Semantic segmentation of high-resolution aerial imagery using a fully convolutional network

Farida Bint Ahmad Nchare, Hippolyte Tapamo

Semantic segmentation of high-resolution aerial imagery using a fully convolutional network

Farida Nchare[1] **and Hippolyte Tapamo**[2]

[1]UMMISCO, University of Yaounde I, Cameroon
[2]UMMISCO, University of Yaounde I, Cameroon

*E-mail : farida.nchare@facsciences-uy1.cm

Abstract

Semantic segmentation applied to aerial imagery allows the extraction of terrestrial objects such as roads, buildings and even vegetation. Having large, detailed datasets of navigable roads, is of paramount importance in several application fields; namely urban planning, automatic navigation, disaster management. To reach this goal, extracting all roads in a given territory area is the first step. This paper presents a modern method to semantically segment aerial images for a road network extraction. We employ an encoder-decoder architecture to approach the problem of disconnected road regions faced by some existing methods. Using an FCN approach, the localization information was combined to the semantic one, to enable the reconstruction of the road by the proposed model, while being consistent with following the spatial alignment. The method was implemented and evaluated on the public dataset Massassuchets Roads. Results appear to be in full agreement with the theorical predictions and a significant improvement in road connectivity over some previous works; the proposed network achieved a precision of 87.86% and a recall of 87.89%.

Keywords

Road Extraction; FCN; Semantic Segmentation; Aerial Imagery

I INTRODUCTION

Roads play an essential role in the economy of a society, allowing the transportation of goods, people and merchandise. Consequently, the possession of large datasets of navigable roads in a given territory area is of crucial importance in many fields such as urban planning, traffic management, creation of land use maps which is carried out at the National Institute of Cartography in Cameroon. These road datasets are usually created manually. The process is thus very slow and more expensive with the increasing volume of aerial images data [14]. There is therefore, a need to automate the process to ensure functional continuity of the underlying systems [14]. The task to accomplish throughout this work is to propose an automatic system, that, given an aerial image, will produce an output image presenting the existing road network on the input image. The associated scientific problem is called semantic segmentation. Many researches have been conducted in this field, which led to the development of several techniques, including the work of **Tapamo et al.** [3], which proposes a texture descriptors extraction for a forest biomass prediction, using several supervised classification methods such as k-NN (k-nearest neighbors), SVM (support vector machines) and Random Forests.

This work is inspired by different previous works using a deep learning approach, in particular, convolutional neural networks that succeeded in obtaining good feature extractors automatically [2, 5, 7]; to solve the road extraction problem.

II STATE-OF-THE-ART

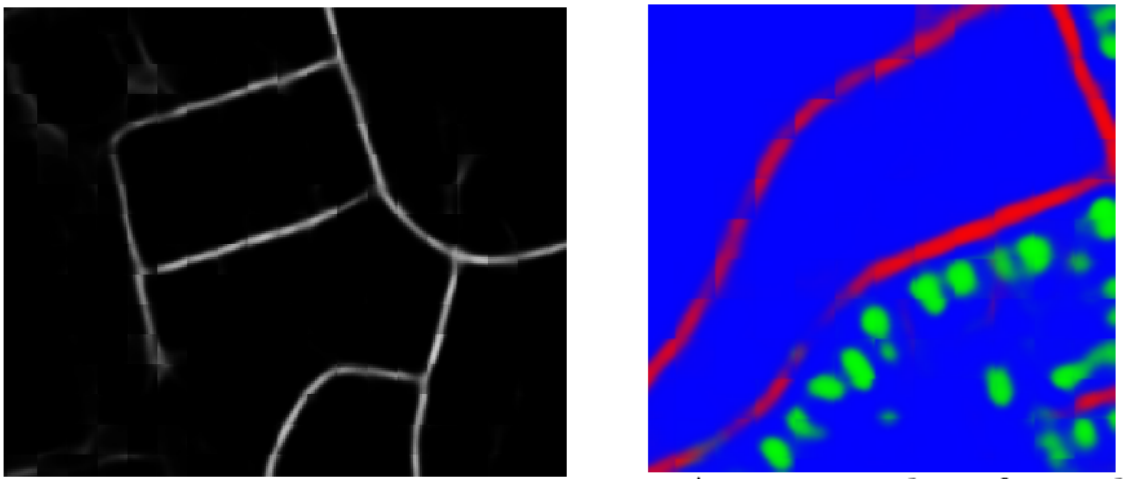
2.1 Related works

Recently, deep learning techniques have excelled in numerous areas, including semantic segmentation. In this section, related works proposing CNN-based approaches that serve as a foundation to understand the topic are presented. It should be noted that CNNs principally learn the contextual information. **Mnih et al.** [2] proposed a single-class prediction system using a patch-based CNN. Thus, to learn and evaluate roads and buildings, they train two CNNs separately. They predict those objects existence probability distribution from aerial imagery and formulate the problem of extracting relative pixels as obtaining a mapping from an aerial image patch to a label image patch. In their method, firstly an input aerial image is divided into 64×64 patches, to better integrate the local context, and normalized using a Gaussian as a preprocessing step. Then, the normalized patches are input into a 3 convolutional layers followed by 2 fully connected layers, which ouput a 256-vector, reshaped as a 16x16 label patch. They have tested their approach with two datasets that consist of large aerial imagery and binary road and buildings labels images. In order to extract road and buildings from Mnih [2] datasets, **Saito et al.** [5] extend the previous problem [2] into a multi-class dense classification problem. They proposed a single CNN approach to predict roads, buildings and the background simultaneously. They used the same CNN architecture as Mnih [2], but instead of a 256-vector, the ouput was reconfigured to produce a 768-vector, reshaped into a 16x16x3 RGB label patch. The model outperforms the baseline [2] both for the road and the building classes. This stems from the fact that the single CNN architecture has the advantage of being able to exploit the correlation when it exists, between a road and a building on the same image. Further on, the authors applied a channel-wise inhibited softmax (CIS) function to suppress the effect of the background.

Although the contextual features are neccessary in a semantic segmentation, object location is of a very importance in the field. The fundamental problem with CNNs for semantic segmentation is that, they can combine pixel location (in their fully connected layers) resulting in a loss of this information. Consequently, a set of new architectures, FCNs, have been proposed to preserve this location information. **Long et al.** [4] introduced the FCN network, which is the first end-to-end architecture to apply CNNs to semantic segmentation in the literal sense. The idea of this method is, instead of creating a new network from scratch, they proposed an algorithm called the *convolutionalization*, to convert a base convolutional neural network (CNN) into a fully convolutional network (FCN). Their method transforms the fully connected layers of a CNN into corresponding convolutions. In addition, they introduced a technique of upsampling using deconvolutional layers, which allows them to output a segmentation map. **Maggiori et al.** [6] addressed the buildings extraction task by suggesting a FCN approach. Their architecture is based on the encoder-decoder concept, in which the input image is compressed using Mnih [2] *convolutionalized* CNN, into a smaller representation before being reconstructed to the label size using upsampling layers in the decoder. The authors succeed in filling the gaps observed in the buildings [2], but some irregularities at the edges of the predicted objects are noticed. **Rasha et al.** [8] also used an FCN architecture to extract roads and buildings from two datasets with

different spatial resolutions and image conditions. They used the normal downsampling process of the base CNN [2] but at the end of the network, the fully connected layers are replaced by a GAP (Global Average Pooling) layer to address the problem of low computing efficiency. Their model outperforms the previous models [2, 5] for roads and buildings by using a multi-class approach.

2.2 Limitation with patch-based CNN



((a)) Mnih [2] ((b)) Saito and Aoki [5]

Figure 1: Examples of predicted road chunks that ignore patch continuity.

The limit that will be the focus of this work is the disconnected road segments observed in **Mnih et al.** [2]. This limitation is visually perceptible in Figure 1.

A close observation of the figure shows that the predictions at the extremities of the patches are not continuous for some road areas. This is due to the fact that not having kept the location information of the pixels, the networks [2, 5] succeed in predicting the road but the expected spatial alignment is not always respected.

2.3 Research question

Is it possible to develop a segmentation model that improves road connectivity using a deep learning approach?

III PROPOSED METHOD

Pixels location is a key element in semantic segmentation. The usual CNN architectures lose this information in the fully connected layers, usually added at the end of the network and designated as the classifier. Due to their fully connected nature, there is no way to determine on these classifiers which input contributed to the prediction result of an output. The network of **Mnih et al.** [2] captures semantics but at its fully connected layers, there is a loss of spatial information. However, the FCN of **Maggiori et al.** [6] manages to combine semantics and pixels location, which allows them to fill in the gaps on the buildings [2] during predictions. In a similar way, this work proposes an FCN approach to solve the problem of disconnected road segments [2].

Since the FCN of **Maggiori et al.** [6] is a conversion of Mnih network [2], the idea of the proposed method is to also convert the base CNN [2] into an FCN. This is what **Long et al.** [4] called a *convolutionalization*. A CNN can hence be converted into an FCN as follows:

1. First, the fully connected layer that performs the classification is rewritten as a convolution. The resulting connections are comparable to a fully connected layer if the chosen convolution kernel has dimensions that match the previous layer. Continiously, the other fully connected layers are turned into convolutions.

2. At this step, the output image is of very low resolution. So to recover the size of the label, a deconvolution is added to the network. This layer will learn filters to increase the size of the output mask. In the case of this work, the deconvolution which is actually a transposed convolution, will receive in an input of size 7x7 and will multiply it by a 4x4 filter and a factor of 2 to produce a patch label of size 16x16.

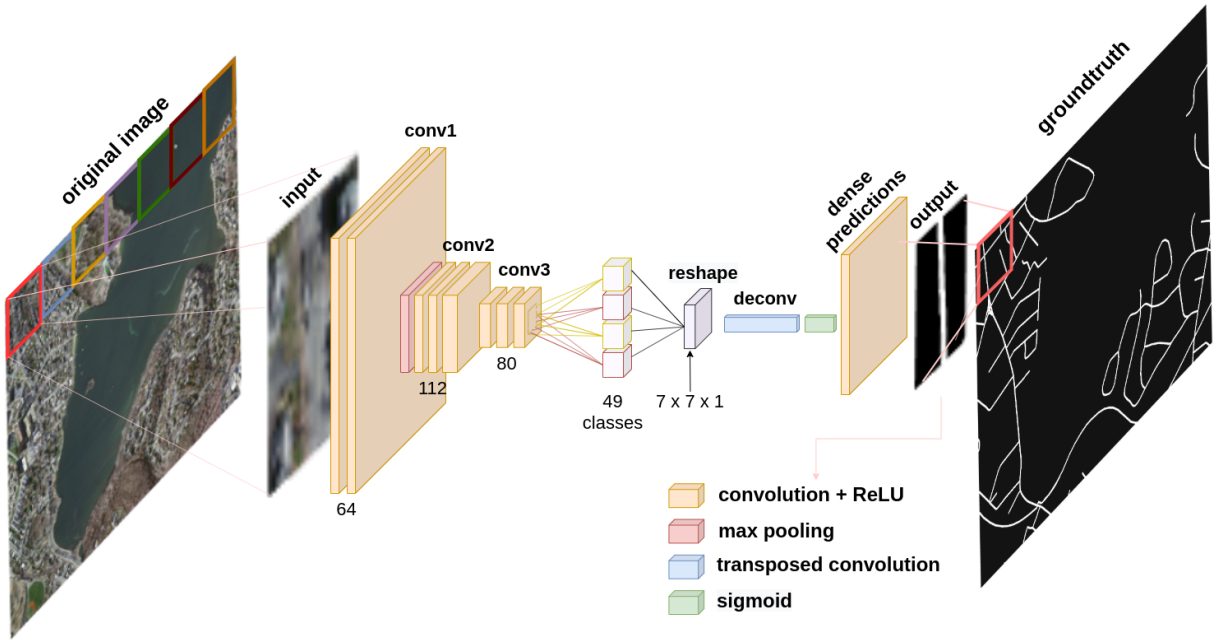


Figure 2: Proposed architecture.

The architecture of the proposed model shown in figure 2 has two modules, an encoder or the convolutional network and a decoder or the deconvolutional network.

The main goal of the first module is to produce a reduced representation of the input image, by analyzing each sub-region of the image and performing mathematical convolution and max-pooling operations on the pixel values [10]. The output produced will represent the probability that a road appears on each of the analyzed sub-regions. Then, the deconvolutive module will transform the probability vector produced by the first module, into a final road topology output.

IV EXPERIMENTATIONS

4.1 Dataset and evaluation metrics

In this work, the public dataset Massachusetts Roads [2] was used. It contains more than 1108 aerial images with a size of 1500x1500, each. This roads dataset consists of 1107 images for training, 49 for the validation and 14 for the testing part. But in order to be consistent with the

previous baseline methods, in this work also, only 137 aerial images and the associated binary label images have been retained for training, 10 for testing and 4 for the validation part.

In order to prepare the data for the network, the images are split into 64x64 patches, which amounts to 8844 patches per image. A rotation of 90° is then applied to each patch, which allows us to increase the dataset.

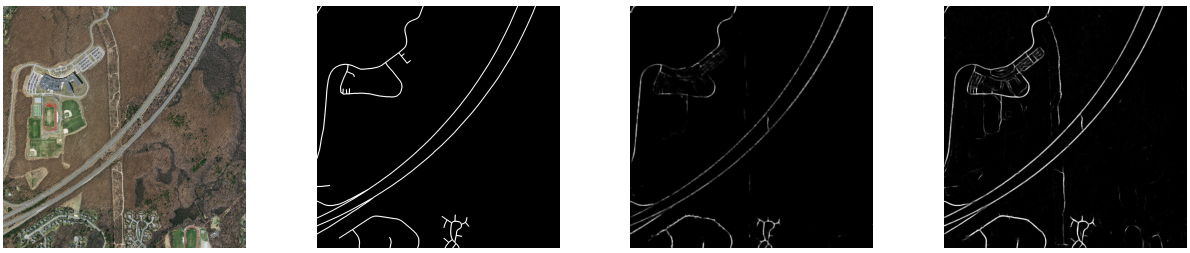
The evaluation of the model was done at two stages: a quantitative evaluation which consists in calculating the precision and the recall of the model and a qualitative evaluation which consists in a visual inspection of the prediction results. Nevertheless, just as **Mnih et al.** [2], the quantitative metrics used are the relaxed precision and relaxed recall.

4.2 Implementation details

The model training was done after initializing several hyperparameters. Indeed, all neural networks were trained for 60 epochs instead of 400 [2, 5, 8] on the training data, by minimizing the binary cross-entropy loss and using a stochastic gradient descent, with a momentum of 0.9, a learning rate that varied between 0.001 or 0.0005, and was reduced by a factor of 0.1 when there were no further improvements in the performances after nearly 5 consecutive epochs. In addition, each network was regularized using the L2 regularization method with a weight decay of 0.0002. To justify the reduced number of epochs, as the necessary resources are not constantly available over 5 consecutive days, which is the minimum required for 400 epochs, we have limited this number to 60, to ensure that the models can be run without an interruption on the workstation used.

4.3 Results

Examples of the results obtained after testing the base CNN and the proposed FCN are shown in figure 3.



((a)) aerial image ((b)) road label ((c)) Mnih et al. [2] ((d)) proposed FCN

Figure 3: Prediction results of the different models

To compare the proposed method in this work with those in the state of the art [2, 5, 8], an implementation of the architecture of each model was done under the same conditions and using the same environments. The results presented in brackets are the results published in their respective works while those outside are those obtained after the conducted experiments (Rasha model [8] was implemented without the post-processing).The models [2, 5, 8] were trained for 60 epochs instead of 400, due to some constraints. This work also compares the proposed model to other models tested on the same dataset and published more recently Cf. Table 1.

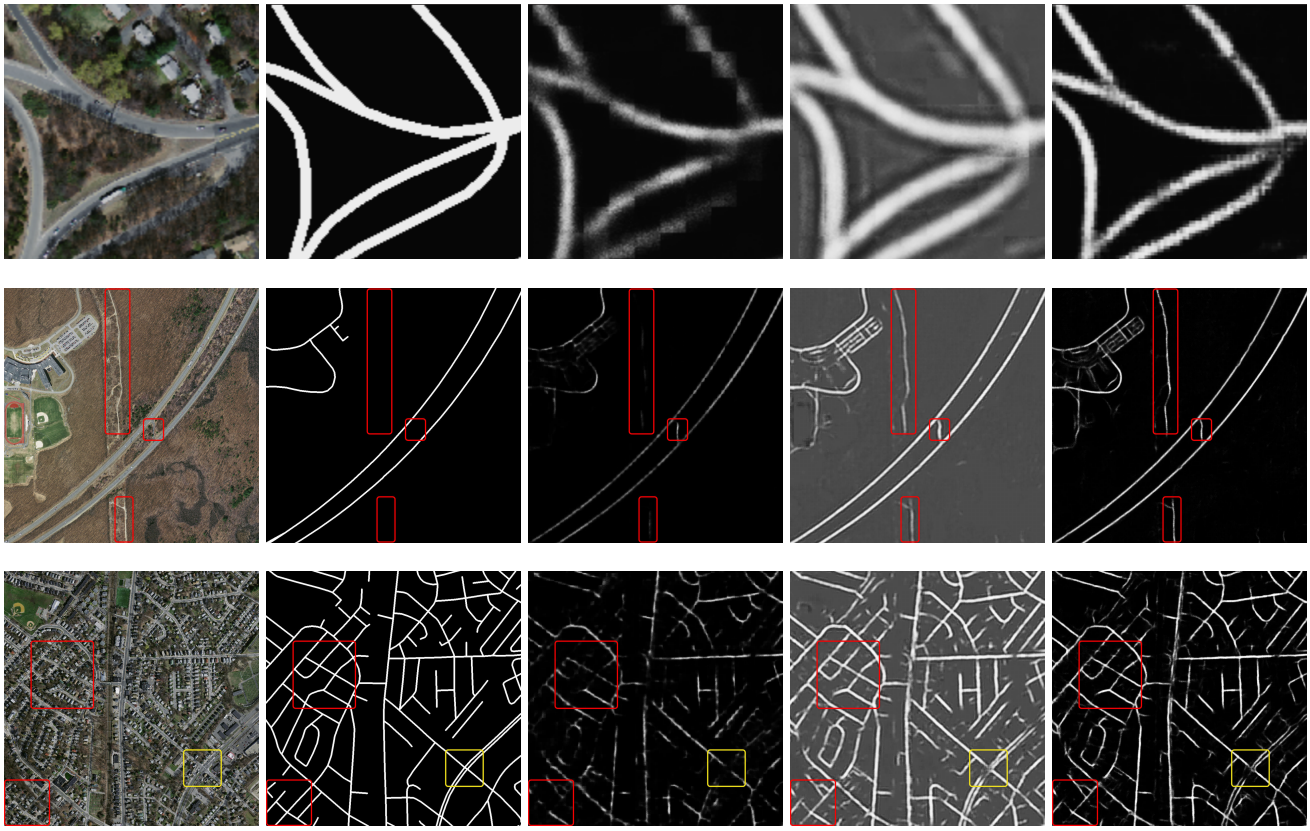
Dataset	Model	Precision	Recall
Massachusetts Roads	**Mnih** [2] (2013)	86.23 (88.73)	86.78
	Saito [5] (2015)	88.66 (90.05)	-
	Rasha [8] (2017)	89.10 (91.7)	-
	Proposed model	87.86	87.89
	ASPP-UNet-SSIM [12] (2019)	*87.10*	*80.50*
	JointNet [13] (2019)	*85.36*	*71.90*

Table 1: Comparison of road extraction models on Massachusetts Roads dataset.

V CONCLUSION AND REFERENCES

5.1 Discussions

Even after a short training time, lasting only one day (rather than a minimum of 5), it can be easily noticed that the results are promising and that the network started to recognize the road network in an accurate way.



((k)) ((l)) ((m)) ((n)) ((o))

Figure 4: Qualitative evaluation on different models. ((k)): Original image. ((l)): Label. ((m)): Mnih et al. [2] ((n)): Rasha et al. [8] ((o)): proposed FCN

From table 1, it can be seen that the proposed model achieves better performance than the baseline model [2]. Firstly, the proposed method shows a 1.63% improvement on precision which is explained by an accuracy of the predicted road pixels at the borders of some road segments; secondly, a 1.11% improvement on recall is observed, which is justified by a more

apparent road pattern. Let us now turn to a qualitative evaluation, that can verify the previous statements. Figure 4 shows for the model of Mnih [2], a disparity of the road network at some edges of segments which is not the case for the proposed FCN. Moreover the road marking of the proposed FCN is more pronounced compared to the base CNN.

Another qualitative evaluation (Cf. Figure 4, line 2) shows that, the proposed model succeeds in predicting road tracks that are not referenced on the labels. This could lead to a satisfaction with the model's performance, but it should be noted that the proposed model will be penalized for correctly predicting a road.

The strength of the proposed method, although it has been trained on a small dataset, reside in overcoming some of the main challenges consisting firstly, in recognizing several road patterns despite having large visual differences and isolating thin objects. Besides, the predictions are visually not coarses, quite smooth, and for the most part of the road network, they are continuous and free of noise.

However, our model had trouble generalizing a wide variety of patterns, predicting unexpected objects like parking spaces and missing many roads segments, mostly the intersections and the two-lane roads (Cf. Figure 4, line 3). The qualitative evaluation of the results shows the limits of the proposed method, as the model does not annotate in prior to the road topology (road corners, complex shapes). Consequently, many prediction failures are due to this shortcoming.

5.2 Conclusion

A road extraction method to solve the problem of disconnected road segments, has been proposed in this work. The FCN architecture of the proposed model consists of two modules, namely the convolutional module and the deconvolutional module. The first module will use contextual information to produce a coarse representation of the road while the 2nd module will use spatial information to reconstruct the existing road topology. The obtained results can be qualified as competitive after a quantitative and qualitative evaluation.

The main advantage of this method is that it increases the connectivity of the road of the base network [2]. Nevertheless, a limit is identified by observing the edges of the predicted roads, which are rough and can be refined.

As perspectives, the execution of the proposed model on another dataset for example images of roads in Cameroon is planned, also an adaptation of this model for the simultaneous prediction of several classes, and finally a future work could focus on the correction of the errors of the Massachusetts Roads dataset.

REFERENCES

Publications

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk. "Slic superpixels compared to state-of-the-art superpixel methods". In: *IEEE Trans. Pattern Anal. Mach. Intell.* (2012), pages 2274–2282.

[2] V. Mnih. "Machine Learning for Aerial Image Labeling". PhD thesis. University of Toronto, 2013.

[3] T. Hippolyte, M. Adamou, N. Blaise, C. Pierre, and M. Olivier. “Linear vs non-linear learning methods A comparative study for forest above ground biomass, estimation from texture analysis of satellite images”. In: *ARIMA* (2014).

[4] L. Jonathan, S. Evan, and D. Trevor. “Fully convolutional networks for semantic segmentation”. In: *IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)* (2015), pages 3431–3440.

[5] S. Saito and Y. Aoki. “Building and road detection from large aerial imagery”. In: *Image Processing: Machine Vision Applications VIII* (2015).

[6] M. Emmanuel, T. Yuliya, C. Guillaume, and A. Pierre. “Fully Convolutional Neural Networks For Remote Sensing Image Classification”. In: *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)* (2016).

[7] S. Shunta, Y. Yakayoshi, and A. Yoshimitsu. “Multiple Object Extraction from Aerial Imagery with Convolutional Neural Networks”. In: *Journal of Imaging Science and Technology* (2016).

[8] R. Alshehhi, P. R. Marpu, W. L. Woon, and M. D. Mura. “Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks”. In: *ISPRS Journal of Photogrammetry and Remote Sensing* (2017).

[9] T. Ali. “Deep Fully Residual Convolutional Neural Network for Semantic Segmentation”. preprint on webpage at https://scholarworks.unist.ac.kr/handle/201301/24592. 2018.

[10] G. Bogdan. “Detecting Roads from Aerial Images using Deep Learning”. preprint on webpage at todaysoftmag.com/article/2428/detecting-roads-from-aerial-images-using-deep-learning. 2018.

[11] X. Yongyang, W. Liang, X. Zhong, and C. Zhanlong. “Building Extraction in Very High Resolution Remote Sensing Imagery Using Deep Learning and Guided Filters”. In: *Remote Sensing* (2018).

[12] H. He, A. Yang, and D. Wang. “Road Extraction by Using Atrous Spatial Pyramid Pooling Integrated Encoder-Decoder Network and Structural Similarity Loss”. In: *Remote Sensing* (2019).

[13] Z. Zhengxin and W. Yunhong. “JointNet: A Common Neural Network for Road and Building Extraction”. In: *Remote Sensing* (2019).

[14] K. Pantelis. “Road Detection from Remote Sensing Imagery”. preprint on webpage at repository.tudelft.nl/islandora/object/uuid:21fc20a8-455d-4583-9698-4fea04516f03?collection=education. 2020.