying the three DL operations, was counterbalanced by the delivery of a significantly higher quality road predictions in a processing procedure that requires fewer debugging hours from part of a human operator (resulting in decreased overall processing times). Therefore, although the total processing time has not significantly decreased, the quality of the results delivered is higher.

Another important advantage of the proposed solution is the easy parallelization of the extraction procedure to the level of the entire national territory, as this processing procedure only needs to be replicated for the rest of 1061 PNOA orthoimages. From a data processing viewpoint, the proposed methodology and the developed components make use of the computing resources available on a machine and scale effectively, as the processing jobs can be run in parallel into several machines using the same processing environment, and the results can be stored in a centralized geospatial repository. However, due to the large processing times and the additional computational power required, processing the whole national territory falls beyond the objective of this research, but the proof-of-concept experiments carried out proved the feasibility of the task.

Finally, the end-to-end solution is versatile, as it enables an easy integration of future development in the field—the featured components can be easily replaced in the workflow by newer models that improve the manipulation of geospatial data. The solution is also capable of delivering results in the GeoTIFF format, compatible with geodatabases, which further increases its applicability.

6.2. Challenges Posed by the Proposed Solution

The proposed processing strategy requires DL model components trained on large datasets—the training operations can take several days of training per model, and it is recommended that several training iterations are run to identify the best possible model. We defer the analysis of additional DL models as DL components and the comparison and explorations of additional road extraction methods to future work, due to the size of the dataset considered for implementing the end-to-end solution and the long training times and computational efforts required. Nonetheless, it is important to note that, once the training is completed, the models perform the evaluation in very short times (by the orders of milliseconds, for any given new tile of 256 $\times$ 256 pixels).

It was noticed that tiles where road elements occupy a very small part of the road (in the corners of tiles, etc.) have a negative influence on the results delivered in the road classification model. Regarding the segmentation operation, significant prediction artefacts in tiles where the representations of the classes were very unbalanced were observed (when more than 95% of the pixels in a tile belonged to a certain class). The IoU score is sensible in

such scenarios, and the lower performance score computed can influence the evolution of the training and the learned road representations. It is advised that we remove this type of image data from the training set. For the final post-processing operation, the main challenge was the unsupervised nature of the conditional GANs, where the learning evolution is difficult to control and the interpretability decreases (several post-processing scenarios where the initial road geometries were counterintuitively deleted were identified).

Important inaccuracies were observed in the road extraction from urban and hilly areas, due to the natural underrepresentation of the mentioned areas in datasets covering extended territories. Furthermore, urban regions feature additional challenges for the road extraction (e.g., materials used in pedestrian pavements have similar spectral signatures, increased presence of occlusions). Future implementation should consider the inclusion of multiple specialized models, to increase the quality of the results from the mentioned regression (for example, an additional semantic segmentation network specialized in extracting road elements from urban areas).

Existing orthophotographs provided by public agencies are not radiometrically and optically homogenous, and some of the artefacts and errors observed can be attributed to imperfections present in the existing data (e.g., differences in brightness levels between scenes). To reduce this effect, the networks were trained with significant data augmentation levels (especially random changes in brightness, contrast and gamma shifts, and grid and optical distortions) to expose the model to more aspects of data.

In addition, the available cartographic support does not have a uniform spatial resolution, and the road representations present in cartographic support do not always cover the full road surface area. Additional research on the influence of a tile's size should also be carried out, as false predictions might be reduced by using higher tile resolutions that contain more road information (e.g., 512 $\times$ 512 pixels, or 1024 $\times$ 1024 pixels). From a representation perspective, a reflection on symbolizing road elements with consistent notations and labels should be carried out to enable a better extraction of the road transport network.

Many imperfections arise from the complex nature of the geospatial object. Roads have different spectral signatures, depending on the material used in pavements and do not feature consistent geometry. Furthermore, there are other similar geospatial objects that can confuse the neural networks trained because of the similar shapes (e.g., rivers), or spectral signatures (e.g., canals, or railroads). This will cause inaccuracies in the extracted elements, even if state-of-the-art model components are trained on large-scale datasets.

Although the results generated are not perfect, they confirm the suitability of the proposed workflow for extracting the road transport network, and the methodology can be applied to achieve significant increases in the quality of results in tasks related to the extraction of geospatial elements. To tackle the above-mentioned challenges, the predictions delivered by the proposed end-to-end solution should still be systematically revised by a specialized operator to remove false predictions from a cloned layer of the results; nonetheless, this task would require far less working time when compared to having to manually generate the road representations from scratch (the current approach implemented by state agencies). This could be solved with the introduction of a module that allows the human factor to intervene in the road mapping process, which is necessary in a future update to solve artefacts predicted in complex areas and to discard false predictions.

7. Conclusions

In this work, a large-scale road mapping procedure based on the consecutive execution of ① road recognition, ② semantic segmentation of road surface areas, and ③ post-processing with conditional generative learning operations, within a common processing environment, was proposed. To the best of our knowledge, it is the first intent of road surface area extraction solution applied and verified at a such a large scale. The research can provide a technical base for other large-scale mapping projects where the requirements are related to assigning pixel information with geographical information.

The results obtained in Sections 4 and 5 and discussed in Section 6 prove the suitability of approaching the road surface area mapping task with the proposed processing workflow. In this regard, the implementation of the proposed solution delivered higher quality road representations from aerial orthophotographs when compared to other DL implementations trained for the same task. Furthermore, the versatility and flexibility of the mapping solution given by the consecutive execution of the three operations proved its effectiveness and enables the integration of an application that alleviates the manipulation of geospatial data, while allowing for an easy integration of future models and algorithms.

Nonetheless, it is important to note that the extracted representations are not perfect, and a fully automated road surface area extraction operation cannot be achieved by means of current state-of-the-art technologies combined with artificial intelligence techniques. For a final debugging of the results, a human operator should be introduced at the end of the process—a task that would require considerably less time. In addition, it was found that more research on (1) the effect of image resolution and image overlap and (2) the application of super-resolution techniques to homogenize the spatial resolution of the orthoimages should be carried out. However, by reducing reliance on human participation in the large-scale mapping of the road transport network from aerial imagery, while obtaining high efficiency levels, state administration can be assisted in automatically monitoring and detecting changes in the road layout, while reducing costs, and consequently, the citizens can be provided with up-to-date cartography faster.

Author Contributions: C.-I.C.: conceptualisation, formal analysis, investigation, methodology, software, validation, visualisation, writing—original draft; M.-Á.M.-C.: conceptualisation, data curation, funding acquisition, investigation, methodology, project administration, resources, software, supervision, validation, visualisation, writing—review and editing; R.A.: resources, supervision, validation, visualisation, writing—review and editing; B.B.S.: validation, visualisation, writing—review and editing; J.G.M.: validation, visualisation, writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research received funding from the "Deep learning applied to the recognition, semantic segmentation, post-processing, and extraction of the geometry of main roads, secondary roads and paths (SROADEX)" project, grant PID2020-116448GB-I00 funded by the AEI.

Data Availability Statement: The SROADEX dataset (approximately 527,000 images) used for training and testing the model components of the end-to-end road surface area extraction solution proposed in this manuscript is openly avail-able under a CC-BY 4.0 license and can be download from the Zenodo data repository using the link: https://zenodo.org/record/6482346, (accessed on 14 April 2023).

Acknowledgments: We thank Irene Martínez Encinas and all other SROADEX participants for their help in the initial phases of the research design, and in generating the dataset.

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

References

1. Instituto Geográfico Nacional (Spain) Especificaciones de Producto de Redes e Infraestructuras del Transporte del Instituto Geográfico Nacional. Available online: http://www.ign.es/resources/IGR/Transporte/20160316_Espec_RT_V0.5.pdf (accessed on 22 March 2023).
2. 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; Conference Track Proceedings. Bengio, Y., LeCun, Y., Eds.;
3. 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, 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.
4. Isola, P.; Zhu, J.-Y.; Zhou, T.; Efros, A.A. Image-to-Image Translation with Conditional Adversarial 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: Washington, DC, USA, 2017; pp. 5967–5976.

5. 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]
6. Zhang, Z.; Liu, Q.; Wang, Y. Road Extraction by Deep Residual U-Net. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 749–753. [CrossRef]
7. 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]
8. Xin, J.; Zhang, X.; Zhang, Z.; Fang, W. Road Extraction of High-Resolution Remote Sensing Images Derived from DenseUNet. *Remote Sens.* **2019**, *11*, 2499. [CrossRef]
9. Cheng, G.; Wang, Y.; Xu, S.; Wang, H.; Xiang, S.; Pan, C. Automatic Road Detection and Centerline Extraction via Cascaded End-to-End Convolutional Neural Network. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3322–3337. [CrossRef]
10. Manso-Callejo, M.A.; Cira, C.-I.; Alcarria, R.; Gonzalez Matesanz, F.J. First Dataset of Wind Turbine Data Created at National Level with Deep Learning Techniques from Aerial Orthophotographs with a Spatial Resolution of 0.5 m/Pixel. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 7968–7980. [CrossRef]
11. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A.A. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; Singh, S.P., Markovitch, S., Eds.; AAAI Press: Washington, DC, USA, 2017; pp. 4278–4284.
12. 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]
13. Mattyus, G.; Wang, S.; Fidler, S.; Urtasun, R. Enhancing Road Maps by Parsing Aerial Images Around the World. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 11–18 December 2015; IEEE Computer Society: Washington, DC, USA, 2015; pp. 1689–1697.
14. Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
15. Wang, J.; Qin, Q.; Gao, Z.; Zhao, J.; Ye, X. A New Approach to Urban Road Extraction Using High-Resolution Aerial Image. *IJGI* **2016**, *5*, 114. [CrossRef]
16. Hutchison, D.; Kanade, T.; Kittler, J.; Kleinberg, J.M.; Mattern, F.; Mitchell, J.C.; Naor, M.; Nierstrasz, O.; Pandu Rangan, C.; Steffen, B.; et al. Learning to Detect Roads in High-Resolution Aerial Images. In *Computer Vision—ECCV 2010*; Daniilidis, K., Maragos, P., Paragios, N., Eds.; Springer: Berlin/Heidelberg, Germany, 2010; Volume 6316, pp. 210–223. ISBN 978-3-642-15566-6.
17. 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]
18. 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]
19. Zhang, Z.; Zhang, X.; Sun, Y.; Zhang, P. Road Centerline Extraction from Very-High-Resolution Aerial Image and LiDAR Data Based on Road Connectivity. *Remote Sens.* **2018**, *10*, 1284. [CrossRef]
20. Abdollahi, A.; Pradhan, B.; Shukla, N.; Chakraborty, S.; Alamri, A. Deep Learning Approaches Applied to Remote Sensing Datasets for Road Extraction: A State-Of-The-Art Review. *Remote Sens.* **2020**, *12*, 1444. [CrossRef]
21. Zhong, Z.; Li, J.; Cui, W.; Jiang, H. Fully Convolutional Networks for Building and Road Extraction: Preliminary Results. In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2016, Beijing, China, 10–15 July 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 1591–1594.
22. Mnih, V. Machine Learning for Aerial Image Labeling. Ph.D. Thesis, University of Toronto, Toronto, ON, Canada, 2013.
23. Wei, Y.; Wang, Z.; Xu, M. Road Structure Refined CNN for Road Extraction in Aerial Image. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 709–713. [CrossRef]
24. Wang, S.; Yang, H.; Wu, Q.; Zheng, Z.; Wu, Y.; Li, J. An Improved Method for Road Extraction from High-Resolution Remote-Sensing Images That Enhances Boundary Information. *Sensors* **2020**, *20*, 2064. [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. Li, P.; Zang, Y.; Wang, C.; Li, J.; Cheng, M.; Luo, L.; Yu, Y. Road Network Extraction via Deep Learning and Line Integral Convolution. In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2016, Beijing, China, 10–15 July 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 1599–1602.
27. Gao, H.; Xiao, J.; Yin, Y.; Liu, T.; Shi, J. A Mutually Supervised Graph Attention Network for Few-Shot Segmentation: The Perspective of Fully Utilizing Limited Samples. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, 1–13. [CrossRef]
28. Li, Y.; Xu, L.; Rao, J.; Guo, L.; Yan, Z.; Jin, S. A Y-Net Deep Learning Method for Road Segmentation Using High-Resolution Visible Remote Sensing Images. *Remote Sens. Lett.* **2019**, *10*, 381–390. [CrossRef]
29. Buslaev, A.; Seferbekov, S.S.; Iglovikov, V.; Shvets, A. Fully Convolutional Network for Automatic Road Extraction From Satellite Imagery. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2018, Salt Lake City, UT, USA, 18–22 June 2018; IEEE Computer Society: Washington, DC, USA, 2018; pp. 207–210.
30. 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, 26–30 June 2016; IEEE: Las Vegas, NV, USA, 2016; pp. 770–778.

31. Xu, Y.; Feng, Y.; Xie, Z.; Hu, A.; Zhang, X. A Research on Extracting Road Network from High Resolution Remote Sensing Imagery. In Proceedings of the 26th International Conference on Geoinformatics, Geoinformatics 2018, Kunming, China, 28–30 June 2018; Hu, S., Ye, X., Yang, K., Fan, H., Eds.; IEEE: Washington, DC, USA, 2018; pp. 1–4.
32. Zhou, L.; Zhang, C.; Wu, M. D-LinkNet: LinkNet With Pretrained Encoder and Dilated Convolution for High Resolution Satellite Imagery Road Extraction. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2018, Salt Lake City, UT, USA, 18–22 June 2018; IEEE Computer Society: Washington, DC, USA, 2018; pp. 182–186.
33. 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]
34. Panboonyuen, T.; Vateekul, P.; Jitkajornwanich, K.; Lawawirojwong, S. An Enhanced Deep Convolutional Encoder-Decoder Network for Road Segmentation on Aerial Imagery. In Proceedings of the Recent Advances in Information and Communication Technology 2017—Proceedings of the 13th International Conference on Computing and Information Technology (IC2IT), Bangkok, Thailand, 6–7 July 2017; Meesad, P., Sodsee, S., Unger, H., Eds.; Springer: Berlin/Heidelberg, Germany, 2017; Volume 566, pp. 191–201.
35. 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]
36. Clevert, D.-A.; Unterthiner, T.; Hochreiter, S. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). In Proceedings of the 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, 2–4 May 2016; Conference Track Proceedings. Bengio, Y., LeCun, Y., Eds.;
37. Doshi, J. Residual Inception Skip Network for Binary Segmentation. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2018, Salt Lake City, UT, USA, 18–22 June 2018; IEEE Computer Society: Washington, DC, USA, 2018; pp. 216–219.
38. 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, 26–30 June 2016; IEEE: Las Vegas, NV, USA, 2016; pp. 2818–2826.
39. He, H.; Yang, D.; Wang, S.; Wang, S.; Liu, X. Road Segmentation of Cross-Modal Remote Sensing Images Using Deep Segmentation Network and Transfer Learning. *Ind. Robot* **2019**, *46*, 384–390. [CrossRef]
40. Xia, W.; Zhang, Y.-Z.; Liu, J.; Luo, L.; Yang, K. Road Extraction from High Resolution Image with Deep Convolution Network—A Case Study of GF-2 Image. *Proceedings* **2018**, *2*, 325. [CrossRef]
41. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 834–848. [CrossRef] [PubMed]
42. Gao, L.; Song, W.; Dai, J.; Chen, Y. Road Extraction from High-Resolution Remote Sensing Imagery Using Refined Deep Residual Convolutional Neural Network. *Remote Sens.* **2019**, *11*, 552. [CrossRef]
43. Xie, Y.; Miao, F.; Zhou, K.; Peng, J. HsgNet: A Road Extraction Network Based on Global Perception of High-Order Spatial Information. *ISPRS Int. J. Geo Inf.* **2019**, *8*, 571. [CrossRef]
44. Chaurasia, A.; Culurciello, E. LinkNet: Exploiting Encoder Representations for Efficient Semantic Segmentation. In Proceedings of the 2017 IEEE Visual Communications and Image Processing (VCIP), St. Petersburg, FL, USA, 10–13 December 2017; pp. 1–4. [CrossRef]
45. Etten, A.V.; Lindenbaum, D.; Bacastow, T.M. SpaceNet: A Remote Sensing Dataset and Challenge Series. *CoRR arXiv* **2018**, arXiv:1807.01232.
46. Demir, I.; Koperski, K.; Lindenbaum, D.; Pang, G.; Huang, J.; Basu, S.; Hughes, F.; Tuia, D.; Raskar, R. DeepGlobe 2018: A Challenge to Parse the Earth Through Satellite Images. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2018, Salt Lake City, UT, USA, 18–22 June 2018; IEEE Computer Society: Washington, DC, USA, 2018; pp. 172–181.
47. 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*, 016020. [CrossRef]
48. 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]
49. Manso-Callejo, M.-Á.; Cira, C.-I.; González-Jiménez, A.; Querol-Pascual, J.-J. Dataset Containing Orthoimages Tagged with Road Information Covering Approximately 8650 Km2 of the Spanish Territory (SROADEX). *Data Brief* **2022**, *42*, 108316. [CrossRef] [PubMed]
50. Varia, N.; Dokania, A.; Jayavelu, S. DeepExt: A Convolution Neural Network for Road Extraction Using RGB Images Captured by UAV. In Proceedings of the IEEE Symposium Series on Computational Intelligence, SSCI 2018, Bangalore, India, 18–21 November 2018; IEEE: Boston, MA, USA, 2018; pp. 1890–1895.
51. 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.

52. Shi, Q.; Liu, X.; Li, X. Road Detection From Remote Sensing Images by Generative Adversarial Networks. *IEEE Access* **2018**, *6*, 25486–25494. [CrossRef]
53. Yang, C.; Wang, Z. An Ensemble Wasserstein Generative Adversarial Network Method for Road Extraction From High Resolution Remote Sensing Images in Rural Areas. *IEEE Access* **2020**, *8*, 174317–174324. [CrossRef]
54. Hartmann, S.; Weinmann, M.; Wessel, R.; Klein, R. StreetGAN: Towards Road Network Synthesis with Generative Adversarial Networks. In Proceedings of the International Conference on Computer Graphics, Visualization and Computer Vision, Marrakesh, Morocco, 23–25 May 2017.
55. Costea, D.; Marcu, A.; Leordeanu, M.; Slusanschi, E. Creating Roadmaps in Aerial Images with Generative Adversarial Networks and Smoothing-Based Optimization. In Proceedings of the 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), Venice, Italy, 27–29 October 2017; IEEE: Venice, Italy, 2017; pp. 2100–2109.
56. Zhang, Y.; Li, X.; Zhang, Q. Road Topology Refinement via a Multi-Conditional Generative Adversarial Network. *Sensors* **2019**, *19*, 1162. [CrossRef] [PubMed]
57. Belli, D.; Kipf, T. Image-Conditioned Graph Generation for Road Network Extraction. *CoRR arXiv* **2019**, arXiv:1910.14388.
58. Liu, L.; Ma, H.; Chen, S.; Tang, X.; Xie, J.; Huang, G.; Mo, F. Image-Translation-Based Road Marking Extraction From Mobile Laser Point Clouds. *IEEE Access* **2020**, *8*, 64297–64309. [CrossRef]
59. Cira, C.-I.; Manso-Callejo, M.-Á.; Alcarria, R.; Fernández Pareja, T.; Bordel Sánchez, B.; Serradilla, F. Generative Learning for Postprocessing Semantic Segmentation Predictions: A Lightweight Conditional Generative Adversarial Network Based on Pix2pix to Improve the Extraction of Road Surface Areas. *Land* **2021**, *10*, 79. [CrossRef]
60. Cira, C.-I.; Kada, M.; Manso-Callejo, M.-Á.; Alcarria, R.; Bordel Sanchez, B.B. Improving Road Surface Area Extraction via Semantic Segmentation with Conditional Generative Learning for Deep Inpainting Operations. *IJGI* **2022**, *11*, 43. [CrossRef]
61. Sirotkovic, J.; Dujmic, H.; Papic, V. Image Segmentation Based on Complexity Mining and Mean-Shift Algorithm. In Proceedings of the 2014 IEEE Symposium on Computers and Communications (ISCC), Funchal, Portugal, 23–26 June 2014; IEEE: Funchal, Madeira, Portugal, 2014; pp. 1–6.
62. PostgreSQL: Documentation. Available online: https://www.postgresql.org/docs/ (accessed on 21 May 2022).
63. Documentation | PostGIS. Available online: https://postgis.net/documentation/ (accessed on 21 May 2022).
64. 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]
65. Cira, C.-I.; Alcarria, R.; Manso-Callejo, M.-Á.; Serradilla, F. A Deep Learning-Based Solution for Large-Scale Extraction of the Secondary Road Network from High-Resolution Aerial Orthoimagery. *Appl. Sci.* **2020**, *10*, 7272. [CrossRef]
66. Manso-Callejo, Miguel Angel; Calimanut-Ionut, Cira SROADEX: Dataset for Binary Recognition and Semantic Segmentation of Road Surface Areas from High Resolution Aerial Orthoimages Covering Approximately 8650 km^2 of the Spanish Territory Tagged with Road Information. *Data Brief* **2022**, *42*, 108316.
67. Instituto Geográfico Nacional Plan Nacional de Ortofotografía Aérea. Available online: https://pnoa.ign.es/caracteristicas-tecnicas (accessed on 25 November 2019).
68. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In Proceedings of the 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, 7–13 December 2015; IEEE Computer Society: Washington, DC, USA, 2015; pp. 1026–1034.
69. 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]
70. 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; Conference Track Proceedings. Bengio, Y., LeCun, Y., Eds.
71. Yakubovskiy, P. Segmentation Models; GitHub, 2019. Available online: https://github.com/qubvel/segmentation_models (accessed on 14 April 2023).
72. 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, 18–23 June 2018; IEEE: Salt Lake City, UT, USA, 2018; pp. 7132–7141.
73. Instituto Geográfico Nacional Centro de Descargas del CNIG (IGN). Available online: http://centrodedescargas.cnig.es (accessed on 3 February 2020).