land

MDPI

Article

Generative Learning for Postprocessing Semantic Segmentation Predictions: A Lightweight Conditional Generative Adversarial Network Based on Pix2pix to Improve the Extraction of Road Surface Areas

Calimanut-Ionut Cira [1], Miguel-Ángel Manso-Callejo [1,*], Ramón Alcarria [1], Teresa Fernández Pareja [1], Borja Bordel Sánchez [2] and Francisco Serradilla [3]

[1] Departamento de Ingeniería Topográfica y Cartografía, E.T.S.I. en Topografía, Geodesia y Cartografía, Universidad Politécnica de Madrid, 28031 Madrid, Spain; ionut.cira@upm.es (C.-I.C.); ramon.alcarria@upm.es (R.A.); teresa.fpareja@upm.es (T.F.P.)
[2] Departamento de Sistemas Informáticos, E.T.S.I. de Sistemas Informáticos, Universidad Politécnica de Madrid, 28031 Madrid, Spain; borja.bordel@upm.es
[3] Departamento de Inteligencia Artificial, E.T.S.I. de Sistemas Informáticos, Universidad Politécnica de Madrid, 28031 Madrid, Spain; fserra@eui.upm.es
* Correspondence: m.manso@upm.es

Abstract: Remote sensing experts have been actively using deep neural networks to solve extraction tasks in high-resolution aerial imagery by means of supervised semantic segmentation operations. However, the extraction operation is imperfect, due to the complex nature of geospatial objects, limitations of sensing resolution, or occlusions present in the scenes. In this work, we tackle the challenge of postprocessing semantic segmentation predictions of road surface areas obtained with a state-of-the-art segmentation model and present a technique based on generative learning and image-to-image translations concepts to improve these initial segmentation predictions. The proposed model is a conditional Generative Adversarial Network based on Pix2pix, heavily modified for computational efficiency (92.4% decrease in the number of parameters in the generator network and 61.3% decrease in the discriminator network). The model is trained to learn the distribution of the road network present in official cartography, using a novel dataset containing 6784 tiles of 256 × 256 pixels in size, covering representative areas of Spain. Afterwards, we conduct a metrical comparison using the Intersection over Union (IoU) score (measuring the ratio between the overlap and union areas) on a novel testing set containing 1696 tiles (unseen during training) and observe a maximum increase of 11.6% in the IoU score (from 0.6726 to 0.7515). In the end, we conduct a qualitative comparison to visually assess the effectiveness of the technique and observe great improvements with respect to the initial semantic segmentation predictions.

Keywords: conditional Generative Adversarial Network; generative learning; postprocessing semantic segmentation predictions; road extraction; road surface areas

Citation: 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. https://doi.org/10.3390/land10010079

Received: 21 December 2020
Accepted: 14 January 2021
Published: 16 January 2021

1. Introduction

Remotely sensed images have been used lately by researchers in machine vision applications such as object identification [1,2], detection [3], or extraction [4]. At the same time, deep learning algorithms proved to be useful for classification tasks and land use analysis [5] in satellite imagery data [6,7]—an important remote sensing application, where semantic segmentation techniques (based on supervised learning) are applied to assign a land cover class to every pixel of an image. This extraction task is generally carried out by means semantic segmentation and can be considered very challenging due to complex nature of geospatial objects, due to defects present in imagery (noise, occlusions, etc.), due to imperfections in the ground-truth segmentation masks or due to particularities of the segmentation algorithms applied.

In one of our previous works [8], we studied the appropriateness of using state-of-the-art segmentation models for extracting the surface areas of secondary roads and conducted a large-scale evaluation on unseen areas, obtaining IoU and F1 scores of 0.5790 and 0.7120, respectively (with 97.87% of the samples being correctly classified). However, even the best performing state-of-the-art segmentation model (U-Net [9] as base architecture with SEResNeXt50 [10] as backbone network) displayed the problem of inaccurate extraction. Many resulting segmentation masks presented discontinuities, and the connection points were often overlooked, resulting in road segments that were unconnected. We also identified higher rates of "false positive labels in areas where the materials used in the road pavement have a similar spectral signature with their surroundings, or areas where geospatial objects with similar features are present (such as dry riverbeds, railroads, or irrigation canals) and higher rates of false negatives in sections where other objects cover large portions of the roads were covered" (page 13 in [8]). Similar problems are still observed in recent works dealing with the road extraction from high-resolution aerial imagery—improving the road extraction task is an active area of research [11–14].

To overcome the deficiencies observed in our previous work [8], we developed a postprocessing technique based on image-to-image translation [15] concepts to operate over the initial semantic segmentation predictions and improve the road surface extraction for automatic mapping purposes. In this work, we apply generative learning techniques and propose a conditional Generative Adversarial Network (cGAN) architecture based on Pix2pix [15], greatly improved for computational efficiency (92.4% decrease in the number of parameters in the generator G network and 61.3% decrease in the discriminator D network, when compared to the original Pix2pix) to improve the initial road surface area extraction. For training and testing the model's performance, we use a novel dataset containing 8480 rasterized masks of roads tagged at pixel level, covering a land area of approximately 181 km^2 from representative areas of Spain.

The goal is to conditionally generate new synthetic images based on the initial segmentation predictions similar to samples belonging to the domain of official cartography, in this way reducing the effect and overcoming the inaccurate extraction. We evaluated the model on unseen data and calculated the Intersection over Union score (IoU score). This metric is a number from 0 to 1 that specifies the amount of overlap between predictions and ground truth masks (or the area of intersection divided with the area of union) for any two sets, M and N; $IoU\ score\ (M, N) = |M \cap N|/|M \cup N|$. We observed average improvements in IoU scores of the order of 11.3% when compared to the initial predictions, and 4.04% when compared to the original Pix2pix model. In the end, we conducted a perceptual validation to assess the quality of the postprocessing operation and observed significant improvements.

Our contributions can be summarized as follows.

- We propose a conditional GAN architecture based on Pix2pix (which we heavily modify for computational efficiency) to postprocess binary semantic segmentation predictions of road surface areas. Our Generator G is based on the U-Net [9] architecture (modified to reduce the number of parameters by 92.4%), while our Discriminator D is a modified version of PatchGAN [15], which allows the processing of larger patches of images (128×128, instead of 32×32), while reducing the number of parameters with 61.3%.
- We train the proposed architecture on a new dataset composed of 6784 real segmentation maps tagged at pixel level (representing our target domain) and their corresponding initial segmentation masks (representing our conditional information) obtained with a state-of-the-art segmentation model, after applying Gaussian noise to the input.
- We study the appropriateness of applying generative learning techniques for postprocessing initial semantic segmentation predictions of road surface areas by conducting a metrical (IoU score) comparison and a perceptual validation on a new test set composed of 1696 real segmentation maps and their correspondent semantic segmentation

predictions (unseen during training). We proceed as follows. In Section 2, we discuss related works. In Section 3, we describe the task from a mathematical perspective. In Section 4, we present the dataset used for training and testing. Details of our proposed model are presented in Section 5. The experiments carried out are described in Section 6. The results obtained in the postprocessing operation are analyzed in Section 8 from a metrical and a perceptual perspective. Finally, Section 8 offers the conclusions.

2. Related Works

Unsupervised learning is a paradigm of learning where only input variables, X (and no output variables, y), are given to the model. The goal is to learn underlying hidden distribution of the data using just the unlabeled data. In [16], unsupervised learning based on grammar-guided genetic programming has been successfully applied to obtain new Convolutional Neural Network (CNN) architectures specialized in road recognition in aerial imagery. In [17], unsupervised training was applied to generate filters that improved the predictions of the road detection model.

Generative models are a class of unsupervised learning, where the goal is to generate new samples from an unlabeled distribution. This task addresses the density estimation in an explicit way (e.g., PixelRNN [18], or variational autoencoders [19]) to learn lower-dimensional feature representations from unlabeled training data), or in an implicit way, with Generative Adversarial Networks (GANs). GANs [20] were introduced by Goodfellow et al. in 2014, and act like a system composed of two networks (called Generator G and Discriminator D) trained simultaneously in an adversarial setting to create variations in data. The goal of the training is to implicitly find the probability density function that best describes the training examples. This way, the model learns to successfully map from random noise z to an output image, y, $G : z \rightarrow y$.

GANs evolved over the following years [21]. Deep Convolutional Generative Adversarial Networks (DCGANs) [22] are an extension of the GAN architecture that use deep convolutional neural networks with certain architectural constraints for both G and D networks (e.g., use of upsampling networks with fractionally-strided convolutions), allowing to learn representations from unlabeled image data. After training, DCGANs are able to generate high-quality synthetic images from the learned underlying distribution.

Conditional Generative Adversarial Networks (cGANs) [23] applied to computer vision tasks make use of image conditional information as additional input to both the generator and the discriminator. The mapping to the output image y is learned from the observed image x and the random noise z, $G : \{x, z\} \rightarrow y$. Popular applications of cGANs in land use analysis are centred on improving the representations obtained from remotely sensed imagery. Some examples are the creation of higher resolution images in image super resolution task [24–27], and texture synthesis and realistic reconstructions [28–31].

Pix2pix [15] was introduced by Isola et al. as an extension of the cGAN architecture for that task referred as image-to-image translation. The model uses an U-Net-based [9] network as G (updated to include the vector distance from the target output image), and the PatchGAN architecture as discriminator network (only penalising structures at the scale of image patches). This task is one of the most important applications of cGANs, directly applicable to land use analysis and geospatial elements detection, and can be used for mapping from one domain to another (e.g., assign a class to every pixel in a remotely sensed image, turn aerial images into maps, etc.), G and D being conditioned during training with additional information. The model evolved in recent years with the introductions of CycleGAN [32], DiscoGAN [33], or DualGAN [34].

Postprocessing semantic segmentation predictions has been traditionally done by means of conditional random fields [35]. In [36], shape filtering is applied to improve the extraction the roads' centerlines by combining high-resolution imagery with LiDAR data and vectorial data from OpenStreetMap. In [37], the authors extract the linear characteristics of selected road segments by constructing a geometric knowledge base of rural roads to

improve the initial results. In [38], a network for road extraction with a final module to highlight high level information and improve the classification is proposed.

Recently, GAN-based approaches for postprocessing road extraction emerged. The authors of [39] tackle the road extraction task by adding the Wasserstein distance and gradient penalty to a standard GAN and applying ensembling techniques to achieve an IoU score of 0.73 and obtain road geometries in Chinese rural areas. In [40], a GAN is trained to synthesize arbitrary-sized road network patches and enrich the attributes in areas where the extraction is difficult (where discontinuities are present or in complex areas, such as intersections or highway ramps). In [41], the authors propose a Multi-conditional Generative Adversarial Network (McGAN) to refine the road topology and obtain complete road networks graphs. McGAN is composed of two discriminators (one to employ the original spectral information, and the other discriminator to refine the road network topology) and a generator. The authors of [42] propose a method for extracting roads consisting in a GAN stage for detecting road edges and a second stage of smoothing-based optimization, postprocessing at pixel level to improve the initial segmentation masks.

3. Problem Description

The goal is to learn a correct transformation from a simple distribution (e.g., Gaussian distribution) to the complex target distribution using a neural network, while applying a condition. In our case, the cGAN will learn a mapping from random Gaussian noise vector z of dimension d, to the target domain containing the real road network features present in official cartography, y, while incorporating conditional information from x (initial segmentation masks) coming from a similar distribution, as proposed in [43]. By adding conditional information, the generation of the output image is conditioned on a source image x, and z will not be completely random anymore (the training will not be m).

The training procedure can be seen as a two-player game, where G tries to "fool" D by producing images that look real, while D trains to distinguish between real and fake images. As proposed in [20], D and G are trained jointly in a Minimax game, where the Minimax objective function $min\ max(D,\ G)$ performs a gradient ascent on D, $max\ \ [log\ D(x,\ y|x) + (1 - log(D(x,\ G(z|x))))]$ and a gradient descent on G, $min\ [1 - log(D(x,\ G(z|x)))]$. This encourages G to produce samples with a low probability of being fake. The objective function of the model can be expressed as $\mathcal{L}\ (\mathrm{G}, \mathrm{D}) = E_{x,y}[\log D(x,\ y|x)] + E_{x,z}(1 - \log D(x,\ G(z|x)))$, where $E_{z,x}$ is the expected value over all generated fake instances $G(z|x)$ given the condition x, and $E_{y,x}$ is the expected value over all real data instances (belonging to the density distribution we wish to replicate) given x [15,44]. The generator and discriminator are therefore two "players" that alternate in updating their model weights with the criteria of maximizing the likelihood of D being wrong, and maximising the likelihood of G being right (maximize the probability of the images produced by G being real) [45]. A simplification of the described cGAN training procedure is presented in Figure 1.

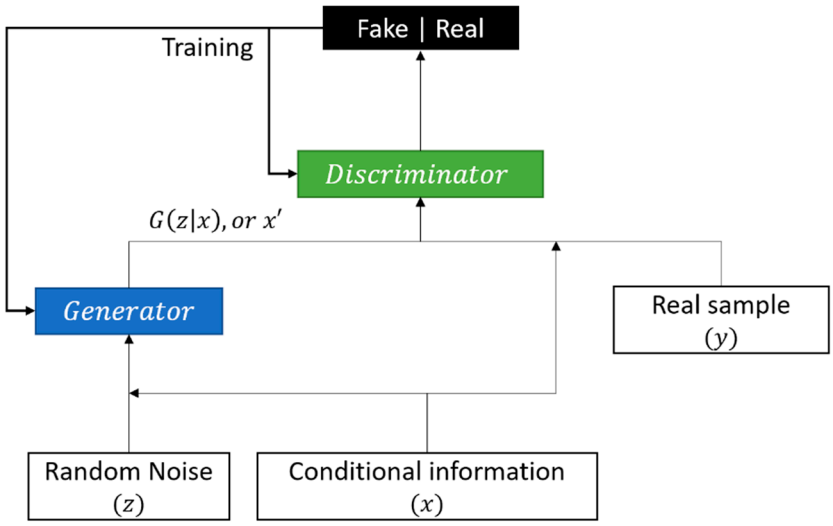

Figure 1. Simplification of the training procedure of a cGAN model. Note: In cGANs, we explicitly define x as an additional input to $G(z)$, resulting $G(z|x)$, and as additional input to $D(y)$, resulting $D(y|x)$.

4. Dataset

First, the target domain (or real samples, y) required was obtained from the National Topographical Map, scale 1:25,000 (Spanish: Mapa Topográfico Nacional 1:25,000, or MTN25), by rasterizing the available openly information containing the road network in vectorial format and dividing the resulting images in tiles of 256 × 256 pixels in size. We binarized the tiles to black and white to represent the classes "Road exists" and "No road" (background). The dataset contains 8480 tiles (covering 181 km^2 of representative areas from the Spanish territory) and was divided by applying the 80:20% criteria (allowing for more training data [8]), resulting in 6784 tiles used for training and 1696 tiles used for testing (data unseen during training). Compared to our previous work [8], we made the dataset bigger, and included more road structures (highways, paved roads, urban roads, etc.). Therefore, the target domain y is represented by the ground-truth tiles of 256 × 256 pixels present in MTN25.

Second, we obtained our source domain (x) by retraining the semantic segmentation model that statistically proved to be the most suitable for road extraction tasks in [8] (U-Net as base architecture and SEResNeXt50 as backbone network). We used the resulting model to evaluate the entire dataset, this way obtaining the initial segmentation masks stored in the PNG (Portable Network Graphics) lossless format. These predictions will represent our conditional information (x). The performance of the segmentation model was evaluated on the corresponding test set (containing tiles our target domain, y), the model achieving an IoU score of 0.6726. As intuition, the IoU score will be 1 if the prediction completely overlaps with the ground-truth mask; a model obtaining an IoU score greater than 0.5 is considered to have a good performance [46]. In Figure 2, we can find the correspondence between the aerial orthoimage, the binarized ground-truth segmentation mask (or target domain, y), and the initial segmentation prediction (or source domain, x) from six random tiles.

In this work, we will use the initial segmentation masks (third row) as conditional information for training and testing the model. The goal is to map this initial semantic segmentation distribution of data to the target domain containing the distribution of the road network in official cartography. The training procedure will enable the generator to synthesize fake examples belonging to targeted data domain, improving this way the initial segmentation predictions. We will evaluate G's performance on the test set and compare its IoU score with the one obtained by the segmentation model on the same set. We are actively increasing the size of this dataset and plan to make it publicly available once we reach around 500,000 tiles.

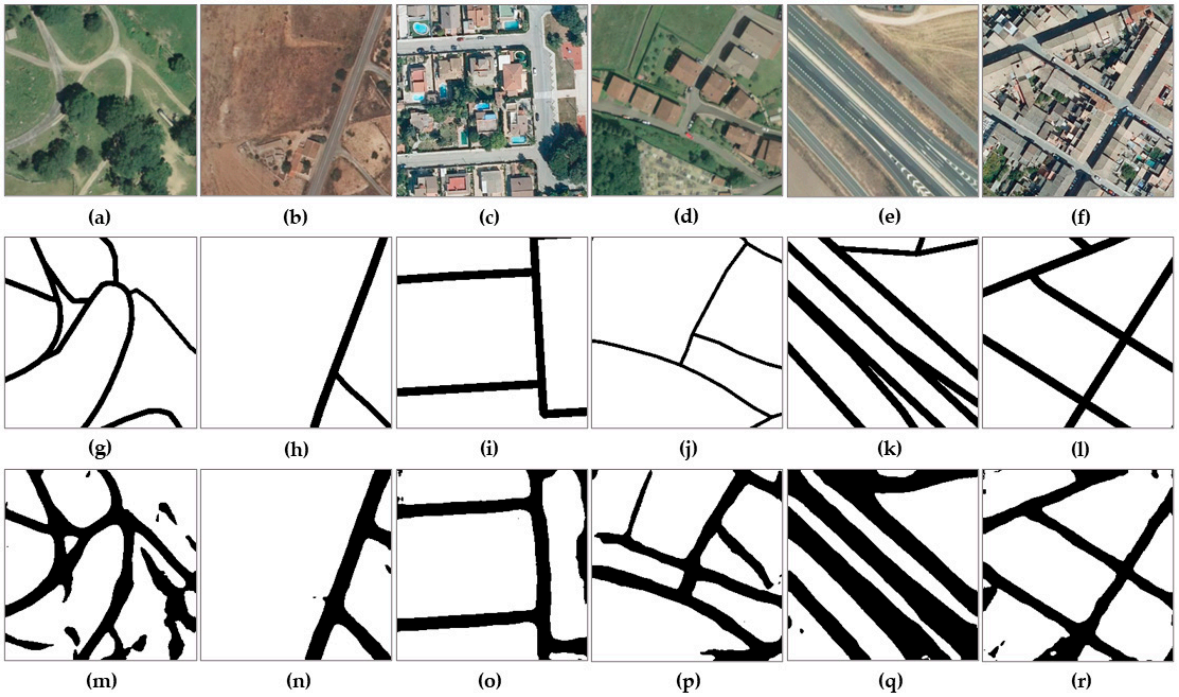

Figure 2. The relation between the aerial orthoimage (first row (**a**–**f**)), the rasterized segmentation mask (ground-truth or real sample coming from MTN25, found in the second row (**g**–**l**)), and the semantic segmentation predictions (conditional information, x, seen in third row (**m**–**r**)). Note: The train set used as conditional information for G contains the same 6784 tiles as the set from the target distribution y. The same is valid for the test set (unseen data).

5. Proposed cGAN Architecture

The cGAN architecture presented in this work is based on Pix2pix [15]. The code from Pix2pix's implementation [47] was used as basis for building the model; however, we changed the original implementation in order to make the computation more efficient and better suited for postprocessing sematic segmentation predictions containing road surface areas. The proposed model was defined and trained using the open source deep learning library TensorFlow version 2.2.0 [48] on a Linux server with a 12-core Intel Core i7 processor and a Nvidia RTX 2060 graphics card with 6 gigabytes of memory.

5.1. Generator

The generator G takes random noise z and a condition x (initial semantic segmentation mask) as input and will output a synthetic image, $G(z|x)$. In our case, z is randomly generated from a normal distribution. The generator will apply a generative function to obtain a new sample, this time from the generative model. The output sample, $G(z|x)$, should be reasonably similar to training data distribution.

G is a fully convolutional network consisting of an encoder with skip connections (introduced in U-Net [9]) and takes as input conditional Gaussian noise (with the condition x applied). In the encoder part, we used strided convolutions (with a stride of 2) instead of pooling layers and applied padding to allow for more space for the kernel to convolute and not lose information near the borders. We used the ReLU [49] activation function in all generator layers, except for the last one, where we uses tanh (instead of sigmoid) and applied Batch Normalisation [50] to all layers, except the input layer, following recommendations in [22].

In the decoder, we used transposed convolutions (with a stride of 2) to upsample and resize the output to the input's dimensions. As a means to avoid overfitting, we applied a dropout operation with the rate of 0.5 between the encoder and decoder. Similarly to U-Net [9], skip connections were added between the encoder and decoder parts, to avoid the

loss of low-level information and enable the share of information between different stages across the network. Details about the generator architecture can be found in Figure 3.

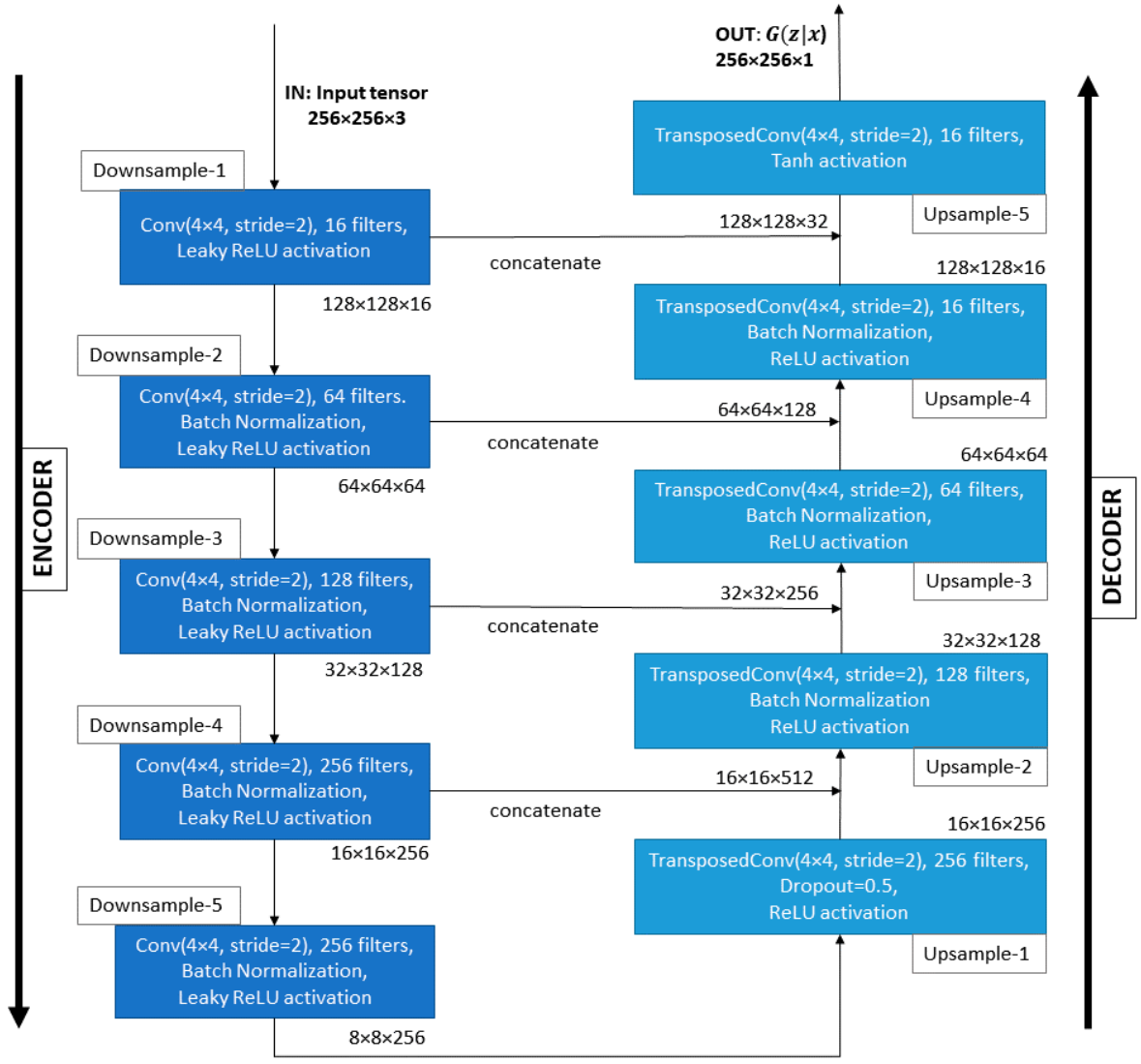

Figure 3. Description of the generator architecture. Note: The input is passed through a series of layers that progressively downsample its size until the bottleneck, where the process is reversed, and the tensor is upsampled.

G takes in the input image (the segmentation mask with Gaussian noise) and passes it through a series of convolution and upsampling layers to produce an output image that has the same size as the input. The generator is trained to produce synthetic outputs undistinguishable from "real" images. By using this architecture, the total number of parameters in *G* decreased from 54,425,859 in the original Pix2pix to 4,117,825 (a 92.4% reduction).

5.2. Discriminator

The discriminator network takes as input both the conditioned real sample from the target domain, $D(x, y|x)$, and the conditioned fake sample generated by *G*, $D(x, G(z|x))$. *D* analyzes the distribution of data and decides whether the data are generated or coming from the target domain data (using a sigmoid function that outputs the probability between 0 and 1). The output of discriminator *D* represents the probability that the sample is coming from the training distribution. The discriminator *D* will take $G(z|x)$ and real y and will output whether the image is real or fake, every input of *D* has a $1/2$ probability of being

real and 1/2 of being fake (acts like a binary classifier for the generated data, training to detect as well as possible the synthetic images produced by G).

Our D is a modified PatchGAN (introduced in [15]). We used Leaky ReLU activation [51] instead of ReLU activation for all D's layers to reduce gradient sparsity and applied Batch Normalization in all layers, except for the input. Instead of polling operations, we used again strided convolutions (with a stride of 2) and applied zero padding to avoid the loss of information near the borders. Details about D's architecture can be found in Figure 4.

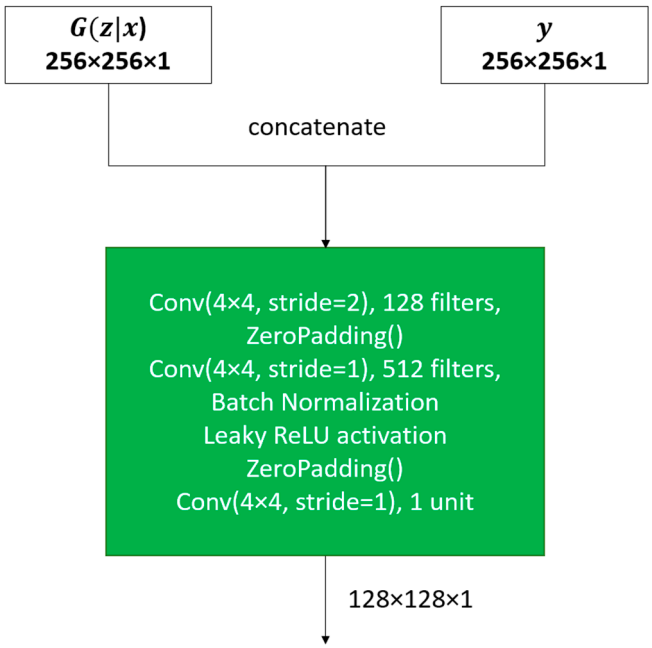

Figure 4. The architecture of the Discriminator (modified PatchGAN [15]). Note: In our case, we will concatenate z and x as input to the generator $G(z, x)$ with the objective of learning the distribution of the target domain,y.

D operates convolutionally over the 256 × 256 tile and is capable of classifying larger image patches (128 × 128, instead of 32 × 32) to decide if the input is real or fake. The discriminator takes four patches of 128×128 as input and classifies from which distribution the input comes from (real or fake). The decision scores are averaged to obtain a final prediction for the input tile (as seen in Figure 5).

By applying these changes, the total number of parameters was reduced from 2,770,433 to 1,071,105, a 61.3% decrease when compared to the original PatchGAN [15] network used in Pix2pix.

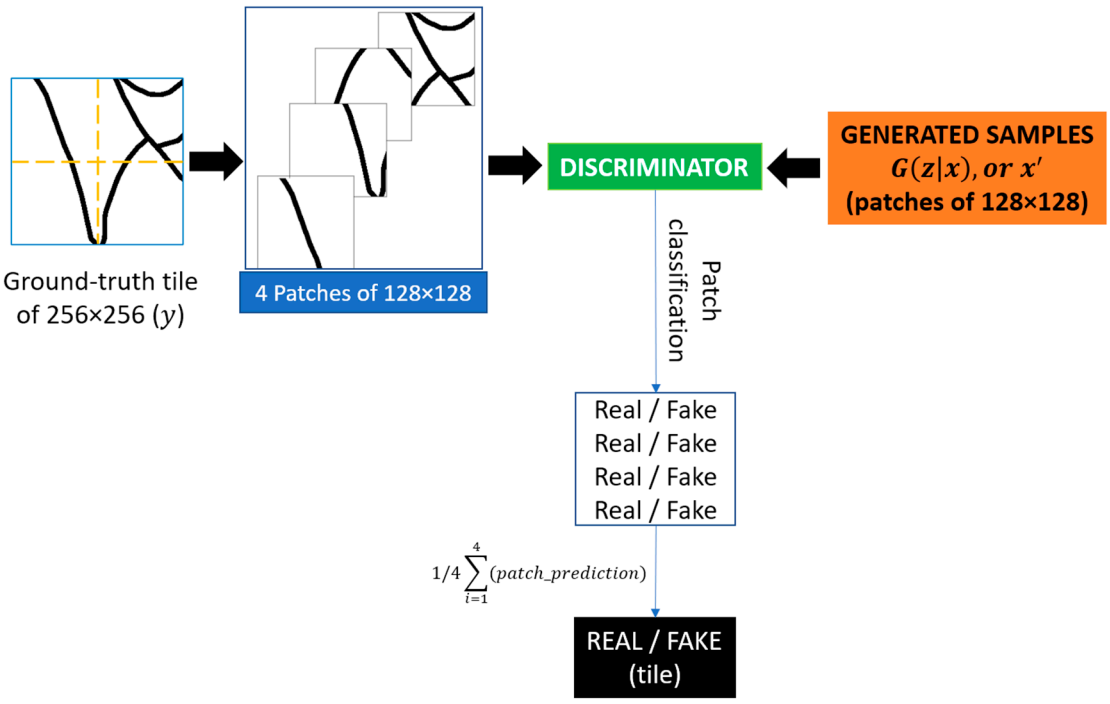

Figure 5. Analysis of the input data distribution with the proposed discriminator D (estimation of the probability that the input is real).

6. Experiments

As explained in Section 3, cGANs [23] take conditional information x (in our case, the initial sematic segmentation results) as an extension to the latent space z. The conditioning is performed by concatenating x into the correspondent tensors of images $G(z|x)$ and y. In the generator, the prior input Gaussian noise and x are combined in a joint hidden representation, the resulting $G(z|x)$ being fed into D, together with its correspondent tile from the target domain. The discriminator tries to identify which images are real and which one comes from G (produces a guess about how realistic they look), while G trains to maximise the log-probability of the discriminator $D(x, G(x|z))$ being mistaken [52]. When D cannot distinguish real images from fake images, the optimal state is reached. The intuition is that over time, the generator will be forced to create synthetic data that comes as closely as possible to the distribution of the target domain, while the discriminator will become better at telling them apart. Details about the training procedure of the proposed cGAN can be found in Figure 6.

As preprocessing for the input data, we applied online data augmentation techniques to rotate the images randomly picked as training sets by 90, 180, and 270 degrees, reducing this way the overfitting behavior.

As for the cost functions, the total generator loss is a modified version of the loss introduced by [15], where we sum the adversarial loss to the mean absolute error (also called L1 regularization) between the generated image and the target image (expected output image), while add a weight $\lambda = 1000$ to the L1_loss, $\mathcal{L}(G) = (1 - \log D(x,\ G(z|x))) + \lambda * L1_{loss}$. The adversarial loss is a cross entropy between 1 (as G tries to fool D, and the real values have the label 1 assigned) and the predicted value from $G(z|y)$ (0, in case D realises the $G(z|x)$ is fake) and encourages the generator to produce images similar to the training domain. In [15], it was shown that the L1 regularization allows $G(z|x)$ to become structurally similar to the target image (y), $\mathcal{L}_{L1}$ (G) $= E_{x,y,z}(|y - G(z|x)|)$, where $E_{x,y,z}$ is the expected value over all generated fake instances $G(z|x)$ and real data instances y given the condition x [15,44]. Therefore, the generator will be updated via a weighted sum of

the adversarial loss and the L1 loss (with a weight of 1000 to 1 in favor of the L1 loss), to encourage G to produce more realistic images (and obtain plausible translations).

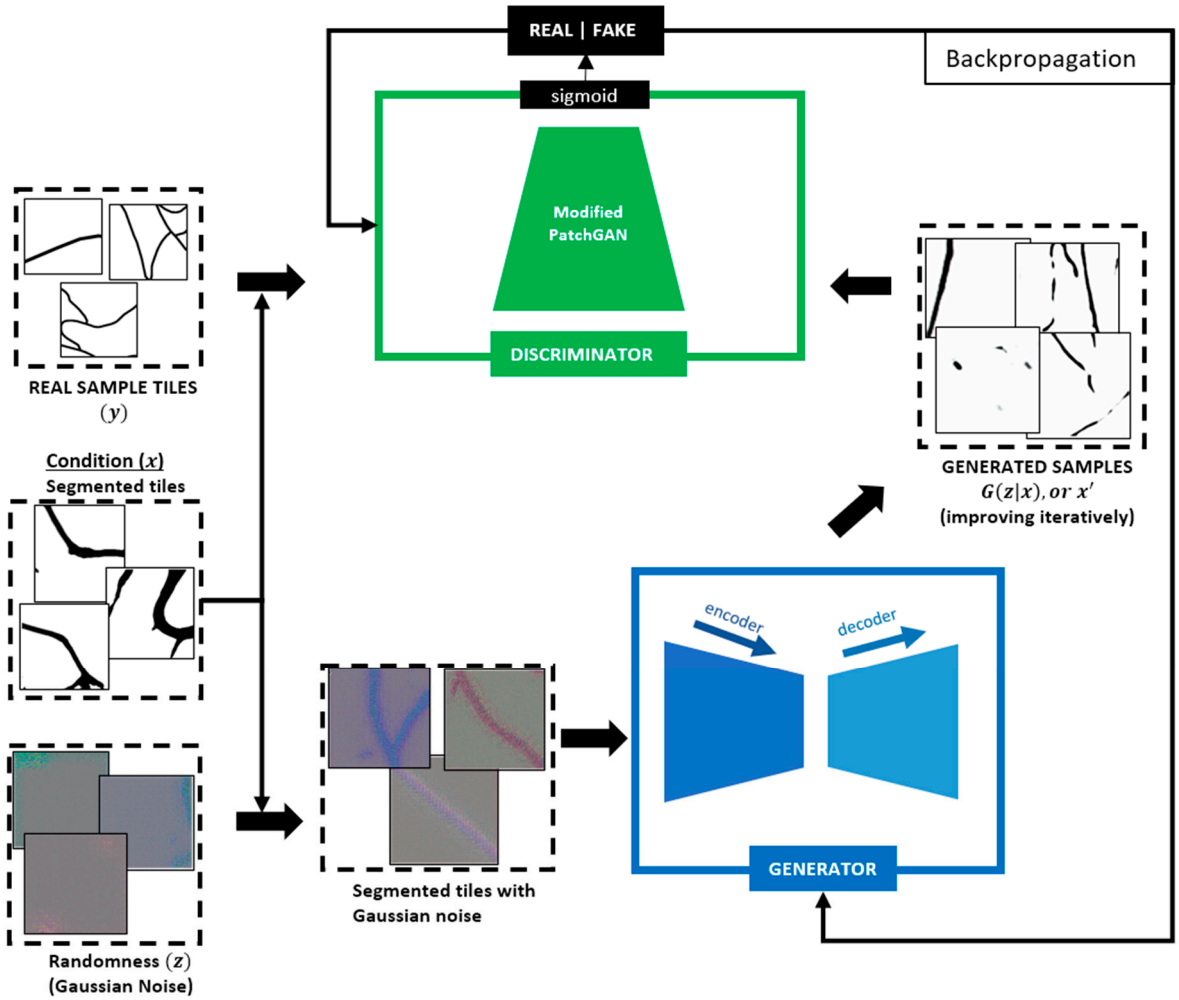

Figure 6. Representation describing the training process of the proposed cGAN architecture based on Pix2pix [15].

Second, D's inputs are x and $G(z|x)$ concatenated in a tensor $(x, G(z|x))$, which should be classified as fake, and the conditional information x concatenated to the target image y, which should be classified as real. When a pixel comes from the target domain (real sample), the true label is 1, whereas when a patch is fake, the true label of the pixels will be 0. The final layer of D is a sigmoid function, with D's cost function being a sum of the cross-entropy between the D's prediction and the actual labels for these two inputs (array of zeros in the case the fake images, and array of ones in the case of the real images).

The model was trained from scratch with minibatch stochastic gradient descent, the batch size being 10, each epoch taking around 350 seconds on the GPU. G's starting weights were randomly initialised from a Gaussian distribution, the mean of the random values to generate being 0 and the standard deviation being 0.04. For training, we used Adam optimizer [53] with learning rates $1 \times e^{-4}$ for G and $2 \times e^{-4}$ for D. The experiments apply a decay rate for the first momentum estimate $\beta_1 = 0.5$, and a decay rate for the second moment estimate $\beta_2 = 0.999$. We adopt different learning rates for generator and the discriminator to improve the convergence of GANs (as proposed in [54]). In our experiments, we found that not adding noise to the discriminator inputs improves the predictions, although some researchers have found it to be a form a regularization to improve the convergence [55].

The training is done via backpropagation, the model alternating between training D and training G. The weights of D are adjusted based on the classification error produced. G is then updated via the discriminator network (during generator training, D does not update its weights, but its gradients are used so that the generator can update its weights) to minimize G's cost function. Over time, G improves its output to better reproduce the real data distribution, iteratively changing the produced synthetic data to make it more realistic.

We repeated the experiments described above five times using random weight initializations. For comparison reasons, we trained the original Pix2pix model as well. Each time, the trained generator was exported to h5 format and used to evaluate the capacity of G in mapping from the source domain to the target domain.

7. Metrical Analysis and Perceptual Validation of the Results

To evaluate the model's performance, we used the test set containing the initial predictions (described in Section 4) unseen during training. These initial semantic segmentation predictions were passed through the trained generator, accumulating the goodness of results in a confusion matrix (by comparing them with the ground truth segmentation masks) to calculate the IoU score with the formula $IoU\ score = True\ Positives/(True\ Positives + False\ Positives + False\ Negatives)$. A comparison between the IoU scores obtained is reported in Table 1.

Table 1. Comparison of IoU score results obtained on the test set ($n = 1696$ titles).

Model	IoU Score (test Set)	Improvement with Respect to the Initial Semantic Segmentation Results
Semantic Segmentation(U-Net [9]–SEResNeXt50 [10])	0.6726 (best model)	-
Original Pix2pix [15]	0.7232 ± 0.006	Average: + 7.25%; Maximum: + 8.18%
Our implementation	0.7530 ± 0.004	Average: + 11.27%; Maximum + 11.62%

We observe a significant increase in IoU score by using the proposed cGAN model, from 0.6726 to an average of 0.7530 (an 11.27% increase over the performance metric obtained by the semantic segmentation model). At the same time, the original Pix2pix model achieved an average IoU score of 0.7232 on the test set containing data unseen during training (an average increase of 7.25%, when compared to initial IoU score of 0.6726). Our model obtained an average improvement of 4.04% when compared to the original Pix2pix model. The maximum IoU scores are 0.7300 in the case of the original Pix2pix, and 0.7556 in the case of our proposed cGAN architecture.

To see what these gains in IoU score mean for the postprocessing operation, we conducted a perceptual validation using the predictions obtained by passing test set (data unseen by the model, described in Section 4) through the trained generator. These predictions were subsequently stored in PNG format. In Figure 7, we can find a comparison between the results from eight random scenes.

The qualitative inspection of the generated images shows clear improvements over the initial segmentation predictions. The trained generator enabled a correct image-to-image translation ($G: \{x, z\} \rightarrow y$) and learned to produce synthetic images similar to those belonging to the target domain. The results from Figure 7 assert our metrical comparison from Table 1.

Compared to the initial semantic segmentation masks, we can observe a thinner road line representation, similar to the representation from the target domain, y. We also observe a consistent elimination of unconnected parts (e.g., Figure 7(c2–b2),(c6–e6)). Compared to the original Pix2pix, we observe smoother road line representations m, and cleaner road predictions (e.g., Figure 7(d1–e1),(d5–e5)). However, we still observe imperfections, mostly in urban areas (e.g., Figure 7(e5,e2)), but we believe they are caused by insufficient urban road data in the training set. We are actively working into solving this drawback and by actively building a bigger dataset. Nonetheless, the predictions show a clear enhancement over the initial data.

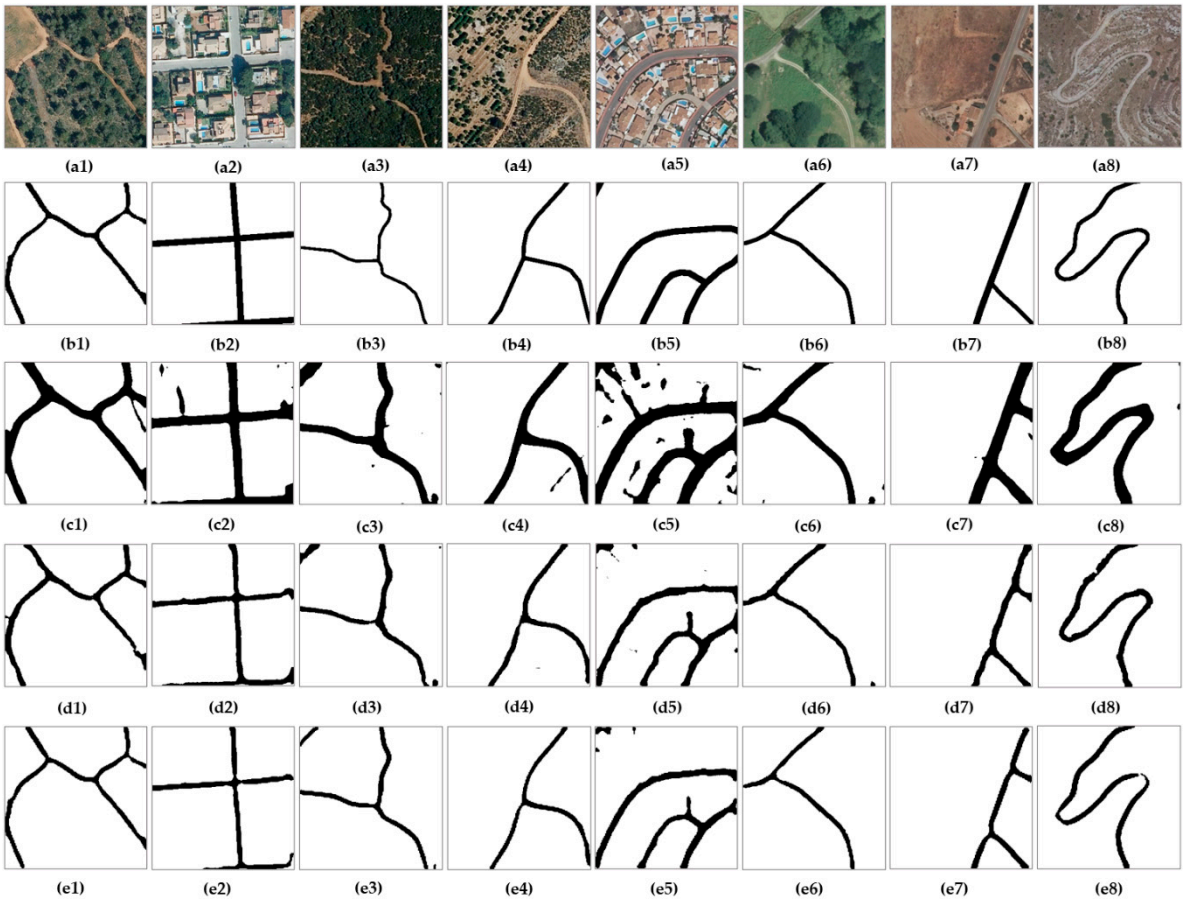

Figure 7. Perceptual validation carried out on eight random tiles from the test set (data unseen during training). The first row (**a1**–**a8**) represents the aerial orthoimage, the second row (**b1**–**b8**) represents the ground truth mask (target domain, y), the third row (**c1**–**c8**) represents the initial semantic segmentation prediction (conditional information, x), the fourth row (**d1**–**d8**) presents the predictions obtained with the original Pix2pix, and the fifth row (**e1**–**e8**) presents the prediction obtained by the model proposed in this paper.

The increases in performance metrics from Table 1 demonstrate the suitability of using conditional Generative Adversarial Networks for postprocessing initial semantic segmentation. Although the results from Figure 7 are not perfect, they represent a great improvement over state-of-the art semantic segmentation models and demonstrate the appropriateness of applying generative learning techniques in postprocessing tasks.

8. Conclusions

We presented an effective approach for postprocessing road segmentation masks in an adversarial way using a cGAN architecture based on Pix2pix, greatly modified for computational efficiency. We demonstrated its effectiveness by training and testing the model on a new dataset containing initial road surface area and observed average increases by the level of 11.3% when compared to a state-of-the-art segmentation model and 4.04% when compared to the original Pix2pix architecture, while achieving a 92.4% reduction in the numbers of parameters in G and a 61.3% decrease in the number of parameters in D.

The proposed architecture delivered significant improvements and can be viewed as a postprocessing technique for semantic segmentation predictions of road surface areas. The metrical comparison presented in Table 1 and the perceptual validation carried out in Section 7 proved the efficacy of the network—we presume that the quality of these predictions can be further improved by using a bigger dataset. We also strongly believe that the proposed architecture can be used to enhance the extraction operation of other continuous geospatial objects such as rivers, railroads and other transportation networks,

or irrigation canals. Furthermore, based on the results obtained in this work, we consider that the proposed procedure can be applied to other land analysis tasks or operations that involve the extraction of geospatial objects from remote sensing images.

As future lines of research, we are actively working to increase the size of the dataset and plan to integrate a complete end-to-end solution capable of large-scale recognizing, segmenting, and postprocessing the road network elements by means of classification, semantic segmentation, and GAN operations. The end goal is to obtain a robust system capable of monitoring and mapping the occurring changes in the road network for the whole national territory, reducing this way the human factor in updating existent road cartography. We also leave for a future study the implementation of the proposed cGAN architecture for postprocessing predictions from urban and rural areas where multiple land cover classes are present.

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

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

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to ongoing efforts to considerably increase the size of the dataset to around 500,000 tiles.

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

Conflicts of Interest: The authors declare no conflict of interest. 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. Albert, A.; Kaur, J.; Gonzalez, M.C. Using Convolutional Networks and Satellite Imagery to Identify Patterns in Urban Environments at a Large Scale. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining—KDD '17*; ACM Press: Halifax, NS, Canada, 2017; pp. 1357–1366.
2. 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]
3. Li, Y.; Zhang, Y.; Huang, X.; Yuille, A.L. Deep Networks under Scene-Level Supervision for Multi-Class Geospatial Object Detection from Remote Sensing Images. *ISPRS J. Photogramm. Remote Sens.* **2018**, *146*, 182–196. [CrossRef]
4. Manso-Callejo, M.-Á.; Cira, C.-I.; Alcarria, R.; Arranz-Justel, J.-J. Optimizing the Recognition and Feature Extraction of Wind Turbines through Hybrid Semantic Segmentation Architectures. *Remote Sens.* **2020**, *12*, 3743. [CrossRef]
5. Vali, A.; Comai, S.; Matteucci, M. Deep Learning for Land Use and Land Cover Classification Based on Hyperspectral and Multispectral Earth Observation Data: A Review. *Remote Sens.* **2020**, *12*, 2495. [CrossRef]
6. Radočaj, D.; Obhođaš, J.; Jurišić, M.; Gašparović, M. Global Open Data Remote Sensing Satellite Missions for Land Monitoring and Conservation: A Review. *Land* **2020**, *9*, 402. [CrossRef]
7. Feltynowski, M.; Kronenberg, J. Urban Green Spaces—An Underestimated Resource in Third-Tier Towns in Poland. *Land* **2020**, *9*, 453. [CrossRef]
8. 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]
9. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*; Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 234–241.

10. 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; pp. 7132–7141.
11. Shan, B.; Fang, Y. A Cross Entropy Based Deep Neural Network Model for Road Extraction from Satellite Images. *Entropy* **2020**, *22*, 535. [CrossRef]
12. Lin, Y.; Xu, D.; Wang, N.; Shi, Z.; Chen, Q. Road Extraction from Very-High-Resolution Remote Sensing Images via a Nested SE-Deeplab Model. *Remote Sens.* **2020**, *12*, 2985. [CrossRef]
13. Hu, F.; Xia, G.-S.; Hu, J.; Zhang, L. Transferring Deep Convolutional Neural Networks for the Scene Classification of High-Resolution Remote Sensing Imagery. *Remote Sens.* **2015**, *7*, 14680–14707. [CrossRef]
14. Senthilnath, J.; Varia, N.; Dokania, A.; Anand, G.; Benediktsson, J.A. Deep TEC: Deep Transfer Learning with Ensemble Classifier for Road Extraction from UAV Imagery. *Remote Sens.* **2020**, *12*, 245. [CrossRef]
15. 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.
16. De la Fuente Castillo, V.; Díaz-Álvarez, A.; Manso-Callejo, M.-Á.; Serradilla García, F. Grammar Guided Genetic Programming for Network Architecture Search and Road Detection on Aerial Orthophotography. *Appl. Sci.* **2020**, *10*, 3953. [CrossRef]
17. 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.
18. Van den Oord, A.; Kalchbrenner, N.; Kavukcuoglu, K. Pixel Recurrent Neural Networks. In Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York, NY, USA, 19–24 June 2016; Balcan, M.-F., Weinberger, K.Q., Eds.; JMLR.org: Brookline, MA, USA, 2016; Volume 48, pp. 1747–1756.
19. Kingma, D.P.; Welling, M. Auto-Encoding Variational Bayes. In Proceedings of the 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, 14–16 April 2014. Conference Track Proceedings; 2014.
20. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.C.; Bengio, Y. Generative Adversarial Nets. In Proceedings of the Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, Montreal, QC, Canada, 8–13 December 2014; pp. 2672–2680.
21. Pan, Z.; Yu, W.; Yi, X.; Khan, A.; Yuan, F.; Zheng, Y. Recent Progress on Generative Adversarial Networks (GANs): A Survey. *IEEE Access* **2019**, *7*, 36322–36333. [CrossRef]
22. Radford, A.; Metz, L.; Chintala, S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In Proceedings of the 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, 2–4 May 2016. Conference Track Proceedings; 2016.
23. Mirza, M.; Osindero, S. Conditional Generative Adversarial Nets. *arXiv* **2014**, arXiv:1411.1784.
24. Liu, X.; Wang, Y.; Liu, Q. Psgan: A Generative Adversarial Network for Remote Sensing Image Pan-Sharpening. In Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; pp. 873–877.
25. Ledig, C.; Theis, L.; Huszar, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 105–114.
26. Wang, X.; Yu, K.; Wu, S.; Gu, J.; Liu, Y.; Dong, C.; Qiao, Y.; Loy, C.C. ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks. In *Computer Vision—ECCV 2018 Workshops*; Leal-Taixé, L., Roth, S., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2019; Volume 11133, pp. 63–79. ISBN 978-3-030-11020-8.
27. Jolicoeur-Martineau, A. The Relativistic Discriminator: A Key Element Missing from Standard GAN. In Proceedings of the 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, 6–9 May 2019.
28. Hu, B.; Yao, P.; Fu, L.; Li, X.; Dong, K.; Zheng, T. Transfer Learning in Remote Sensing Images with Generative Adversarial Networks. In Proceedings of the 2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS), Beijing, China, 17–19 June 2019; pp. 124–129.
29. Jetchev, N.; Bergmann, U.; Vollgraf, R. Texture Synthesis with Spatial Generative Adversarial Networks. *arXiv* **2016**, arXiv:1611.08207.
30. Li, C.; Wand, M. Precomputed Real-Time Texture Synthesis with Markovian Generative Adversarial Networks. In *Proceedings of the Computer Vision—ECCV 2016—14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016*; Proceedings Part III; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer: Berlin/Heidelberg, Germany, 2016; Volume 9907, pp. 702–716.
31. Bergmann, U.; Jetchev, N.; Vollgraf, R. Learning Texture Manifolds with the Periodic Spatial GAN. In Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6–11 August 2017; Volume 70, pp. 469–477, PMLR, 2017.
32. Zhu, J.-Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2242–2251.
33. Kim, T.; Cha, M.; Kim, H.; Lee, J.K.; Kim, J. Learning to Discover Cross-Domain Relations with Generative Adversarial Networks. In Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6–11 August 2017; Volume 70, pp. 1857–1865, PMLR, 2017.

34. Yi, Z.; Zhang, H. (Richard); Tan, P.; Gong, M. DualGAN: Unsupervised Dual Learning for Image-to-Image Translation. In Proceedings of the IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, 22–29 October 2017; IEEE Computer Society: Washington, DC, USA, 2017; pp. 2868–2876.
35. 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]
36. 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]
37. 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]
38. 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]
39. 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]
40. Hartmann, S.; Weinmann, M.; Wessel, R.; Klein, R. StreetGAN: Towards Road Network Synthesis with Generative Adversarial Networks. In Proceedings of the 25th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision, Pilsen, Czech Republic, 29 May–2 June 2017.
41. Zhang, Y.; Li, X.; Zhang, Q. Road Topology Refinement via a Multi-Conditional Generative Adversarial Network. *Sensors* **2019**, *19*, 1162. [CrossRef] [PubMed]
42. 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, 22–29 October 2017; pp. 2100–2109.
43. Wang, X.; Gupta, A. Generative Image Modeling Using Style and Structure Adversarial Networks. In *Proceedings of the Computer Vision—ECCV 2016—14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016*; Proceedings Part IV; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer: Berlin/Heidelberg, Germany, 2016; Volume 9908, pp. 318–335.
44. He, H.; Wang, H.; Lee, G.-H.; Tian, Y. Bayesian Modelling and Monte Carlo Inference for GAN. In Proceedings of the ICML 2018: Theoretical Foundations and Applications of Deep Generative Models, Stockholm, Sweden, 10 July 2018; p. 13.
45. Li, F.-F.; Johnson, J.; Yeung, S. Lecture 13: Generative Models. 2017. Available online: https://cse.iitkgp.ac.in/~{}sudeshna/courses/DL18/Generative-Models-27Mar-18.pdf (accessed on 7 November 2020).
46. Forczmański, P. Performance Evaluation of Selected Thermal Imaging-Based Human Face Detectors. In *Proceedings of the 10th International Conference on Computer Recognition Systems CORES 2017*; Kurzynski, M., Wozniak, M., Burduk, R., Eds.; Springer International Publishing: Cham, Switzerland, 2018; Volume 578, pp. 170–181. ISBN 978-3-319-59161-2.
47. Isola, P.; Zhu, J.-Y.; Zhou, T.; Efros, A.A. Image-to-Image Translation with Conditional Adversarial Networks. GitHub repository. Available online: https://phillipi.github.io/pix2pix/ (accessed on 12 May 2020).
48. Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; et al. TensorFlow: A System for Large-Scale Machine Learning. In The Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI '16), Savannah, GA, USA, 2–4 November 2016; p. 21.
49. Nair, V.; Hinton, G.E. Rectified Linear Units Improve Restricted Boltzmann Machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10), Haifa, Israel, 21–24 June 2010*; Fürnkranz, J., Joachims, T., Eds.; Omnipress: Madison, WI, USA, 2010; pp. 807–814.
50. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6–11 July 2015; Bach, F.R., Blei, D.M., Eds.; JMLR.org: Brookline, MA, USA, 2015; Volume 37, pp. 448–456.
51. Xu, B.; Wang, N.; Chen, T.; Li, M. Empirical Evaluation of Rectified Activations in Convolutional Network. *arXiv* **2015**, arXiv:1505.00853.
52. Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; Chen, X.; Chen, X. Improved Techniques for Training GANs. In *Proceedings of the Advances in Neural Information Processing Systems*; Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2016; Volume 29, pp. 2234–2242.
53. 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; 2015.
54. Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Proceedings of the Advances in Neural Information Processing Systems*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30, pp. 6626–6637.
55. Arjovsky, M.; Bottou, L. Towards Principled Methods for Training Generative Adversarial Networks. In Proceedings of the 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, 24–26 April 2017. Conference Track Proceedings.