AUTOMATIC OBJECT EXTRACTION FROM HIGH RESOLUTION AERIAL IMAGERY WITH SIMPLE LINEAR ITERATIVE CLUSTERING AND CONVOLUTIONAL NEURAL NETWORKS

A. C. Carrilho [1, *], M. Galo [1,2]

[1] Graduate Program in Cartographic Sciences - PPGCC, São Paulo State University – UNESP, Presidente Prudente, São Paulo, Brazil - (andre.carrilho, mauricio.galo)@unesp.br
[2] Dept. of Cartography, São Paulo State University - UNESP, Presidente Prudente, São Paulo, Brazil

KEY WORDS: Object extraction, Simple Linear Iterative Clustering, Convolutional Neural Networks

ABSTRACT:

Recent advances in machine learning techniques for image classification have led to the development of robust approaches to both object detection and extraction. Traditional CNN architectures, such as LeNet, AlexNet and CaffeNet, usually use as input images of fixed sizes taken from objects and attempt to assign labels to those images. Another possible approach is the Fast Region-based CNN (or Fast R-CNN), which works by using two models: (i) a Region Proposal Network (RPN) which generates a set of potential Regions of Interest (RoI) in the image; and (ii) a traditional CNN which assigns labels to the proposed RoI. As an alternative, this study proposes an approach to automatic object extraction from aerial images similar to the Fast R-CNN architecture, the main difference being the use of the Simple Linear Iterative Clustering (SLIC) algorithm instead of an RPN to generate the RoI. The dataset used is composed of high-resolution aerial images and the following classes were considered: house, sport court, hangar, building, swimming pool, tree, and street/road. The proposed method can generate RoI with different sizes by running a multi-scale SLIC approach. The overall accuracy obtained for object detection was 89% and the major advantage is that the proposed method is capable of semantic segmentation by assigning a label to each selected RoI. Some of the problems encountered are related to object proximity, in which different instances appeared merged in the results.

1. INTRODUCTION

Automatic object detection and extraction from high resolution aerial images in urban regions is a challenging task due to the complexity of the scene (Gonzalo-Martin et al., 2016). Recent advances in machine learning techniques for image interpretation have led to the development of robust approaches to both object detection and extraction. Most methods proposed recently rely on Deep Learning approaches by using Convolutional Neural Networks (CNN or ConvNets).

Traditional CNN architectures, such as LeNet (LeCun et al., 1998), AlexNet (Krizhevsky et al., 2012) and CaffeNet (Jia et al., 2014), usually capture as input images of fixed sizes taken from objects and attempt to assign labels to those images. More sophisticated architectures, such as Fully Convolutional Networks (FCN) proposed by Long et al. (2015) and its extension, U-Net (Ronneberger et al., 2015), are capable of dealing with different input sizes and of performing image segmentation. Another suitable approach is the Fast Region-based CNN (or Fast R-CNN), which works by using two models: (i) a Region Proposal Network (RPN) which generates a set of potential Regions of Interest (RoI) in the image; and (ii) a traditional CNN which assigns labels to the proposed RoI.

In general, the main disadvantage of using CNNs is the necessity of large datasets, thousands of images per class, to ensure class generalization during training, and also the increasing network complexity and number of parameters as more layers are introduced. According to Ronneberger et al. (2015), training FCN can be more difficult than traditional architectures since the training images must contain segmentation maps, which is more time consuming to produce. The other main problem of such methods is the loss of spatial resolution for boundary delineation due to the pooling layers (Yang et al., 2019). As an alternative, this study proposes a new approach to automatic object extraction from high resolution aerial images which is based on the Simple Linear Iterative Clustering (SLIC) algorithm to generate RoI that are then inferenced with a simple CNN architecture derived from CaffeNet.

2. IMAGE CLASSIFICATION

2.1 Image Classification Using Neural Networks

Traditional machine learning techniques (support vector machines - SVM, multi-layer perceptron - MLP, etc.) employ shallow features (geometrical, textural and contextual information) for low resolution image classification (Lv et al., 2018). The scene complexity associated with high resolution aerial imagery requires more powerful pattern recognition models. A straightforward approach is to associate every pixel of the image to a neuron at the input layer of the neural network, assuming that the connection weights within the hidden layers are capable of detecting the relevant aspects that make it possible to distinguish the class of each pixel.

The concept of Convolutional Neural Networks (CNNs) was originally proposed by Fukushima (1980), and then improved by LeCun et al. (1998) and Krizhevsky (2012). This research topic had a slow pace of development during the first decades due to the lack of processing power required to train the models. Nowadays this field has regained attention due to

* Corresponding author

powerful and affordable graphics processing units (GPUs) allied to better algorithms for training the networks. There were two main breakthroughs on the algorithm side of the model: (i) the adoption of a simpler activation function (the rectified linear unit – ReLU), which, according to Glorot et al. (2010), can speed up the training process, aiming at faster convergence; and (ii) the adoption of the dropout strategy (Hilton et al., 2012) to minimize the effects of overfitting. According to Jiang et al. (2018), the ReLU function is:

$$f(x) = \max(0, x) \tag{1}$$

The advantage of CNNs over traditional techniques is their ability to learn and extract their own features. The main idea is to simulate the process within the visual cortex of the brain (Fukushima, 1980). However, as the network is deeper (several convolution and pooling layers, as well as fully-connected layers, for instance), the number of parameters increase, thus requiring powerful hardware and large datasets for training (Amirkolaee and Arefi, 2019).

2.2 Object Localization Problem

Traditional CNN models are only capable of assigning a label to the image, i.e. the object localization problem remains. The Region-based Convolutional Neural Networks (R-CNN) instead attempt to solve the localization problem by using regions. According to Girshick et al. (2014), this kind of technique can solve both object detection and semantic segmentation by generating approximately 2000 category-independent region proposals which are then resized (with an affine transformation) to 227 by 227 pixels and used as input in the AlexNet. The original paper from Girshick et al. (2014) adopted the selective search as the region proposal method, however, they state that R-CNN is agnostic in this aspect.

The problem with R-CNN is that the method is not optimal as it requires the execution of inference with AlexNet 2000 times per image, which might be a bottleneck for real-time applications. According to Girshick (2015), the Fast R-CNN model attempts to increase the performance by using a Region Proposal Network (RPN). The RPN is a fully convolutional network used to acquire object bounds, generating high-quality region proposals. This approach was later refined by the Faster R-CNN (Ren et al., 2015) and the Mask R-CNN (He et al., 2017).

3. PROPOSED METHOD

The proposed method is similar to the Fast R-CNN architecture, the main difference being the use of the Simple Linear Iterative Clustering (SLIC) algorithm instead of an RPN to generate the RoI, as illustrated in Figure 1. This approach is also similar to the one presented in Chen et al. (2019) and Chen and Ming (2019), where the authors describe a multi-scale per-superpixel CNN (MCNN) based on the SLIC algorithm.

3.1 Generating Regions of Interest

Among the several image segmentation algorithms available in the literature, Simple Linear Iterative Clustering (SLIC) is regarded as most suitable for image interpretation due to the characteristics of its results (Achanta et al., 2012). In addition to its simplicity, this algorithm can cluster pixels into segments of similar size and shape, and is compared to state-of-the-art superpixels generation algorithms.

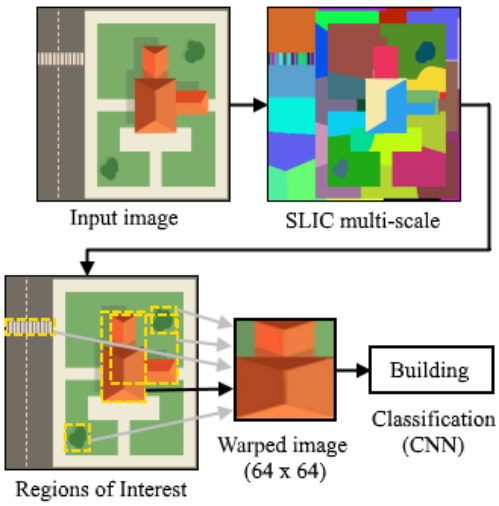

Figure 1 – Proposed approach for object extraction and classification from high resolution aerial images.

Assuming an image with N pixels, the first step of the algorithm works by selecting a predefined number (k) of regularly-spaced seed points over the image which are then disturbed (i.e. moved to the lowest gradient pixel inside a 3x3 neighboring window) to avoid object edges and noise. The seed points must be spaced within about $S=[N / k]^{1/2}$ pixels of each other, and they are defined as:

$$C_i = [l_i \;\; a_i \;\; b_i \;\; x_i \;\; y_i]^T \quad i = 1 \ldots k \tag{2}$$

In this vector C_i the first three elements correspond to CIELAB color space components, while the other two are the pixel position. The SLIC algorithm uses an adaption of the k-means clustering to aggregate neighboring pixels to each seed point. This iterative step is repeated until there are no further changes to the clusters.

As shown in Achanta et al. (2012), the combined metric (D) that takes both spatial (d_s) and color distances (d_c) is used in order to identify the seed point which is going to receive the current pixel:

$$D = \sqrt{d_c^2 + \left(\frac{d_s}{S}\right)^2 m^2} \tag{3}$$

where:

$$d_c = \sqrt{(l_j - l_i)^2 + (a_j - a_i)^2 + (b_j - b_i)^2} \tag{4}$$

and

$$d_s = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2} \tag{5}$$

are, respectively, the color and spatial distances between two pixels i and j, and $m \in [1,40]$ is a constant that weights the importance between the spatial and color distances. Achanta et al. (2012) emphasizes that the spatial distances might outweigh the color difference for large superpixels, so this metric has been shown to be useful.

3.2 CNN Architecture

The adopted CNN model illustrated in Figure 2 is a simplified version of CaffeNet, a framework derived from AlexNet (Hu et al., 2015). It takes RGB images of 64 x 64 pixels as input and assigns the object label (class) to each one. The CNN was composed of three nodes followed by a dense layer (fully connected) with 256 neurons. Each node consisted of a convolution layer followed by a max pooling layer, with increasing dropout on each node (25%, 30% and 40% respectively) to avoid overfitting. The number of kernels on each node was 32, 64 and 96 respectively. All convolution kernels were 3x3, and all layers of the network considered ReLU as the activation function.

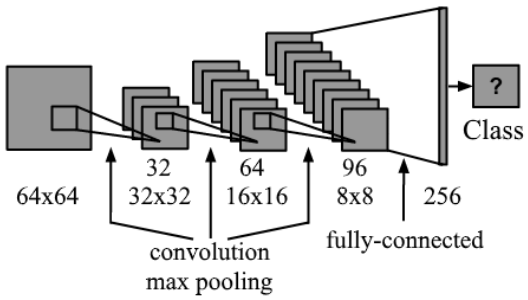

Figure 2 – Illustration of the adopted CNN architecture.

4. EXPERIMENTS AND RESULTS

The dataset used for training the CNN was structured in a similar manner to the UC Merced Land Use dataset (Yang and Newsan, 2010), but with fewer classes: house, sport court, hangar, building, swimming pool, tree, and street/road. Approximately 100 to 200 RGB image samples with 64 x 64 pixels were collected for each class, depending on their availability in the original images. In Figure 3 it is possible to see 10 examples of objects for each of the 7 classes considered.

Figure 3 – Sub images extracted from aerial images showing urban objects for CNN training.

The original aerial images come from the Unesp Photogrammetry Data Set (Tommaselli et al., 2018) collected from flights over the urban region of Presidente Prudente/Brazil in 2014. The digital images of 10328 x 7760 pixels (pixel size of 5.2 µm) were acquired by a Phase One iXA 180 digital camera, whose Charge-Couple Device (CCD), size 53.7 mm by 40.4 mm, registers RGB data. The Ground Sample Distance (GSD) of the images is approximately 12 cm.

4.1 Data Augmentation

Each image sample was subjected to a data augmentation process in order to achieve better generalization. Since the objects from different classes are arranged in different rotations in the urban area and direction of the flight lines changes, each image sample was rotated by 90º, 180º and 270º, so the CNN would be more robust to the object orientation. They were also subjected to random cropping in order to deal with partially occluded objects, as some low buildings might appear behind others due to the camera view point and perspective projection geometry. Applying the data augmentation process to the original image samples resulted in about 800 to 1200 images of the size 64 x 64 pixels for each class.

4.2 CNN Training

The dataset was divided into 80% for training and 20% for validation, and the selected CNN model achieved 96.7% accuracy. This result is similar to the accuracy achieved with the framework proposed in Jiang et al. (2018). An external validation was conducted with image samples collected with different sensors in other years, and the details are described in Section 5.

4.3 Assessment of the Proposed Method

Three experiments were conducted: (I) to identify the desired range of sizes for the RoI; (II) to assess the accuracy of object detection by inferencing the RoI with the CNN; and (III) analysis of the semantic segmentation.

The SLIC algorithm was used to generate superpixels (image segments considered as RoI by the CNN) with approximately the same predefined size. This characteristic is important for high resolution aerial image interpretation since the Ground Sample Distance (GSD) is known. The proposed method can generate RoI with different sizes by running a multi-scale SLIC approach as shown in Figure 4.

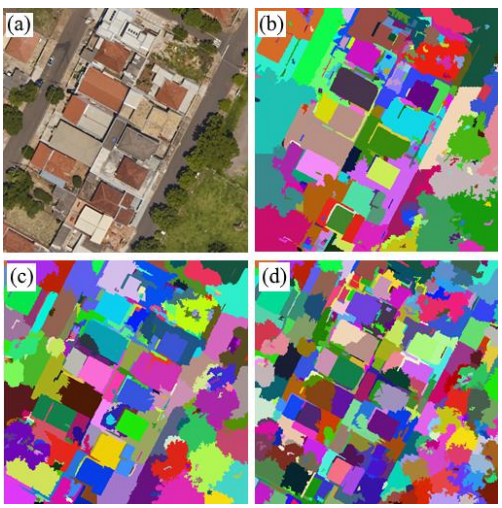

Figure 4 - Multi-scale SLIC approach. (a) Original image, SLIC superpixels with (b) k=50 (c) k=80 and (d) k=150.

4.4 Experimental Results

The range of sizes to be considered on the multi-scale SLIC approach depends mostly on the studied scene. A range of 5 – 150 m^2 was selected for residential areas, whereas industrial regions with large hangars or shopping malls achieved better

results within a range of 30 – 500 m^2. Using predefined scales seems to provide good results as emphasized in Chen and Ming (2019).

The overall accuracy for object detection with the proposed approach was 89%. Most of the confusion consisted of building roofs whose colors are similar to the street pavement. The method has also proven robust when detecting ceramic roofs as well as trees and swimming pools. The major advantage is that the proposed method is capable of semantic segmentation by assigning labels to the proposed RoI. In Figure 5 it is possible to see the bounding rectangle for some objects in the original images, for two residential areas.

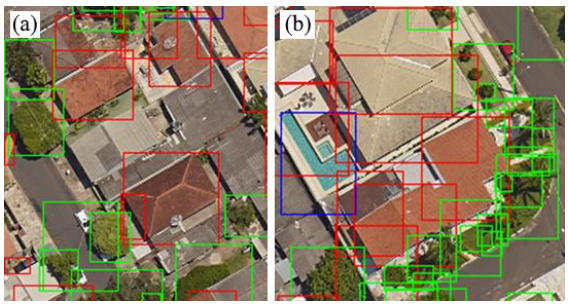

Figure 5 - Object detection for 2 regions. Legend: buildings (red), trees (green), swimming pool (blue).

As can be seen in Figure 5, some RoI intersect each other, thus, the same pixel might end up being processed several times by the CNN. This is more severe for the building roofs, due mostly to the use of axis-aligned bounding boxes (AABB) with the image coordinate system instead of oriented bounding boxes (OBB) when computing the warped image of the superpixels.

The semantic segmentation results were only assessed by visual inspection as no segmentation maps were generated for the training dataset. In Figure 6 some results can be seen for a residential area of Presidente Prudente. Some of the problems encountered are related to object proximity, in which different instances appeared merged in the results.

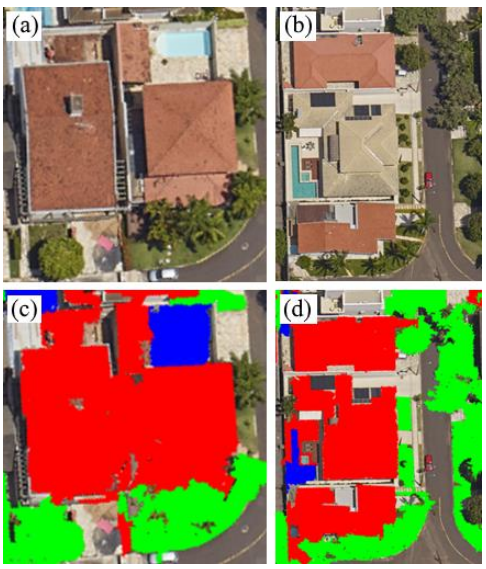

Figure 6 - Object extraction for regions (a) and (b) and the respective results in (c) and (d). Legend: buildings (red), trees (green), swimming pool (blue).

In Figure 6 it can be seen that the objects appear well delineated, as they come from the superpixel edges. This is an advantage over fully-convolutional approaches, since the feature maps in FCN are sub-sampled in the pooling layers, thus loosing spatial resolution.

5. EXTERNAL VALIDATION

A last analysis was conducted to provide an external validation of the selected CNN model. Three image samples for each class were collected from Google Earth and from a Quick Bird image, as depicted in Figure 7. The Quick Bird image was acquired in 2007. The original bands were combined considering pansharpening using the HSV color space from the RGB bands. The multispectral bands have a spatial resolution of 2.4 m, whereas the panchromatic have a 0.6 m GSD.

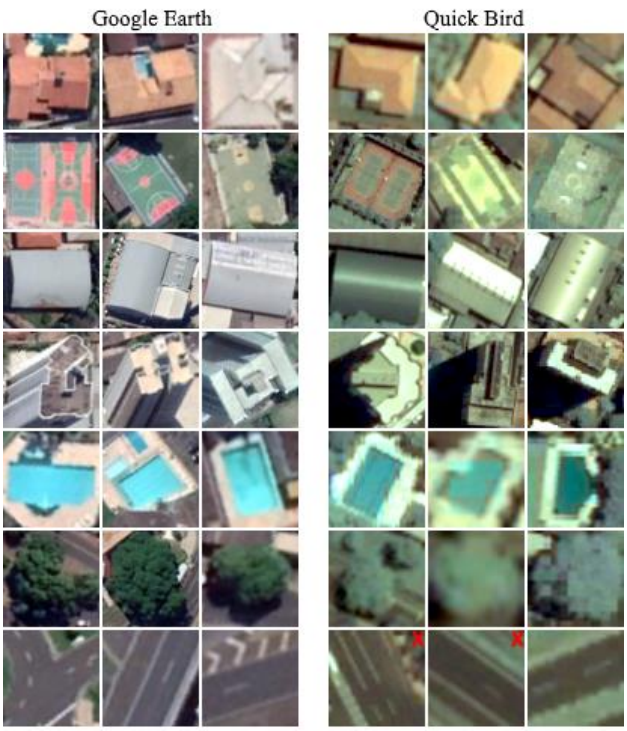

Figure 7 – External validation of the CNN training using Google Earth and Quick Bird images. The ‘x’ mark in red indicates misclassification.

Although the images used to train the CNN have a higher spatial resolution (12 cm GSD as stated before), the model was capable of correctly inferencing most of the images from the external validation set. The misclassification shown in Figure 7 (indicated by the ‘x’ mark) occurred in the street/road class for the Quick Bird images. This was expected for the Quick Bird image since the colors have some issues due to the pansharpening procedure. Only two of the samples were misclassified from the total of 42, that is, the CNN achieved 95.2% accuracy with the images from other sensors.

6. CONCLUSIONS

The proposed method was capable of solving object detection and segmentation from high resolution aerial images with satisfactory accuracy (89%). Even with a modest size (up to 200 samples per class), the dataset used in this paper was capable of training the selected CNN model without significant overfitting.

The data augmentation process was fundamental to ensuring a better generalization of the image samples.

The two main advantages are: (1) the good delineation of segmented objects; and (2) capability of object segmentation without using segmentation maps in the CNN training. The first advantage (1) requires further assessment and comparison with other models, such as Mask R-CNN, for instance. The second advantage (2) is interesting since the segmentation maps for the training dataset are time consuming to produce.

Future research might focus on the following aspects: (1) development of better region proposal techniques for object detection using variants of the SLIC algorithm and also variants of the suggested architecture; (2) application and validation of this technique for datasets with different characteristics.

ACKNOWLEDGEMENTS

The authors would like to acknowledge support from Coordination for the Improvement of Higher Level Personnel - CAPES, and the National Council for Scientific and Technological Development – CNPq (process 304189/2016-2). We also would like to acknowledge the São Paulo Research Foundation - FAPESP (process 05/01652-5) for providing the Quick Bird image. Finally, we would like to acknowledge the anonymous paper reviewers for their constructive comments.

REFERENCES

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S., 2012. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34, 2274-2282. doi.org/10.1109/TPAMI.2012.120

Amirkolaee, H.A., Arefi, H., 2019. Height estimation from single aerial images using a deep convolutional encoder-decoder network. *ISPRS Journal of Photogrammetry and Remote Sensing*, 149, 50-66. doi.org/10.1016/j.isprsjprs.2019.01.013

Chen, Y., Ming, D., Lv, X., 2019. Superpixel based land cover classification of VHR satellite image combining multi-scale CNN and scale parameter estimation. *Earth Science Informatics*, 1-23. doi.org/10.1007/s12145-019-00383-2

Chen, Y., Ming, D., 2019. Superpixel classification of high spatial resolution remote sensing image based on multi-scale CNN and scale parameter estimation. In: *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2-W13, 681-685. doi.org/10.5194/isprs-archives-XLII-2-W13-681-2019

Fukushima, K., 1980. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4): 193-202. doi.org/10.1007/BF00344251

Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 580–587.

Girshick, R., 2015. Fast R-CNN. In: *Proceedings of the IEEE International Conference on Computer Vision*, Santiago, Chile, 7–13; pp. 1440–1448.

Glorot, X., Bordes, A., Bengio, Y., 2010. Deep Sparse Rectifier Neural Networks. *J. Mach. Learn. Res*. 15, 315–323.

Gonzalo-Martin, C., Garcia-Pedrero, A., Lillo-Saavedra, M., Menasalvas, E., 2016. Deep learning for superpixel-based classification of remote sensing images. In: *Proceedings of the GEOBIA 2016: Solutions and Synergies*, Enschede, The Netherlands, 14–16.

He, K., Gkioxari, G., Dollar, P., Girshick, R., 2017. Mask R-CNN. In: *2017 IEEE International Conference on Computer Vision (ICCV)*, 2980–2988.

He, S., Lau, R.W.H., Liu, W., Huang, Z., Yang, Q., 2015. SuperCNN: A superpixelwise convolutional neural network for salient object detection. *International Journal of Computer Vision*. 115(3), 330-344. doi.org/10.1007/s11263-015-0822-0

Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R. R., 2012. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv Computer Science*, arXiv:1207.0580.

Hu, F., Xia, G.-S., Hu, J., Zhang, L., 2015. Transferring Deep Convolutional Neural Networks for the scene classification of high-resolution Remote Sensing Imagery. *Remote Sensing*, 7, 14680-14707. doi.org/10.3390/rs71114680

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T., 2014. Caffe: Convolutional Architecture for Fast Feature Embedding. In: *Proceedings of the ACM International Conference on Multimedia*, Orlando, FL, USA, 3–7 November 2014.

Jiang, S., Zhao, H., Wu, W., Tan, Q., 2018. A novel framework for remote sensing image scene classification. *Int. Arch. Photogram. Remote Sens. Spat. Inf. Sci.*, 42, 657–663. doi.org/10.5194/isprs-archives-XLII-3-657-2018

LeCun, Y., Bottou, Y., Bengio, Haffner, P., 1998. Gradient-Based Learning Applied to Document Recognition. In: *Proc. of the IEEE*, 86(11), 2278–2324. doi.org/10.1109/5.726791

Long, J., Shelhamer, E., Darrell, T., 2015. Fully Convolutional Networks for Semantic Segmentation. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3431–3440. doi.org/10.1109/CVPR.2015.7298965

Lv, X., Ming, D., Chen, Y., Wang, M., 2018. Very high resolution remote sensing image classification with SEEDS-CNN and scale effect analysis for superpixel CNN classification. *International Journal of Remote Sensing*, 40(2), 506-531. doi.org/10.1080/01431161.2018.1513666

Krizhevsky, A., Sutskever, I., Hinton, G., 2012. ImageNet Classification with Deep Convolutional Neural Networks. In: *Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS).*

Noh, H., Hong, S., Han, B., 2015. Learning deconvolution network for semantic segmentation. In: *IEEE International Conference on Computer Vision*, 1520-1528.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: *MICCAI 2015 – Medical Image Computing and Computer-Assisted Intervention*. doi.org/10.1007/978-3-319-24574-4_28

Ren, S., He, K., Girshick, R., Sun, J.,2015. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39 (6), 1137-1149.
doi.org/10.1109/TPAMI.2016.2577031

Tommaselli, A.M.G., Galo, M., Reis, T.T., Ruy, R.S., Moraes, M.V.A., Matricardi, W.V., 2018. Development and assessment of a data set containing frames images anddense airborne LASER scanning point clouds. *IEEE Geoscience and Remote Sensing Letters*, 15(2), 192-196.
doi.org/10.1109/lgrs.2017.2779559

Yang, Y., Newsam, S., 2010. Bag-of-visual-words and spatial extensions for land-use classification. In: *Sigspatial International Conference on Advances in Geographic Information Systems*. ACM. 70-279.

Yang, C., Rottensteiner, F., Heipke, C., 2019. Towards better classification of land cover and land use based on convolutional neural networks. In: *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2-W13, 139-146.
doi.org/10.5194/isprs-archives-XLII-2-W13-139-2019