Semantic Segmentation of Remote Sensing Images through Fully Convolutional Neural Networks and Hierarchical Probabilistic Graphical Models

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

▶ **To cite this version:**

Martina Pastorino, Gabriele Moser, Sebastiano B Serpico, Josiane Zerubia. Semantic Segmentation of Remote Sensing Images through Fully Convolutional Neural Networks and Hierarchical Probabilistic Graphical Models. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60, pp.1-16. 10.1109/TGRS.2022.3141996 . hal-03534026

HAL Id: hal-03534026

https://inria.hal.science/hal-03534026

Submitted on 19 Jan 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Semantic Segmentation of Remote Sensing Images through Fully Convolutional Neural Networks and Hierarchical Probabilistic Graphical Models

Martina Pastorino, *Student Member, IEEE,* Gabriele Moser, *Senior Member, IEEE,*
Sebastiano B. Serpico, *Fellow, IEEE,* and Josiane Zerubia, *Fellow, IEEE*

***Abstract*—Deep learning is currently the dominant approach to image classification and segmentation, but the performances of deep learning methods are remarkably influenced by the quantity and quality of the ground truth (GT) used for training. In this paper, a deep learning method is presented to deal with the semantic segmentation of very high resolution (VHR) remote sensing data in the case of scarce GT. The main idea is to combine a specific type of deep convolutional neural networks (CNNs), namely, fully convolutional networks (FCNs), with probabilistic graphical models (PGMs). Our method takes advantage of the intrinsic multiscale behavior of FCNs to deal with multiscale data representations and to connect them to a hierarchical Markov model (e.g., making use of a quadtree). As a consequence, the spatial information present in the data is better exploited, allowing a reduced sensitivity to GT incompleteness to be obtained. The marginal posterior mode criterion is used for inference in the proposed framework. To assess the capabilities of the proposed method, the experimental validation is conducted with the ISPRS 2D Semantic Labeling Challenge datasets on the cities of Vaihingen and Potsdam, with some modifications to simulate the spatially sparse GTs that are common in real remote sensing applications. The results are quite significant, as the proposed approach exhibits a higher producer accuracy than the standard FCNs considered and especially mitigates the impact of scarce GTs on minority classes and small spatial details.**

***Index Terms*—remote sensing, semantic segmentation, probabilistic graphical model (PGM), hierarchical Markov model, convolutional neural network (CNN), fully convolutional network (FCN), multiscale analysis.**

I. Introduction

SPACE missions currently allow VHR satellite imagery to reach spatial resolutions as fine as 30 cm [1]. Acquisitions from airborne platforms (e.g., airplanes or unmanned aerial vehicles – UAVs) can reach spatial resolutions of even a few centimeters. The data acquired can be optical (e.g., panchromatic, multispectral, and hyperspectral images) or radar, with different synthetic aperture modalities (e.g., stripmap, spotlight, ScanSAR) with various trade-offs between resolution and coverage [2]. This offers great application potential in the field of remote sensing. In particular, image classification techniques in this sector can be used for land cover mapping applications in areas such as urban planning, precision agriculture, inventory and monitoring of forest species, identifying and discriminating between the categories of pixels or objects. An important role in this context is played by dense supervised classification – or semantic segmentation – of remote sensing images, whose purpose is to assign each pixel in an image to a semantic class, typically related to land cover or land use. The development of classifiers that can benefit from the detailed spatial information conveyed by the above input imagery constitutes, however, a major challenge [3], [4].

Currently, deep learning (DL) techniques are the dominant methods for image classification and segmentation [5], as they can reach very high accuracies and even faithfully reproduce the shapes of the segmented objects, and are employed in remote sensing applications, too [6], [7]. Notable architectures are fully convolutional networks (FCNs) [8], e.g., U-Net [9], SegNet [10], and HRNet [11], which have been proven to exhibit outstanding performances [12], [13]. This is because the upper layers of such models can capture shape statistics and inject them in the output maps [14]. However, to correctly model those statistics, a DL architecture requires a large dataset with densely labeled ground truths that accurately represent objects with their boundaries. These exhaustive ground truths, whose production is very time consuming, are typical of benchmark datasets: dense pixel level ground truths are rarely available in real-world mapping applications [14], [15], [16]. Real-world ground truths are typically spatially sparse and do not represent the spatial borders among the classes. This is known to significantly affect the accuracy of the resulting maps and is a major challenge in the development of deep networks for remote sensing and in taking advantage of their potential. The proposed work addresses the challenge of semantic segmentation [17], [18] of VHR images based on deep learning methods in the case of realistic scarce ground truths by introducing a novel combination of hierarchical probabilistic graphical models (PGMs) [19], [20] and deep neural networks.

Thanks to the ever-growing availability of VHR images, as mentioned above, interest in structured output learning and PGMs has intensified in the past few years. PGMs, in particular Bayesian networks and random fields, are structured prediction models that use graph-based representations to express the conditional dependence structure between random variables

M. Pastorino, G. Moser, and S. B. Serpico are with the Department of Electrical, Electronic, Telecommunications Engineering and Naval Architecture, University of Genoa, 16145 Genoa, Italy (e-mail: gabriele.moser@unige.it, martina.pastorino@edu.unige.it).

M. Pastorino and J. Zerubia are with the Ayana Research Group, Institut National de Recherche en Informatique et en Automatique (INRIA), Université Côte d'Azur (UCA), 06902 Sophia Antipolis, France.

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

over a multidimensional space. Within the family of structured output learning methods, Markov models on planar or multilayer graphs are flexible and powerful stochastic models to incorporate spatial and possibly multiscale/multiresolution information [21]. For most categories of Markov random fields (MRFs), Markovianity is formulated with respect to a neighborhood of each node of the related graph.

Examples of causal Markov models with feasible and efficient inference algorithms for 2-dimensional image processing are the Markov mesh random fields (MMRFs) on planar lattices [22] and the hierarchical MRFs on quadtrees [23], [24]. These two architectures have complementary properties: an MMRF describes spatial interactions among the pixels but is a single scale model, while a hierarchical MRF captures relations among sites located at different scales through the use of a Markov chain, but does not explicitly characterize spatial dependencies within the layer at each scale [23]. These two models have been combined in recent approaches [25], ensuring Markovianity across the scales of a quadtree (interlayers) and with respect to the neighborhood system of pixels associated with each layer of the tree (intralayer).

The processing operations executed by a convolutional neural network (CNN) [26] involve several multiscale processing stages, both through convolutions with given window sizes and through pooling layers. These processes naturally match the structure of multiscale graph topologies on which probabilistic models can be efficiently and effectively formulated [19], [24].

The objective of the present work is to develop a remote sensing image semantic segmentation method that, by leveraging the combination of deep learning techniques and PGMs, is able to obtain good results even in the case of a poor training dataset, thus expanding the applicability of CNN-based approaches to situations in which ground truth is scarce – a common scenario in many land cover mapping applications of remote sensing.

The contribution of this paper is the formalization of a new method in which, first of all, an FCN is trained and applied to the input imagery; then, the activations of the network at different scales are used to fill the layers of a quadtree, to develop the hierarchical Markov model. Markov chains are formulated both across the scales of the quadtree and with respect to a 1D scan of the pixel lattice of each layer. This joint strategy benefits from both the spatial information within each layer and from the multiscale information carried by the activations of the network at different scales.

Decision tree ensembles [27] are employed to compute the pixelwise posterior probabilities necessary for the inference on the PGM, which is accomplished through the marginal posterior mode (MPM) criterion. This criterion is especially advantageous for classification and segmentation methods associated with multiresolution/multiscale hierarchical Markov models [23]. The integration of these three methodological components – FCN, hierarchical PGM, and tree ensemble – allows exploiting the representations extracted by the FCN across all its layers within the final pixelwise classification process, thus incorporating prior information on the spatial behavior and the structure of the prediction output. In the proposed method, this is aimed at mitigating the limitations of the FCN in the learning of spatial relations from non-exhaustive training maps, in which spatial class boundaries may not be present or may be poorly represented.

The paper is organized as follows. Section II provides an overview of the state of the art on PGMs, deep learning models, and their combination for remote sensing applications. Section III is devoted to the presentation of the proposed methodology. The results of the experiments conducted with the proposed method and the comparison with the results obtained with some standard FCNs and previous approaches are presented and discussed in Section IV. Finally, conclusions and perspectives of the proposed technique are reported in Section V.

II. Previous Work

In this section, some of the previous approaches to remote sensing image classification that benefit from methodological concepts drawn from both MRFs – or PGMs at large –, deep neural networks, and their combination are reviewed.

In particular, among the currently recognized approaches of great effectiveness in the supervised classification of remote sensing images, methods based on MRFs play a primary role. Different techniques based on MRFs have been applied to the problem of spatial-contextual classification of land cover, also using hierarchical [19], multiresolution, and multiscale models [28]. For example, the approaches introduced in [25] to address the problem of the joint classification of multiple images acquired on the same scene at different spatial resolutions involve the use of a probabilistic graphical approach with a hierarchical Markov mesh framework that models the spatial-contextual classification of multiresolution and possibly multisensor images. In [25], Markovianity is postulated on a quadtree, both inter-scale and intra-scale. This joint strategy benefits from the spatial information within each layer and inherently supports multiresolution fusion. The MPM criterion is used [19], instead of the common maximum a posteriori (MAP), because it is especially advantageous for classification and segmentation methods associated with multiresolution/multiscale models [23].

Additionally, numerous attempts have been made to design deep learning architectures for semantic segmentation. In this context, a special attention has been devoted to CNN architectures [20], [29], which have achieved state-of-the-art classification performance on various datasets [20], [29]. Despite their success, their design and optimal configuration often require dedicated trial-and-error procedures. In addition, these models involve a large number of parameters that are learned with the help of powerful computational resources (which may limit their practical application [30]) and require large training datasets with exhaustive ground truths, which may not be available for many remote sensing applications. Classification algorithms based on convolutional networks typically use the output of the last layer for feature extraction, since it is the most sensitive to category-level semantic information and the most invariant to "nuisance" effects. However, the information in this layer may be too coarse to allow for precise results. In contrast, the information is represented at finer scales in

earlier layers, but they do not capture semantics. To fully exploit the intrinsic multiscale representation capabilities of neural networks, a "hypercolumn" may be assigned to each pixel, i.e., a column vector made up of all the activations of the CNN units above that pixel [31].

Concerning, in particular, the combination of DL techniques (such as CNNs) and PGMs (such as MRFs or conditional random fields, CRFs), the latter are mostly used as post-processing steps [32], [33], [34], [35] to refine the classification results obtained by the neural networks. In [32], a context-guided VHR image classification method that combines semantic-free segments and CNN-based deep features is presented; it uses semantic segments to overcome possible oversegmentation and undersegmentation phenomena due to the heterogeneity of VHR images. A CRF is successively applied to capture the contextual information and refine the classification map. In [33] neural networks are first used to obtain the initial classification result, which is then improved with a higher-order co-occurrence CRF model. In [36], CNNs and handcrafted features are applied to dense image patches to produce per-pixel class probabilities postprocessed by a CRF. In [37], a DL model using a long short-term memory (LSTM) recurrent neural network [26] and a CRF in cascade-parallel form is proposed to make landslide susceptibility predictions based on remote sensing images and on a geographic information system (GIS). In [38], a CRF is applied again as a postprocessing technique to the classification results obtained by FCNs.

The idea developed in [39] is to integrate the results of CNN classification, characterised by prediction uncertainty, especially when classifying spatially and spectrally complicated VHR imagery, into a "regional decision fusion", to increase classification accuracy. For this purpose, a multilayer perceptron-based Markov random field classifier is used to provide crisp and accurate boundary delineation. The proposed MRF-CNN fusion decision strategy exploits the complementary characteristics of the two classifiers based on variable precision rough set uncertainty description and classification integration [39]. In [40], an MRF fully convolutional network (M-FCN) is proposed. M-FCN uses a cascade strategy that consists of an FCN-based coarse candidate extraction stage, a multi-MRF-based region proposal (RP) generation stage, and a final classification stage for airplane detection. After the first stage, the coarse candidate map is used as the initial labeling field for a multi-MRF algorithm, and RPs are generated according to this multi-MRF output. Finally, PGMs have been used to simulate ground truth maps for the training of a CNN in a semisupervised manner [34] and to combine probabilistic predictions obtained by different sources (for example neural networks and probabilistic classifiers) [35].

III. Methodology

A. Overview of the proposed approach

The proposed method for semantic segmentation combines an FCN, the hierarchical Markov model defined in [25], and a Random Forest (RF) classifier [41]. The network is trained with a training set made up of a set of remotely sensed images and their available densely labeled ground truth maps (see Section III-B). In real-world land cover mapping applications, these ground truth data are usually spatially non-exhaustive and most often do not provide examples of spatial class borders. The rationale of the proposed method is precisely to leverage the spatial modeling capabilities of the MRFs – and especially the multiscale modeling capability of hierarchical MRFs (see Section III-C) – to mitigate the impact of this suboptimal ground truth on the semantic segmentation results.

In the training phase of the proposed method, the activations of the network at several different blocks (i.e., at several different spatial resolutions, all of them with a $\times 2$ relationship with respect to the previous one) and the channels of the original image are used to build a training quadtree. If L is the number of considered spatial resolutions in addition to the original resolution of the input images, then this quadtree is made of $(L + 1)$ levels. The resulting quadtree topology allows the features extracted by the FCN to be naturally integrated into a hierarchical Markov model associated with the quadtree itself. In particular, the FCN is combined with the hierarchical MRF approach in [25]. The inference equations of this model are fed with pixelwise posteriors associated with all nodes in the hierarchical graph, i.e., all pixel locations in the training quadtree. The RF classifier is applied to predict these posteriors on all levels of the training quadtree. For this purpose, the labeled ground truths, which are defined on the same pixel lattice as the original image, are resized at the same scale as the upper L levels of the quadtree and inserted in a label quadtree used to train the RF classifier.

In general, the proposed approach can be combined with an arbitrary FCN model. In particular, U-Net [9] (see Section III-B) is used as the reference model, since it is widely employed and has been found to be effective in applications to remote sensing imagery. SegNet [10] and HRNet are also considered further alternatives in the experiments.

B. Deep Learning Architecture

FCNs are convolutional neural networks that do not contain any fully connected layer, also called dense layer, and, consequently, have the possibility to yield an output with the same size of the input (which can be of arbitrary size), a very useful feature in the case of semantic segmentation tasks [8].

Here, this property is used to exploit the intrinsically multiscale behavior of an FCN and merge it with a hierarchical MRF for multiscale image processing. Multiple scales allow the model to capture spatial relationships for objects of different sizes, from large arrangements of buildings to individual trees, allowing for a better analysis of the scene.

As mentioned above, the FCN used in this framework is primarily a U-Net. However it is important to emphasize that this choice is not an intrinsic limitation of the proposed approach, which can be applied in conjunction with any FCN. The U-Net architecture consists of an encoder-decoder scheme, also called contracting and expansive paths. Each convolutional block of the encoder contains a double convolutional layer with a 3×3 kernel and a zero padding of dimension 1, followed by ReLU activation and batch normalization. The

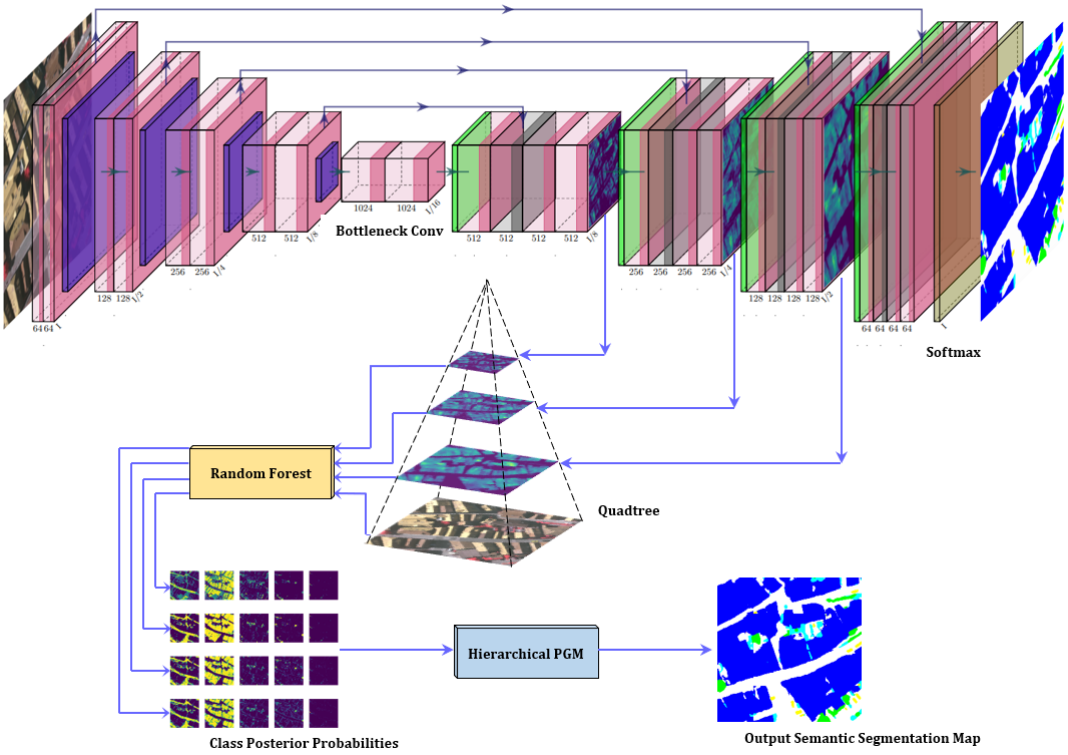


Fig. 1. Overall architecture of the proposed method.

convolutional blocks are followed by max pooling layers of size 2×2. At each downsampling step, the number of filters is doubled. The decoder, symmetrical with respect to the encoder, has the task of performing upsampling and classification. It consists of upsamplings and concatenations followed by regular convolution operations.

Three skip connections starting from the first three deconvolution blocks of the decoder allow to collect the activations of the network at different resolutions. These activations are inserted into the quadtree to connect the FCN to the hierarchical PGM.

C. Hierarchical Markov Model and MPM criterion

The idea is to exploit the advantages of information extracted at different scales. In [25], a causal hierarchical framework that postulates Markovianity not only among pixels belonging to pixel lattices with different resolutions but also among pixels in the same lattice is described, which considers both cross-layer and intralayer dependencies. In this method, Markov chains are formulated both across the scales of a quadtree and with respect to a 1D scan of the pixel lattice of each layer. This joint strategy benefits from the spatial information within each layer and inherently supports multiresolution fusion, which makes it an interesting model to combine with the multiresolution behavior of CNNs.

In a quadtree, each site in a grid corresponds to a group of 2 $\times$ 2 sites in the grid below it [19]. Consider $\{S^0, S^1, \ldots, S^L\}$, $S^\ell \subset \mathbb{Z}^2$ $(\ell = 0, 1, \ldots, L)$, 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 $(\ell = L)$, not having any children site, and the root layer $(\ell = 0)$, not having a parent site. A hierarchy on the tree $S = \bigcup_{\ell=0}^{L} S^\ell$ from the root to the leaves is determined. 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 $(\ell \geqslant 1)$ [21], [23]:

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

where $\mathcal{X}^\ell = \{x_s\}_{s \in S^\ell}$ $(\ell = 0, 1, \ldots, L)$, and Markovianity holds across the scales. In this hierarchical model, these transition probabilities are also represented by [23]:

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

thus removing the contextual dependency within $\mathcal{X}^\ell$. Conditional independence is another important concept, as it can be used to decompose complex probability distributions into a product of factors, each consisting of a subset of corresponding random variables. For the observation model $P(\mathcal{Y}|\mathcal{X})$, where $\mathcal{Y} = \{y_s\}_{s \in S}$ is the random field of the observations associated with all the pixels in the quadtree, a standard pixelwise factorization is assumed:

$$P(\mathcal{Y}|\mathcal{X}) = \prod_{s \in S} P(y_s | x_s) = \prod_{\ell=0}^{L} \prod_{s \in S^\ell} P(y_s | x_s). \tag{3}$$

The quadtree is extended to incorporate spatial information while keeping the causality of the hierarchical model. Consider

a rectangular lattice R and an order relation $\prec$ in the grid of pixels, such that $r \prec s$ indicates the sites $r \in R$ located before a given site $s \in R$. A neighborhood relation is assumed in R consistently with this order relation, and $r \precsim s$ indicates that r is a causal neighbor of s. Hence, spatial Markovianity is expressed as [22], [42]:

$$P(x_s|x_r, r \prec s) = P(x_s|x_r, r \precsim s) \tag{4}$$

$$P(\mathcal{X}) = \prod_{s \in R} P(x_s|x_r, r \precsim s) \tag{5}$$

In the hierarchical Markov model integrated with the aforementioned FCN in the proposed method, the observations are extracted from the network activations and, similar to [25], the dependencies in Equation (4) are taken into account so that Equation (1) can be modified to express both the cross-layer and the intralayer dependencies. The model is defined by the following assumptions:

(i) the $(L-\ell)$-th lattice $S^{L-\ell}$ in the quadtree is the pixel grid shared by both the ℓ-th and the $(N-\ell)$-th blocks of FCN activations in the network, where N represents the last layer of the FCN ($\ell = 0, 1, \ldots, L$). Hence, the leaf level S^L corresponds to the lattice of the original input image, S^{L-1} is the pixel grid on which the activations after the first pooling layer of the FCN are defined, together with the activations of the third block of the decoder, etc.

(ii) for each $s \in S^\ell$, the vector y_s collects the activations of all convolutional layers in block ℓ ($\ell = 0, 1, \ldots, L$) for pixel s. In the case of the leaf level of the quadtree ($\ell = L$), the channels of the input image are used. Hence, $\mathcal{Y}$ is the multiresolution random field of the network activations at all scales. These activations are collected after the nonlinearity of the network and stacked together through the skip-connections from the encoder to the decoder and from the decoder to the quadtree (see Fig. 1).

(iii) for each $s \in S$, x_s again indicates the class label. The corresponding random field $\mathcal{X}$ satisfies the hierarchical Markovianity in Equation (1);

(iv) $\mathcal{X}^0$ is an MMRF on the root lattice S^0;

(v) the following proportionality holds for $\ell = 1, 2, ..., L$:

$$P(\mathcal{X}^\ell|\mathcal{X}^{\ell-1}) \propto \prod_{s \in S^\ell} P(x_s|x_r, r \precsim s) P(x_s|x_s^-) \tag{6}$$

(vi) $\mathcal{Y}$ satisfies the conditional independence in Equation (3).

As already stated in [25], Assumption (v) and Equation (6) combine into a unique model the Markovianity conditions that are separately formalized by Equations (1)-(2), and (4) with regard to hierarchical and spatial models, respectively. In particular, it can be proven that $P(\mathcal{X}, \mathcal{Y})$ is entirely defined by the parent-child transition probabilities $P(x_s|x_s^-)$, the causal sibling transition probabilities $P(x_s|x_r, r \precsim s)$ and the data conditional likelihood $P(y_s|x_s)$ (the proof can be found in [25]). This factorization implies that $(\mathcal{X}, \mathcal{Y})$ is a Markov random field with respect to the causal neighborhood of pixels, which determines the causality of the hierarchical PGM framework [25]. This causality property is important because it allows the use of sequential inference algorithms, which are remarkably time-efficient.

As pointed out in [23], the conventional MAP estimate is not satisfying since its cost function would assign equal cost to a single mislabeled pixel at the finest scale or to the mislabeling of hundreds of pixels at the coarsest scale. In contrast, the causality of the proposed framework, both spatially and across scales, allows an efficient recursive algorithm to be formulated for the MPM criterion. MPM assigns each $s \in S$ the class label x_s that maximizes $P(x_s|\mathcal{Y})$ [21] and penalizes errors according to the scale where they are made. Accordingly, as proven in [25], MPM inference in the considered framework is accomplished through the following recursive steps:

$$P(x_s) = \sum_{x_{s^-} \in \Omega} P(x_s|x_{s^-}) P(x_{s^-}), \tag{7}$$

$$P(x_s|y_s^d) \propto P(x_s|y_s) \prod_{t \in s^+} \sum_{x_t \in \Omega} \frac{P(x_t|y_t^d) P(x_t|x_s)}{P(x_t)}, \tag{8}$$

$$P(x_s^c|x_s, y_s^d) \propto \frac{P(x_s|y_s^d) P(x_s|x_{s^-}) P(x_{s^-})}{P(x_s)^{n_s}} \cdot \cdot \prod_{r \precsim s} P(x_s|x_r) P(x_r), \tag{9}$$

$$P(x_s|\mathcal{Y}) = \sum_{x_s^c \in \Omega^{n_s}} P(x_s^c|x_s, y_s^d) P(x_{s^-}|\mathcal{Y}) \prod_{r \precsim s} P(x_r|\mathcal{Y}), \tag{10}$$

where y_s^d collects the observations of all descendants of s in the tree (including s), x_s^c the labels of all sites connected to s (x_{s^-} and $\{x_r\}_{r \precsim s}$), and n_s is the number of such sites. First, (7) calculates $P(x_s)$ at all sites through a top-down pass from the root to the leaves. For the root layer, these probabilities are initialized as the relative frequency of the classes in the training set. Then, (8) and (9) compute $P(x_s|x_s^c, y_s^d)$ through a bottom-up pass from the leaves to the root. Finally, (10) derives $P(x_s|\mathcal{Y})$ through a second top-down pass. More details on the mathematical formulation of the Markovian framework can be found in [25].

The "past" of a pixel is determined on a 1D basis according to a case-specific way of scanning all the pixels in a lattice. On each layer S^ℓ of the quadtree ($\ell = 0, 1, .., L$), the sequence that visits every pixel once and moves from one pixel to one of its neighbors consists of a combination of four zigzag scans and two Hilbert space-filling curves (details can be found in [25]). Each site is visited multiple times in a symmetric manner, which prevents the risk of geometrical artifacts that may occur while integrating contextual information in a 1D scan on a 2D pixel lattice [25].

The transition probability $P(x_s|x_s^-)$ across consecutive scales is defined through a parametric stationary model [43] in which $P\{x_s = \omega | x_{s^-} = \omega\} = \vartheta$ for all $\omega \in \Omega$, where ϑ is a hyperparameter of the method, and $P\{x_s = \omega | x_{s^-} = \omega'\}$ is constant over all $\omega \neq \omega' (\omega, \omega' \in \Omega)$ [25]. $P(x_s|x_r)$ ($r \precsim s$) is modeled analogously with a parameter ψ.

The data representation extracted by the FCN, formalized through the random field $\mathcal{Y}$ of multiresolution network activations, affects the behavior of the described probabilistic graphical model through the pixelwise posteriors $P(x_s|y_s)$

in Equation (8). The RF classifier [41] is used to estimate these posteriors from the training samples of the classes. A separate RF is trained on each ℓ-th level of the quadtree ($\ell = 0, 1, \ldots, L$), using the training quadtree described in Section III-A.

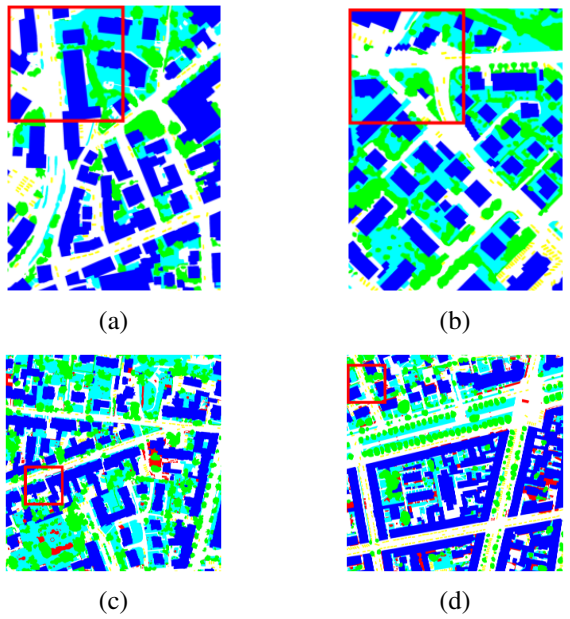


Fig. 2. Highlights of the crops of the images used to train and test the proposed architecture: (a) Vaihingen training tile, (b) Vaihingen test tile, (c) Potsdam training tile, (d) Potsdam test tile. Classes: buildings (blue), impervious (white), vegetation (cyan), trees (green), cars (yellow).

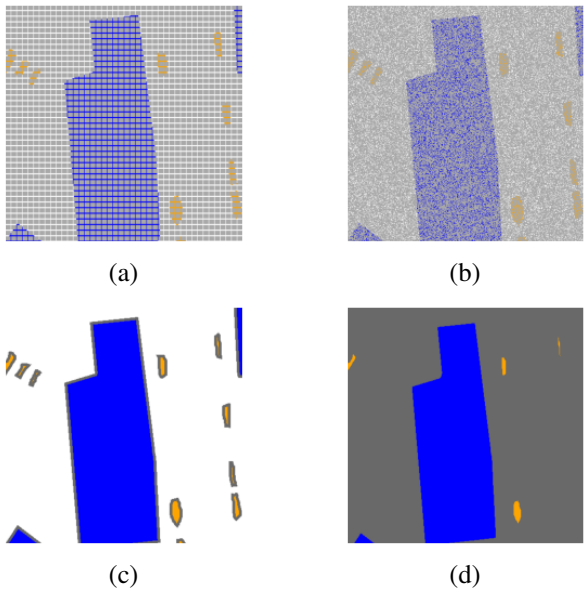


Fig. 3. Different modifications of a dense ground truth used to approximate sparse ground truths: (a) removed labeled pixels in rectangular blocks, according to a regular grid, (b) randomly removed labeled pixels, (c) morphological erosion, and (d) removal of connected components. Classes: buildings (blue), impervious (white), cars (orange); removed labeled pixels are shown in gray.

The method is summarized in Algorithm 1.

Algorithm 1 FCN + MPM on the Hierarchical PGM

1: Training of the FCN with the input VHR dataset
2: Input to the MPM: $L-$levels quadtree containing, in the random field $\mathcal{Y} = \{y_s\}_{s \in S}$ of the observations, the network activations and the original channels of the image to classify
3: First top-down pass: estimation of the priors $P(x_s)$ via (7)
4: Estimation of the posterior probabilities $P(x_s|y_s)$ in (8) through the RF classifier
5: Bottom-up pass: estimation of $P(x_s|y_s^d)$ and $P(x_s^c|x_s, y_s^d)$ via (8) and (9)
6: Second top-down pass: estimation of $P(x_s|\mathcal{Y})$ at each level of the quadtree via (10)
7: Output: maximization of $P(x_s|\mathcal{Y})$

D. Computational Complexity

The proposed method combines an underlying FCN with the described hierarchical PGM and with RF. Indeed, the technique can integrate any kind of FCN in a flexible manner, and the contribution of the FCN to the total complexity of the method obviously depends on the specifics of the FCN itself. RF is known to be a computationally efficient algorithm, hence it does not critically contribute to the overall computational burden of the proposed approach. Here, we focus on the computational complexity of the PGM component of the developed method.

The computational complexity of the hierarchical PGM employing the MPM criterion is determined by the four equations formalizing its inference process (Equations (7)-(10)). The complexity can be evaluated as:

$$\mathcal{O}\left(\frac{p}{3}(4 - 4^{-L}) \cdot (M^2 + M^2 q + 2M^{n+1})\right), \qquad (11)$$

where p is the number of pixels in the finest-resolution lattice S^L of the quadtree, q is the number of children of a generic site, and n is the number of parent and past neighbor sites connected to a generic site. L and M indicate again the numbers of levels in the quadtree and of classes, respectively. Specifically, in the considered PGM topology, $q = 4$, because each site has four children, and $n = 2$, because each site is connected to one past neighbor in the 1D scan and to one parent. In the present evaluation of complexity, we are not considering the special cases of the root and leaf levels of the quadtree, whose sites have no parent and children, respectively. Furthermore, in (11), we focus on the complexity associated with Eqs. (7)-(10) *per se* without considering that, within each lattice of the quadtree, the Hilbert and zigzag scans visit each pixel multiple times (see Section III-C). Indeed, these approximations do not significantly affect the order of magnitude of the complexity of the technique.

In (11), the term $p(4-4^{-L})/3$ is the total number of sites in the quadtree. It derives from the power-of-2 relation between the pixel grids at different levels: since p is the number of pixels in the leaf level, the number of sites in the ℓ-th level is

TABLE I
TEST-SET ACCURACIES OF THE PROPOSED METHOD ON THE VAIHINGEN DATASET COMPARED TO THE STANDARD FCNS. "70% BU" AND "80% RU" STAND FOR "70% OF UNLABELED PIXELS IN RECTANGULAR BLOCKS" AND "80% OF RANDOMLY UNLABELED PIXELS", RESPECTIVELY.

	Architecture	buildings	impervious	vegetation	trees	cars	overall acc.	recall	precision	Cohen's κ	F1 score
Full dataset	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
	HRNet [11]	0.89	0.81	0.42	0.92	0.63	0.77	0.73	0.76	0.68	0.74
	Proposed method, "IRRG-RF" (U-Net)	0.85	0.82	0.69	0.92	0.38	0.81	0.73	0.78	0.74	0.75
	Proposed method, "PGM+Net" (U-Net)	0.87	0.84	0.71	0.92	0.89	0.82	0.81	0.78	0.75	0.80
	Proposed method, "Net for cars" (U-Net)	0.84	0.81	0.68	0.92	0.86	0.81	0.82	0.72	0.75	0.77
	Proposed method, "resize" (U-Net)	0.84	0.82	0.69	0.92	0.88	0.81	**0.83**	0.77	0.75	0.80
	Proposed method, "IRRG-RF" (SegNet)	0.81	0.74	0.68	0.91	0.29	0.77	0.69	0.74	0.69	0.71
	Proposed method, "PGM+Net" (SegNet)	0.82	0.74	0.72	0.93	0.94	0.80	**0.83**	0.73	0.73	0.78
	Proposed method, "Net for cars" (SegNet)	0.71	0.73	0.66	0.91	**0.99**	0.74	0.80	0.67	0.66	0.73
	Proposed method, "resize" (SegNet)	0.80	0.74	0.67	0.91	0.88	0.78	0.80	0.73	0.70	0.76
	Proposed method, "IRRG-RF" (HRNet)	0.73	0.79	0.52	0.86	0.17	0.72	0.62	0.66	0.62	0.64
	Proposed method, "PGM+Net" (HRNet)	0.77	0.76	0.22	**0.98**	0.93	0.68	0.73	0.65	0.58	0.69
	Proposed method, "Net for cars" (HRNet)	0.70	0.76	0.48	0.87	0.93	0.70	0.74	0.62	0.61	0.68
	Proposed method, "resize" (HRNet)	0.73	0.80	0.53	0.86	0.62	0.72	0.70	0.67	0.63	0.69
	LWN-Attention [44]	0.91	**0.97**	0.61	0.88	0.64	**0.85**	0.80	0.82	**0.80**	0.81
Erosion	Standard U-Net	0.94	0.84	0.67	0.72	0.26	0.81	0.68	**0.81**	0.74	0.74
	Proposed method, "IRRG-RF" (U-Net)	0.91	0.81	0.66	**0.79**	0.02	0.80	0.64	0.75	0.72	0.69
	Proposed method, "PGM+Net" (U-Net)	0.93	0.81	**0.74**	0.69	0.70	**0.82**	**0.77**	0.77	**0.75**	**0.77**
	Proposed method, "Net for cars" (U-Net)	0.87	0.81	0.65	0.78	**0.76**	0.79	**0.77**	0.70	0.71	0.73
	Proposed method, "resize" (U-Net)	0.91	0.81	0.66	**0.79**	0.50	0.80	0.73	0.80	0.73	0.76
	LWN-Attention [44]	**0.95**	**0.85**	0.38	0.75	0.12	0.75	0.61	0.79	0.66	0.69
70% BU	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, "IRRG-RF" (U-Net)	0.80	0.83	0.68	0.89	0.36	0.79	0.71	0.70	0.72	0.70
	Proposed method, "PGM+Net" (U-Net)	0.82	0.85	**0.76**	0.86	0.96	0.82	**0.85**	0.72	0.76	0.78
	Proposed method, "Net for cars" (U-Net)	0.72	0.82	0.67	0.88	**0.99**	0.76	0.82	0.68	0.69	0.74
	Proposed method, "resize" (U-Net)	0.78	0.83	0.68	0.89	0.90	0.79	0.81	0.71	0.71	0.76
	FESTA [45]	0.80	**0.92**	0.67	**0.91**	0.63	**0.83**	0.79	**0.83**	**0.77**	**0.81**
	LWN-Attention [44]	**0.90**	0.88	0.74	0.72	0.25	0.82	0.70	0.80	0.75	0.75
80% RU	Standard U-Net	0.90	**0.87**	0.72	**0.91**	0.67	**0.85**	0.82	**0.84**	**0.80**	**0.83**
	Proposed method, "IRRG-RF" (U-Net)	0.84	0.82	0.69	0.90	0.35	0.81	0.72	0.74	0.74	0.73
	Proposed method, "PGM+Net" (U-Net)	0.83	**0.87**	**0.80**	0.88	0.87	0.84	**0.85**	0.77	0.79	0.81
	Proposed method, "Net for cars" (U-Net)	0.82	0.82	0.69	0.90	**0.91**	0.80	0.83	0.71	0.74	0.77
	Proposed method, "resize" (U-Net)	0.83	0.82	0.69	0.90	0.83	0.81	0.82	0.76	0.74	0.79
	LWN-Attention [44]	**0.91**	0.83	0.56	0.86	0.47	0.80	0.73	0.75	0.73	0.74
Connected components	Standard U-Net	**0.96**	0.65	0.47	0.89	0.48	0.76	0.69	**0.81**	0.66	**0.75**
	HRNet [11]	0.84	0.77	**0.68**	0.80	0.29	**0.77**	0.68	0.74	**0.71**	0.71
	Proposed method, "IRRG-RF" (U-Net)	0.94	0.63	0.25	**0.96**	0.13	0.71	0.58	0.75	0.59	0.65
	Proposed method, "PGM+Net" (U-Net)	0.94	0.68	0.49	0.86	0.74	0.76	**0.74**	0.75	0.67	**0.75**
	Proposed method, "Net for cars" (U-Net)	0.92	0.62	0.25	0.95	**0.79**	0.70	0.71	0.67	0.60	0.69
	Proposed method, "resize" (U-Net)	0.94	0.63	0.25	**0.96**	0.65	0.71	0.68	0.77	0.60	0.72
	Proposed method, "IRRG-RF" (HRNet)	0.81	**0.79**	0.63	0.88	0.09	0.76	0.63	0.76	0.68	0.69
	Proposed method, "PGM+Net" (HRNet)	0.87	0.76	0.44	0.94	0.56	0.75	0.71	0.68	0.67	0.70
	Proposed method, "Net for cars" (HRNet)	0.88	0.76	0.43	0.94	0.48	0.76	0.70	0.68	0.67	0.69
	Proposed method, "resize" (HRNet)	0.81	**0.79**	0.63	0.88	0.10	**0.77**	0.64	0.78	0.68	0.70
	LWN-Attention [44]	0.91	0.75	0.44	0.87	0.51	0.76	0.70	0.74	0.67	0.72

$p \cdot 4^{-(L-\ell)}$ ($\ell = 0, 1, \ldots, L$). Therefore, the total number of sites in the quadtree is:

$$\sum_{\ell=0}^{L} p{\cdot}4^{-(L-\ell)} = p \sum_{i=0}^{L} \left(\frac{1}{4}\right)^i = p\frac{1-(1/4)^{L+1}}{1-1/4} = \frac{p}{3}(4-4^{-L}). \tag{12}$$

Equations (7)-(10) are applied on every site s of the quadtree. For each site, (7) is evaluated for all of the M possible labels x_s by looping over the M possible labels x_{s^-} (neglecting the special case of the root level). This yields the term M^2 in (11). Similarly, (8) is evaluated for all labels x_s by looping over all q children and all labels x_t of each children (neglecting the special case of the leaf level), thus contributing the term M^2q in (11). Finally, (9) is evaluated for every label configuration (x_s, x_s^c), while (10) is evaluated for every x_s by looping over x_s^c. Since x_s^c includes n labels, each taking on M values, both (9) and (10) have complexities of M^{n+1}, hence the term $2M^{n+1}$ in (11).

According to (11), the overall complexity of the MPM formulation of the proposed method is linear in the number p of pixels of the input image – which is a desirable property – and grows with the number of classes as M^3 (since $n = 2$).

IV. EXPERIMENTS

The proposed method was experimentally validated with the ISPRS 2D Semantic Labeling Challenge Vaihingen and Potsdam datasets[1]. These datasets consist of very high-resolution aerial images provided by the German Society for Photogrammetry, Remote Sensing, and Geoinformation (DGPF).

A. Datasets

Both the ISPRS Vaihingen and Potsdam datasets are characterized by six classes: impervious surfaces, buildings, low vegetation, trees, cars, and clutter.

The Vaihingen dataset has a spatial resolution of 9 cm with images (tiles) with different sizes $N_h \times N_v$, with N_h in the range [1388, 3816] and N_v in the range [1281, 3313]. Each image contains three channels: near infrared (NIR), red, and green (IRRG). There are 33 images, 16 of which have a public ground truth. Twelve tiles were chosen for training the network (tiles 1, 3, 7, 11, 13, 17, 23, 26, 28, 32, 34, and 37) and four

[1]https://www2.isprs.org/commissions/comm2/wg4/benchmark/semantic-labeling/

TABLE II
TEST-SET ACCURACIES OF THE PROPOSED METHOD ON THE POTSDAM DATASET COMPARED TO THE STANDARD FCNS. "70% BU" AND "80% RU" STAND FOR "70% OF UNLABELED PIXELS IN RECTANGULAR BLOCKS" AND "80% OF RANDOMLY UNLABELED PIXELS", RESPECTIVELY.

	Architecture	buildings	impervious	vegetation	trees	cars	clutter	overall acc.	recall	precision	Cohen's κ	F1 score
Full GT	Standard U-Net	**0.94**	**0.99**	0.85	0.80	0.83	0.26	**0.91**	**0.88**	**0.90**	**0.88**	**0.89**
	HRNet [11]	0.90	0.95	0.79	0.64	0.69	0.13	0.80	0.80	0.83	0.76	0.82
	Proposed method, "IRRG-RF" (U-Net)	0.92	**0.99**	0.81	**0.82**	0.87	0.33	0.90	**0.88**	0.89	0.87	0.88
	Proposed method, "PGM+Net" (U-Net)	0.87	**0.99**	0.83	0.81	0.89	0.30	0.89	**0.88**	0.86	0.86	0.87
	Proposed method, "Net for cars" (U-Net)	0.90	**0.99**	0.81	0.81	0.89	**0.38**	0.90	**0.88**	0.87	0.86	0.87
	Proposed method, "resize" (U-Net)	0.93	**0.99**	0.81	**0.82**	0.86	0.24	0.90	**0.88**	0.88	0.87	0.88
	Proposed method, "IRRG-RF" (HRNet)	0.89	0.95	0.72	0.72	0.90	0.09	0.82	0.84	0.81	0.78	0.82
	Proposed method, "PGM+Net" (HRNet)	0.82	0.88	**0.86**	0.43	**0.96**	0.16	0.72	0.79	0.73	0.67	0.76
	Proposed method, "Net for cars" (HRNet)	0.85	0.94	0.72	0.71	**0.96**	0.13	0.80	0.83	0.76	0.76	0.80
	Proposed method, "resize" (HRNet)	0.91	0.95	0.72	0.72	0.82	0.15	0.82	0.82	0.82	0.78	0.82
Erosion	Standard U-Net	**0.97**	**0.98**	**0.72**	**0.68**	0.36	0.01	**0.85**	0.74	**0.87**	**0.79**	**0.80**
	Proposed method, "IRRG-RF" (U-Net)	**0.97**	**0.98**	0.63	0.67	0.30	0.00	0.83	0.71	0.85	0.76	0.77
	Proposed method, "PGM+Net" (U-Net)	0.96	**0.98**	**0.72**	**0.68**	0.41	**0.04**	**0.85**	**0.75**	0.86	**0.79**	**0.80**
	Proposed method, "Net for cars" (U-Net)	**0.97**	**0.98**	0.64	0.67	**0.52**	0.01	0.83	**0.75**	0.83	0.78	0.79
	Proposed method, "resize" (U-Net)	**0.97**	**0.98**	0.63	0.67	0.42	0.00	0.83	0.73	0.85	0.77	0.79
70% BU	Standard U-Net	**0.94**	**0.99**	**0.85**	0.78	0.82	0.18	**0.90**	0.87	**0.89**	**0.87**	**0.88**
	Proposed method, "IRRG-RF" (U-Net)	**0.94**	**0.99**	0.78	0.81	0.86	0.20	**0.90**	0.87	0.88	0.86	0.87
	Proposed method, "PGM+Net" (U-Net)	0.90	**0.99**	0.81	**0.84**	0.86	0.28	**0.90**	**0.88**	**0.89**	**0.87**	**0.88**
	Proposed method, "Net for cars" (U-Net)	0.91	**0.99**	0.78	0.81	**0.89**	**0.39**	**0.90**	**0.88**	0.88	0.86	**0.88**
	Proposed method, "resize" (U-Net)	**0.94**	**0.99**	0.78	0.81	0.84	0.15	**0.90**	0.87	0.88	0.86	0.87
80% RU	Standard U-Net	**0.94**	**0.99**	0.85	0.82	0.80	0.21	**0.90**	0.88	**0.91**	**0.87**	**0.89**
	Proposed method, "IRRG-RF" (U-Net)	0.93	**0.99**	0.80	**0.85**	0.82	0.18	**0.90**	0.88	0.89	**0.87**	0.88
	Proposed method, "PGM+Net" (U-Net)	0.85	**0.99**	**0.86**	0.83	**0.91**	0.26	**0.90**	**0.89**	0.87	0.86	0.88
	Proposed method, "Net for cars" (U-Net)	0.91	**0.99**	0.79	0.83	0.89	**0.29**	**0.90**	0.88	0.88	**0.87**	0.88
	Proposed method, "resize" (U-Net)	0.93	**0.99**	0.80	0.84	0.82	0.13	**0.90**	0.88	0.89	**0.87**	0.88
Connected components	Standard U-Net	**0.93**	0.94	0.63	**0.75**	0.86	0.00	0.71	0.73	0.75	0.64	0.74
	HRNet [11]	0.91	0.80	0.45	0.67	0.67	0.00	0.72	0.70	**0.77**	0.65	0.73
	Proposed method, "IRRG-RF" (U-Net)	0.86	0.96	0.46	0.70	0.98	0.00	0.76	0.79	0.74	0.71	0.76
	Proposed method, "PGM+Net" (U-Net)	0.77	0.95	0.73	0.68	**0.99**	0.00	**0.78**	**0.82**	0.74	**0.73**	**0.78**
	Proposed method, "Net for cars" (U-Net)	0.84	0.95	0.46	0.68	**0.99**	0.00	0.75	0.78	0.71	0.70	0.74
	Proposed method, "resize" (U-Net)	0.89	**0.97**	0.46	0.71	0.57	0.00	0.76	0.72	0.75	0.71	0.73
	Proposed method, "IRRG-RF" (HRNet)	0.89	0.95	0.58	0.62	0.91	0.00	0.76	0.79	0.75	0.71	0.77
	Proposed method, "PGM+Net" (HRNet)	0.85	0.89	**0.75**	0.41	0.97	0.00	0.71	0.77	0.71	0.65	0.74
	Proposed method, "Net for cars" (HRNet)	0.85	0.94	0.57	0.60	0.94	0.00	0.74	0.78	0.70	0.69	0.74
	Proposed method, "resize" (HRNet)	0.90	0.95	0.58	0.62	0.85	0.00	0.76	0.78	0.76	0.71	0.77

TABLE III
COMPUTATIONAL TIME OF THE PROPOSED METHOD BASED ON U-NET OVER THE TWO DATASETS, COMPARED TO U-NET, SEGNET, HRNET, AND FESTA.

	Proposed method, "IRRG-RF"	Proposed method, "PGM+Net"	Proposed method, "Net for cars"	Proposed method, "resize"	U-Net	SegNet	HRNet	FESTA
Vaihingen	11258 s	11297 s	11265 s	11315 s	5674 s	5203 s	$\approx$ 60000 s	800376 s
Potsdam	11197 s	11243 s	11238 s	11286 s	5698 s	5342 s	$\approx$ 60000 s	800117 s

TABLE IV
TEST-SET ACCURACIES OF THE PROPOSED METHOD APPLIED WITH DIFFERENT CONFIGURATIONS TO THE POTSDAM DATASET (RESIZED TO 2000 $\times$ 2000 PIXELS) AND COMPARED TO FESTA. THE GROUND TRUTH HAVE 70% OF UNLABELED PIXELS IN RECTANGULAR BLOCKS.

70% of unlabeled pixels in rectangular blocks	buildings	impervious	vegetation	trees	cars	overall acc.	recall	precision	F1 score
Standard U-Net	**0.96**	**0.93**	0.81	0.80	0.83	**0.89**	0.86	**0.88**	**0.87**
Proposed method, "IRRG-RF" (U-Net)	0.95	0.92	0.81	**0.81**	0.87	**0.89**	**0.87**	0.86	0.86
Proposed method, "PGM+Net" (U-Net)	0.95	0.88	**0.84**	0.79	0.86	0.87	0.86	0.84	0.85
Proposed method, "Net for cars" (U-Net)	0.95	0.91	0.81	**0.81**	**0.89**	**0.89**	**0.87**	0.85	0.86
Proposed method, "resize" (U-Net)	**0.96**	0.92	0.82	**0.81**	0.86	**0.89**	**0.87**	0.86	**0.87**
FESTA [45]	0.95	0.86	0.82	0.78	0.57	0.84	0.80	0.83	0.81

were chosen for testing the network (tiles 5, 15, 21, and 30). Only a subsection of 1280 $\times$ 1280 pixels of the training tile number 1 (1919 $\times$ 2569 pixels) and of the test tile number 30 (1934 $\times$ 2563 pixels), shown in Fig. 2(a) and Fig. 2(b), were selected to train the RF and to apply the hierarchical PGM, respectively, since it was impossible to perform the analysis on larger patches because of RAM limitations (all experiments were run on an Alienware Aurora R11 with a RAM of 16 GB and a GPU NVIDIA GeForce RTX 2080 Ti). This does not represent a drawback of the proposed technique, since any image can be divided into separate or overlapping patches, which can be processed singularly, and then reconstructed. The images to which the PGM was applied have instances of all the classes except for the clutter, which was then excluded from the experimentation on this dataset.

The ISPRS Potsdam dataset contains 38 images with IRRG channels at a spatial resolution of 5 cm and of size 6000 $\times$ 6000 pixels. Some of these 38 images are labeled. The training set for the FCN consisted of 11 images (tiles 3_11, 4_11, 5_10, 6_7, 6_8, 6_9, 7_7, 7_8, 7_9, 7_10, 7_12) and the test set contained 5 images (tiles 3_12, 4_10, 4_12, 5_11, 6_12). A subsection of 1024 $\times$ 1024 pixels of tile 4_11 and one of tile 5_11 were used to train the RF classifier and to obtain the output classification map from the hierarchical Markov model, respectively.

As previously done by other authors in [46], [47], [48], the results for the clutter class, which includes all covers that are not attributed to the other classes and mixes water

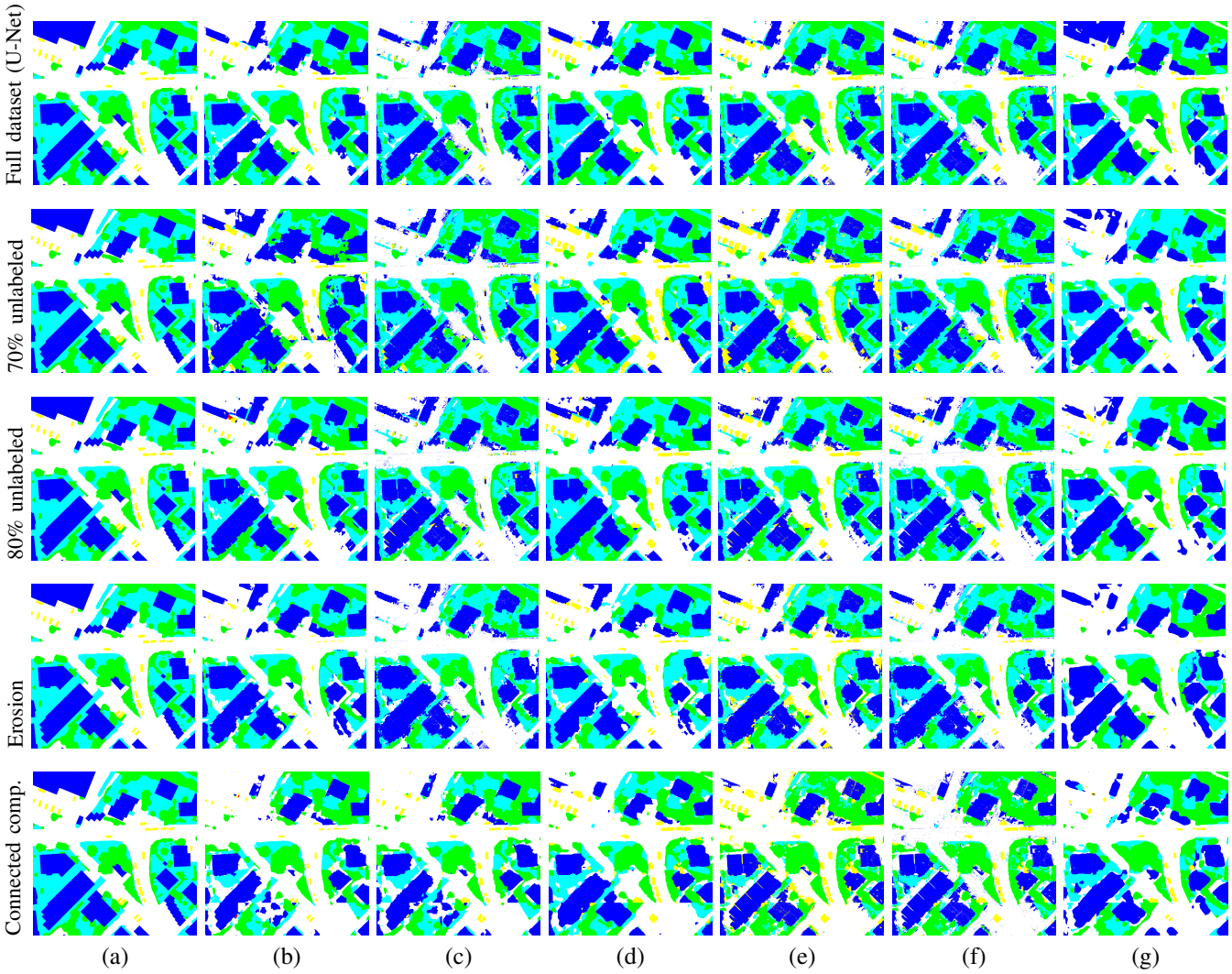


Fig. 4. **Ground truth and classification maps for the Vaihingen dataset:** (a) ground truth and classification maps obtained with (b) U-Net, the proposed method in versions (c) "IRRG-RF", (d) "PGM+Net", (e) "Net for cars", (f) "resize", and (g) LWN-Attention [44]. Classes: buildings (blue), impervious (white), vegetation (cyan), trees (green), cars (yellow).

bodies, background, and others in the same thematic class, were excluded from the averaged metrics. This class accounts for only a small percentage of pixels. However, for the sake of completeness, the classwise recalls of class "clutter" are reported in Table II.

B. Experimental setup

The DL models were trained with the learning rate fixed to 0.01 and a decay rate of 0.0005. The optimizer employed was the Adam algorithm [49]. As mentioned above, the backbones used to experimentally validate the proposed method were U-Net, SegNet, and HRNet[2]. The last network consists of multiresolution subnetworks connected in parallel. We trained and validated U-Net and SegNet with patches of size 256×256, (512×512 for HRNet) obtained through a sliding-window approach and a batch size of 10. The three networks were pretrained on ImageNet[3], in particular U-Net and SegNet were initialized with the weights of VGG-16 [50]. Since U-Net obtained the best results, it was selected to conduct most of the experiments. Three skip connections starting from the three central deconvolution blocks of the decoder to the output layer allow to collect the activations of the network at three different resolutions (128×128, 64×64, 32×32 pixels), which are then inserted into the quadtree used to connect the FCN to the hierarchical PGM, to exploit the information hidden at different layers. For the Vaihingen dataset, a group of 25 subsequent patches belonging to the original images used for training and testing (see Section IV-A) were merged to obtain a final image of dimension 1280×1280 (see Fig. 2(a)-(b) and Fig. 4(a)). It was not possible to create larger images because of the aforementioned RAM limitations. The quadtree was built with 4 layers, the highest 3 depending on the activation output from the network, while the base corresponded to the IRRG channels of the original image. In the case of the Potsdam dataset, it was necessary to include among the features the activations of the network in the base of the pyramid of the quadtree as well, to obtain fine results.

[2]https://github.com/HRNet/HRNet-Semantic-Segmentation
[3]https://image-net.org/

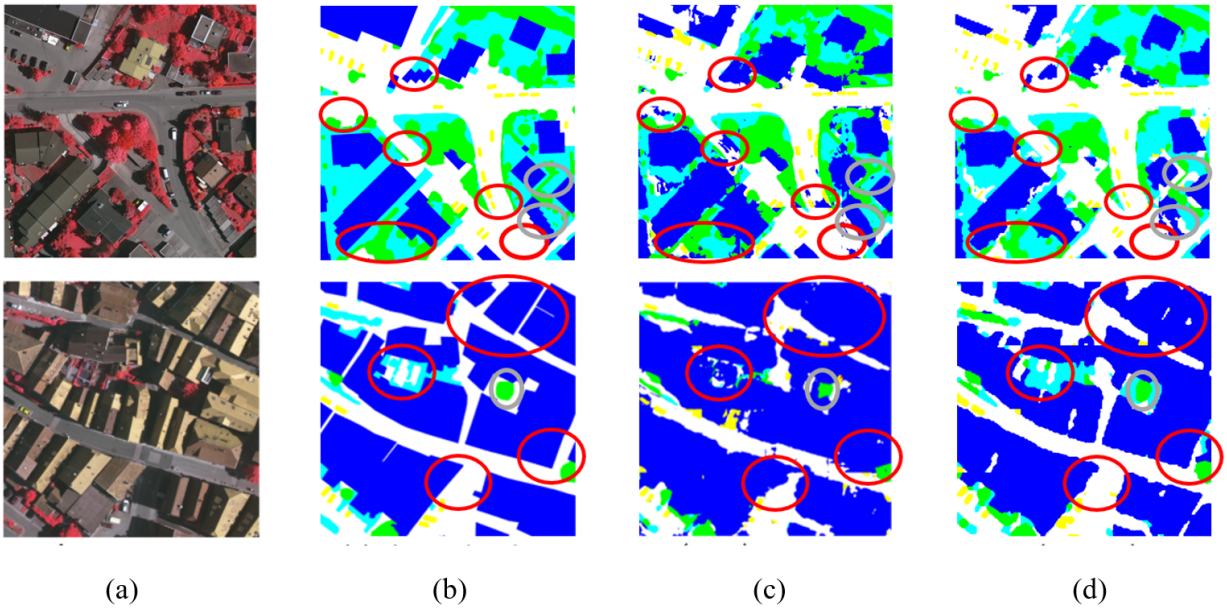


Fig. 5. **Ground truth and classification maps for the Vaihingen dataset, highlight on some of the classification errors (gray) and improvements (red):** (a) IRRG image, (b) ground truth, and classification maps obtained with: (c) U-Net and (d) the proposed method. Classes: buildings (blue), impervious (white), vegetation (cyan), trees (green), cars (yellow).

Indeed, since the spatial resolution of the Potsdam dataset is higher than that of the Vaihingen dataset, the patches used to train the RF classifier and to apply the hierarchical PGM showed more spatial details and fewer entire objects in the image. Therefore, RF, which is non-contextual, was able to discriminate every little detail in the image (e.g., windows on the rooftops, gutters, shadows on the streets), assigning each of these objects a class label that did not correspond to the semantic macrostructure of the ground truth available (for example, the windows on the rooftops were not considered parts of the class building). The addition of the network activations of the last layer allowed to smooth this issue out.

Due to the presence of another layer of activations and memory limitations, the subsequent patches considered for the Potsdam dataset were 16, thus creating a cropped image of dimension 1024×1024 (see Fig. 2(c)-(d) and Fig. 6(a)). The cropped images obtained from training tiles 1 and 4_11 for the Vaihingen and Potsdam datasets, respectively, were used to train the RF classifier. The cropped images of testing tiles 30 and 5_11 were processed by the RF classifier to obtain the posterior probabilities used by the PGM to perform the classification.

The PGM depends on two hyperparameters, i.e., the transition probabilities between the scales (ϑ) and intrascale (ψ). For the Vaihingen dataset, preliminary experiments (not reported for brevity) demonstrated that values of ψ and ϑ equal to 0.82 brought the best results (as ψ and ϑ varied in $[0, 1]$) with respect to the average accuracy metrics (e.g., overall accuracy (OA), precision, recall, Cohen's κ, and F1 score). For the Potsdam dataset, a lower value of spatial transition probability (ψ equal to either 0.20 or 0.50) was obtained through this hyperparameter tuning procedure, thus requiring less smoothing. As mentioned above, the average results reported in the tables were computed without including the pixels belonging to the class "clutter".

The proposed architecture was compared with the results of the aforementioned standard U-Net, Segnet, and HRNet architectures. However, the best results were obtained with U-Net; hence, it was selected for the rest of the experimentation (see also Section IV-C).

The training set of the Vaihingen dataset is an "ideal" one, where the true label is known for all pixels in the training tiles. This is feasible in such a dataset designed for an international contest but is very seldom available in real-world scenarios, where the goal is to generate accurate land cover or land use maps using fewer training samples, which are most often arranged in homogeneous patches and do not include spatial class borders. Therefore, each backbone network was further trained with "deteriorated" training sets, to assess whether the full pipeline could provide improvements in accuracy. These modifications involved the degradation of the ground truths of the training set: (i) with a percentage of removed training pixels (either randomly or in rectangular blocks); (ii) by morphological operators; or (iii) by first removing entire connected components from the GT map and then applying mathematical morphology (see Fig. 3 and Section IV-C). Options (ii) and (iii) are meant as approximations of the ground truths that are usually found in real applications, i.e., maps with isolated patches of labeled pixels associated with different classes. They are approximations because, in these cases, the sections of labeled pixels are regular, with a preservation of the prior probabilities of the classes: the relative frequency of the classes in the training set is maintained.

The pixelwise posterior probabilities were inserted in the hierarchical PGM, as mentioned before. However, concerning the experimentation conducted on the Vaihingen dataset, the

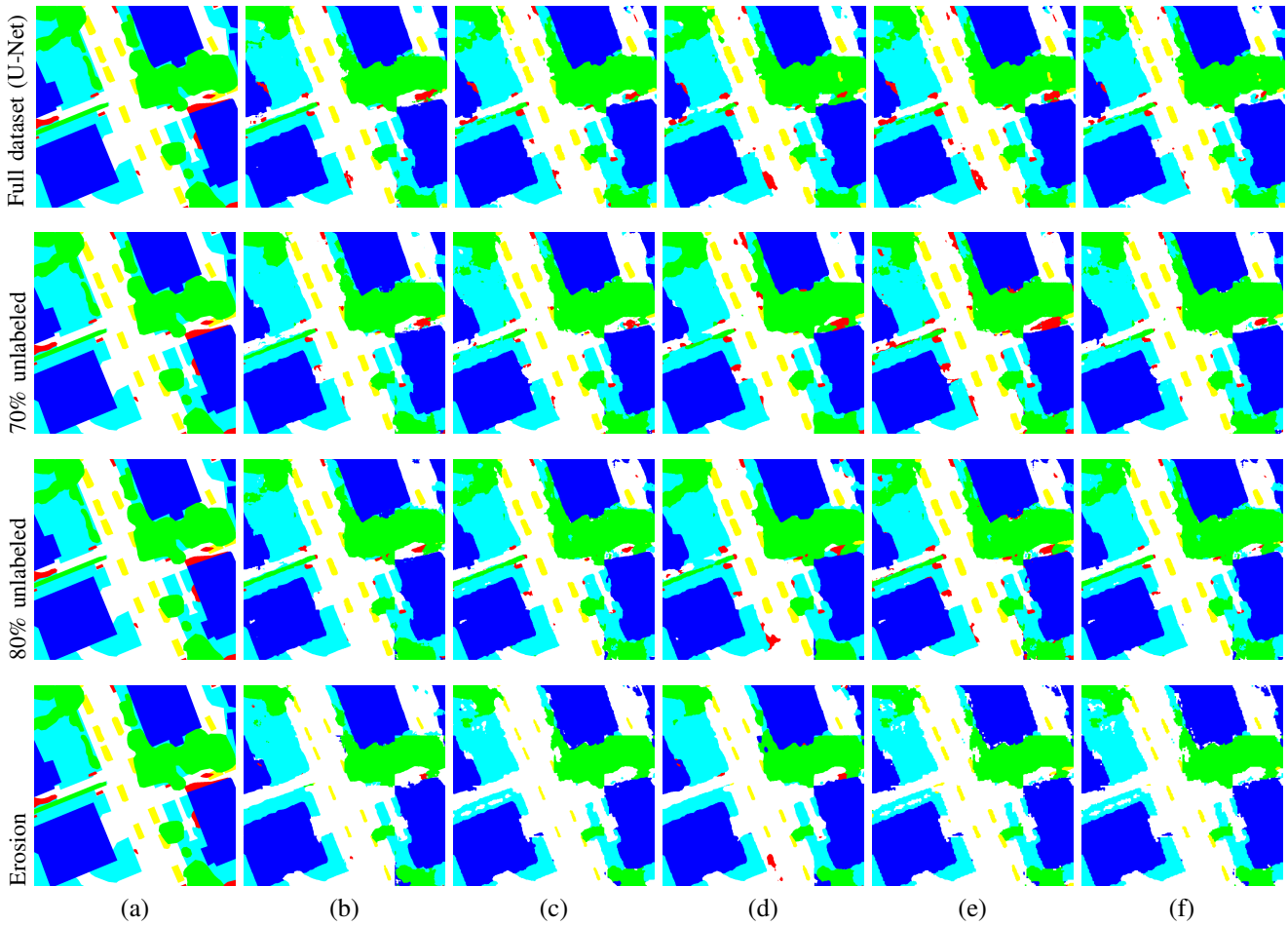


Fig. 6. **Ground truth and classification maps for the Potsdam dataset:** (a) ground truth and classification maps obtained with (b) U-Net, and the proposed method in versions (c) "IRRG-RF", (d) "PGM+Net", (e) "Net for cars", and (f) "resize". Classes: buildings (blue), impervious (white), vegetation (cyan), trees (green), cars (yellow), clutter (red).

posteriors obtained by RF for class 5, "cars", on the finest lattice (1024 × 1024 pixels) did not appear to be detailed enough. This result is related to the fact that on this lattice, there are no network activations, and the RF classifier was only applied to the original image data (thus, only three image channels – a configuration indicated hereafter as "IRRG-RF"). One way to overcome this problem is to substitute these posterior probabilities with those obtained in the output layer of the network ("PGM+Net"), or, focus only on the posterior probabilities of class 5, either substituting the RF estimation with the network posteriors ("Net for cars"), or upscale the same class in lattice 512 × 512 ("resize"). The rationale of the last two formulations is to allow the proposed approach to focus on the discrimination of the minority classes (e.g., trees and cars in this dataset).

This problem occurred significantly less in the case of the Potsdam dataset, given that the quadtree presents the network activations even in the base (finest) layer. However, the same tests were performed on this dataset for completeness.

Focusing on the case of 70% of removed labeled pixels in rectangular blocks, we also performed a further comparison with the results of the recently proposed "FESTA"[4] method, where the problem of FCN training with a "scribbled" GT is addressed through a loss term that favors regularization in the spatial and feature domains [45]. It was not feasible to train the architecture developed for FESTA on a large dataset such as Potsdam, using the aforementioned equipment. Consequently, we proceeded with a nearest neighbor resize of the 6000 × 6000 images to 2000 × 2000 pixels. To ensure consistency in the results, the proposed architecture was trained on these resized images. The results of the comparison are shown in Table IV. The hyperparameters of FESTA were set to the values indicated by the authors in the publicly available code of the method.

Finally, focusing on the case of the Vaihingen dataset, the results of the developed method were also compared with those obtained by the light-weight attention network (LWN-Attention) in [44]. This technique is based on a multiscale feature fusion approach and makes use of multiscale information through the concatenation of feature maps associated with different scales [44]. An additional experimental comparison was also performed with DeepLabV3+ [51] in the case of

[4]https://github.com/Hua-YS/Semantic-Segmentation-with-Sparse-Labels

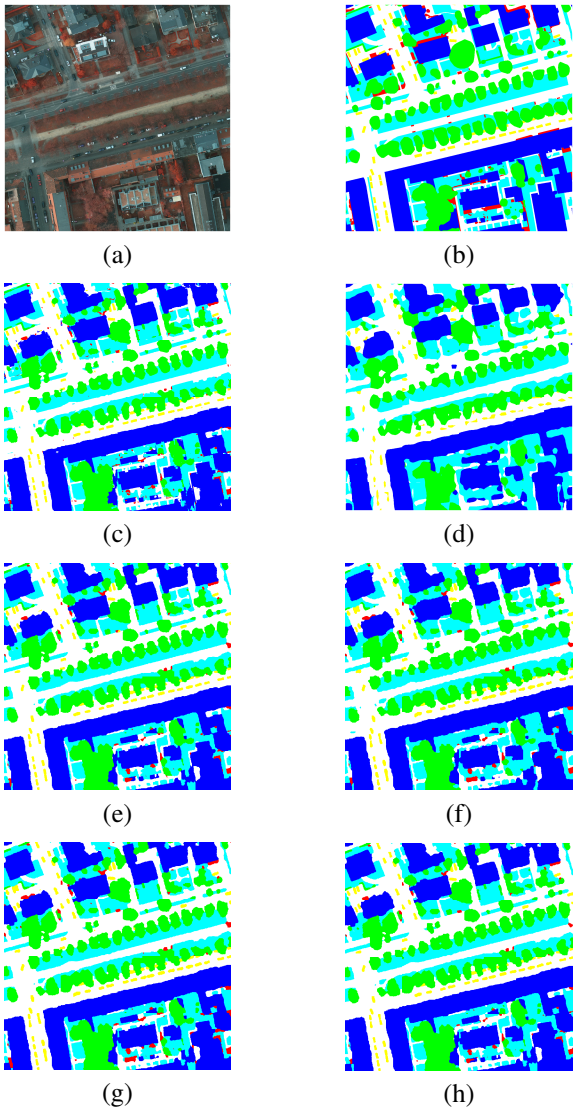


Fig. 7. **Ground truth and classification maps for the Potsdam dataset:** (a) false color composition of the IRRG channels, (b) original ground truth; classification maps: (c) U-Net, (d) FESTA [45] and the proposed method in version (e) "IRRG-RF", (f) "PGM+Net", (g) "Net for cars", (h) "resize". Classes: buildings (blue), impervious (white), vegetation (cyan), trees (green), cars (yellow), clutter (red).

the Vaihingen dataset and of the full GT. The results of this experiment with DeepLabV3+, which were similar to those achieved by U-Net and SegNet, are omitted for brevity and can be found in [52].

C. Results

The results shown in this Section refer to the experiments conducted with the following training conditions: 70% of removed labeled pixels in rectangular blocks; 80% of randomly removed labeled pixels; morphological erosion; and an incomplete GT where 70% of the connected components of the training map were removed and the remaining components were further morphologically eroded. The case in which the full exhaustive ground truth was used was also considered as a baseline. These different training maps brought different results. For example, in the case of the removal of a percentage of labeled pixels randomly, the relative frequency of all the classes and some of the spatial borders between the classes were maintained: the standard U-Net trained with such maps was still able to characterize the shape of the objects in the images but tended to assign almost random class labels to some pixels. Morphological erosion especially affects the minority classes. Accordingly, U-Net trained on morphologically eroded ground truths failed to recognize some of the pixels belonging to classes "cars" and "trees" in the input images. A part of the spatial information between class borders was lost in the GTs where labeled pixels were removed in rectangular blocks, and this was further intensified in the case of the removal of the connected components. Therefore, the output maps generated by U-Net, when trained on these labeled maps, tended to lose some class boundaries and did not appear completely visually regular.

The quantitative results for the Vaihingen dataset reported in Table I show that the proposed approach, when applied with U-Net as the backbone, exhibits remarkable improvements in the classification of minority classes (for example the class "cars"), especially when the input training data approach typical sparse GTs available for real-world land cover mapping applications. In the case of morphological erosion, for example, while U-Net scored a recall of 0.26 in the classification of the cars, the proposed method was able to reach 0.76. Moreover, the recalls attained by the proposed approach were generally higher than those of the standard FCNs. For example, with 70% of unlabeled pixels there is an overall improvement in accuracy for the classification, especially noticeable for the class "vegetation", with an increase of 21% with respect to the results obtained with the simple U-Net. This suggests the capability of the hierarchical PGM to mitigate the impact of sparse GTs by considering the multiscale and long-range spatial dependencies among the pixels.

When full GT is provided, the F1 scores of the proposed method tend to be lower than those of the standard U-Net. Indeed, due to the aforementioned RAM limitations, the part of the proposed technique that consists of the hierarchical PGM and the ensemble learning component is trained and applied only on sections of the input images, while the neural network is trained on the whole dataset. Therefore, when full GT information is provided, the standard U-Net can take benefit from a larger input dataset than the proposed technique. From this perspective, it is interesting to note that, notwithstanding these computational RAM limitations, the proposed approach obtained improved performance in terms of recall and of discrimination of minority classes. This is interpreted as due to the modeling of spatial-contextual information and long-range dependencies through the proposed PGM approach.

Fig. 4 shows the ground truth used as the test set and the mapping results generated by U-Net [9] and by our model with its variants. The results of U-Net are depicted in Fig. 4(b), while those from the proposed approach in its variants are in Figs. 4(c)-(f). For example, with 70% of unlabeled pixels in rectangular blocks, the comparison between the images obtained with the network and with the proposed method shows that the full proposed architecture does a better job in

the semantic segmentation of the edges between the classes, recovering spatial information that was lost in the prediction of U-Net through the hierarchical Markov model. In particular, the output maps generated by the proposed algorithm are quite visually regular without exhibiting spatial oversmoothing. The method also recovers class boundaries that were lost in the prediction of U-Net – and this further confirms its capability to mitigate the impact of scarce GT. In Fig. 5, it is possible to see a highlight of the classification errors and improvements made by the proposed technique compared to the standard U-Net.

In general terms, the two variants of the method focusing on the smallest class, "Net for cars" and "resize", exhibit some advantages in the discrimination of the minority classes (see Tables I, II). The approach employing the posteriors computed by the FCN on the base layer of the quadtree, "PGM+Net", shows some consistent advantages in the discrimination of the class "vegetation", since it is the only variant containing information extracted by the network for this class on the base layer of the quadtree. Therefore, thanks to the data representation extracted by the FCN trained on multiple patches (and not only one, as the RF classifier), the proposed method is able to improve the discrimination of pixels belonging to low vegetation from those belonging to trees.

Concerning the Potsdam dataset, the overall results (shown in Table II) of U-Net by itself are more precise than those obtained on the Vaihingen dataset (comparable with the results recently shown in [53], [54]). However, as the GTs become scarcer, the proposed architecture progressively obtains higher recalls for the smallest classes (e.g., "trees" and "cars"). In all the considered situations, the proposed approach surpasses or equals the accuracy of the standard FCN as per the recall.

In Fig. 6, the comparison on the Potsdam dataset between the images obtained using U-Net (Fig. 6(b)) and with the proposed method (Figs. 6(c)-(f)) clearly shows that the full proposed approach does a better job in the segmentation of the small classes, such as "cars", once more suggesting the effectiveness of incorporating the spatial information obtained through the hiearchical architecture in the original predictions of the FCNs.

In the case of the full GT and of the Vaihingen dataset, the use of SegNet was effective, both *per se* and in conjunction with the proposed approach, although with generally lower accuracies than when U-Net was employed. Accordingly, for the sake of brevity, the accuracies obtained using SegNet are shown only in this case and not in the case of the degraded training GTs or of the Potsdam dataset. Similarly, when using the full GT, the results reached by HRNet [11], both by itself and as a backbone in the proposed approach, are slightly less accurate than those obtained using U-Net. Tables I and II also show the results achieved using HRNet in the case of the removal of the connected components. In this case as well, the use of HRNet led to quite accurate results, although the use of U-Net allowed for slightly better performance. The same comments apply as above, as the addition of the PGM ameliorates the classification results for the minority classes thanks to the integration of spatial information, however with slightly lower effectiveness than in the results attained using U-Net. On one hand, these results confirm the discrimination capability of HRNet within the addressed semantic segmentation problem. On the other hand, the improvements obtained by the proposed approach also with HRNet confirm its effectiveness in modeling spatial information and long-range dependencies as well as its applicability in conjunction with multiple underlying FCN architectures.

The proposed technique was compared to the recently proposed "FESTA" method. The results suggest that FESTA also mitigated the impact of suboptimal GT on the training process (see Tables I and IV), although the proposed approach, across its variants, obtained higher or the same per-class accuracy (especially for "cars") and required significantly shorter computation times (see Table III). The results confirm that the proposed approach tends to improve the discrimination of the small classes and the average recall.

The results of the comparison between the proposed architecture and the FESTA model on the Potsdam dataset in the case of 70% of removed labeled pixels in rectangular blocks are shown in Table IV. As mentioned in Section IV-B, to perform this comparison, it was necessary to resize the images to 2000×2000 pixels (i.e., to a spatial resolution of 15 cm) due to memory limitations. For the sake of fairness, the results of both FESTA and the proposed technique refer to these resized images. In general, on this resized dataset, the proposed method provided semantic segmentation maps with higher accuracy than both U-Net [9] and FESTA [45] for almost all of the classes. Consequently, the values for the averaged metrics for the proposed architecture were the highest among the considered approaches. The classification maps are shown in Fig. 7

These results confirm the capability of the proposed approach to exploit the spatial modeling ability of hierarchical Markov models to mitigate the limitations of DL approaches in the case of poor training data. In this challenging case, the developed method outperforms the state-of-the-art DL methods in the discrimination of small classes, while maintaining quite accurate classification results for all classes.

The comparison with the LWN-Attention multiscale feature fusion method in [44], which is aimed at the semantic segmentation of VHR images, corroborates the comments made above. This previous approach provided accurate classification results, as well. Indeed, in the case of scarcer GTs, the proposed technique, leveraging both hierarchical and long-range information through the combination of hierarchical and planar MRFs, obtained generally higher average classification results. This suggests that, at least on the considered datasets, incorporating multiscale information through the proposed integration of FCN and hierarchical PGM can be advantageous as compared to the direct concatenation of the feature maps.

V. Conclusion

In this paper, we have addressed the critical problem of the impact of spatially sparse ground truth on semantic segmentation maps obtained by fully convolutional networks by proposing a novel approach to the semantic segmentation of remote sensing images based on CNNs, hierarchical PGMs,

and decision tree ensembles. The rationale is to leverage the spatial modeling capabilities of hierarchical PGMs and the multiscale data representation extracted by FCNs to mitigate the impact of incomplete ground truth and obtain accurate classification results in scenarios where exhaustive ground truth does not exist. The integration was performed through a quadtree topology.

The reported results show that the proposed approach, combined with different backbone networks, surpasses the accuracy of the standard FCNs as per the recall, thus suggesting the capability of this approach to exploit spatial information through the hierarchical Markov model and mitigate the limitations of the neural networks trained with a scarce dataset. The new method outperforms the state of the art especially in the discrimination of minority classes, while maintaining quite accurate classification results for all classes. For example, the proposed method guarantees an improvement of 21% for the classification of class "vegetation" in the case of training with ground truths with 70% of removed labeled pixels in rectangular blocks (for the Vaihingen dataset, in combination with U-Net). For the Potsdam dataset, for instance in the case of morphological erosion, the proposed method was able to recover 16% more pixels belonging to class "cars" compared to a U-Net backbone.

All these advantages are more remarkable the more the training set approaches a realistic scenario of spatially sparse ground truth. Compared to a previous multiscale feature fusion approach, the proposed technique obtains improved performances in the case of scarcer training set. Compared to the previous FESTA algorithm, the proposed method is also advantageous in terms of computational burden.

Perspectives for future work involve the introduction of feed-forward layers to compute the pixelwise posterior probabilities instead of the RF classifier and develop an end-to-end neural architecture without ensemble learning components. Furthermore, it would be interesting to integrate the proposed method with transfer learning [55] and test it with various datasets related to real-world applications, such as disaster management [56], with different complexity and features, to further investigate its generalization capabilities and compare them to those of a standard FCN in the same framework.

ACKNOWLEDGMENT

The authors would like to thank the authors of [45] for making available the code of the "FESTA" method to perform the comparisons and the authors of the codes of HRNet [11] and LWN-Attention [44]. The Vaihingen and Potsdam datasets were provided by the German Society for Photogrammetry, Remote Sensing and Geoinformation (DGPF) [57]: http://www.ifp.uni-stuttgart.de/dgpf/DKEP-Allg.html

REFERENCES

[1] N. Longbotham, F. Pacifici, S. Malitz, W. Baugh, and G. Camps-Valls, "Measuring the spatial and spectral performance of worldview-3," in *Hyperspectral Imaging and Sounding of the Environment (HISE)*, 2015.

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

[3] C. Pohl and J. van Genderen, "Remote sensing image fusion: an update in the context of digital earth," *International Journal of Digital Earth*, vol. 7, no. 2, pp. 158–172, 2014.

[4] D. Marmanis, J. Wegner, S. Galliani, K. Schindler, M. Datcu, and U. Stilla, "Semantic segmentation of aerial images with an ensemble of cnns," *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. III-3, pp. 473–480, 06 2016.

[5] 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.

[6] 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.

[7] N. Audebert, A. Boulch, H. Randrianarivo, B. Saux, M. Ferecatu, S. Lefèvre, and R. Marlet, "Deep learning for urban remote sensing," in *Proceedings of the 2017 Joint Urban Remote Sensing Event (JURSE)*, Dubai, United Arab Emirates, pp. 1–4, 2017.

[8] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 3431–3440, 2015.

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

[10] 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.

[11] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5686–5696.

[12] M. Kampffmeyer, A. Salberg, and R. Jenssen, "Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks," in *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Los Alamitos, CA, USA: IEEE Computer Society, pp. 680–688, 2016.

[13] 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.

[14] L. Maggiolo, D. Marcos, G. Moser, and D. Tuia, "Improving maps from cnns trained with sparse, scribbled ground truths using fully connected crfs," in *Proceedings of the 2018 IEEE International Geoscience and Remote Sensing Symposium*, Valencia, Spain, pp. 2099–2102, 2018.

[15] X. Sun, B. Wang, Z. Wang, H. Li, H. Li, and K. Fu, "Research progress on few-shot learning for remote sensing image interpretation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 2387–2402, 2021.

[16] Y. Li, T. Shi, W. Chen, Z. Wang, and H. Li, "Learning deep semantic segmentation network under multiple weakly-supervised constraints for cross-domain remote sensing image semantic segmentation," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 175, pp. 20–33, 05 2021.

[17] X. Wang, "Deep learning in object recognition, detection, and segmentation," *Foundations and Trends in Signal Processing*, vol. 8, no. 4, pp. 217–382, 2016.

[18] A. Arnab, S. Zheng, S. Jayasumana, B. Romera-Paredes, M. Larsson, A. Kirillov, B. Savchynskyy, C. Rother, F. Kahl, and P. Torr, "Conditional random fields meet deep neural networks for semantic segmentation: Combining probabilistic graphical models with deep learning for structured prediction," *IEEE Signal Processing Magazine*, vol. 35, pp. 37–52, 2018.

[19] 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.

[20] S. Nowozin and C. H. Lampert, "Structured learning and prediction in computer vision," *Foundations and Trends in Computer Graphics and Vision*, vol. 6, no. 3–4, p. 185–365, 2011.

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

[22] 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.

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

[24] 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.

[25] 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, p. 849, 2021.

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

[27] M. C. Zhang, Y., *Ensemble Machine Learning: Methods and Applications*, 1st ed. New York: Springer-Verlag, 2012.

[28] 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.

[29] X. X. Zhu, D. Tuia, L. Mou, G. Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4, pp. 8–36, 2017.

[30] Q. He, X. Sun, Z. Yan, and K. Fu, "Dabnet: Deformable contextual and boundary-weighted network for cloud detection in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022.

[31] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Hypercolumns for object segmentation and fine-grained localization," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 447–456, 2015.

[32] W. Zhao, S. Du, Q. Wang, and W. Emery, "Contextually guided very-high-resolution imagery classification with semantic segments," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 132, pp. 48–60, 2017.

[33] W. Zhao, W. J. Emery, Y. Bo, and J. Chen, "Land cover mapping with higher order graph-based co-occurrence model," *Remote Sensing*, vol. 10, 1713, no. 11, 2018.

[34] D. Buscombe and A. C. Ritchie, "Landscape classification with deep neural networks," *Geosciences*, vol. 8, 244, no. 7, 2018.

[35] Y. Liu, S. Piramanayagam, S. Monteiro, and E. Saber, "Semantic segmentation of multisensor remote sensing imagery with deep convnets and higher-order conditional random fields," *Journal of Applied Remote Sensing*, vol. 13, pp. 1–23, 2019.

[36] S. Paisitkriangkrai, J. Sherrah, P. Janney, and A. van den Hengel, "Semantic labeling of aerial and satellite imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 7, pp. 2868–2881, 2016.

[37] L. Zhu, L. Huang, L. Fan, J. Huang, F. Huang, J. Chen, Z. Zhang, and Y. Wang, "Landslide susceptibility prediction modeling based on remote sensing and a novel deep learning algorithm of a cascade-parallel recurrent neural network," *Sensors*, vol. 20, no. 6, 2020.

[38] G. Fu, C. Liu, R. Zhou, T. Sun, and Q. Zhang, "Classification for high resolution remote sensing imagery using a fully convolutional network," *Remote Sensing*, vol. 9, no. 5, 2017.

[39] C. Zhang, I. Sargent, X. Pan, A. Gardiner, J. Hare, and P. M. Atkinson, "Vprs-based regional decision fusion of cnn and mrf classifications for very fine resolution remotely sensed images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 8, pp. 4507–4521, 2018.

[40] Y. Yang, Y. Zhuang, F. Bi, H. Shi, and Y. Xie, "M-fcn: Effective fully convolutional network-based airplane detection framework," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 8, pp. 1293–1297, 2017.

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

[42] A. Dunmur and D. Titterington, "Computational bayesian analysis of hidden markov mesh models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 11, pp. 1296–1300, 1997.

[43] 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.

[44] S. Liu, C. He, H. Bai, Y. Zhang, and J. Cheng, "Light-weight attention semantic segmentation network for high-resolution remote sensing images," in *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, 2020, pp. 2595–2598.

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

[46] N. Audebert, B. Le Saux, and S. Lefèvre, "Beyond RGB: Very High Resolution Urban Remote Sensing With Multimodal Deep Networks," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 140, pp. 20–32, 2018.

[47] Q. Liu, M. Kampffmeyer, R. Jenssen, and A.-B. Salberg, "Dense dilated convolutions' merging network for land cover classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 9, pp. 6309–6320, 2020.

[48] L. Lv, Y. Guo, T. Bao, C. Fu, H. Huo, and T. Fang, "Mfalnet: A multiscale feature aggregation lightweight network for semantic segmentation of high-resolution remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, pp. 1–5, 2020.

[49] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 2014.

[50] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *ArXiv:1409.1556*, 2015.

[51] 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.

[52] M. Pastorino, G. Moser, S. B. Serpico, and J. Zerubia, "Hierarchical probabilistic graphical models and deep convolutional neural networks for remote sensing image classification," in *2021 29th European Signal Processing Conference (EUSIPCO)*, 2021, pp. 1740–1744.

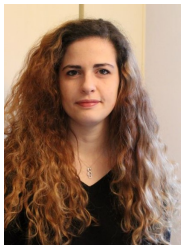
[53] A. Li, L. Jiao, H. Zhu, L. Li, and F. Liu, "Multitask semantic boundary awareness network for remote sensing image segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–14, 2021.

[54] X. Yang, S. Li, Z. Chen, J. Chanussot, X. Jia, B. Zhang, B. Li, and P. Chen, "An attention-fused network for semantic segmentation of very-high-resolution remote sensing imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 177, pp. 238–262, 2021.

[55] O. Tasar, S. L. Happy, Y. Tarabalka, and P. Alliez, "Semi2i: Semantically consistent image-to-image translation for domain adaptation of remote sensing data," in *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, pp. 1837–1840, 2020.

[56] S. B. Serpico, S. Dellepiane, G. Boni, G. Moser, E. Angiati, and R. Rudari, "Information extraction from remote sensing images for flood monitoring and damage evaluation," *Proceedings of the IEEE*, vol. 100, no. 10, pp. 2946–2970, 2012.

[57] M. Cramer, "The dgpf-test on digital airborne camera evaluation overview and test design," *Photogrammetrie - Fernerkundung - Geoinformation*, vol. 2010, no. 2, pp. 73–82, 2010.



Martina Pastorino received the B.Sc. degree in electronic engineering and information technologies in 2018 from the University of Genoa, Italy and a double M.Sc. degree in telecommunication engineering from the University of Genoa, Italy, and IMT Atlantique, Brest, France in 2020. She is currently pursuing a joint Ph.D. degree in science and technology for electronic and telecommunication engineering between the University of Genoa, Italy and Inria d'Université Côte d'Azur, Sophia-Antipolis, France. In 2021 she received the Best Student Paper Award at IGARSS 2021 and the Prix d'excellence d'Université Côte d'Azur. Her research activity is focused on the combination of stochastic models and deep learning techniques for remote sensing image analysis.

Gabriele Moser (S'03–M'05–SM'14) received the Laurea (M.Sc. equivalent) degree in telecommunications engineering, and the Ph.D. degree in space sciences and engineering from the University of Genoa, Italy, in 2001 and 2005, respectively. He is a Full Professor of Telecommunications at the University of Genoa. Since 2001, he has cooperated with the Image Processing and Pattern Recognition for Remote Sensing laboratory of the University of Genoa. Since 2013, he has been the Head of the Remote Sensing for Environment and Sustainability laboratory at the Savona Campus of the University of Genoa. From January to March 2004, he was a Visiting Student with the Institut National de Recherche en Informatique et en Automatique (INRIA), Sophia Antipolis, France. From 2012 to 2016, he was an external collaborator of the Ayin laboratory at INRIA. In 2016, he spent a period as Visiting Professor at the Institut National Polytechnique de Toulouse, France. His research activity is focused on pattern recognition and image processing methodologies for remote sensing and energy applications. He has been an Associate Editor of the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS since 2008. He was an Area Editor of PATTERN RECOGNITION LETTERS (PRL) from 2015 to 2018, an Associate Editor of PRL from 2011 to 2015, and a Guest Co-Editor of the September 2015 special issue of the IEEE GEOSCIENCE AND REMOTE SENSING MAGAZINE. He served as Chair of the IEEE GRSS Image Analysis and Data Fusion Technical Committee (IADF TC) from 2013 to 2015, and as IADF TC Co-Chair from 2015 to 2017. He was Publication Co-Chair of the 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Technical Program Co-Chair of the IEEE GRSS EARTHVISION workshop at the 2015 IEEE/CVF Computer Vision and Pattern Recognition conference (CVPR), and Co-Organizer of the second edition of EARTHVISION at CVPR 2017. He received the Best Paper Award at the 2010 IEEE Workshop on Hyperspectral Image and Signal Processing and the Interactive Symposium Paper Award at IGARSS 2016. Since 2019, he has been the Head of the M.Sc. program in Engineering for Natural Risk Management at the University of Genoa.

Sebastiano B. Serpico Fellow of the IEEE, full professor of telecommunications at the Polytechnic School of the University of Genoa, he received the Laurea (M.S.) degree in electronic engineering and the Ph.D. degree in telecommunications from the University of Genoa, Italy. He is the Coordinator of the research group on Signal Processing and Recognition Methods and Systems of the Department of Electrical, Electronic, Telecommunications Engineering, and Naval Architecture (DITEN) of the University of Genoa; he is the Coordinator of the Information and Communication Technologies section of DITEN. His current research interests are mainly in pattern recognition for remote sensing image analysis. He was the Chairman of the Institute of Advanced Studies in Information and Communication Technologies (ISICT) from 2003 to 2019. He has been the project manager of numerous research projects and an evaluator of project proposals for various programs of the European Union, Italian Space Agency, Italian Ministry of Education and Research, etc. He is the author (or coauthor) of over 250 scientific articles published in journals and conference proceedings. He received the Education Award from the IEEE Geoscience and Remote Sensing Society in 2019, the Interactive Symposium Paper Award at the IEEE IGARSS in 2016, and the Best Paper Award at the IEEE Workshop on Hyperspectral Image and Signal Processing in 2010. He is an associate editor of the international journal IEEE Transactions on Geoscience and Remote Sensing (TGRS). He coedited two Special Issues of TGRS on Analysis of Hyperspectral Image Data (July 2001) and on Advances in Techniques for Analysis of Remotely Sensed Data (March 2005), respectively, and a special issue of the Proceedings of the IEEE on Remote Sensing of Natural Disasters. From 1998 to 2002, he was the chairman of the SPIE/EUROPTO series of conferences on Signal and Image Processing for Remote Sensing. He was Co-Chair of the IEEE International Geoscience and Remote Sensing Symposium in 2015 (Milan, Italy).

Josiane Zerubia has been a permanent research scientist at INRIA since 1989 and director of research since July 1995 (DR 1st class since 2002). She received the MSc degree from the Department of Electrical Engineering at ENSIEG, Grenoble, France in 1981, the Doctor of Engineering degree, her PhD and her 'Habilitation', in 1986, 1988, and 1994 respectively, all from the University of Nice, France. She was head of the PASTIS remote sensing laboratory (INRIA Sophia-Antipolis) from mid-1995 to 1997 and of the ARIANA research group (INRIA/CNRS/University of Nice), which worked on inverse problems in remote sensing and biological imaging, from 1998 to 2011. From 2012 to 2016, she was head of AYIN research group (INRIA-SAM) dedicated to models of spatio-temporal structure for high-resolution image processing with a focus on remote sensing and skincare imaging. She is head of AYANA exploratory research group since 2020. AYANA is an interdisciplinary project using knowledge in stochastic modeling, image processing, artificial intelligence, remote sensing and embedded electronics/computing. She was full professor (PR1) at SUPAERO (ISAE) in Toulouse from 1999 to 2020. She received a Doctor Honoris Causa degree from the University of Szeged in Hungary in November 2021, and 3 times the Excellence Award from University of Nice (now called Université Côte d'Azur) in 2016, 2019 and 2020. She is a Fellow of the IEEE (2003-), the EURASIP (2019-) and the IAPR (2020-), and IEEE SP Society Distinguished Lecturer (2016-2017). She has been a member of the editorial boards of the Foundation and Trends in Signal Processing since 2007 and of the IEEE Signal Processing Magazine since September 2018. She has been member-at-large of the Awards Board of the IEEE SP Society since 2020 and of the IAPR Fellow Committee since 2021. Finally, she has been a member of the Best Paper Award Committee for EURASIP JIVP in 2021. She was part of the organizing committees of the workshop EarthVision (co-chair) at IEEE CVPR 2017 (Honolulu, USA) and GRETSI 2017 symposium (Juan les Pins, France). She is scientific advisor and co-organizer of ISPRS 2020 (virtual), 2021 (virtual) and 2022 congress (Nice, France) and technical co-chair of IEEE-EURASIP EUSIPCO 2021 (virtual, Dublin, Ireland). Her main research interest is in image processing using probabilistic models. She also works on parameter estimation, statistical learning and optimization techniques, and artificial intelligence. For more information see http://www-sop.inria.fr/members/Josiane.Zerubia/index-eng.html