Learning to Semantically Segment High-Resolution Remote Sensing Images

Keiller Nogueira[1], Mauro Dalla Mura[2], Jocelyn Chanussot[2], William Robson Schwartz[1], Jefersson A. dos Santos[1]

[1]*Department of Computer Science, Universidade Federal de Minas Gerais, Brazil*
{keiller.nogueira,william,jefersson}@dcc.ufmg.br,
[2]*GIPSA-lab, Grenoble Institute of Technology, France*
{mauro.dalla-mura,jocelyn.chanussot}@gipsa-lab.grenoble-inp.fr

***Abstract*—Land cover classification is a task that requires methods capable of learning high-level features while dealing with high volume of data. Overcoming these challenges, Convolutional Networks (ConvNets) can learn specific and adaptable features depending on the data while, at the same time, learn classifiers. In this work, we propose a novel technique to automatically perform pixel-wise land cover classification. To the best of our knowledge, there is no other work in the literature that perform pixel-wise semantic segmentation based on data-driven feature descriptors for high-resolution remote sensing images. The main idea is to exploit the power of ConvNet feature representations to learn how to semantically segment remote sensing images. First, our method learns each label in a pixel-wise manner by taking into account the spatial context of each pixel. In a predicting phase, the probability of a pixel belonging to a class is also estimated according to its spatial context and the learned patterns. We conducted a systematic evaluation of the proposed algorithm using two remote sensing datasets with very distinct properties. Our results show that the proposed algorithm provides improvements when compared to traditional and state-of-the-art methods that ranges from 5 to 15% in terms of accuracy.**

***Index Terms*—Land-cover Mapping; Pixel-wise Classification; Semantic Segmentation; Deep Learning; Remote Sensing; Feature Learning; High-resolution Images;**

I. Introduction

Fully scene understanding is a primary task in a wide range of applications and one of the most important in the remote sensing community. It is typically referred as land cover classification and is essential in a wide range of fields, such as urban planning [1], crop and forest management [2], and disaster relief [3].

The improvements in sensor technologies have significantly increased the accessibility to high spatial resolution images. Consequently, there is an increasing demand for new classification techniques able to better exploit these data. In the last decade, many approaches employing image segmentation were proposed to better exploit the spatial information present in high resolution images [4]. Accordingly, many works have discussed the advantages of region-based classification against the classical pixel-wise approach. However, classification considering segmentation-based techniques is still an open task [5], [6]. The main challenges are related to the semantic aspects inherent to the extraction of objects of interest, which is directly related to the selection of the most suitable segmentation scale. Tuning the segmentation parameters to effectively delineate objects of interest is application-dependent and usually needs some type of supervised learning or user interaction [7]. In the attempt of improving the precision of the results and reduce the impact of a non-optimal parametric configuration, many segmentation approaches based on multiple scales have been proposed [2], [8].

State-of-the-art methods [4] for land-cover classification rely on supervised learning based on pre-designed features extracted, for instance, from segmented regions of the original image, which may contain enough information. However, in a typical scenario, different descriptors may produce distinct results depending on the data. Thus, it is imperative to design and evaluate multiple feature descriptor approaches to find the most suitable ones for each particular application [6]. This process is also expensive and does not guarantee an accurate descriptive representation. A recent approach, called deep learning [9], overcome this limitation, since it can learn specific and adaptable spatial features and classifiers for the images, all at once. In this paper, we propose a supervised method to perform semantic segmentation on remote sensing images based on deep learning.

Deep learning, a branch of machine learning that refers to multi-layered neural networks, aims at learning features and classifiers at once from raw input data, i.e., a unique network may be able to learn features and classifiers (in different layers) and adjust the parameters, at running time, based on accuracy, giving more importance to one layer than another depending on the problem. When compared to previous state-of-the-art methods [10] (such as mid-level (BoVW) and low-level color and texture descriptors), deep learning presents a great advantage of end-to-end feature learning (e.g., from image pixels to semantic labels), which allows the method to effectively and adaptively learn features based on the input.

Amongst all deep learning-based networks, a specific type, called Convolutional (Neural) Networks, ConvNets or CNNs [9], is the most popular for learning visual features in

computer vision applications, including remote sensing. This type of network relies on the natural stationary property of an image, i.e., the statistics of one part of the image are assumed to be the same as those of any other part. In this way, information extracted at one part of the image can also be employed to describe other parts. Furthermore, deep ConvNets can be considered as an inherently multiscale approach since they usually obtain different levels of abstraction for the data, ranging from local low-level information in the initial layers (e.g., corners and edges), to more semantic descriptors, mid-level information (e.g., object parts) in intermediate layers and high level information (e.g., whole objects) in the final layers.

ConvNets have recently become the new state-of-the-art solution for visual recognition applications, such as image classification [11] and segmentation [1], depth estimation and object detection. Given their success, they have been intensively employed in several distinct tasks of different domains [9], including remote sensing [12], [13]. In this domain, the use of deep learning is growing very quickly, since it has a natural ability to effectively encode spectral and spatial information based on the data itself. Methods based on deep learning have obtained state-of-the-art results in many different remote sensing applications, such as agriculture [14], oil spill [15], and poverty mapping [16].

In this work, we propose a novel technique to automatically perform pixel-wise land cover classification. To the best of our knowledge, there is no other work in the literature that perform pixel-wise semantic segmentation based on data-driven feature descriptors for high-resolution remote sensing images. The main idea is to exploit the power of convolutional network feature representations to learn how to semantically segment remote sensing images. First, our method learns each label in a pixel-wise manner by taking into account the spatial context of each pixel. In a predicting phase, the probability of a pixel belonging to a class is also estimated according to its spatial context and the learned patterns.

In practice, we claim the following benefits and contributions over existing solutions:

- Our main contribution is a novel approach to create land-cover mapping for remote sensing domains.
- Two new ConvNet architectures used to perform pixel-wise image segmentation.

II. Related Work

Deep learning is making its way to the remote sensing community based on the success in several computer vision tasks, mainly due to the possibility of learning specific and adaptable features and classifiers for the images, all at once. Previous works using deep learning for remote sensing classification can be arranged in two main groups: (1) pixel-based methods for hyperspectral image classification; and (2) approaches for entire high-resolution image scene classification.

Scene classification is the task of assigning a label to a patch or to the entire image. In its context, Penatti *et al.* [12] show that features extracted from ConvNets pre-trained on general datasets (such as the one of The ImageNet Large Scale Visual Recognition Challenge – ILSVRC) can successfully be used to classify aerial and remote sensing images, outperforming state-of-the-art descriptors. These results corroborate with the suggestion of Razavian *et al.* [17], that features obtained from deep learning should be the primary candidates in most visual recognition tasks. Marmamis *et al.* [18] propose a two-stage method that combines features extracted from a pre-trained ConvNet and a supervised method trained over these descriptors. They successfully deal with the limited-data problem in an end-to-end processing scheme, mitigating overfitting and achieving excellent results. Finally, Nogueira *et al.* [19] evaluate three strategies to exploit ConvNets considering the remote sensing domain: (i) using ConvNet as feature extractor, (ii) fine-tuning a pre-trained network, and (iii) trained them from scratch (with randomly initialized weights). Six well-succeeded ConvNets were evaluated in each of these strategies with the goal to classify aerial and remote sensing images. They conclude that fine-tuning a pre-trained ConvNet is the best strategy for aerial and remote sensing domain, achieving state-of-the-art results in three datasets.

Girshick *et al.* [20] propose a method of object detection and segmentation based on features extracted from pre-trained ConvNets. In this work, which is one of the first to achieve state-of-the-art by using ConvNets to perform image segmentation, they classify patches extracted from the images (using selective search algorithm).Firat *et al.* [21] propose a method that combines Markov Random Fields with convolutional auto-encoders for object detection and classification in high-resolution remote sensing images. Chen *et al.* [22] perform pixel classification of hyperspectral and spatial data by combining stacked auto-encoders and Principal Component Analysis. Zhang *et al.* [23] propose a deep network based on stacked auto-encoders that is used to classify patches extracted from the images based on the saliency map, which increases the overall accuracy. In [24], the authors perform pixel labeling by combining ConvNets, hand-crafted (non data-driven) descriptors, and conditional random fields. The segmentation is carried out by classifying patches extracted from the images.

Our work differs from literature since there is no other one for land-cover mapping based on pixel-wise semantic segmentation for high-resolution remote sensing images.

III. The Proposed Approach

Our method uses the idea of *context window*, which is described in Section III-A. In Section III-B, we present the ConvNets we have proposed to perform this task and also explain how they could be trained by using a set of labeled pixels. Finally, in Section III-C, we explain how to employ the ConvNets to create land-cover maps.

A. Context Windows

The proposed land-cover mapping method assigns an object class to each pixel of an image by using sliding input *context windows*, which produces a label hypothesis for each pixel. Example of a context window is presented in Figure 1.

The overlapping windows centered on each pixel helps to understand the spatial patterns around them.

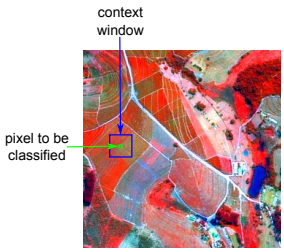


Fig. 1. Example of a context window. The pattern is represented by a large window that is centered on the pixel of interest in order to include the context of its neighborhood.

B. Learning to semantically segment

ConvNets [9] are deep learning architectures typically composed of several layers (each layer composed of processing units, also known as neurons) that can learn data-driven features and classifiers at the same time adjusting the learning, in processing time, based on the accuracy of the network. In this paper, we propose ConvNets architectures to learn the specific patterns from training pixels, which are represented by their context windows.

The feature learning step may be stated as a technique that learns a transformation of raw data input to a representation that improves the class separability [9]. Since encoding the spatial features in an efficient and robust fashion is the key for generating discriminatory models, the feature learning step is a great advantage of ConvNets when compared to conventional methods. In fact, the multiple layers (responsible for encoding spatial features automatically) learn adaptable and specific feature representations in a data-dependent hierarchical way. Thus, low-level descriptors are learned in initial layers of the network and high-level features in the deeper ones. This process learn all feasible information from the data, which creates robust features and classifiers.

The process for **learning to semantically segment** remote sensing images works as follows: given a set of labeled sample pixels and their contextual windows, a ConvNet is trained to learn the feature patterns that compose the class of regions of interest. One can note that the ConvNet architecture depends on the contextual window size. Applications with objects with more complex patterns may require large window size. Consequently, large window size requires more complex ConvNets, i.e., more layers, filtering and pooling operations. In this paper, we have proposed two networks named *Small* and *Large* ConvNet architectures. They are presented in Figures 2 and 3, respectivelly. The *Small ConvNet* receives as input 7×7 pixels context windows and has two convolutional layers and one fully-connected ones. The *Large ConvNet* receives as input 25×25 pixels context windows and has three convolutional layers and two fully-connected ones.

We used different techniques between some of layers, as can be seen in Figures 2 and 3, such as dropout regularization [25]

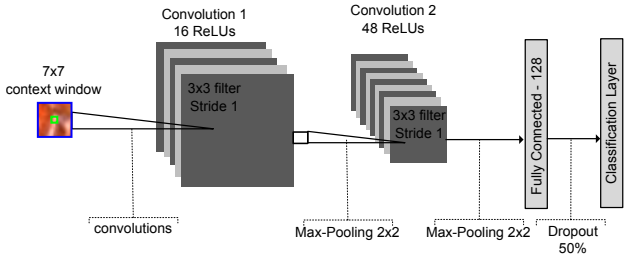


Fig. 2. Small ConvNet architecture: 7×7 context windows as input.

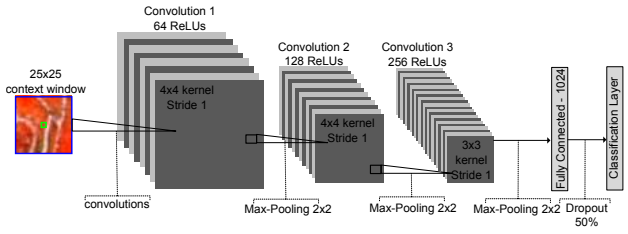


Fig. 3. Large ConvNet architecture: 25×25 context windows as input.

and max-polling. The main difference between them is that the *Large ConvNet* is used when larger context brings more useful information and less noise, while the *Small ConvNet* may be used in cases when larger context brings too much noise, disturbing the results. It is important to emphasize that Rectified Linear Unit (ReLU) [26] was the processing units used in all layers of the proposed ConvNets because of its advantages when compared to others, including: (i) prevents saturation during the learning, (ii) induces the sparsity in the hidden units, and (iii) does not face gradient vanishing problem [1].

C. Creating land-cover maps

Given an input image, the process of creating a land-cover map consists in classifying the context windows of each pixel by using the trained ConvNet. It is important to note that this process also works for a set of non-contiguous pixels. A overview of the predicting process is presented in Figure 4.

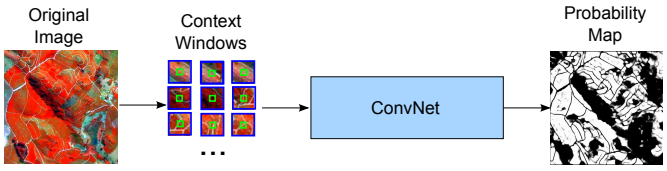


Fig. 4. A set of context windows are classified by an already trained ConvNet, generating the probability map, which are pos-processed generating the segmented image.

When a context window is classified by a ConvNet, the probability function generated by this classification is, in fact, associated with the pixel at the center of that window. This process allows the method to create a probability map over the

[1]The gradient vanishing problem occurs when the propagated errors become too small and the gradient calculated for the backpropagation step vanishes

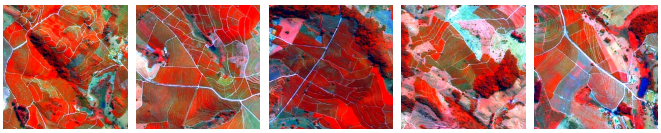
entire image, which, after some pos-processing method, results in a semantic segmented image. After obtaining the probability map, several methods can be used to create the final segmented image, such as classify the pixel with the highest probability class, which was the method employed in this paper.

IV. Experimental Setup

A. Dataset

To better evaluate the robustness and effectiveness of the proposed method, we carry out experiments on datasets with very distinct properties. The first one, named AGRICULTURE dataset, is a multispectral high-resolution scenes of coffee crops and non-coffee areas. The second, referred as URBAN dataset, is a very high spatial resolution in the visible spectrum and the objective is to map urban targets, such as roads and buildings.

1) AGRICULTURE dataset: This dataset is a composition of five images taken by the SPOT sensor in 2005 over Monte Santo, a coffee grower county in the State of Minas Gerais, Brazil. Each image has 500×500 pixels with green, red, and near-infrared bands, which are the most useful and representative ones for discriminating vegetation areas. More specifically, the dataset has 1,250,000 pixels with 637,544 (51%) coffee pixels and 612,456 (49%) non-coffee pixels annotated by specialists. Figure 5 shows each image and the respective ground-truths.



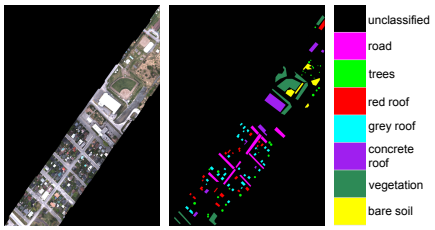
(a) Image

(b) Ground-Truth

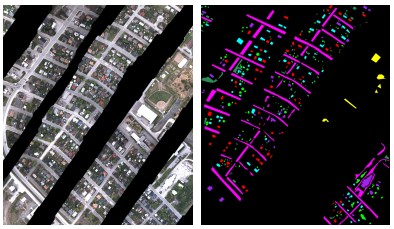
Fig. 5. The AGRICULTURE dataset. Multispectral images and ground-truth. Black regions represent non-coffee areas while white pixels represent coffee crops.

This is a very challenging dataset since intraclass variance is high due to different crop management techniques. Furthermore, coffee is an evergreen culture and the South of Minas Gerais is a mountainous region, which means that this dataset includes scenes with plants at different stages of growth and/or with spectral distortions caused by shadows.

2) URBAN dataset: Proposed in the IEEE GRSS Data Fusion Contest 2014, this dataset is composed of a fine-resolution visible image that covers an urban area near Thetford Mines in Quebec, Canada, containing seven different classes: trees, vegetation, road, bare soil, red roof, gray roof, and concrete roof. In this work, we do not consider the hyperspectral data available in this dataset. The training image has 2830×3989 pixels while the testing one has 3769×4386 pixels of resolution. The images, as well as the respective ground-truths, are presented in Figure 6.



(a) Training



(b) Test

Fig. 6. The URBAN dataset. Training and test data and their respective ground-truths.

B. Baselines

For the AGRICULTURE dataset, we consider a region-based classification [27], which is the most common strategy for high-resolution remote sensing. It is composed by the following steps: (i) segmentation; (ii) feature extraction of the regions; and (iii) supervised training with shallow learning method. In this case, we have used SLIC [28], which usually achieves good results for remote sensing images [29]. We employed BIC [30] as feature extraction algorithm, which is the most suitable descriptor to describe coffee crops, as pointed out by [12]. We used Support Vector Machine with Radial Basis Kernel (RBF-SVM) to perform the supervised learning [31].

For the same dataset, we also considered the method proposed by Nogueira et al. [14], called Cascade Convolutional Neural Network (CCNN). This method employs a multi-scale strategy allied to ConvNet to perform classification of fixed size tiles towards a final segmentation of the image.

For the URBAN dataset, we used a diversity-based fusion framework (DFF) proposed by Faria et al. [32] as implemented in [33]. DFF combines image characterization and learning methods using meta-learning approach, that is responsible for assessing which methods contribute more towards the solution of a given problem. DFF uses a selection strategy to pick the less correlated classifier, yet effective, classifiers according to a series of diversity measures analysis.

C. Experimental Protocol

For the AGRICULTURE dataset, we conduct a five-fold cross validation to assess the overall accuracy of the proposed algorithm. The results reported are the average accuracy of the five runs followed by the standard deviation. For the URBAN dataset, as introduced in Section IV-A2, an image is used for

training while the other is used for test. For both datasets, the Kappa index [34], which measures the agreement between the reference data and the classifier result, was reported. In general, negative Kappa means that there is no agreement between classified data and reference data while Kappa value equals to 1 means "perfect agreement".

The proposed method was built by using a framework called Torch[2]. This framework is more suitable due to its support to parallel programming using CUDA, a NVidia parallel programming based on graphics processing units. Thus, in this paper, Torch was used along with libraries as CUDA and CuDNN[3]. All experiments were performed on a 64 bits Intel i7 4960X machine with 3.6GHz of clock and 64GB of RAM memory. Two GPUs were used: a GeForce GTX770 with 4GB of internal memory and a GeForce GTX Titan X with 12GB of memory, both under a 7.5 CUDA version. Ubuntu version 14.04.3 LTS was used as operating system.

The ConvNet and its parameters were adjusted by considering a full set of experiments guided by [35]. First, several ConvNets with different layers and neurons were tested. Then, the best one was selected, and a full experimental setup was performed in order to choose the best parameters, such as learning rate and weight decay. Finally, for the best parameters, we also evaluated new architectures, close to the initial one, to select the best architecture. After all the setup experiments, the best values for the learning rate, weight decay, momentum and number of epochs, for both architectures were 0.01, 0.0005, 0.9 and 20, respectively.

V. Results and Discussion

In this section, we present and discuss the obtained results of the proposed method. The proposed ConvNets (*Small* and *Large*) were tested in all datasets, but only the best result were reported for each dataset.

A. AGRICULTURE dataset

The results for the AGRICULTURE dataset are presented in Table I. Although high standard deviation, all results were verified by a fold-by-fold paired test with confidence level of 95%, presenting statistical difference between the proposed method and the baselines. In this case, the *Large ConvNet*, that takes as input context windows of 25×25, achieve better results, since coffee crops present homogeneous regions, and larger context windows may bring more useful information.

TABLE I
Results for the AGRICULTURE dataset.

	Accuracy	Kappa
SLIC+BIC+SVM-RBF	84.14±15.50	0.70±0.25
CCNN [14]	85.25±7.41	0.72±0.14
Proposed Method	89.51±4.18	0.75±0.08

[2]Torch is a scientific computing framework with wide support for machine learning algorithms available at http://torch.ch/.

[3]It is a GPU-accelerated library of primitives for deep neural networks

According to Table I, the proposed method outperforms the baselines. It is important to emphasize that the baselines that uses BIC [30] and SVM-RBF require more efforts, since features need to be extracted first to be, then, used with some machine learning technique, which is the opposite direction of the proposed network, that learns all at once. CCNN method follows the same direction of the proposed algorithm, since it uses multi-scale cascade composed of three ConvNets to perform classification of patches generating, at the end, a segmented image. Although CCNN method yield good results, the proposed method is better and faster, since it only uses one ConvNet. Furthermore, it is worth mentioning that agricultural scenes are very hard to classify since the method must differentiate among different vegetation types.

Figure 7 presents the probability maps obtained for each image of AGRICULTURE dataset. Observe that most of coffee pixels are corrected labeled in comparison with the ground truth (Figure 5).

Fig. 7. Probability maps for the AGRICULTURE dataset. Black regions represent non-coffee areas while white pixels represent coffee crops.

B. URBAN Dataset

The results for the 2014 IEEE GRSS Data Fusion Contest Dataset are presented in Table II. In this case, the *Small ConvNet*, that takes as input context windows of 7×7, achieves better results, since urban regions present more shuttered areas, and smaller context windows may bring more pure information, i.e., less noise. The proposed ConvNet outperforms both baselines. BIC and SVM-RBF is outperformed in, at least, 20% in terms of overall accuracy, and 30% in terms of kappa index, while the DFF [32] method is outperformed in, at least, 10% in terms of overall accuracy and 15% in terms of kappa index. Figure 8 presents the probability maps obtained by the proposed method for the URBAN dataset. Most pixels are corrected labeled in comparison with the ground truth (Figure 6).

TABLE II
Results for the URBAN dataset.

	Accuracy	Kappa
SLIC+BIC+SVM-RBF	72.15	0.58
DFF [32]	76.01	0.67
Proposed Method	91.27	0.88

VI. Conclusion

In this paper, we propose a new approach based on Convolutional Neural Networks to perform pixel labeling in remote sensing scenes. Experimental results show that our

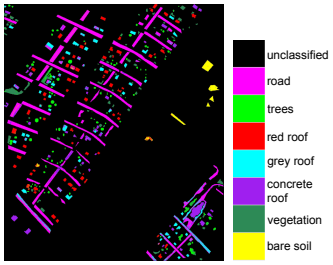


Fig. 8. Probability maps for the URBAN dataset.

method is effective and robust. We have achieved state-of-the-art performance, in terms of Kappa, for two different application datasets, outperforming all baselines. Our method also presented suitable results for aerial and remote sensing datasets, which shows the robustness of our approach.

As future work, we intend to use different post-processing methods, such as Conditional Random Fields, in order to exploit the contextual information. Also, we intend to evaluate the robustness of the proposed method by testing it in other remote sensing applications.

ACKNOWLEDGMENT

This work was partially financed by CNPq (grant 449638/2014-6), CAPES, and Fapemig (APQ-00768-14).

REFERENCES

[1] M. Volpi and V. Ferrari, "Semantic segmentation of urban scenes by learning local class interactions," in *CVPRW*, 2015, pp. 1–9.

[2] J. A. d. dos Santos, P.-H. Gosselin, S. Philipp-Foliguet, R. d. S. Torres, and A. X. Falao, "Multiscale classification of remote sensing images," *Transactions on Geoscience and Remote Sensing*, vol. 50, no. 10, pp. 3764–3775, 2012.

[3] D. Fustes, D. Cantorna, C. Dafonte, B. Arcay, A. Iglesias, and M. Manteiga, "A cloud-integrated web platform for marine monitoring using gis and remote sensing," *Future Generation Computer Systems*, vol. 34, pp. 155–160, 2014.

[4] V. Dey, Y. Zhang, and M. Zhong, *A review on image segmentation techniques with remote sensing perspective*, 2010.

[5] J. Benediktsson, J. Chanussot, and W. Moon, "Advances in very-high-resolution remote sensing [scanning the issue]," *Proceedings of the IEEE*, vol. 101, no. 3, pp. 566–569, 2013.

[6] J. A. dos Santos, O. A. B. Penatti, P.-H. Gosselin, A. X. Falcão, S. Philipp-Foliguet, and R. d. S. Torres, "Efficient and effective hierarchical feature propagation," *Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 12, pp. 4632–4643, 2014.

[7] H. Hichri, Y. Bazi, N. Alajlan, and S. Malek, "Interactive segmentation for change detection in multispectral remote-sensing images," *Geoscience and Remote Sensing Letters*, vol. 10, no. 2, pp. 298–302, 2013.

[8] Y. Tarabalka, J. C. Tilton, J. A. Benediktsson, and J. Chanussot, "A marker-based approach for the automated selection of a single segmentation from a hierarchical set of image segmentations," *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, vol. 5, no. 1, pp. 262–272, 2012.

[9] I. Goodfellow, Y. Bengio, and A. Courville, "Deep learning," 2016, book in preparation for MIT Press. [Online]. Available: http://www.deeplearningbook.org

[10] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[12] O. Penatti, K. Nogueira, and J. dos Santos, "Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?" in *CVPRW*, 2015, pp. 44–51.

[13] K. Nogueira, W. O. Miranda, and J. A. Dos Santos, "Improving spatial feature representation from aerial scenes by using convolutional networks," in *SIBGRAPI*, 2015, pp. 289–296.

[14] K. Nogueira, W. R. Schwartz, and J. A. dos Santos, "Coffee crop recognition using multi-scale convolutional neural networks," in *CIARP*, 2015, pp. 67–74.

[15] M. Fingas and C. Brown, "Review of oil spill remote sensing," *Marine pollution bulletin*, vol. 83, no. 1, pp. 9–23, 2014.

[16] M. Xie, N. Jean, M. Burke, D. Lobell, and S. Ermon, "Transfer learning from deep features for remote sensing and poverty mapping," *arXiv preprint arXiv:1510.00098*, 2015.

[17] A. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: an astounding baseline for recognition," in *CVPRW*, 2014, pp. 806–813.

[18] D. Marmanis, M. Datcu, T. Esch, and U. Stilla, "Deep learning earth observation classification using imagenet pretrained networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 1, pp. 105–109, 2016.

[19] K. Nogueira, O. A. Penatti, and J. A. d. Santos, "Towards better exploiting convolutional neural networks for remote sensing scene classification," *arXiv preprint arXiv:1602.01517*, 2016.

[20] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*, 2014, pp. 580–587.

[21] O. Firat, G. Can, and F. T. Y. Vural, "Representation learning for contextual object and region detection in remote sensing," in *ICPR*. IEEE, 2014, pp. 3708–3713.

[22] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu, "Deep learning-based classification of hyperspectral data," *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, vol. 7, no. 6, pp. 2094–2107, 2014.

[23] F. Zhang, B. Du, and L. Zhang, "Saliency-guided unsupervised feature learning for scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 4, pp. 2175–2184, 2015.

[24] S. Paisitkriangkrai, J. Sherrah, P. Janney, and A. Hengel, "Effective semantic pixel labelling with convolutional networks and conditional random fields," in *CVPRW*, 2015, pp. 36–43.

[25] S. Wager, S. Wang, and P. Liang, "Dropout training as adaptive regularization," in *Advances in Neural Information Processing Systems*, 2013, pp. 351–359.

[26] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *ICML*, 2010, pp. 807–814.

[27] T. Blaschke, "Object based image analysis for remote sensing," *ISPRS journal of photogrammetry and remote sensing*, vol. 65, no. 1, pp. 2–16, 2010.

[28] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *TPAMI*, vol. 34, no. 11, pp. 2274–2282, 2012.

[29] J. E. Vargas, P. T. Saito, A. X. Falcao, P. J. de Rezende, and J. A. dos Santos, "Superpixel-based interactive classification of very high resolution images," in *SIBGRAPI*, 2014, pp. 173–179.

[30] R. de O. Stehling, M. A. Nascimento, and A. X. Falcao, "A compact and efficient image retrieval approach based on border/interior pixel classification," in *CIKM*, 2002, pp. 102–109.

[31] G. Mountrakis, J. Im, and C. Ogole, "Support vector machines in remote sensing: A review," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 66, no. 3, pp. 247–259, 2011.

[32] F. A. Faria, J. A. Dos Santos, A. Rocha, and R. d. S. Torres, "A framework for selection and fusion of pattern classifiers in multimedia recognition," *Pattern Recognition Letters*, vol. 39, pp. 52–64, 2014.

[33] E. F. Andrade, A. A. de Araújo, and J. A. dos Santos, "A multiclass approach for land-cover mapping by using multiple data sensors," in *CIARP*, 2015, pp. 59–66.

[34] R. G. Congalton, "A review of assessing the accuracy of classifications of remotely sensed data," *Remote sensing of environment*, vol. 37, no. 1, pp. 35–46, 1991.

[35] Y. Bengio, "Practical recommendations for gradient-based training of deep architectures," in *Neural Networks: Tricks of the Trade*. Springer, 2012, pp. 437–478.