SCENE CONTEXT IMPORTANCE FOR URBAN ROAD NETWORK DETECTION BY DEEP LEARNING

A. P. Dal Poz [1] *, A. G. G. Filho [2], M H. Shimabukuro [3]

[1] Dept. of Cartography, São Paulo State University, College of Sciences and Technology, Presidente Prudente-SP, Brazil - aluir.dal-poz@unesp.br
[2] Brazilian Arm, Brasília-DF, Brazil - guimaraes.filho@unesp.br@unesp.br
[3] Dept. of Mathematics and Computer Sciences, São Paulo State University, College of Sciences and Technology, Presidente Prudente-SP, Brazil - milton.h.shimabukuro@unesp.br

KEY WORDS: Road Detection, High-Resolution Images, Deep Learning, Unet, Scene Context.

1. INTRODUCTION

Since 2014, Deep Learning (DL) methods have been applied to numerous tasks in the Remote Sensing (RS) field. DL can extract features at different depth levels to produce mixed representations to pursue a specific purpose. Besides, among RS subareas and topics, road extraction has been a relevant research issue with significant challenges since the 70s. After the Fully Convolutional Network (FCN) (Long et al., 2015), some authors (as e.g Zhu et al., 2021) seek strategies to reduce discontinuity and omission caused by building and tree shadows, climatic conditions, and spectral similarity with other objects in the urban environment. However, despite several papers highlighting the scene context in segmentation, they did not explain how the context effectively impacts road detection. In this paper, we investigate scene context significance in road network detection.

2. PROPOSED METHOD

For this purpose, we chose uncomplicated Unet (pre-trained ImageNet encoder VGG-16) with some modifications in architecture and unbalanced losses (Dice (Milletari et al., 2016) and Focal Tversky (Abraham and Khan, 2018)) for road network detection in the Massachusetts dataset benchmark (Mnih, 2013). The chosen optimizer was Stochastic Gradient Descent (SGD) with 0.9 momentum, initial and final learning rates of 0.01 and 0.0001, regularization by Gradient Centralization (Yong et al. 2020), batch size of 4, and the following data augmentation techniques: 270º rotation, random vertical and horizontal mirror and random crop 512x512 size, also we used 224x224. In the 512x512 case, the benchmark has 8037, 126, and 441 for training, validation, and test tiles, respectively. The Unet was trained by 120 epochs (241,110 iterations), an average time of 13 hours per model to training performed in NVIDIA RTX 3080 10GB graphics processing unit (GPU).

3. EXPERIMENTAL RESULTS

The results showed that there are impacts that depend on the loss choice and size of training tiles. For example, the best model Unet dice 512 (optimization using dice loss and 512x512 size for training) reaches IoU = 65.62 and F1 = 79.05 in the average of 49 test samples. It was superior on IoU at about 3.5% and F1 by 2.1% compared with the model optimized by Tversky loss. Also, the Unet dice 512 was superior to the model trained with size 224x224 by 1.8% on IoU and 1.1% on F1. The impact of tile size is related to local and global context role in road network extraction, where local context allows the detection of road segments and the global context to infer the detection of the road network in occluded areas. Thus, the superior performance of the encoder is associated with the context presented in the training phase — more context, more class discrimination, and, consequently, a reduction of discontinuities. Moreover, we investigated two test samples (best and worst metric performance) to understand the disparity in results. For this, the Score-CAM method (Wang et al., 2020) was applied to support our analyses of road segmentation with the convolutional network.

In image 1 (Fig. 1), the road class was detected only at the highest abstraction encoder level (block5 conv). At lower levels, the encoder seeks to identify vegetation (canopy and hedge plants) and roofs at first and then detects roads. It indicates that the encoder requires the context of other objects to determine the road network in the urban environment. At the most abstract decoder level, the network switches the activation for the non-road class. The skip connections assist in delimitation because they bring non-road information. Then, at the level of lower abstraction and higher spatial resolution, the network shows greater activation and relevance in the boundary regions of the road network.

The residual detection characteristic of the encoder for roads is relevant for understanding the inferior performance (Fig. 2). At the two shallow levels (block1 conv and block2 conv), map activation is related to identifying classes of vegetation, water bodies, and roofs, as well as geometric characteristics such as edges and textures. From the 3rd level onwards, the network partially identifies the road class within the parking area. Finally, at the most abstract level (block5 conv), activating a vast area in the scene demonstrates the uncertainty in identifying the object. Nevertheless, even with the indecision of detection, the region where the road network has delimitation by vegetation, water, and roof classes presented the correct identification.

4. CONCLUSION

Therefore, the urban context for road network segmentation has a delimitation role. The detection of the road is residual, where non-road classes are identified at first. Thus, global context is essential for suitable segmentation. In State-of-the-art, our metric results overpass more complex networks, such as the JointNet (Zhang and Wang, 2019) (F1=79.35/IoU=64.00).

* Corresponding author

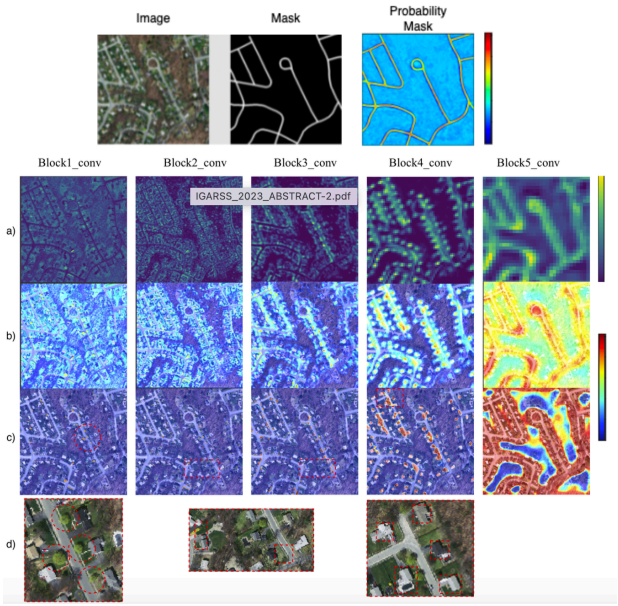

Figure 1. Class activation maps (CAM) from encoder Image 1. a) best score CAM of block; b) image overlaid with score CAM; c) image overlaid with emphasized score CAM; d) Red dashed circles and rectangles at line c).

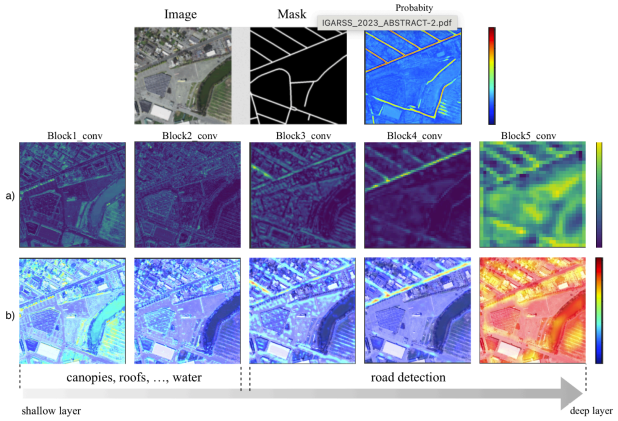

Figure 2. Class activation maps (CAM) from encoder Image 2. a) best score CAM of block; b) image overlaid with score CAM.

REFERENCES

Abraham, N., Khan, N. M., 2018. A Novel Focal Tversky loss function with improved Attention U-Net for lesion segmentation.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2015, 3431–3440.

Milletari, F., Navab, N., Ahmadi, S. A., 2016. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. *4th International Conference on 3D Vision,* 565–571.

Mnih, V., 2013. *Machine Learning for Aerial Image Labeling*, Ph.D. thesis.

Yong, H., Huang, J., Hua, X., Zhang, L., 2020. Gradient Centralization: A New Optimization Technique for Deep Neural Networks. *Lecture Notes in Computer Science* 12346, 635–652.

Zhang, Z., Wang, Y., 2029. JointNet: A Common Neural Network for Road and Building Extraction. *Remote Sensing* 11(6), 696, 2019.

Zhu, Q., Zhang, Y., Wang, L., Zhong, Y., Guan, Q., Lu, X. , Zhang, L., Li, D., 2021. A Global Context-aware and Batch-independent Network for road extraction from VHR satellite imagery. *ISPRS Journal of Photogrammetry and Remote Sensing* 175, 353–365.

Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., Hu, X., 2020. Score-CAM: Score-weighted visual explanations for convolutional neural networks. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 111–119.