Hierarchical Probabilistic Graphical Models and Deep Convolutional Neural Networks for Remote Sensing Image Classification

Martina Pastorino, Gabriele Moser, Sebastiano B Serpico, Josiane Zerubia

Hierarchical Probabilistic Graphical Models and Deep Convolutional Neural Networks for Remote Sensing Image Classification

Martina Pastorino
DITEN department
University of Genoa, Italy
Inria, Université Côte d'Azur
Sophia-Antipolis, France
email: martina.pastorino@edu.unige.it

Gabriele Moser, Sebastiano B. Serpico
DITEN department
University of Genoa, Italy
email: gabriele.moser@unige.it

Josiane Zerubia
Inria, Université Côte d'Azur
Sophia-Antipolis, France
email: josiane.zerubia@inria.fr

***Abstract*—The method presented in this paper for semantic segmentation of multiresolution remote sensing images involves convolutional neural networks (CNNs), in particular fully convolutional networks (FCNs), and hierarchical probabilistic graphical models (PGMs). These approaches are combined to overcome the limitations in classification accuracy of CNNs for small or non-exhaustive ground truth (GT) datasets. Hierarchical PGMs, e.g., hierarchical Markov random fields (MRFs), are structured output learning models that exploit information contained at different image scales. This perfectly matches the intrinsically multiscale behavior of the processes of a CNN (e.g., pooling layers). The framework consists of a hierarchical MRF on a quadtree and a planar Markov model on each layer, modeling the interactions among pixels and accounting for both the multiscale and the spatial-contextual information. The marginal posterior mode criterion is used for inference. The adopted FCN is the U-Net and the experimental validation is conducted on the ISPRS 2D Semantic Labeling Challenge Vaihingen dataset, with some modifications to approach the case of scarce GTs and to assess the classification accuracy of the proposed technique. The proposed framework attains a higher recall compared to the considered FCNs, progressively more relevant as the training set is further from the ideal case of exhaustive GTs.**

***Index Terms*—Remote sensing, semantic segmentation, multiresolution, hierarchical PGM, CNN.**

I. Introduction

Semantic labeling of remote sensing images aims at assigning each pixel in an image to a semantic class, typically related to land cover or land use. A challenging problem in this field is the possibility to exploit information from multimodal data, (e.g., multiview, multiscale, and multiresolution information) [1]. Techniques based on deep learning reach high performances, with very high per-pixel accuracies and a good reproduction of the shapes of the object segmented [2], [3]. Fully convolutional networks (FCNs) [4], for example the U-Net [5], are currently the most popular models. Their encoder-decoder architecture allows to perform pixelwise image classification very efficiently [6]. This is because the upper layers of such models can capture shape statistics and inject them in the output maps [7]. However, to correctly model those statistics, neural networks require big datasets with exhaustive ground truths, often available in benchmark datasets and rarely for real-world mapping applications [7]. Furthermore, their generation requires a lot of work and is time consuming and expensive.

On one hand, models trained with small datasets and sparse ground truths usually produce outputs with poor geometrical fidelity [7]. On the other hand, structured output learning models, such as probabilistic graphical models (PGMs) [8] are able to exploit spatial information contained in the images. Markov models defined on planar or multilayer graphs [9], are examples of PGMs. For most categories of Markov random fields (MRFs), Markovianity is formulated with respect to a neighborhood of each node (which corresponds to a pixel) of the related graph. This determines a difference with respect to the well-known Markov chains for one dimensional data analysis, for which Markovianity is expressed with respect to the past of each pixel, which implies a causal behavior. This causality property, when satisfied, is computationally advantageous because it makes it possible to formulate efficient inference procedures. Markov models for two-dimensional image analysis such as hierarchical MRFs on quadtrees [10], [11] are proven to be causal and able to capture relations among pixels located in images at different resolutions through the use of a Markov chain. But this model does not capture spatial contextual information among the pixels inside each lattice [10], which could be interesting to characterize when dealing with neural networks trained with sparse ground truths, thus having problems in segmenting object shapes. In recent approaches [12], this hierarchical model is combined with a Markov chain on each of its layers, postulating Markovianity with respect to the neighborhood of pixels defined by a scanning trajectory.

The use of multiresolution information is proven to favor accuracy and spatial precision for the task of image classification thanks to the robustness to noise and outliers of the coarser

University of Genoa and Université Côte d'Azur (UCA) are part of the Ulysseus Alliance (European University). https://ulysseus.eu/

resolutions and the spatial details of the finer ones [1]. In this particular application, multiresolution information is brought by the intermediate layers of an FCN [13], which involve several multiscale processing stages, through convolutions and pooling operations. This structure matches perfectly the topology of a hierarchical Markov model built on quadtrees [8], [11].

The contribution of this paper is twofold: firstly, a new method involving hierarchical PGMs and FCNs is developed to perform semantic segmentation of multiresolution remote sensing images by exploiting the intrinsically hierarchical behavior of CNNs; secondly, it is shown that structured output learning methods, when combined with neural networks, can guarantee improvements in regularity and shape segmentation in the case of sparse ground truths, where spatial class boundaries may not be present.

The first step involves the implementation of the U-Net. Its activations at different resolutions are inserted in a quadtree with the multispectral channels of the original image, to develop the hierarchical Markov model. Multiresolution and spatial information is kept into account by Markov chains formulated across the levels of the quadtree and within each layer. This joint strategy benefits from the spatial information within each layer, useful for semantic segmentation of remote sensing images, and from the multiscale information carried by the activations of the network at different resolutions. The model is combined with decision tree ensembles, such as random forest (RF) [14], to estimate the pixelwise posterior probabilities necessary for the inference on the PGM, which is accomplished through the marginal posterior mode (MPM) criterion. The architecture is shown in Fig. 1.

II. Methodology

A. Hierarchical Markov model

The proposed PGM consists of a hierarchical MRF which models multiresolution information across the different layers of the quadtree through a Markov chain, combined with a planar Markov model based on a Markov chain with respect to a 1D scan of the pixel lattice, which models contextual spatial information. Consider $\{S^0, S^1, \ldots, S^L\}$, $S \subset \mathbb{Z}^2$, as a set of pixel lattices organized as a quadtree, where each pixel $s \in S^\ell$ has a parent site $s^- \in S^{\ell-1}$ and four children sites $s^+ \subset S^{\ell+1}$ ($\ell = 0, 1, \ldots, L$), with the exception of the leaf layer, not having any children site, and the root layer, not having a parent site.

Each pixel $s \in S$ is associated with a discrete class label x_s in a finite set Ω of M classes ($x_s \in \Omega$, $s \in S$), thus $\mathcal{X} = \{x_s\}_{s \in S}$ is a hierarchical MRF if [9], [10]

$$P(\mathcal{X}^\ell | \mathcal{X}^{\ell-1}, \mathcal{X}^{\ell-2}, \ldots, \mathcal{X}^0) = P(\mathcal{X}^\ell | \mathcal{X}^{\ell-1}) \tag{1}$$

where $\mathcal{X}^\ell = \{x_s\}_{s \in S^\ell}$, ($\ell = 1, 2, \ldots, L$), and Markovianity holds across the scales. In this hierarchical model, the transition probabilities are also assumed to factorize [10], so the result is

$$P(\mathcal{X}^\ell | \mathcal{X}^{\ell-1}) = \prod_{s \in S^\ell} P(x_s | x_{s^-}). \tag{2}$$

To model spatial information in each pixel lattice, consider a rectangular lattice R and an order relation $\prec$ which defines a neighborhood in R introducing the concept of "past". The pixels $r \in R$ respecting the relation $r \prec s$ are the causal past neighbors of pixel $s \in R$. The relation $r \precsim s$ indicates that r is a past neighbor of s. Formally, $\{r \in R : r \precsim s\} \subsetneq \{r \in R : r \prec s\}$, which means that the past neighbors of $s \in R$ are included in the past of s but constitute a strict subset of its entire past. The Markovianity constrained to the past of each pixel holds for $\mathcal{X}$ if and only if [15]–[17]:

$$P(x_s | x_r, r \prec s) = P(x_s | x_r, r \precsim s). \tag{3}$$

In the proposed approach, the total order relation is defined by a pixel visiting scheme involving the combination of four zig-zag scans and two Hilbert curve scans. More details on the pixel scan trajectory can be found in [11], [12]. Equations (1)-(3) are assumed to hold, thus defining a framework modeling the inter-layer and intra-layer dependencies between pixels. Conditional independence is another important concept as it can be used to decompose complex probability distributions into a product of factors, each consisting of the subset of corresponding random variables. Consider $\mathcal{Y} = \{y_s\}_{s \in S}$ as the random field of the observations associated with all the pixels in the quadtree, the observation model $P(\mathcal{Y}|\mathcal{X})$ is defined by a pixelwise factorization

$$P(\mathcal{Y}|\mathcal{X}) = \prod_{s \in S} P(y_s | x_s) = \prod_{\ell=0}^{L} \prod_{s \in S^\ell} P(y_s | x_s). \tag{4}$$

B. Fully convolutional network

FCNs are neural networks that do not contain any dense layer [13], thus having the possibility to obtain an output with the same size of the input. A very popular example of FCN is the U-Net [5], which uses a bottleneck architecture, with pooling and un-pooling layers performing downsampling and upsampling processes, respectively, in order to achieve a pixelwise semantic segmentation of the original input image, with the same resolution. The multiscale formulation of a hierarchical MRF on a quadtree with pixel lattices with a power-of-2 relationship between layers directly matches the resolutions obtained in the intermediate layers of a convolutional network through pooling layers of size 2×2. The architecture is made up by 5 convolutional blocks, each with a double convolutional layer with kernel of size 3×3 and zero padding of dimension 1, followed by ReLU activation, batch normalization and pooling layers. At each downsampling step, the number of filters is doubled. A pixelwise softmax [13] combined with the cross-entropy loss function [18] is employed over the final feature map.

The activations of the network at three different resolutions are inserted in the quadtree through three skip connections, to be connected to the hierarchical Markov model in order to exploit the multiscale information.

Any kind of FCN can be combined with the PGM, thus guaranteeing the flexibility of the proposed model. Other

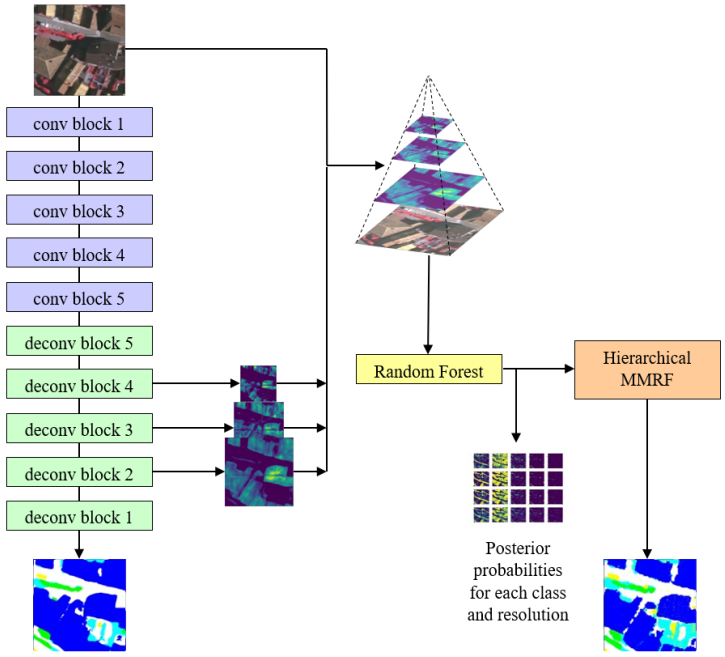

Fig. 1. Overall architecture of the proposed method.

architectures such as SegNet [19] or DeepLabV3+ [20] were considered, but the results obtained were comparable to the ones of U-Net, which was therefore chosen for the experimentation.

C. Inference with the MPM criterion

Since the proposed framework is causal both across the scales and within each scale, it is possible to use the MPM criterion, which is recursive when formulated on quadtrees [10]. This criterion assigns each pixel $s \in S$ the class label x_s that maximizes $P(x_s|\mathcal{Y})$ [9] through the following steps

$$P(x_s) = \sum_{x_{s^-} \in \Omega} P(x_s|x_{s^-})P(x_{s^-}), \tag{5}$$

$$P(x_s|y_s^d) \propto P(x_s|y_s) \prod_{t \in s^+} \sum_{x_t \in \Omega} \frac{P(x_t|y_t^d)P(x_t|x_s)}{P(x_t)}, \tag{6}$$

$$P(x_s|x_s^c, y_s^d) \propto \frac{P(x_s|y_s^d)P(x_s|x_{s^-})P(x_{s^-})}{P(x_s)^{n_s}} \cdot \prod_{r \precsim s} P(x_s|x_r)P(x_r), \tag{7}$$

$$P(x_s|\mathcal{Y}) = \sum_{x_s^c} P(x_s|x_s^c, y_s^d)P(x_{s^-}|\mathcal{Y}) \prod_{r \precsim s} P(x_r|\mathcal{Y}), \tag{8}$$

with y_s^d containing the observations of all descendants of s in the tree (including s itself), x_s^c the labels of all pixels connected to s (i.e., x_{s^-} and $\{x_r\}_{r \precsim s}$), and n_s being the number of such pixels. Details regarding the recursive steps are shown in [11], [12]. Equation (5) is an application of the total probability theorem, and the proof of (6) is in [10] for a hierarchical MRF. Finally, (7) and (8) hold with the following conditional independence assumptions:

$$\text{A1}: \quad P(x_s|x_s^c, \mathcal{Y}) = P(x_s|x_s^c, y_s^d) \tag{9}$$

$$\text{A2}: \quad P(x_s^c|\mathcal{Y}) = P(x_{s^-}|\mathcal{Y}) \prod_{r \precsim s} P(x_r|\mathcal{Y}) \tag{10}$$

$$\text{A3}: \quad P(x_s^c|x_s, y_s^d) = P(x_s^c|x_s) = P(x_{s^-}|x_s) \prod_{r \precsim s} P(x_r|x_s) \tag{11}$$

More details can be found in [11], [12]. The first assumption (A1) implies that, given the parent and sibling labels, the label of s only depends on the observation of its descendants. Given the observation field, the parent and the sibling labels of s are conditionally independent, and the parent and sibling labels of s, when conditioned to the label of s, are independent on the observations of the descendants of s and mutually independent, according to assumptions A2 and A3, respectively. The pixelwise posteriors $P(x_s|y_s)$ in (6) represent the observations in the recursion and are estimated by the RF [14] classifier.

III. Experimental Results

All the experiments described in this section were run on a graphics processing unit (GPU) from Google Colab.

The method was tested on the Vaihingen image set of the ISPRS 2D Semantic Labeling Challenge[1] which consists of very high-resolution (VHR) aerial images over the city of Vaihingen in Germany and was provided by the German Society for Photogrammetry, Remote Sensing, and Geoinformation

[1]https://www2.isprs.org/commissions/comm2/wg4/benchmark/2d-sem-label-vaihingen/

(DGPF). The resolution is 9 cm/pixel and the classes are six: buildings, impervious surfaces (e.g., roads), low vegetation, tree, car, and clutter. The last class comprehends all the instances not belonging to the other classes, thus being highly heterogeneous and of relatively limited interest. For this reason it was excluded from the experimentation.

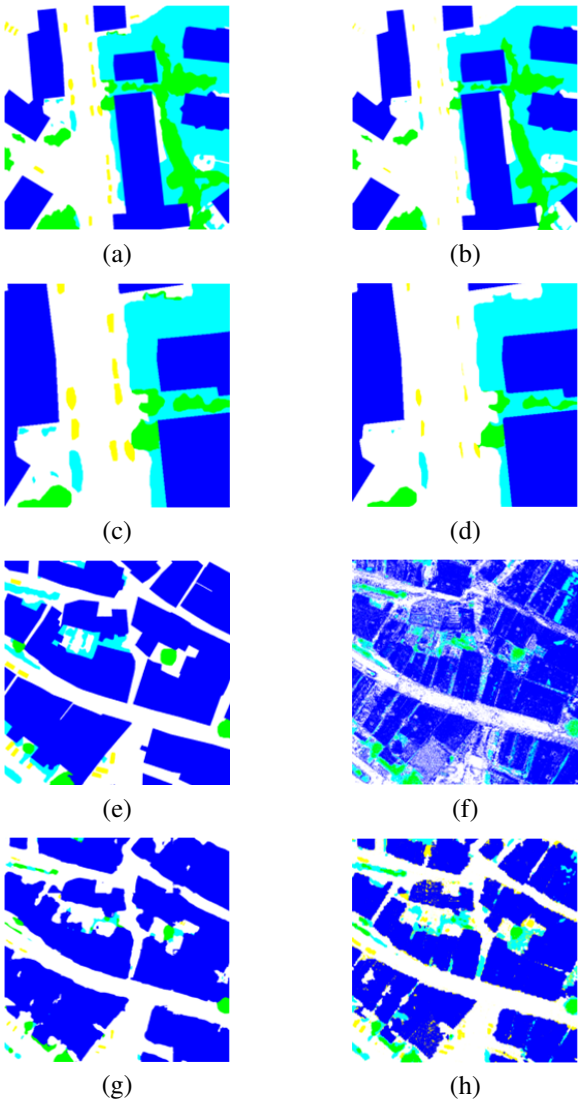

Fig. 2. **Ground truth and classification maps:** (a) original training set, (b) eroded training set, (c) detail of the original training set, (d) detail of the eroded training set, (e) test set; classification maps: (f) RF, (g) U-Net, and (h) the proposed method ("Ver. 3"). Classes: buildings, impervious, vegetation, tree, car.

Tiles consist of near-InfraRed-Red-Green (IRRG) images and digital surface model (DSM) data extracted from a LiDAR point cloud. Within these 16 images, 12 have been chosen for training and 4 for testing the network. The accuracy metrics examined in the following are: overall accuracy (OA), precision, recall [22], Cohen's Kappa coefficient [23], and F1 score.

The proposed method depends on three parameters: the across-scale transition probability ϑ, the spatial transition probability ψ (both with values in the range $[0, 1]$, defined by the Bouman's model [12], [24]), and the number of resolutions L. ϑ and ψ were fixed to 0.82, since preliminary experiments (omitted for brevity) suggested that higher or lower values attained worst results, and $L = 4$.

In order to test the pipeline of the proposed approach and compare its results with the ones of an FCN used by itself, several training conditions were considered. The dataset chosen is an ideal one, with exhaustive pixelwise ground truths. Some modifications were made to the training dataset, involving morphological operators or a percentage of unlabeled pixels, to approximate the ground truths available for real applications.

The posterior probabilities input to the hierarchical PGM are estimated by an RF classifier that works on the activations of the network. In the case of the finest resolution, i.e., the pixel lattice at the base of the quadtree, four variants of the proposed method have been formulated and experimented upon. In the first option, RF is used to predict pixelwise posteriors from the IRRG channels of the input image directly (called "Ver. 1" in the following). However, the posterior probabilities obtained by RF for class 5, "car", at resolution 0, do not appear to be defined well enough to provide an appropriate estimation of the instances of class 5 in the resulting images. This is consistent with the difficulty to discriminate the object class "car" using a purely pixelwise classifier, such as RF, fed with only three visible channels. As a variant, the posterior probabilities of resolution 0, estimated by the RF, are substituted with the ones obtained in the output layer of the network ("Ver. 2"). The two other approaches only focused on the posterior probabilities of class "car", either substituting the RF estimation with the network posteriors ("Ver. 3"), or with a nearest neighbour resampling of the same class in the lattice above in the quadtree ("Ver. 4"). The objective of these two formulations is to focus on the discrimination of the minority classes (e.g., "tree" and "car").

The quantitative results shown in Table I confirm that the proposed approach obtains higher accuracies for the aforementioned minority classes, and these improvements are more remarkable the more the input data approach the sparsely annotated datasets available for land-cover mapping applications. For example, with a 70% of unlabeled pixels there is an overall improvement in accuracy for the classification, especially noticeable for the class "vegetation", with an increase of 50% with respect to the results obtained with the simple U-Net. In all the considered situations, the recall of the proposed framework is higher than the one of the standard FCNs. In Fig. 2, the comparison between the images obtained with the network (Fig. 2(g)) and with the proposed method (Fig. 2(h)) clearly shows that the full proposed pipeline does a better job in the semantic segmentation of the edges between classes. Furthermore, the representation of small classes, such as "car", appears to have improved as well.

The proposed method was compared to the recently proposed "FESTA" method, where a network is trained with

TABLE I

RESULTS OF THE PROPOSED METHOD APPLIED ON THE TEST SET WITH THE FOUR DIFFERENT CONFIGURATIONS AND COMPARED TO OTHER FCNS.

Full dataset	**building**	**impervious**	**vegetation**	**tree**	**car**	**overall acc.**	**recall**	**precision**	**Cohen's κ**	**F1 score**
Standard U-Net	**0.92**	0.83	0.71	**0.92**	0.74	**0.85**	**0.83**	**0.84**	**0.79**	**0.83**
Standard SegNet	0.90	0.73	**0.73**	**0.92**	0.67	0.83	0.79	0.81	0.76	0.80
DeepLabV3+ [20]	0.91	**0.87**	0.66	0.90	0.80	0.84	**0.83**	0.82	0.78	0.82
Proposed method, "Ver. 1" (U-Net)	0.85	0.82	0.69	**0.92**	0.38	0.81	0.73	0.78	0.74	0.75
Proposed method, "Ver. 2" (U-Net)	0.87	0.84	0.71	**0.92**	**0.89**	0.82	0.81	0.78	0.75	0.80
Proposed method, "Ver. 3" (U-Net)	0.84	0.81	0.68	**0.92**	0.86	0.81	0.82	0.72	0.75	0.77
Proposed method, "Ver. 4" (U-Net)	0.84	0.82	0.69	**0.92**	0.88	0.81	**0.83**	0.77	0.75	0.80
70% of unlabeled pixels	**building**	**impervious**	**vegetation**	**tree**	**car**	**overall acc.**	**recall**	**precision**	**Cohen's κ**	**F1 score**
Standard U-Net	**0.83**	**0.92**	0.55	0.87	0.80	0.80	0.79	0.74	0.72	0.76
Proposed method, "Ver. 1" (U-Net)	0.80	0.83	0.68	0.89	0.36	0.79	0.71	0.70	0.72	0.70
Proposed method, "Ver. 2" (U-Net)	0.82	0.85	**0.76**	0.86	0.96	**0.82**	**0.85**	0.72	**0.76**	0.78
Proposed method, "Ver. 3" (U-Net)	0.72	0.82	0.67	0.88	**0.99**	0.76	0.82	0.68	0.69	0.74
Proposed method, "Ver. 4" (U-Net)	0.78	0.83	0.68	0.89	0.90	0.79	0.81	0.71	0.71	0.76
FESTA [21]	0.80	0.91	0.67	**0.91**	0.63	**0.82**	0.77	**0.82**	0.73	**0.79**

a "scribbled" GT and a loss term that favors regularization in the spatial and feature domains [21]. The results show that both approaches mitigated the effects of sparse GT, although the proposed method obtained higher or the same per-class accuracy (especially for "cars") and required shorter computation times.

IV. CONCLUSION

In this paper, a new framework to tackle semantic segmentation of remote sensing images based on FCNs, hierarchical Markov models, and random forest has been proposed. The results are interesting: our approach surpasses in recall U-Net, thus suggesting the capability of this approach to exploit spatial information through the hierarchical Markov model and mitigating the limitations of the neural networks trained with a small dataset. The proposed pipeline outperforms the state-of-the-art especially in the classification of minority classes.

Perspectives for future work involve the substitution of the random forest classifier with feed-forward neural networks to compute the pixelwise posterior probabilities. Moreover, the proposed method could be tested with various datasets related to real-world applications, such as disaster management, with different complexity and features, so to further investigate its generalization capabilities.

REFERENCES

[1] L. Gómez-Chova, D. Tuia, G. Moser, and G. Camps-Valls, "Multimodal classification of remote sensing images: a review and future directions," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1560–1584, 2015.

[2] K. Nogueira, O. Penatti, and J. dos Santos, "Towards better exploiting convolutional neural networks for remote sensing scene classification," *Pattern Recognition*, vol. 61, pp. 539–556, 2017.

[3] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez, "A review on deep learning techniques applied to semantic segmentation," *ArXiv:1704.06857*, 2017.

[4] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440, 2015.

[5] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-assisted Intervention*, ser. LNCS, vol. 9351. Springer, pp. 234–241, 2015.

[6] N. Audebert, B. Saux, and S. Lefèvre, "Semantic segmentation of earth observation data using multimodal and multi-scale deep networks," in *Proceedings of the Asian Conference on Computer Vision*, pp. 180–196, 2016.

[7] L. Maggiolo, D. Marcos, G. Moser, and D. Tuia, "Improving maps from CNNs trained with sparse, scribbled ground truths using fully connected CRFs," in *IGARSS 2018 - IEEE International Geoscience and Remote Sensing Symposium*, pp. 2099–2102, 2018.

[8] Z. Kato and J. Zerubia, "Markov random fields in image segmentation," *Foundations and Trends in Signal Processing*, vol. 5, no. 1-2, pp. 1–155, 2012.

[9] S. Li, *Markov random field modeling in image analysis*, 3rd ed. Springer, 2009.

[10] J. Laferté, P. Pérez, and F. Heitz, "Discrete Markov image modeling and inference on the quadtree," *IEEE Transactions on Image Processing*, vol. 9, no. 3, pp. 390–404, 2000.

[11] I. Hedhli, G. Moser, S. B. Serpico, and J. Zerubia, "Classification of multisensor and multiresolution remote sensing images through hierarchical Markov random fields," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 12, pp. 2448–2452, 2017.

[12] M. Pastorino, A. Montaldo, L. Fronda, I. Hedhli, G. Moser, S. B. Serpico, and J. Zerubia, "Multisensor and multiresolution remote sensing image classification through a causal hierarchical markov framework and decision tree ensembles," *Remote Sensing*, vol. 13, no. 5, 2021.

[13] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. Boston, Massachussetts: USA: MIT Press, 2016.

[14] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[15] K. Abend, T. J. Harley, and L. N. Kanal, "Classification of binary random patterns," *IEEE Transactions on Information Theory*, vol. 11, no. 4, pp. 538–544, 1965.

[16] P. A. Devijver, "Hidden Markov mesh random field models in image analysis," *Journal of Applied Statistics*, vol. 20, no. 5-6, pp. 187–227, 1993.

[17] A. S. Willsky, "Multiresolution Markov models for signal and image processing," *Proceedings of the IEEE*, vol. 90, no. 8, pp. 1396–1458, 2002.

[18] G. Nasr, E. Badr, and C. Joun, "Cross entropy error function in neural networks: forecasting gasoline demand," in *Proceedings of the Fifteenth International Florida Artificial Intelligence Research Society Conference*, Pensacola Beach, Florida, USA, pp. 381–384, 2002.

[19] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: a deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[20] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, pp. 833–851, 2018.

[21] Y. Hua, D. Marcos, L. Mou, X. X. Zhu, and D. Tuia, "Semantic segmentation of remote sensing images with sparse annotations," *IEEE Geosci. Remote Sens. Letters*, pp. 1–5, 2021.

[22] J. A. Richards, *Remote sensing digital image analysis: an introduction*, 5th ed. Springer, 2013.

[23] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, pp. 37–46, 1960.

[24] C. A. Bouman and M. Shapiro, "A multiscale random field model for Bayesian image segmentation," *IEEE Transactions on Image Processing*, vol. 3, no. 2, pp. 162–177, 1994.