CRFNet: A Deep Convolutional Network to Learn the Potentials of a CRF for the Semantic Segmentation of Remote Sensing Images

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

***Abstract*— This article presents a method for the automatic learning of the potentials of a stochastic model, in particular a conditional random field (CRF), in a non-parametric fashion. The proposed model is based on a neural architecture, in order to leverage the modeling capabilities of deep learning (DL) approaches to directly learn semantic and spatial information from the input data. Specifically, the methodology is based on fully convolutional networks and fully connected neural networks. The idea is to access the multiscale information intrinsically extracted in the intermediate layers of a fully convolutional network through the integration of fully connected neural networks at different scales, while favoring the interpretability of the hidden layers as posterior probabilities. The potentials of the CRF are learned through an additional convolutional layer, whose kernel models the local spatial information considered. The loss function is computed as a linear combination of cross-entropy losses, accounting for the multiscale and the spatial information. To evaluate the capabilities of the proposed approach for the semantic segmentation of remote sensing images, the experimental validation was conducted with the ISPRS 2-D semantic labeling challenge Vaihingen and Potsdam datasets and with the IEEE GRSS data fusion contest Zeebruges dataset. As the ground truths of these benchmark datasets are spatially exhaustive, they have been modified to approximate the spatially sparse ground truths common in real remote sensing applications. The results are significant, as the proposed approach obtains higher average classification accuracies than recent state-of-the-art techniques considered in this article. The code is available at https://github.com/Ayana-Inria/CRFNet-RS.**

***Index Terms*— Conditional random fields (CRFs), convolutional neural network (CNN), fully convolutional network (FCN), remote sensing, semantic segmentation.**

I. Introduction

Semantic segmentation—or dense image classification—is the task of assigning a label category to each pixel in an image. When dealing with very high-resolution (VHR) remote sensing images, the results of semantic segmentation tasks may be of use in several real-world applications [1], such as land cover mapping [2], urban planning and management [3], [4], and traffic monitoring [5], [6], [7]. In this framework, several techniques, ranging from stochastic models to deep learning (DL) architectures, have been employed.

DL techniques are currently the state of the art for semantic segmentation tasks. Several DL methods have been proposed to perform remote sensing image classification [8] and, among these architectures, fully convolutional networks [9] play a primary role. These networks can produce classification maps for inputs of arbitrary size and are able to recover the spatial information lost along the downsampling operations through the addition of upsampling blocks to the standard convolutional neural network (CNN), built through the combination of unpooling and deconvolution layers [8], [9]. Fully convolutional networks are defined by an architecture that combines semantic information from deep coarse layers with appearance information to produce precise segmentation maps [8], [9].

However, these models usually require large datasets with exhaustively labeled ground truths in order to obtain accurate segmentation results. These spatially dense ground truths are very costly and time-consuming to produce. They are rarely feasible for real-world remote sensing mapping applications [10], [11], [12], where the available information is usually spatially sparse, not representing spatial borders between different semantic entities. This lack of exhaustive ground-truth maps strongly affects the classification results of DL techniques and is a major challenge in the development of supervised classification methods for remote sensing image analysis.

At the same time, stochastic models such as probabilistic graphical models (PGMs) are popular and powerful tools for computer vision and image processing tasks, as they can be employed for structured prediction problems, such as image restoration [13], [14], image denoising [15], [16], semantic segmentation, and object detection [17], [18], [19], [20]. These models, for example, Bayesian networks [21] and random fields [22], [23], [24], [25], characterize a structured output based on graph representations to express a dependency structure between random variables over a multidimensional space.

In particular, for 2-D image analysis, random fields, such as Markov random fields (MRFs) and conditional random fields (CRFs) [26] are capable of modeling spatial and

Manuscript received 25 March 2024; revised 24 July 2024; accepted 23 August 2024. Date of publication 30 August 2024; date of current version 19 September 2024. *(Corresponding author: Gabriele Moser.)*

Martina Pastorino is with the Department of Electrical, Electronic, Telecommunications Engineering and Naval Architecture, University of Genoa, 16145 Genoa, Italy, and also with the Institut National de Recherche en Informatique et en Automatique (INRIA), Université Côte d'Azur, 06902 Sophia Antipolis, France (e-mail: martina.pastorino@edu.unige.it).

Gabriele Moser and Sebastiano 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).

Josiane Zerubia is with Institut National de Recherche en Informatique et en Automatique (INRIA), Université Côte d'Azur, 06902 Sophia Antipolis, France.

Digital Object Identifier 10.1109/TGRS.2024.3452631

multiresolution information, according to the underlying graph topology. For these models, in general, the concept of spatial information (which can be planar [13], multiresolution [24], [27], or possibly both [17], [28]) is formulated in relation to a neighborhood system of each node in the graph.

In this article, we present a technique to automatically learn the unary and pairwise potentials of a CRF model through a deep convolutional network. As the methodology is completely based on DL architectures, it is non-parametric and capable of directly learning the statistics of the stochastic model from the input data. The family of CRF models that can be learned through the proposed approach is broad and flexible, as it includes all CRFs with up to pairwise potentials and a local smoothing condition, without restricting to any parametric family. In particular, the proposed architecture leverages the definition of convolutional layers and their kernels to formalize the relation between a deep convolutional network and a CRF with pairwise potentials and a smoothing term. This relation is analytically proven in terms of an equivalence theorem under suitable assumptions.

The model also explicitly takes into account multiresolution information, through the manipulation of the features extracted by the neural network at different resolutions in a sort of global-to-local information pyramid managing both coarse semantic knowledge and fine details [9], thus mining the diverse semantics typical of VHR remote sensing images [29]. The multiscale component of the model, introduced with the addition of fully connected layers at different blocks of a fully convolutional network, thus at different scales, favors the interpretability of the hidden layers of the fully convolutional network itself as posterior probabilities. The goal of the automatically learned CRF is to address semantic segmentation tasks on remote sensing images.

In this respect, the main novel contributions of this article are threefold: 1) the definition of an end-to-end fully neural architecture to automatically learn up to the second-order potentials of a CRF model for semantic segmentation; 2) the development of a semantic segmentation algorithm for remote sensing images, which integrates the automatically learned CRF potentials with the multiscale—and possibly complementary—information extracted by a deep convolutional network; and 3) the analytical proof of the equivalence theorem relating the network output and the CRF.

The article is organized in the following way. Section II provides an overview of the state of the art with respect to DL methods for the definition of PGMs, focusing in particular on the models for semantic segmentation. Section III presents the proposed methodology. The results of the experimental validation conducted with the proposed framework and the comparison with the results obtained by state-of-the-art techniques are described and discussed in Section IV. Finally, conclusions and perspectives of the proposed technique are reported in Section V.

II. Previous Work

Previous approaches to learn a PGM through a neural architecture are reviewed in this section.

Many techniques combining these two different frameworks, probabilistic graphical and DL models, have been presented [30], [31] for various image processing tasks. For example, some methods have been developed for image-denoising applications [32]. According to [32], CNNs can be viewed as a generalization of MRFs in applications to image restoration, where the objective is to reconstruct an original image starting from a noisy measurement, typically characterized by additive Gaussian noise with zero mean.

Methodologies were also developed in the framework of multiclass classification. For example, in [33], a generic maximum likelihood estimation procedure is proposed for MRFs, whose potential functions are modeled by neural networks, in particular fully convolutional networks.

In [34], a deep architecture to label 3-D shape parts by considering both spectral and geometric features via a framework consisting of a CNN and a CRF was implemented. First, low-level features are used to learn deep features using a CNN model, and then formulate the deep CRF model to effectively extract the semantic correlations between adjacent triangles on the mesh. In this case, the unary energies are obtained from the CNN and the pairwise term of the CRF are formulated based on geodesic distances and angles between surface normal. In this approach, the 2-D CNN model is used to predict the probability distribution of each face independently of its neighbors, and the CRF inference using mean-field approximation makes a refinement by taking the output probabilities of the CNN as the unary term [34].

Various techniques have been proposed for the joint training of a CRF and a CNN, for example, by embedding the CRF in memory networks, such as recurrent neural networks (RNNs) [30], [35], [36]. In [35], a multiobject tracking framework aiming to model the assignment costs as unary potentials and the long-term dependencies among detection results as pairwise potentials of a deep CRF is presented. The CRF inference is defined as an RNN, and the unary and pairwise components are pretrained separately.

Several models were also proposed in the framework of semantic segmentation applications [37]. As it was demonstrated in [38], the performances of semantic segmentation algorithms can be improved by using the output of a fully convolutional network as the unary potentials of a CRF model characterized by Gaussian pairwise potentials [39]. In this method, however, the CRF is still used as a post-processing technique, whose parameters are selected with cross-validation [38].

Different approaches allowing to learn a CRF have been developed for semantic segmentation tasks, for instance, through a piecewise CNN-based training [40] or embedding the CRF in RNNs [36]. In [40], two different networks performing multiscale feature fusion (MFF) compute the unary and pairwise potentials of the CRF, in a non-parametric fashion, to improve segmentation results in applications with patch-patch and patch-background information between image regions. Conversely, the technique in [36] shows how mean-field inference of the CRF in [39] can be modeled as an RNN and incorporated into the neural network itself, enabling the joint end-to-end training of both the CNN and

CRF parameters by backpropagation. As the mean-field inference is iterative, it can be unrolled across its time-steps to form the RNN [41]. As compared to the technique proposed in [36], which aims to compute Gaussian CRF potentials up to the second order through a CNN and an RNN approximating the mean-field inference of a fully connected CRF, the present paper introduces a method to compute general non-parametric unary and pairwise potentials. In particular, the proposed method leverages on the definition of the last convolutional layer of a convolutional network and on the impulse response of its kernel. The pairwise potentials are fully non-parametric and include a spatially smoothing term, which favors well-defined properties of the local spatial context in remote sensing images.

Other techniques propose the development of deep continuous or discrete PGMs after passing through a previous over-segmentation, e.g., via superpixels [42], [43], [44]. In [42], a fully connected CRF is modeled through CNNs for both continuous and discrete tasks. After an over-segmentation of the original input images in superpixels, these are used as input of two networks: a unary and a pairwise network, to perform the final prediction. A continuous CRF based on superpixel decomposition was proposed in [43] for saliency detection, where parameters for both unary and pairwise potentials are jointly learned. An input image is first over-segmented into superpixels and a graph is built to capture intrinsic image context. The continuous CRF is defined over this graph. Another method making use of a superpixel decomposition for the semantic segmentation of hyperspectral images considering both spectral and spatial information through CNNs and CRFs is presented in [44]. A mean-field approximation algorithm for CRF inference is used and formulated with Gaussian pairwise potentials as RNNs. This combined network is then plugged into the CNN.

The semantic segmentation of hyperspectral images through spectral and spatial information via a framework consisting of CNN and CRF is addressed in [45], as well. The deep CRF is formulated with two 3-D CNNs computing unary and pairwise potential functions to effectively extract the semantic correlations between patches consisting of 3-D data cubes.

III. Methodology

A. Overview of the Proposed Method

As mentioned in Section I, the idea of the proposed technique is twofold. First, it is to benefit from the flexibility of deep neural architectures to automatically learn the potentials of a CRF in a non-parametric fashion. The focus is on categorical-valued CRF models for semantic segmentation with up to pairwise non-zero potentials. Therefore, the proposed method is aimed at learning the first (unary) and second-order (pairwise) potentials. Second, the architecture is integrated with multiscale information extraction to address semantic segmentation.

In particular, the overall diagram of the proposed method is shown in Fig. 1. The neural network employed is based on the family of the fully convolutional networks, thus intrinsically dealing with multiscale information. In the proposed approach, the architecture of a standard fully convolutional network (represented in block A of Fig. 1) is integrated with additional layers whose goal is to explicitly address the multiscale information, enforcing the interpretability of the network, and to automatically compute the first- and second-order statistics of a PGM. It is important to mention that the fully connected layers (represented in the block B of Fig. 1) are employed to process multiscale information through the computation of the multiscale loss terms during the training phase, but they are inactive during the prediction phase. The classification map is obtained in a fully convolutional manner. In the following, the proposed approach will be denoted as CRFNet. After a brief recall of the basics of CRF models in Section III-B, the detailed description of the proposed neural architecture and its explanation as a CRF are reported in Sections III-C and III-D, respectively.

B. CRF Models for Semantic Segmentation

In the framework of semantic segmentation, CRFs represent a family of PGMs capable to characterize both the spatial dependencies between neighboring pixels and the pixelwise class distributions [46]. Let us consider an image, defined on a rectangular pixel lattice $S \subset \mathbb{Z}^2$, and let us assume that each pixel belongs to one out of M classes. Let $\Omega = \{\omega_1, \omega_2, \ldots, \omega_M\}$ be the set of classes, and let $\boldsymbol{x}_i \in \mathbb{R}^d$ and $y_i \in \Omega$ be the d-dimensional feature vector and the class label of pixel $i \in S$, respectively. Considering $\mathcal{X} = \{\boldsymbol{x}_i\}_{i \in S}$ and $\mathcal{Y} = \{y_i\}_{i \in S}$ as the random fields of the observations and of the class labels, respectively, the random field $\mathcal{Y}$ is a CRF if the following conditional Markovianity property holds:

$$P(y_i|y_j, j \neq i, \mathcal{X}) = P(y_i|y_j, j \in \partial i, \mathcal{X}) \tag{1}$$

where ∂i represents a neighborhood of pixel i; and if $P(\mathcal{Y}|\mathcal{X})$, the global posterior distribution, is strictly positive [23]. The energy function of a CRF, $\mathcal{U}(\mathcal{Y}|\mathcal{X})$, is defined according to the corresponding neighborhood system. For models considering up to the second-order potentials—hence, models considering at most interactions between pairs of pixels—it can be written as

$$\mathcal{U}(\mathcal{Y}|\mathcal{X}) = \sum_{i \in S} D_i(y_i|\mathcal{X}) + \sum_{\substack{j \in \partial i \\ i \in S}} V_{ij}(y_i, y_j|\mathcal{X}) \tag{2}$$

where $D_i(y_i|\mathcal{X})$ is the unary potential associated with the statistics of the label y_i of each pixel i, given the random field of the observations, and $V_{ij}(y_i, y_j|\mathcal{X})$ is the pairwise potential that defines the spatial relations among neighboring pixels i and j (i.e., $i \in S$, $j \in \partial i$, with $\partial i \subset S$).

We also recall that, thanks to the Hammersley–Clifford theorem, the energy function of a CRF model is related to the global posterior distribution by $P(\mathcal{Y}|\mathcal{X}) \propto \exp[-U(\mathcal{Y}|\mathcal{X})]$ [26], [47]. Similarly, the local posterior distribution on pixel $i \in S$, conditioned on the labels of the neighboring pixels, can be written as ($k = 1, 2, \ldots, M$)

$$P(y_i = \omega_k|\boldsymbol{y}_{\partial i}, \mathcal{X}) \propto \exp[-\mathcal{U}_i(\omega_k|\boldsymbol{y}_{\partial i}, \mathcal{X})] \tag{3}$$

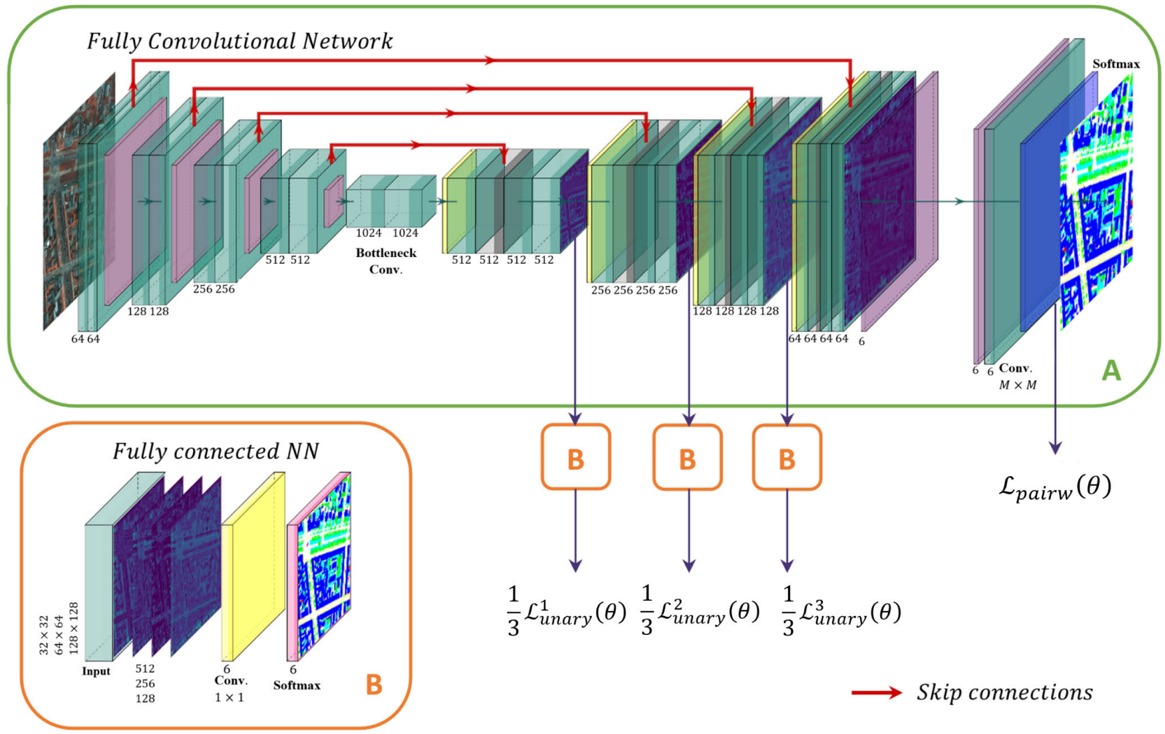

Fig. 1. Overall architecture of the proposed CRFNet approach. The arrows in dark green represent the connections between consecutive layers. The arrows in violet represent the links with the fully connected layers and the corresponding loss functions. The arrows in red represent the skip connections.

where

$$\mathcal{U}_i(\omega_k|\boldsymbol{y}_{\partial i}, \mathcal{X}) = D_i(\omega_k|\mathcal{X}) + \sum_{j \in \partial i} V_{ij}(\omega_k, y_j|\mathcal{X}) \tag{4}$$

where $\boldsymbol{y}_{\partial i}$ is the vector collecting the labels y_j of all pixels j that neighbor i $(j \in \partial i)$.

In particular, in this article, the focus is on the subfamily of CRFs where the pairwise potential is characterized by

$$V_{ij}(y_i, y_j|\mathcal{X}) = E_{ij}(y_i, y_j|\mathcal{X})\delta(y_i, y_j) \tag{5}$$

i.e., it combines multiplicatively a Kronecker impulse term $\delta(y_i, y_j)$ and a further generic function $E_{ij}(y_i, y_j|\mathcal{X})$. Indeed, the Kronecker delta is a desired term in order to favor that homogeneous regions are labeled consistently, in comparison to the surrounding regions. This desired behavior is the reason why, although a generic CRF with up to second-order potentials may not necessarily include this Kronecker dependence in its own pairwise terms, in the proposed approach, we focus on CRF models whose energies belong to the family in (5). In this respect, the term $E_{ij}(y_i, y_j|\mathcal{X})$ is aimed at capturing other, possibly arbitrary, spatial-contextual behaviors (e.g., contrast-sensitive or edge-preserving), in addition to the smoothness characterized by $\delta(y_i, y_j)$. Accordingly, (5) represents a flexible and rather general model for a family of pairwise potentials for the semantic segmentation of remote sensing images.

C. Proposed CRFNet Model

The architecture employed to learn the unary and binary potentials of the aforementioned CRF is based on a fully convolutional network. These networks can take input images of arbitrary size and generate output results with the same size [9], thanks to the adoption of an encoder–decoder architecture. They are characterized by several multiscale processing stages (e.g., convolutional and pooling layers), thus allowing the manipulation of multiscale information.

For semantic segmentation purposes, in order to exploit this information, available in the activations of the feature maps of the hidden layers of the network, the backbone of a simple fully convolutional network, such as the U-Net [48], is modified in the proposed CRFNet with the addition of fully connected networks. These networks are inserted at each convolutional block in the decoder of the original fully convolutional architecture, and linked to the encoder by skip connections, in order to fuse coarse, semantic, and local information [9] and to favor the modeling of long-range spatial dependences. Each fully connected network (see Fig. 1, block B) is built as a CNN with a single convolutional layer with kernels of size 1×1 and a softmax nonlinear activation.

A convolutional layer with a softmax activation is also added at the end of the original fully convolutional network (see Fig. 1, block A). The rationale of this further layer is twofold. First, from a semantic segmentation perspective, it allows integrating further spatial information. Then, as detailed later, the introduction of this layer contributes to relating the network output to a CRF model, and vice versa, to learn the CRF model (specifically, its pairwise potential) through the network. Indeed, we shall discuss later that the size of the kernel of this additional layer (see Fig. 1, block A) defines the number of neighboring pixels that influence the prediction on each pixel i [41], [47]. For example, a 3×3 kernel relates to a first- or a second-order neighborhood system (in the first case, some of its weights need to be set to zero, in order to take into account only the four adjacent pixels to pixel i, see Fig. 2) [23], [47]. This influences the span of the spatial-contextual information modeled by the pairwise term learned automatically by the proposed approach. The last layer of the U-Net is an image composed of M channels, each representing a feature map associated with one of the M classes. Then, the additional layer includes M filters, each acting on one of these M channels.

As usual in the case of CNNs, rectangular image patches, composed of $a \times b$ pixels and drawn from the input image data, are fed as input to the network. We denote as $\boldsymbol{X} \in \mathbb{R}^{a\times b\times d}$ the tensor collecting the feature vectors $\boldsymbol{x}_i$ of all pixels i belonging to a generic patch. As usual, the entries in $\boldsymbol{X}$ are modeled as random variables. We also collect all the parameters of the network in a vector $\boldsymbol{\theta}$.

Let $S^l \subset \mathbb{Z}^2$ be the pixel grid at resolution $l = 1, 2, \ldots, L$ in the network, where $l = L$ corresponds to the pixel lattice of the input image data (i.e., $S^L = S$) and S^1 is the coarsest-resolution grid in the network. Given a pixel $s \in S^l$ in the grid at resolution l ($l = 1, 2, \ldots, L-1$), let z_s be the vector containing the activations obtained on this pixel in the fully connected layer at resolution l when the patch $\boldsymbol{X}$ is fed as input to the network. Since the proposed approach is supervised, the ground truth (which is defined on the original pixel lattice S) is downsampled on each grid S^l. Let us define t_{sk} as the one-hot encoding of the ground-truth labels on pixel $s \in S^l$ [41], hence $t_{sk} = 1$ if and only if s belongs to class ω_k in the training set, otherwise $t_{sk} = 0$ ($k = 1, 2, \ldots, M$).

The fully connected layer at each scale l ($l = 1, 2, \ldots, L-1$) is pushed to align with the ground truth through a cross-entropy loss function. Specifically, the kth output on pixel $s \in S^l$ is obtained through a softmax ($k = 1, 2, \ldots, M$) [41]

$$\hat{P}_{sk} = \frac{\exp(z_{sk})}{\sum_{m=1}^{M} \exp(z_{sm})}. \tag{6}$$

In the proposed CRFNet approach, this output is aimed at estimating the probability that the pixel $s \in S^l$ belongs to class ω_k, conditioned on the activations z_s obtained on that fully connected layer when $\boldsymbol{X}$ is fed to the network ($k = 1, 2, \ldots, M$).

For this purpose, the loss function of the proposed method includes a linear combination of $(L-1)$ weighted cross-entropy losses, associated with the $(L-1)$ fully connected neural networks that have been introduced at the $(L-1)$ different scales in the decoder of CRFNet

$$\mathcal{L}_{\text{unary}}(\boldsymbol{\theta}) = \frac{1}{L-1} \sum_{l=1}^{L-1} \mathcal{L}^l(\boldsymbol{\theta}) \tag{7}$$

where

$$\mathcal{L}^l(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{X}} \left\{ -\sum_{k=1}^{M} \frac{\hat{P}_{\max}}{\hat{P}_k} \sum_{s\in S^l} t_{sk} \ln \hat{P}_{sk} \right\}. \tag{8}$$

Here, $\hat{P}_k$ is the prior probability of ω_k, estimated as its relative frequency in the training set ($k = 1, 2, \ldots, M$), and $\hat{P}_{\max} = \max_k \hat{P}_k$. Indeed, a weighting factor is introduced in (8) in the cross-entropy losses at each scale. It is inversely proportional to the number of training samples of each class, in order to take into account the presence of imbalanced training data—a commonly encountered situation in the semantic segmentation of remote sensing images. In (7) and (8), the dependence of the loss on the network parameters $\boldsymbol{\theta}$ and the fact that the expectation operator is taken with respect to the distribution of the input patch $\boldsymbol{X}$ are emphasized explicitly.

It is worth noting that the additional fully connected layers, whose goal is to analyze information at different scales and guarantee high accuracy at the multiscale level, favor the interpretability of CRFNet in the hidden layers in terms of classwise posterior probabilities.

This loss function $\mathcal{L}_{\text{unary}}$ in (7) is integrated with an explicitly pairwise term, deriving from the cross-entropy loss function over the pixelwise softmax of the aforementioned additional convolutional layer. Specifically, this layer includes M convolutional filters, which operate on the original pixel lattice S (i.e., at the original resolution) and share the same spatial support. The input feature map and the output of each filter are scalar-valued. Let f_k be the input feature map of the kth filter when the patch $\boldsymbol{X}$ is fed to CRFNet, and let h_k be its impulse response (or kernel; $k = 1, 2, \ldots, M$). Given the convolution $h_k * f_k$, the pairwise loss function is defined as

$$\mathcal{L}_{\text{pairw}}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{X}} \left\{ -\sum_{k=1}^{M} \frac{\hat{P}_{\max}}{\hat{P}_k} \sum_{i\in S} t_{ik} \ln \hat{P}_{ik} \right\} \tag{9}$$

where the output estimated probabilities $\hat{P}_{ik}$ ($k = 1, 2, \ldots, M$) on pixel $i \in S$ are obtained through a pixelwise softmax

$$\hat{P}_{ik} = \frac{\exp[(h_k * f_k)_i]}{\sum_{m=1}^{M} \exp[(h_m * f_m)_i]} \tag{10}$$

where t_{ik} indicates the one-hot encoding of the training set on the original pixel lattice S [41]. The same comments we have made in the case of (8) about the weights to mitigate class imbalance, about the dependence on $\boldsymbol{\theta}$, and about the expectation over the distribution of $\boldsymbol{X}$ hold with regard to (9) as well.

The total loss function is defined as the sum of the two aforementioned terms

$$\mathcal{L}(\boldsymbol{\theta}) = \mathcal{L}_{\text{unary}}(\boldsymbol{\theta}) + \mathcal{L}_{\text{pairw}}(\boldsymbol{\theta}). \tag{11}$$

As usual, within the training of the network, the expectations in (8) and (9) are estimated as sample means [41].

D. Interpretation as a CRF

The proposed formulation of a fully convolutional network-based framework automatically learning the first- and second-order potentials of a CRF leverages the definition of the last convolutional layer of CRFNet. According to the training of the network, the output of this layer, followed by a softmax, provides an estimate $\hat{P}_{ik}$ of the posterior probability that pixel $i \in S$ belongs to class ω_k ($k = 1, 2, \ldots, M$), conditioned on the input observations. In the following, we shall use the explicit notation $\hat{P}_{ik}(\mathcal{X})$ to acknowledge the dependence on the random field $\mathcal{X}$ of the observations, which is the input of CRFNet.

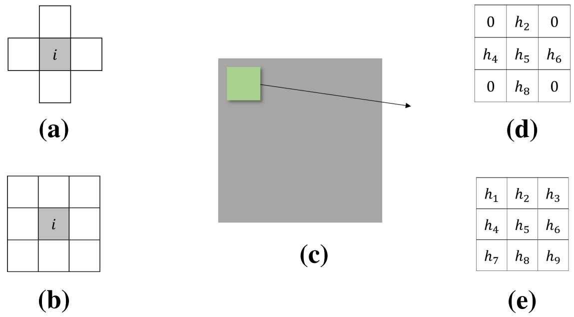

Fig. 2. Example of integration of spatial information. (a) First-order and (b) second-order neighborhoods; (c) image (*gray*) convolved by a kernel (*green*); 3×3 kernel encoding, (d) first-order, and (e) second-order neighborhoods.

Specifically, the estimated probabilities are obtained by a softmax operating on the output of the aforementioned convolutional layer, i.e., ($i \in S; k = 1, 2, \ldots, M$)

$$\hat{P}_{ik}(\mathcal{X}) \propto \exp\left[\sum_{j \in \Delta + i} h_{i-j}(\omega_k) f_j(\omega_k|\mathcal{X})\right] \tag{12}$$

where $\Delta \subset \mathbb{Z}^2$ is the common support of all the M convolutional filters, $\Delta + i = \{j + i \in S : j \in \Delta\}$ indicates its translation centered on pixel i, $h_j(\omega_k)$ is the value of the impulse response h_k on pixel $j \in \Delta$, and $f_i(\omega_k|\mathcal{X})$ is the value of the feature map f_k on pixel i. Here, the explicit notations $h_j(\omega_k)$ and $f_i(\omega_k|\mathcal{X})$ are used to emphasize the dependence on the class ω_k and on the pixel location i.

We assume that the support Δ of the considered convolutional filters is symmetric with respect to the origin of the 2-D pixel lattice (i.e., $i = (m, n) \in \Delta$ also implies $(-m, -n) \in \Delta$). This is a commonly satisfied assumption for 2-D convolutional filters with finite impulse responses (e.g., supported on a 3×3 window) [41]. This symmetry property is desired to align with the properties of CRF neighborhoods.

Equation (12) can be equivalently rewritten

$$\begin{aligned}\ln \hat{P}_{ik}(\mathcal{X}) &= h_0(\omega_k) f_i(\omega_k|\mathcal{X}) \\ &\quad + \sum_{\substack{j \in \Delta + i \\ j \neq i}} h_{i-j}(\omega_k) f_j(\omega_k|\mathcal{X}) + \phi_i(\mathcal{X})\end{aligned} \tag{13}$$

where $\phi_i(\mathcal{X})$ is an additive term that does not depend on ω_k ($k = 1, 2, \ldots, M; i \in S$) and thus does not affect the resulting decision. Up to this additive contribution, the last line of (13) is the sum of two terms. The first one is related to the value of the kth feature maps in the location $i \in S$ and only depends on the central value of the kernel h_0. The second term relates two distinct pixels i and j located within the span of the kernel employed. In the proposed approach, we exploit this interpretation to introduce a CRF model whose potentials are learned by the modeling capabilities of a DL approach directly applied to the input data. Specifically, we relate the first and second contributions in (13) to unary and pairwise potentials, respectively.

Let us introduce a neighborhood system $\{\partial i\}_{i \in S}$ on the pixel lattice S by defining the neighborhood of pixel $i \in S$ as

$$\partial i = (\Delta + i) \setminus \{i\} = \{j \in S : j - i \in \Delta, j \neq i\}. \tag{14}$$

This definition is well-posed because it satisfies the properties that characterize a neighborhood system associated with a CRF model, i.e., (a) $i \notin \partial i$ and (b) $i \in \partial j$ if and only if $j \in \partial i$ [49] The latter derives from the aforementioned symmetry assumption on the support Δ.

Then, we define a CRF model whose potentials are expressed as ($i, j \in S; j \in \partial i; k, m = 1, 2, \ldots, M$)

$$D_i(\omega_k|\mathcal{X}) = -h_0(\omega_k) f_i(\omega_k|\mathcal{X}) \tag{15}$$

$$V_{ij}(\omega_k, \omega_m|\mathcal{X}) = -h_{i-j}(\omega_k) f_j(\omega_k|\mathcal{X}) \delta(\omega_k, \omega_m). \tag{16}$$

Equivalently, taking into account the properties of the Kronecker impulse, the energy function is

$$\begin{aligned}\mathcal{U}(\mathcal{Y}|\mathcal{X}) = \sum_{i \in S} \Bigg[&- h_0(y_i) f_i(y_i|\mathcal{X}) + \\ &- \sum_{j \in \partial i} h_{i-j}(y_j) f_j(y_j|\mathcal{X}) \delta(y_i, y_j) \Bigg].\end{aligned} \tag{17}$$

The relation between this CRF model and the network architecture is twofold. First, the CRF potentials are explicitly defined according to the feature map f_k and to the convolutional kernels h_k ($k = 1, 2, \ldots, M$), which are automatically learned through the training of the network. Second, the following theorem holds.

Theorem 1: The pixelwise probability distribution (12), predicted on the output of CRFNet, is equal to the local posterior distribution of the CRF model defined by (15)–(17), in the particular case of neighboring pixels sharing the same label: for each pixel $i \in S$ and each class ω_k ($k = 1, 2, \ldots, M$), if $y_j = \omega_k$ for all $j \in \partial i$, then

$$\hat{P}_{ik}(\mathcal{X}) = P^{(\mathrm{crf})}(y_i = \omega_k | \mathbf{y}_{\partial i}, \mathcal{X}). \tag{18}$$

The proof is reported in the Appendix. In (18), the superscript "(crf)" emphasizes that the distribution on the right-hand side is modeled by the CRF established by (15)–(17). The theorem implies that the connection between the network architecture and the energy function of the CRF model is not only formal but also rooted in their probabilistic interpretation. This connection is especially relevant since the case of homogeneous neighborhoods sharing the same semantic class label usually covers the majority of the pixels in a natural image.

We also note that, within the family of CRF models in (5), the network output—and in particular, the feature maps f_k and the kernels h_k—fully determine the unary and pairwise potentials of the CRF. In this respect, the proposed approach learns a CRF model in a non-parametric manner through a deep fully convolutional network, while taking into account a prior on the desired spatial-contextual regularization, encoded by the presence of the Kronecker impulse term [23], [50]. Indeed, this term favors well-defined properties of the local spatial context in remote sensing images—and in natural images at large—i.e., the fact that neighboring pixels within homogeneous image segments are often characterized by similar intensities and are likely to belong to the same semantic class [49].

It is also worth noting that the CRF model learned through the proposed approach is generally non-homogeneous [23],

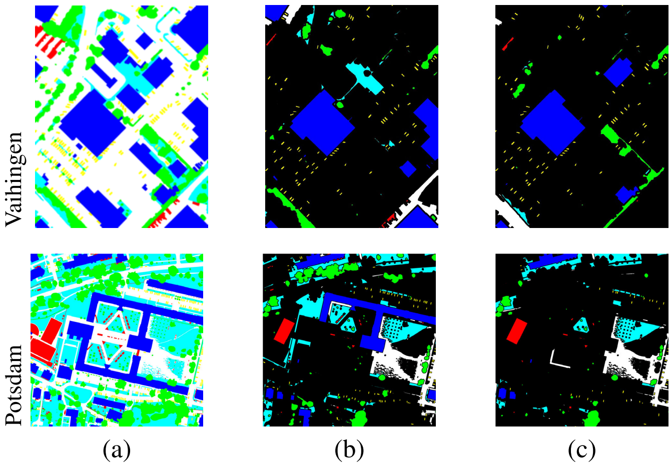

Fig. 3. Ground truths. (a) Full, (b) sparse with 30% of the annotated labels, and (c) sparse with 10% of the annotated labels. Color legend for the classes: buildings (blue), impervious (white), low vegetation (cyan), trees (green), cars (yellow), clutter (red). Black in (b) and (c) indicates unlabeled pixels, i.e., pixels removed from the ground truth.

since its potentials in (15) explicitly depend on the pixel locations i and j. This is a desirable property because it makes it possible for the resulting 2-D PGM to capture the generally non-stationary (or piecewise stationary) behavior that is usually observed in natural image data.

IV. Experimental Results

The experimental validation was conducted with the two ISPRS 2-D semantic labeling challenge datasets,[1] collected by the German Society for Photogrammetry, Remote Sensing, and Geoinformation (DGPF) over the cities of Vaihingen and Potsdam, in Germany; and with the 2015 IEEE GRSS data fusion contest (DFC) Zeebruges dataset[2] [51], [52]. These datasets consist of very high-resolution aerial images.

A. Datasets

The two ISPRS datasets are characterized by the same class encoding, which includes six different semantic classes: impervious surfaces, buildings, low vegetation, trees, cars, and clutter. The Zeebruges dataset presents the same classes and two additional ones: water and boats. Clutter is highly mixed, as it comprises all the surface covers that are not attributed to the other five classes, and accounts for only a small percentage of pixels. Particularly, for the Vaihingen dataset, this class appears only in a few of the training tiles, hence it was discarded from the experimental validation on this dataset (see Table I). Concerning the Potsdam dataset, following the work done previously by other authors in Audebert et al. [53], Liu et al. [54], and Lv et al. [55] the results for the clutter class, which is of relatively limited interest according to the aforementioned comments, were excluded from the average accuracy metrics. The classwise accuracies for this class are nevertheless reported in Table II.

The ISPRS datasets contain three-channel images—near-infrared (NIR), red, and green—and are made of multiple tiles. In the case of Vaihingen, the spatial resolution is 9 cm and the average size of each tile is 2000 $\times$ 2000 pixels. In the case of Potsdam, the spatial resolution is 5 cm and the average tile size is 6000 $\times$ 6000 pixels. For the ISPRS Vaihingen dataset, 12 tiles were chosen for training (tiles 1, 3, 7, 11, 13, 17, 23, 26, 28, 32, 34, and 37) and four for testing (tiles 5, 15, 21, and 30), while for the ISPRS Potsdam dataset, the training set consisted of ten tiles (3_11, 4_11, 5_10, 6_7, 6_8, 6_9, 7_7, 7_8, 7_9, 7_10) and the test set of five tiles (3_12, 4_10, 4_12, 5_11, 6_12).

In the case of the Zeebruges dataset, each tile contains an RGB image, with a spatial resolution of 5 cm and a size of 10 000 $\times$ 10 000 pixels, and a digital surface model (DSM) with a spatial resolution of 10 cm. In order to work with all the information available, the RGB tiles were downsampled at 10 cm of spatial resolution. Five tiles are endowed with public ground truth. These five tiles were used for experiments, applying the aforementioned downsampling of the RGB images. Three tiles were used for training (tiles 315 130_56 865, 315 130_56 870, and 315 140_56 865) and two were employed for testing (tiles 315 135_56 870 and 315 150_56 865).

B. Experimental Setup

1) Setup and Training of CRFNet: The experiments were run on an Alienware Aurora R11 with a RAM of 16 GB and a GPU NVIDIA GeForce RTX 2080 Ti. The network was trained for 30 epochs with patches of size 256 $\times$ 256 pixels, obtained through a sliding-window approach, a batch size of 10, and pretrained on ImageNet.[3] The learning rate was fixed to 0.01, with a decay rate of 0.0005, and the optimizer employed was the Adam algorithm [56].

To implement the proposed CRFNet approach, three fully connected neural networks were added to the three central deconvolution blocks of the decoder, already linked with the ones of the encoder through skip connections, to integrate multiscale information available in the hidden layers of the fully convolutional network at three different spatial resolutions. These resolutions are twice, four times, and eight times coarser than the spatial resolution of the input image and the corresponding ground truth (i.e., if s is the pixel size in meters in the lattice of the input image, then the pixel sizes in the three coarser-resolution grids are 2, 4, and 8 s). Therefore, for this experimental validation, L, the number of multiscale pixel lattices considered, is equal to 4. In general, L is a hyperparameter of the proposed approach. Its value can be chosen as a function of the spatial resolution of the input data and of the multiscale processing operations executed by the chosen fully convolutional network. In our experiments, $L = 4$ allows us to take advantage of the multiple spatial resolutions modeled by the decoder of the network and simultaneously maintain a high spatial resolution even at the coarsest scale taken into account. This information is inserted in the overall loss function of the neural model after a resampling of the ground truth at the considered scales, as mentioned in Section III-C.

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

[2] http://dase.grss-ieee.org/

[3] https://image-net.org/

The additional convolutional layer is responsible for the learning of the potentials of the CRF and models the spatial-contextual information. The neighborhood system on which the CRF is defined and the relative spatial information analyzed depends on the characteristics of the kernel of this layer. Indeed, as mentioned in Section III, the size of the kernel defines the neighborhood system and it is a hyperparameter of the proposed method. In particular, two different kernels were employed: 1) a 3×3 kernel whose weights were learned during the training of the network (i.e., $\Delta = \{(m, n) \in \mathbb{Z}^2 : |m| \leq 1, |n| \leq 1\}$), defining a second-order neighborhood system (an eight-pixel neighborhood) and 2) a 3×3 kernel with the four weights at the corners set to 0 and the remaining five learned during training (i.e., $\Delta = \{(m, n) \in \mathbb{Z}^2 : |m| + |n| \leq 1\}$), defining a first-order neighborhood system (taking into account only the four adjacent pixels). The latter is meant as a symmetric approximation of a 2×2 kernel, which was proven to provide interesting results in the literature [57]. Some additional experiments were conducted with larger kernel sizes in order to assess the sensitivity of the method to additional spatial information and the effects on its performances (see Section IV-E).

As mentioned in Section IV-A, the average accuracy results reported in Tables I–VII for the Vaihingen and Potsdam datasets were computed without considering the pixels belonging to the class "clutter."

2) Training Sets Used for Experiments: Since all three datasets are used for benchmark competitions, their ground-truth information is an "ideal" one, where the true label is known for all pixels in the training maps. This is generally unfeasible in datasets related to real-world remote sensing applications, where the objective is to generate accurate classification maps using fewer training samples, often arranged in homogeneous patches, not including spatial class borders. Hence, three training conditions were considered: 1) the full dataset with exhaustive ground truths [shown in Fig. 3(a)]; 2) a dataset with scarce ground truths, obtained by removing entire connected components from the original exhaustive ground truth and then applying morphological erosion (as in [17]), until only 30% of the labeled pixels were left [see Fig. 3(b)]; and 3) a further dataset with sparse ground truth, obtained as in 2) but leaving only 10% of the labels (see Fig. 3(c)). The versions 2) and 3) are approximations of the ground truths usually found in realistic remote-sensing applications, usually involving maps with isolated patches of labeled pixels associated with different classes.

The results presented in Sections IV-C and IV-D refer to the experiments conducted with the three aforementioned training conditions. These training datasets, differing for the amount of input training samples, led to different results. For example, and expectedly, the values of the averaged accuracy metrics (overall accuracy, recall, precision, F1-score) tend to be higher for both of the considered datasets when more training ground-truth samples are available. In most of the cases, this phenomenon is reflected also on the classwise accuracies.

3) Experimental Comparisons: In the proposed approach, a CRF model is automatically learned, and simultaneously, the semantic segmentation of the input image is addressed by exploiting multiscale information. Accordingly, experimental comparisons have been performed from both viewpoints. First, the results of the proposed approach were compared with those obtained by the previous technique in [36], which, in the framework of semantic segmentation, aims at defining a CRF through a neural learning process. As recalled in Section II, this method defines an end-to-end CRF-RNN formulation, in which a mean-field approximate inference for a CRF with Gaussian pairwise potentials is formulated as an RNN. In this benchmark approach, the approximation developed in [39] for a fully connected CRF is used for the initialization of the CRF parameters. To ensure consistency with the results obtained by the proposed CRFNet method, which was formulated in the experiments using U-Net as a backbone, and to guarantee a coherent experimental comparison, the selected backbone is a U-Net [48] for CRF-RNN, as well.

Then, further comparisons were performed with state-of-the-art techniques for semantic segmentation tasks based on multiscale information, such as HRNet [58], a network consisting of multiresolution subnetworks connected in parallel, and an MFF method, namely the lightweight attention network (LWN-Attention) in [59]. The latter approach is capable of exploiting multiscale information through the concatenation of feature maps associated with different scales [59]. These methods were considered as recent benchmarks taking into account multiscale information in fully convolutional network-based approaches to semantic segmentation. A comparison with the results of the U-Net backbone *per se* [48] was also conducted to verify the benefit deriving from the additional layers that characterize CRFNet. Finally, an additional comparison with a recent state-of-the-art technique for the semantic segmentation based on vision transformers [60] was included for the Vaihingen dataset. Even though the rationale of transformers is quite far from the one of CRFNet and the other benchmark methods, this last comparison has been added to show the performance of the proposed model as compared to a very recent and popular technique.

An ablation study to evaluate the effectiveness of CRFNet and its layers is also reported in Section IV-E.

The training and inference times, in the case of full and scarce ground truths, of the proposed method and the comparison techniques are presented in Table IV, together with the GFLOPs—meant as giga floating point operations—of the different architectures. The times are reported for the Zeebruges dataset. Those for the other two datasets, omitted for brevity, are similar, as the training conditions are the same. As indicated in Table IV, the proposed method has a comparable computational burden with U-Net.

The McNemar's test was employed to validate whether the difference in accuracy between the proposed method and the techniques considered for comparison was statistically significant. The test computes the following asymptotically normal statistics:

$$Z = \frac{f_{12} - f_{21}}{\sqrt{f_{12} + f_{21}}} \tag{19}$$

TABLE I

TEST-SET ACCURACIES ON THE VAIHINGEN DATASET. RECALL, PRECISION, AND F1-SCORE ARE AVERAGED OVER THE CLASSES. THE CLASSWISE ACCURACIES IN THE TABLE ARE THE CORRESPONDING INDIVIDUAL RECALLS

	Architecture	buildings	impervious	vegetation	trees	cars	overall acc.	recall	precision	F1 score
Full dataset	U-Net [48]	0.97	0.84	0.82	0.89	0.95	0.88	0.89	0.89	0.89
	HRNet [58]	0.89	0.89	0.50	0.89	0.84	0.79	0.80	0.81	0.80
	MFF [59]	0.98	0.84	0.76	0.85	0.81	0.85	0.85	0.87	0.86
	CRF-RNN [36]	0.97	0.84	0.81	0.88	0.93	0.87	0.89	0.89	0.89
	DC-Swin [60]	0.98	0.85	0.84	0.89	0.96	**0.90**	**0.90**	**0.90**	**0.90**
	Proposed, (conv. 3×3, 4 connected)	0.97	0.84	0.83	0.88	0.96	0.88	**0.90**	**0.90**	**0.90**
	Proposed, (conv. 3×3, 8 connected)	0.96	0.83	0.81	0.90	0.95	0.88	0.89	**0.90**	0.89
30% labels	U-Net [48]	0.87	0.93	0.64	0.87	0.76	0.82	0.81	0.84	0.82
	HRNet [58]	0.84	0.75	0.82	0.69	0.49	0.77	0.72	0.79	0.75
	MFF [59]	0.95	0.82	0.65	0.85	0.60	0.81	0.78	0.83	0.80
	CRF-RNN [36]	0.86	0.92	0.63	0.87	0.70	0.81	0.80	0.84	0.82
	DC-Swin [60]	0.80	0.91	0.50	0.88	0.86	0.77	0.79	0.75	0.77
	Proposed, (conv. 3×3, 4 connected)	0.93	0.82	0.79	0.87	0.93	**0.85**	**0.87**	**0.86**	**0.86**
	Proposed, (conv. 3×3, 8 connected)	0.94	0.80	0.78	0.88	0.94	**0.85**	**0.87**	0.85	**0.86**
10% labels	U-Net [48]	0.91	0.86	0.46	0.88	0.89	0.78	0.80	0.74	0.77
	HRNet [58]	0.82	0.92	0.18	0.97	0.80	0.72	0.74	0.74	0.74
	MFF [59]	0.88	0.88	0.44	0.90	0.56	0.77	0.73	0.74	0.73
	CRF-RNN [36]	0.91	0.69	0.76	0.76	0.66	0.78	0.80	0.76	0.78
	DC-Swin [60]	0.81	0.90	0.50	0.88	0.84	0.77	0.79	0.76	0.77
	Proposed, (conv. 3×3, 4 connected)	0.95	0.83	0.57	0.89	0.90	**0.81**	0.83	0.80	0.81
	Proposed, (conv. 3×3, 8 connected)	0.94	0.87	0.58	0.90	0.91	**0.81**	**0.84**	**0.82**	**0.83**

with f_{12} the number of test samples misclassified by the proposed method and not by the comparison method, and f_{21} its opposite [61], [62], [63]. Using the common 5% level of significance, the difference in results is statistically significant if $|Z| > 1.96$: in particular, a negative value of Z indicates that the proposed method is more accurate than the other used for comparison [62], [63]. The results of the test are reported in Section IV-E.

C. Results and Comparisons About CRF Model Learning

The discussion in this section refers to the experimental comparison carried out between the results of the proposed technique, CRFNet, and of the previously mentioned CRF-RNN in [36]. Since both aim at learning CRF models for a semantic segmentation task, their comparison is addressed in terms of their classification accuracies on this task.

The quantitative results for the Vaihingen, Potsdam, and Zeebruges datasets are reported in Tables I–III, respectively, in terms of classwise accuracies and averaged metrics. The two methods were compared in the case of the two scarce ground-truth configurations to assess the spatial modeling capabilities of the CRF model.

1) ISPRS Vaihingen Dataset: Concerning the results on the Vaihingen dataset, the proposed method attains the best performances in terms of all the averaged accuracy metrics. This trend is generally confirmed by the classwise accuracies, with the exception of "impervious surfaces" and "low vegetation," in the case of 30% and 10% of ground-truth labels, respectively, where the comparison technique in [36] has better performances. CRF-RNN also provides accurate results. CRFNet reaches the second highest values in terms of the F1-score in the case of 30% of ground-truth labels and in terms of overall accuracy in the case of 10% of ground-truth labels. In the other cases, it reaches at least the third highest value among all considered approaches. These results confirm the potential of the integration between CRF and DL concepts and the opportunity to exploit the latter to learn the spatial modeling structure of the former. Indeed, the classification results show the effectiveness of both approaches and suggest that, at least on the considered dataset, the proposed method allowed higher accuracies to be obtained.

The classification maps obtained with the Vaihingen dataset are presented in Fig. 4, where the contours of the homogeneous regions of the ground truth are superimposed to both the ground truth and the classification maps produced by the various methods to make the comparison easier. We focus here on two test tiles, however, the behavior on the other test tiles is similar. As expected, the less scarce the input ground truth, the smoother the classification maps are. Nonetheless, the maps deriving from CRFNet and from CRF-RNN, i.e., from two techniques modeling the spatial information through CRFs [see Fig. 4(c) and (f)], exhibit more homogeneous zones and sharper edges than the other semantic segmentation approaches used for comparison [U-Net, HRNet, and the MFF method, see Fig. 4(b), (d), (e)]. The classification results of the proposed method are particularly remarkable for the class "building" with 30% of training labels, as shown in the first row of Fig. 4.

2) ISPRS Potsdam Dataset: The comments on the quantitative results of Table II, relative to the Potsdam dataset, are similar to the ones discussed for the images of Vaihingen, but with generally lower values in terms of overall accuracy and higher recall and precision. This overall behavior may generally be due to its finer spatial resolution. The comparison technique CRF-RNN reaches at least the third highest value for all the averaged performance metrics in both training configurations with 30% and 10% of the ground-truth pixels. In particular, in the case of 10% of training labels, it reaches the second highest value in terms of recall. The proposed

TABLE II

TEST-SET ACCURACIES ON THE POTSDAM DATASET. RECALL, PRECISION, AND F1-SCORE ARE AVERAGED OVER THE CLASSES. THE CLASSWISE ACCURACIES IN THE TABLE ARE THE CORRESPONDING INDIVIDUAL RECALLS

	Architecture	buildings	impervious	vegetation	trees	cars	clutter	overall acc.	recall	precision	F1 score
Full dataset	U-Net [48]	0.93	0.92	0.88	0.87	0.86	0.45	0.87	**0.91**	**0.92**	**0.91**
	HRNet [58]	0.89	0.95	0.87	0.85	0.91	0.43	0.86	0.89	0.90	0.89
	MFF [59]	0.96	0.87	0.78	0.85	0.82	0.44	0.83	0.86	0.89	0.87
	CRF-RNN [36]	0.96	0.87	0.84	0.91	0.91	0.44	0.86	0.89	0.91	0.90
	Proposed, (conv. 3×3, 4 connected)	0.96	0.88	0.83	0.92	0.92	0.44	0.87	**0.91**	**0.92**	**0.91**
	Proposed, (conv. 3×3, 8 connected)	0.96	0.89	0.84	0.91	0.92	0.51	**0.88**	**0.91**	**0.92**	**0.91**
30% labels	U-Net [48]	0.94	0.88	0.71	0.90	0.92	0.19	0.81	**0.88**	0.87	0.87
	HRNet [58]	0.76	0.92	0.48	0.89	0.88	0.08	0.70	0.79	0.79	0.79
	MFF [59]	0.92	0.86	0.69	0.85	0.80	0.25	0.79	0.83	0.94	0.83
	CRF-RNN [36]	0.94	0.83	0.66	0.84	0.94	0.20	0.79	0.85	0.83	0.84
	Proposed, (conv. 3×3, 4 connected)	0.93	0.87	0.69	0.93	0.93	0.30	0.81	0.87	0.87	0.87
	Proposed, (conv. 3×3, 8 connected)	0.94	0.87	0.73	0.90	0.92	0.28	**0.82**	**0.88**	**0.88**	**0.88**
10% labels	U-Net [48]	0.93	0.82	0.66	0.83	0.94	0.21	0.77	0.85	0.82	0.83
	HRNet [58]	0.85	0.71	0.75	0.65	0.86	0.02	0.71	0.76	0.76	0.76
	MFF [59]	0.91	0.78	0.54	0.85	0.79	0.29	0.73	0.78	0.79	0.78
	CRF-RNN [36]	0.91	0.86	0.46	0.89	0.85	0.21	0.78	0.85	0.82	0.83
	Proposed, (conv. 3×3, 4 connected)	0.94	0.80	0.73	0.85	0.96	0.25	**0.80**	**0.86**	**0.84**	**0.85**
	Proposed, (conv. 3×3, 8 connected)	0.92	0.82	0.68	0.86	0.95	0.33	0.79	0.85	0.83	0.84

TABLE III

TEST-SET ACCURACIES ON THE ZEEBRUGES DATASET. RECALL, PRECISION, AND F1-SCORE ARE AVERAGED OVER THE CLASSES. THE CLASSWISE ACCURACIES IN THE TABLE ARE THE CORRESPONDING INDIVIDUAL RECALLS

	Architecture	buildings	impervious	vegetation	trees	cars	clutter	water	boats	overall acc.	recall	precision	F1 score
Full dataset	U-Net [48]	0.80	0.90	0.96	0.62	0.99	0.41	0.92	0.33	0.95	0.74	0.74	0.74
	HRNet [58]	0.82	0.92	0.69	0.83	0.92	0.34	0.93	0.41	0.92	0.73	0.73	0.73
	MFF [59]	0.66	0.97	0.99	0.38	0.67	0.57	0.90	0.39	0.96	0.68	0.76	0.72
	CRF-RNN [36]	0.80	0.89	0.97	0.62	0.98	0.34	0.92	0.32	0.95	0.73	0.73	0.73
	Proposed, (conv. 3×3, 4 connected)	0.82	0.89	0.97	0.72	0.99	0.44	0.92	0.27	0.96	**0.76**	0.75	0.76
	Proposed, (conv. 3×3, 8 connected)	0.77	0.98	0.96	0.62	1.0	0.39	0.92	0.17	**0.97**	0.74	**0.79**	**0.77**
30% labels	U-Net [48]	0.69	0.98	0.98	0.33	0.86	0.70	0.85	0.21	0.95	0.71	0.73	0.72
	HRNet [58]	0.49	0.97	0.73	0.56	0.45	0.34	0.88	0.02	0.92	0.61	0.69	0.65
	MFF [59]	0.49	0.98	0.98	0.32	0.76	0.69	0.74	0.31	0.95	0.59	0.79	0.68
	CRF-RNN [36]	0.69	0.97	0.98	0.32	0.75	0.68	0.74	0.20	0.94	0.68	0.70	0.69
	Proposed, (conv. 3×3, 4 connected)	0.71	0.98	0.92	0.40	0.99	0.49	0.91	0.28	0.95	0.72	0.76	0.74
	Proposed, (conv. 3×3, 8 connected)	0.68	0.98	0.95	0.45	0.93	0.73	0.92	0.23	**0.96**	**0.74**	**0.82**	**0.78**
10% labels	U-Net [48]	0.81	0.45	0.80	0.30	0.84	0.69	0.88	0.49	0.86	**0.68**	0.49	0.57
	HRNet [58]	0.90	0.29	0.99	0.07	0.67	0.53	0.90	0.45	0.86	0.62	0.49	0.55
	MFF [59]	0.84	0.41	0.94	0.12	0.63	0.65	0.73	0.52	0.86	0.60	0.47	0.53
	CRF-RNN [36]	0.81	0.43	0.78	0.32	0.78	0.69	0.88	0.49	0.85	0.67	0.49	0.57
	Proposed, (conv. 3×3, 4 connected)	0.92	0.32	0.98	0.12	0.94	0.59	0.92	0.61	0.88	**0.68**	**0.56**	**0.62**
	Proposed, (conv. 3×3, 8 connected)	0.86	0.54	0.99	0.10	0.96	0.62	0.90	0.49	**0.90**	**0.68**	0.54	0.60

method, as for the Vaihingen dataset, reaches the most accurate average results, tendentially confirmed even for the classwise accuracies.

The visual qualitative results on the Potsdam dataset are shown in Fig. 5, in which we focus again on two test tiles, while the behavior on the other tiles is analogous. The classification maps, obtained with training conditions keeping only 30% or 10% of label information, reproduce quite faithfully the original ground truth. As for the Vaihingen dataset, the ones deriving from the techniques employing the CRF model [see Fig. 5(c), (f)] appear to be more visually regular than the ones obtained by the semantic segmentation techniques leveraging only on multiresolution information [see Fig. 5(d) and (e)]. The proposed method [see Fig. 5(c)] is the most effective, among the considered ones, at discriminating the two vegetated classes, "low vegetation" and "trees."

3) IEEE GRSS DFC Zeebruges Dataset: The quantitative results obtained in the case of the Zeebruges dataset are reported in Table III. Here again, CRF-RNN proves effective, but the proposed technique reaches generally more accurate values for all the averaged classification metrics. The results in terms of OA are comparable with the ones of the baseline U-Net and of CRF-RNN [36] in the case of full ground truth and 30% of ground-truth labels. Accuracy differences become larger when the ground-truth labels available are 10%. In this case, the proposed technique provides an improvement of 4% for OA, 7% for precision, and 5% for F1-score, thus confirming its relevance especially when scarce ground-truth data are available.

The classification maps obtained on the two test tiles for the Zeebruges dataset are shown in Fig. 6. In particular, this figure presents a zoom-in of an area of the original segmented image in case of 30% and 10% of ground-truth labels. From these maps, it is possible to appreciate how the proposed technique, as compared to U-Net, HRNet, and CRF-RNN [see Fig. 6(b), (d), (f)], more accurately captures spatial details, in particular for the classes "buildings" and "water" [see Fig. 6(c)], while limiting the confusion with the other classes [see Fig. 6(b), (e), (f)].

TABLE IV

TRAINING TIMES, TESTING TIMES, AND GFLOPS OF THE PROPOSED METHOD AND THE COMPARISON TECHNIQUES IN THE CASE OF FULL AND SCARCE GROUND TRUTH

	U-Net [48]	HRNet [58]	MFF [59]	CRF-RNN [36]	Proposed, 4 connected	Proposed, 8 connected
Training time *(full GT)* [s]	10590	9792	5966	17747	12975	13494
Training time *(scarce GT)* [s]	9272	9423	5786	16144	11247	11834
Inference time [s]	482	417	218	609	451	490
GFLOPs	61.52	51.24	18.31	74.4	61.56	61.58

TABLE V

Z STATISTICS OF THE MCNEMAR TEST TO VALIDATE WHETHER THE DIFFERENCES IN ACCURACY ARE STATISTICALLY SIGNIFICANT: FOR EACH DATASET, THE VALUE OF Z CORRESPONDS TO THE RESULT OF THE COMPARISON BETWEEN THE PROPOSED METHOD AND EACH BENCHMARK TECHNIQUE

	U-Net [48]	HRNet [58]	MFF [59]	CRF-RNN [36]	DC-Swin [60]	Proposed, 4 connected	Proposed, 8 connected
Vaihingen *(scarce)*	-52.52	-204.36	-108.08	-65.09	-185.97	–	–
Potsdam *(scarce)*	-16.99	-315.43	-120.93	-34.20	–	–	–
Zeebruges *(scarce)*	-31.89	-261.78	-58.60	-64.25	–	–	–

These experiments confirm the capabilities of reaching accurate and visually smooth classification maps with methods that model spatial information, such as CRFs. The proposed technique, leveraging both multiresolution and spatial information, is capable of attaining more accurate classification performances thanks to the exploitation of the diverse semantics contained at different resolutions and of the spatial-contextual information. Comparisons with some state-of-the-art semantic segmentation methods especially based on the modeling of multiresolution information are presented in Section IV-D.

4) Statistical Significance of the Differences Among the Results: According to McNemar's test, reported in Table V for all three datasets in the case of scarce ground truth, the differences between the result of CRFNet and the outputs of CRF-RNN and the baseline U-Net are statistically significant. In particular, $Z < -50$ in the comparisons with the baseline U-Net [48] for the Vaihingen dataset and with the CRF-RNN method [36] for the Vaihingen and Zeebruges datasets. Large values of $|Z|$ are also obtained in the comparisons with the U-Net and the CRF-RNN for the Potsdam dataset. On one hand, these previous techniques generated accurate results on the considered datasets. On the other hand, the outcome of McNemar's test confirmed the significance of the accuracy gains obtained by the proposed method as compared to these techniques in the challenging case of scarce ground truth.

D. Results and Comparisons About Multiscale Semantic Segmentation

This section aims to discuss the experimental results and comparisons from the point of view of the semantic segmentation based on multiscale information, thus comparing the proposed approach, CRFNet, to the two aforementioned multiscale state-of-the-art techniques, HRNet [58] and the MFF method [59].

1) ISPRS Vaihingen Dataset: The quantitative results for the Vaihingen dataset are again reported in Table I. For all the global performance metrics in the three training conditions, CRFNet is capable of achieving higher accuracies than the methods used for comparison, with the exception of the overall accuracy in the case where the full training ground truth is used. In this case, the transformer-based method (dc-Swin [60]) attains the highest accuracy, and the performances of the U-Net and the proposed method are equal. In general, the proposed CRFNet architecture presents slightly higher classwise accuracies, and these improvements are progressively more remarkable the scarcer is the input ground-truth information. Particularly, compared to HRNet, CRFNet shows an improvement of about 8%–9% for all the averaged accuracy metrics, and generally more accurate results compared to the MFF technique. This confirms the effectiveness of the proposed approach in exploiting both the multiscale information extracted by the network and spatial-contextual information to address semantic segmentation in the challenging case of spatially scarce training data.

The classification results related to the Vaihingen dataset, shown in Fig. 4, confirm the spatial modeling capabilities of CRFNet, as the output maps obtained with scarce input training sets appear to be visually smoother and less noisy than the ones generated by the techniques used for comparison [see Fig. 4(b)-(e)]. In fact, the predictions of the proposed method exhibit a better discrimination of the classes in the dataset—especially for those related to the vegetated areas, "low vegetation" and "trees"—compared to HRNet and the MFF method. The maps obtained by CRFNet also exhibit generally more regular edges than U-Net and HRNet. These remarks are especially evident for the class "building," for which the proposed method is capable of recovering both instances and boundaries lost by the other methods, in particular by U-Net. Concerning the transformer-based architecture, dc-Swin [60], as expected, in the case of fully exhaustive ground truth, its performances are the most accurate. However, as the ground truth approaches the more realistic case, with a lower number of training labels and with spatially sparser training data, the performances of the transformer are suboptimal, with lower classification accuracies than CRFNet, U-Net, and CRF-RNN.

2) ISPRS Potsdam Dataset: The results obtained with the Potsdam dataset (see again Table II) are again comparable to those with the Vaihingen dataset described above, again with generally lower values in terms of overall accuracy and higher recall and precision. As above, the proposed methodology

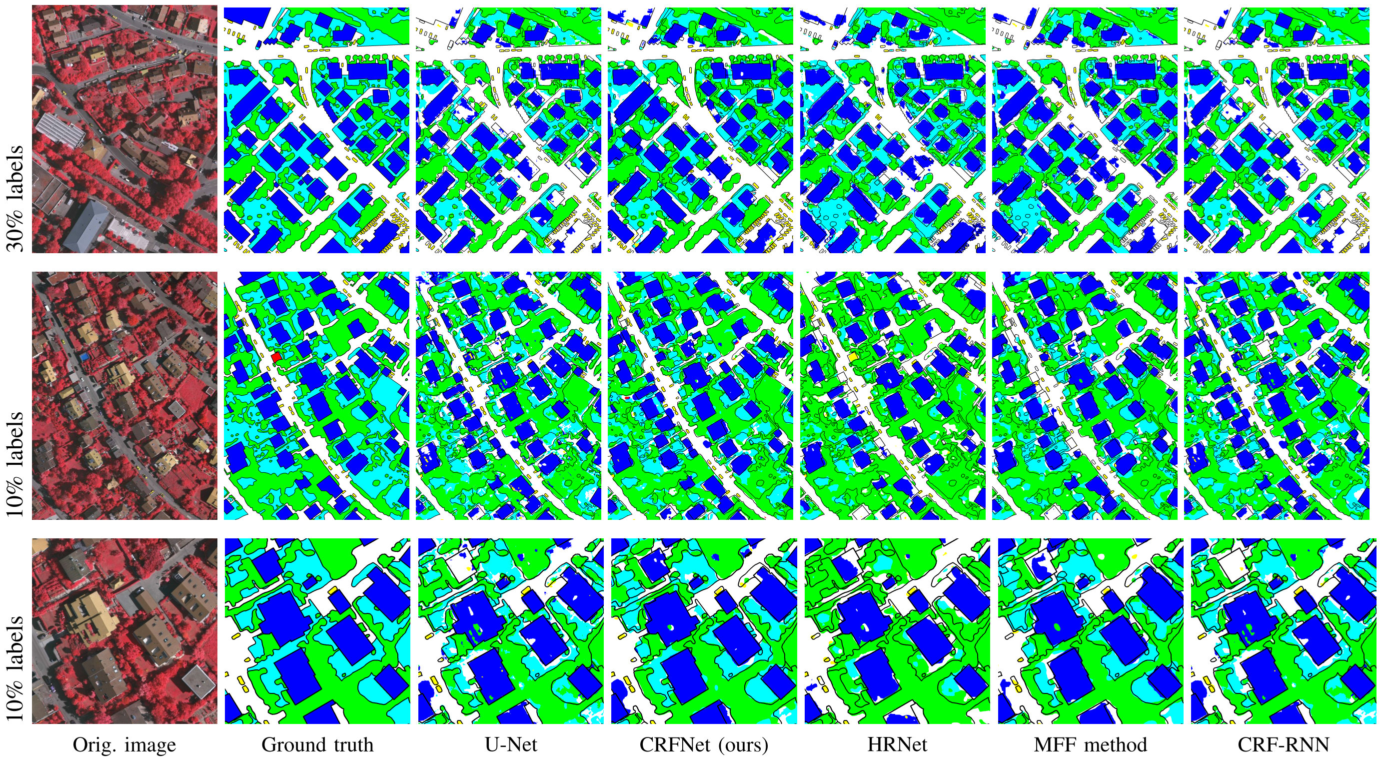

Fig. 4. Test ground truths and classification maps for two test tiles and a zoom-in of the Vaihingen dataset (with 30% and 10% of the training set). Color legend for the classes: buildings (blue), impervious (white), low vegetation (cyan), trees (green), cars (yellow).

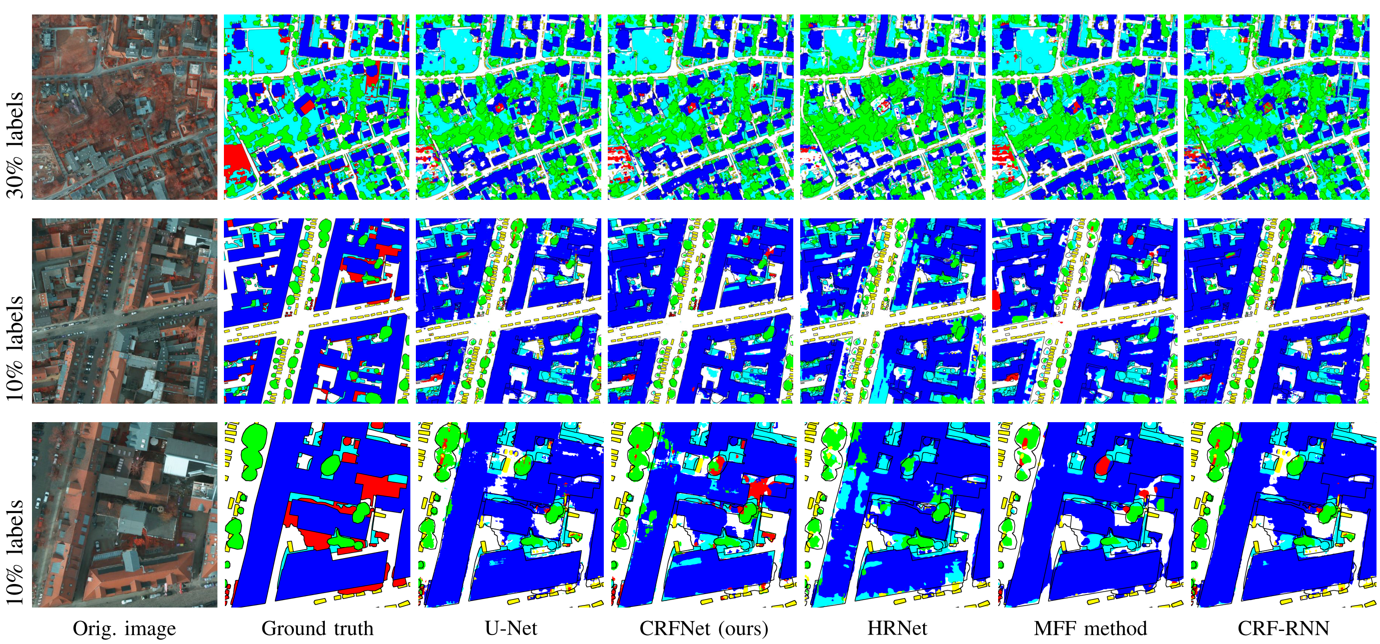

Fig. 5. Test ground truths and classification maps for two test tiles in the Potsdam dataset (with 30% and 10% of the training set). Color legend for the classes: buildings (blue), impervious (white), low vegetation (cyan), trees (green), cars (yellow), clutter (red).

generally reaches higher classification accuracies for what it concerns all the averaged performance metrics (overall accuracy, recall, precision, and F1-score) in all the training configurations, but in particular for the case of ground truths with 10% of annotations. The U-Net is capable of attaining similar performances in terms of precision. Regarding the classwise scores, the proposed approach obtains more accurate results for all the minority classes, “trees,” “cars,” and even “clutter,” while maintaining comparable results with the other reported state-of-the-art techniques for the remaining classes, thus confirming the opportunity to exploit spatial-contextual and multiresolution information for semantic segmentation purposes and the effectiveness of the proposed approach for this task.

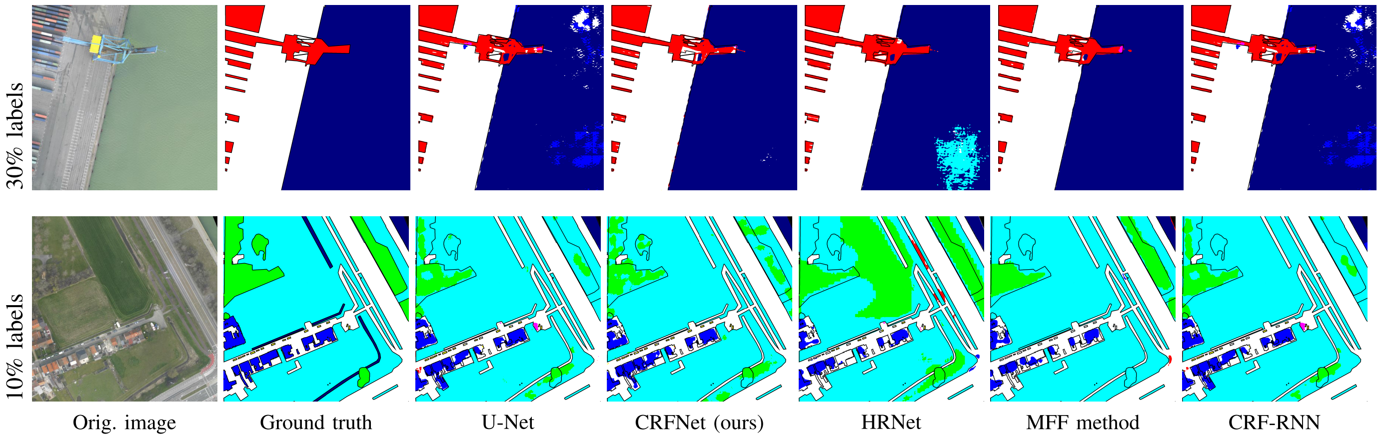

Fig. 6. Zoom-in of the test ground truths and classification maps for two tiles in the Zeebruges dataset (with 30% and 10% of the training set). Color legend for the classes: buildings (blue), impervious (white), low vegetation (cyan), trees (green), cars (yellow), clutter (red), water (dark blue), boats (pink).

TABLE VI
ABLATION STUDY ON THE LOSS FUNCTIONS OF CRFNET FOR THE VAIHINGEN DATASET IN THE CASE OF SCARCE GROUND TRUTH

Setting				**buildings**	**impervious**	**vegetation**	**trees**	**cars**	**overall acc.**	**rec.**	**prec.**	**F1**
$\mathcal{L}^1$	$\mathcal{L}^2$	$\mathcal{L}^3$	$\mathcal{L}_{\text{pairw}}$									
✓	✓	✓	✓	**0.94**	**0.87**	0.58	**0.90**	**0.91**	**0.81**	**0.84**	**0.82**	**0.83**
✗	✓	✓	✓	0.93	0.83	**0.61**	0.84	0.82	0.80	0.80	0.78	0.79
✓	✗	✓	✓	0.93	0.83	0.52	0.87	0.84	0.78	0.80	0.76	0.78
✓	✓	✗	✓	0.92	0.81	0.60	0.84	0.86	0.79	0.81	0.75	0.78
✓	✓	✓	✗	0.93	0.81	0.57	0.85	0.82	0.78	0.80	0.77	0.78
✗	✗	✓	✓	0.93	0.82	0.60	0.85	0.79	0.79	0.80	0.77	0.78
✗	✓	✗	✓	0.92	0.81	0.50	0.87	0.83	0.77	0.79	0.76	0.77
✗	✓	✓	✗	0.93	0.80	0.47	0.88	0.84	0.77	0.78	0.76	0.77
✓	✗	✗	✓	0.92	0.80	0.56	0.85	0.89	0.78	0.80	0.76	0.78
✓	✗	✓	✗	0.92	0.81	0.53	0.85	0.84	0.78	0.79	0.77	0.78
✓	✓	✗	✗	**0.94**	0.82	0.60	0.85	0.84	0.80	0.81	0.78	0.79
✓	✗	✗	✗	0.93	0.83	0.53	0.85	0.87	0.78	0.80	0.77	0.78
✗	✓	✗	✗	0.93	0.83	0.59	0.82	0.80	0.78	0.79	0.77	0.78
✗	✗	✓	✗	0.90	0.82	0.54	0.87	0.80	0.78	0.78	0.76	0.77
✗	✗	✗	✓	0.93	0.82	0.43	0.88	0.76	0.76	0.76	0.74	0.75

TABLE VII
BEHAVIOR OF THE PROPOSED METHOD ON THE VAIHINGEN DATASET AS A FUNCTION OF THE KERNEL SIZE

	Architecture	**buildings**	**impervious**	**vegetation**	**trees**	**cars**	**overall acc.**	**recall**	**precision**	**F1 score**
Full dataset	Conv. 3×3 (4 connected pixels)	0.93	0.91	0.80	0.95	0.86	**0.90**	**0.89**	0.90	**0.89**
	Conv. 3×3 (8 connected pixels)	0.92	0.92	0.77	0.96	0.86	0.89	**0.89**	0.90	**0.89**
	Conv. 5×5	0.93	0.92	0.78	0.96	0.82	**0.90**	0.88	**0.91**	**0.89**
	Conv. 7×7	0.91	0.93	0.77	0.96	0.82	**0.90**	0.88	0.90	**0.89**
	Conv. 9×9	0.91	0.92	0.77	0.96	0.85	0.89	0.88	0.89	0.88
Scarce dataset	Conv. 3×3 (4 connected pixels)	0.89	0.89	0.74	0.94	0.86	**0.87**	**0.87**	**0.85**	**0.86**
	Conv. 3×3 (8 connected pixels)	0.89	0.89	0.72	0.94	0.87	0.86	0.86	**0.85**	0.85
	Conv. 5×5	0.87	0.89	0.72	0.94	0.82	0.86	0.85	0.84	0.84
	Conv. 7×7	0.90	0.87	0.71	0.95	0.84	0.86	0.85	**0.85**	0.85
	Conv. 9×9	0.88	0.89	0.72	0.94	0.83	0.86	0.85	**0.85**	0.85

The classification maps obtained for the Potsdam dataset (see Fig. 5) appear to be consistent with the original ground truth, in the case of both training configurations with 30% and 10% of ground-truth labels. They appear more visually regular than the ones obtained by U-Net and HRNet, and generally more faithful to the original ground truth than the maps generated by all the techniques used for comparison, particularly for what concerns the separation of the areas belonging to the two classes "low vegetation" and "trees."

3) IEEE GRSS DFC Zeebruges Dataset: The quantitative results of the Zeebruges dataset (reported in Table III) are slightly more accurate than those of the ISPRS datasets in terms of overall classification accuracies. CRFNet generally reaches high values for what it concerns all the averaged

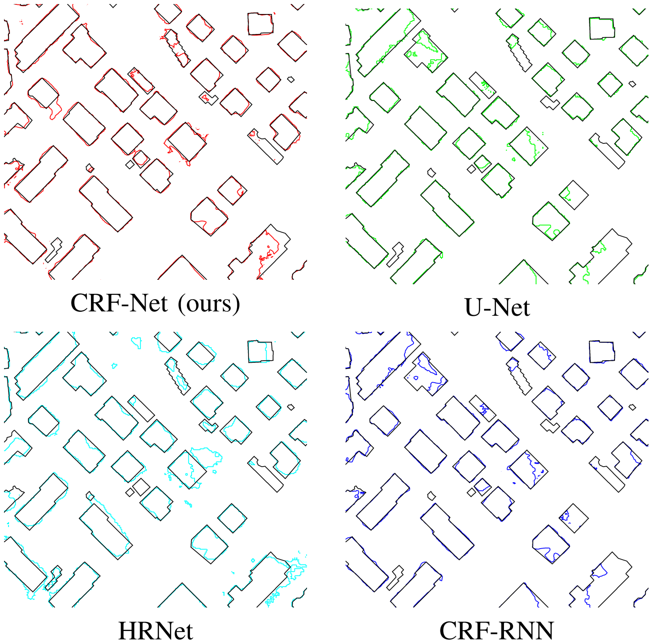

Fig. 7. Building edges extracted by the different methods on the Vaihingen scarce dataset. Color legend for the classes: ground truth (black), proposed method (red), U-Net (green), HRNet (cyan), CRF-RNN (blue).

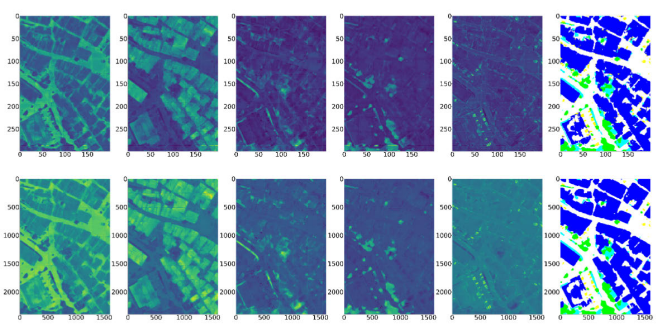

Fig. 8. Classwise posterior probabilities and associated classification maps obtained: by the fully connected neural network connected to the first convolutional block of the decoder of the fully convolutional network (at a spatial resolution eight times coarser than the original image, *top*); and at its output layer (same resolution of the original image, *bottom*).

performance metrics (overall accuracy, recall, precision, and F1-score). The MFF method obtains similar averaged metrics, except in the case of 10% of training labels, where the proposed method achieves higher accuracy values. Concerning the classwise scores, CRFNet attains more accurate results the scarcer is the training information, in particular for the classes "cars" and "boats."

The classification maps related to the Zeebruges dataset (see Fig. 6) also appear consistent with the original ground truth, in the case of both training configurations with 30% and 10% of ground-truth labels. They are more visually regular than the ones obtained by U-Net, HRNet, and CRF-RNN, as it is particularly remarkable in the case of 30% of ground-truth labels, but slightly less visually smooth than the results achieved by the MFF method [see Fig. 6(e)].

Concerning the methods used for comparison, HRNet achieves generally lower average performance metrics than the other techniques taken into consideration and then the proposed approach, reaching similar results to the ones obtained by the other methods only in the case of the Potsdam dataset with full ground truth. As for the classwise results, it obtains lower accuracies for "buildings" in the case of both he ISPRS datasets and for "cars" in the case of Vaihingen; comparable or slightly higher classwise accuracies for "impervious surfaces" in the case of Vaihingen; and average results for "trees" and "cars" in the case of Potsdam. Again, for the Vaihingen dataset, the classification of the vegetated areas, divided into "low vegetation" and "trees," appears to be confused, with several misclassification errors between the two in the various training configurations. On the contrary, for the Zeebruges dataset, the performances of HRNet are more accurate than the other techniques for the classification of the class "trees," albeit slightly poorer for "low vegetation."

The results of the MFF approach for both datasets tend to be slightly less accurate than the ones provided by U-Net and by the proposed approach. Specifically, the MFF method used here for comparison, LWN-Attention, makes use of spatial and channel attention layers and the multiresolution information is taken into account through upsampling stages to stack feature maps at the same resolution, hence possibly also incorporating estimation errors and consequently attaining slightly lower average classification performances. In particular, the MFF method obtains lower accuracies for "low vegetation" and "cars" with all the different training conditions in both the ISPRS datasets, and lower accuracies for "trees," "cars," and "water" for the Zeebruges dataset. However, it is able to reach results comparable to those of the other techniques for "impervious surfaces." Regarding "buildings," one of the majority classes in the considered datasets, the proposed methodology attains the best results in the Vaihingen dataset in the cases of full and 30% of ground truth, and comparable results with the other approaches in the other case.

Spatial details of the results obtained by the proposed technique and the comparison methods are shown in Fig. 7, which shows the perimeters of the buildings in the classification maps for the Vaihingen dataset, obtained by morphological erosion. The proposed method is visually compared to U-Net, HRNet, and CRF-RNN, and obtained more defined and sharper contours, without exceeding the limits of the buildings of the original ground truth. This last issue mostly happens, on the contrary, in the results of HRNet [see Fig. 7(c)].

4) Statistical Significance of the Differences Among the Results: The results of the McNemar's test, reported again in Table V with regard to the case of scarce ground truth, show that the differences between the accuracies of CRFNet and the previous algorithms addressing multiscale information are statistically (very) significant. Very large values of $|Z|$ were obtained when comparing the proposed method with HRNet [58] and MFF [59] for all the considered datasets. Remarks similar to those reported in Section IV-C hold in this case as well. The same comments also apply to the comparison with the transformer-based dc-Swin method [60] on the Vaihingen dataset.

5) Interpretability of Intermediate Layers of CRFNet: As mentioned in Section III, the introduction of multiscale fully connected neural networks at different blocks of the fully convolutional network decoder, and their associated cross-entropy loss terms, promotes a partial interpretability of the intermediate layers of the CRFNet model as posterior probabilities. This assumption is confirmed by Fig. 8, which shows the feature maps and the associated classification map of the fully connected neural network related to the first convolutional block of the decoder of CRFNet (thus, at a spatial resolution eight times coarser than the original image, see the top row of Fig. 8) and of its output layer (at the same resolution of the original image, see the bottom row of Fig. 8). The additional fully connected neural networks, whose aim is to incorporate information contained at different scales of the hidden layer of the fully convolutional network into the class discrimination directly, effectively allow to obtain feature maps that, even at a coarser resolution level, resemble the classwise posterior probabilities normally found at the output layer of the whole architecture.

E. Ablation Study and Sensitivity to Neighborhood Size

An ablation study was performed in order to assess the relevance of each component of the developed architecture. The test was carried out by removing, one after the other, the terms related to the four contributions to the loss function: $\mathcal{L}^l$ for $l = 1, 2, 3$, which, we recall, correspond to the loss terms associated with the fully connected neural networks; and $\mathcal{L}_{\text{pairw}}$, i.e., the term related to the last convolutional layer of CRFNet. The results are reported in Table VI and confirm the importance of all the components of CRFNet. When using the whole loss function, the proposed method attains the most accurate performances in terms of classwise and average accuracy metrics, with the exception of the class "vegetation." Nevertheless, the other cases allow for accurate classification results.

Further tests were also conducted with different kernel sizes in the convolutional layer at the end of the CRFNet architecture. We recall that this kernel size determines the neighborhood system of the CRF learned by the proposed approach. These experiments were aimed at exploring the possible interest of taking into account different ranges of spatial contextual information. The outcomes of the experiments carried out with kernel sizes equal to $5 \times 5, 7 \times 7$, and 9×9 in the case of the Vaihingen dataset are reported in Table VII (similar results, omitted for brevity, were obtained in the case of the other two datasets as well). The results herein presented suggest the limited sensitivity of the developed methodology to changes in the size of the kernel, as they appear to be comparable and not strictly affected by the variations in the kernel itself. In both cases, full and scarce dataset, the 3×3 window defining the four-pixel neighborhood kernel attains the best results in terms of overall accuracy, recall, and F1-score. This suggests that, in the case of this dataset, the manipulation of longer-range spatial information within this kernel does not provide critical advantages from the viewpoint of the classification results. Indeed, this is consistent with the fact that the proposed approach is sensitive to long-range spatial information through its own multiscale structure.

V. Conclusion

In this article, we have proposed a novel method to automatically learn the potentials of a CRF model in a non-parametric fashion through a DL architecture for the supervised semantic segmentation of remote sensing images. The idea is to leverage the modeling capabilities of neural networks to directly learn spatial and semantic information from the input data, to be encoded in the unary and pairwise potentials of a CRF model. Specifically, the relation between such potentials and the output of the proposed network is analytically proven (see the Appendix). Therefore, from the viewpoint of semantic segmentation, the proposed method is aimed at emulating the capabilities of stochastic models based on random fields to integrate spatial information, thus limiting the impact of scarce ground-truth data, a common scenario in remote sensing tasks. To favor accurate class discrimination, multiscale information is also incorporated in the proposed approach, by integrating the network activations extracted at different spatial resolutions into the loss function directly through the introduction of suitable fully connected layers.

The experimental results demonstrate the effectiveness of the proposed DL approach reproducing a CRF model in mitigating the shortcomings of scarce training datasets usually available in land-cover mapping applications, as its classification accuracies are higher than or comparable to the ones of other state-of-the-art techniques. In particular, the proposed approach is capable to produce smoother, more homogeneous classification maps with sharper edges. The comparisons were performed with several state-of-the-art semantic segmentation methods, including: a technique defining an end-to-end trainable CRF, such as the CNN-CRF model where the CRF is formulated as an RNN and learned through an approximate mean-field approach; architectures based on multiscale information extraction, such as HRNet and an MFF technique; and the dc-Swin transformer-based approach.

The presented methodology, integrating multiresolution information, thanks to the fully connected neural networks, and spatial-contextual information, related to a CRF model, is particularly more effective as the training set approaches a realistic scenario of spatially sparse ground truth. In comparison to the other approaches, the proposed technique is capable of reaching higher values of recall and precision, thus reducing the commission and omission errors, and the relative amount of false negatives and false positives. Furthermore, the introduction of loss terms at different scales favors the interpretability of the corresponding hidden layers in terms of posterior probabilities estimated at the various resolutions—a positive by-product of the proposed approach from the viewpoint of the explainability of its processing scheme.

The proposed approach non-parametrically learns a rather general CRF model, although with two limitations: the CRF is supposed to include up to pairwise potentials and a smoothing Kronecker delta term. On one hand, this term may generally favor oversmoothing and does not take into account texture information. On the other hand, the experimental validation did not point out any spatial oversmoothing, at least on the considered datasets. Moreover, the texture information is captured through the unary potentials, which result from a multiscale convolutional architecture. Pairwise potentials can effectively model local context but lack longer range spatial information. From this viewpoint, a possible future generalization of the CRFNet approach could involve expanding it to higher order potentials, for instance, by integrating it with superpixel segmentation methods [64], [65]. Indeed, the proposed method is aimed at the semantic segmentation of single-resolution images and does not address the case of input multiresolution data. This case may be relevant especially when dealing with multisensor data with different spatial resolutions. In this respect, another potential extension of the approach could involve tackling multiresolution classification by integrating the considered planar CRF with hierarchical PGMs [17] in order to model interactions and dependencies between pixels at different spatial resolutions through Bayesian reasoning.

Future work could also focus on integrating the proposed approach with domain adaptation and transfer learning [66], [67], [68], [69], [70]. This integration could favor applicability in contexts that lack substantial ground-truth data, such as those associated with natural disaster response [71]. For instance, domain adaptation techniques have been successfully used to adapt models trained on labeled datasets to perform well on different but related unlabeled datasets. Methods based on generative networks for image-to-image translation, such as generative adversarial networks (GANs) [41], [72], [73], CycleGANs [74], variational autoencoders (VAEs) [75], might facilitate the extension of CRFNet to highly diverse and multimodal datasets. These may involve multisensor and multimission imagery (e.g., optical and radar data) across various domains.

Further generalizations of the proposed technique could also include the application to coarser resolution imagery (e.g., satellite images), which would involve the adaptation of the number and set of scales analyzed by the proposed model (e.g., addition or reduction of convolutional blocks in the fully convolutional network, and of multiscale terms in the loss function). The methodology could also be extended to SAR data, thanks to its non-parametric formulation, with the aforementioned modifications due to the spatial resolution, or to hyperspectral imagery. In the latter case, the proposed approach could be combined with methods for feature reduction, possibly integrated within the network, such as autoencoders [76], spectral convolutional networks [77], and RNNs [78]. Moreover, an important evolution of CRFNet could be its integration in operational workflows for the generation of thematic products from input high-resolution remote sensing images (e.g., high-resolution land cover mapping in climate change monitoring applications). In this case, it may be necessary to scale up to larger datasets with the aid of tiling strategies, for example combining the proposed classification method with the tiling algorithm developed in [79] for large-scale image registration and applied in [80] to multimission SAR data classification.

Appendix
Proof of Theorem 1

Let us focus on pixel $i \in S$ and class ω_k $(k = 1, 2, \dots, M)$. According to (13) and (14), the probability $\hat{P}_{ik}(\mathcal{X})$ of ω_k predicted on pixel i in the output of the network is given by

$$\ln \hat{P}_{ik}(\mathcal{X}) = h_0(\omega_k) f_i(\omega_k|\mathcal{X}) + \sum_{j\in\partial i} h_{i-j}(\omega_k) f_j(\omega_k|\mathcal{X}) + \phi_i(\mathcal{X}). \tag{20}$$

The proposed CRF model is defined by (15) and (17). Therefore, plugging the potentials of (15) into (4), the local posterior energy associated with the CRF is given by

$$\begin{aligned}\mathcal{U}_i(\omega_k|\mathbf{y}_{\partial i}, \mathcal{X}) &= D_i(\omega_k|\mathcal{X}) + \sum_{j\in\partial i} V_{ij}(\omega_k, y_j|\mathcal{X}) \\ &= -h_0(\omega_k) f_i(\omega_k|\mathcal{X}) \\ &\quad - \sum_{j\in\partial i} h_{i-j}(\omega_k) f_j(\omega_k|\mathcal{X})\delta(\omega_k, y_j).\end{aligned} \tag{21}$$

According to (3), this local posterior energy is, up to an additive constant, the negative logarithm of the local posterior distribution determined by the CRF,[4], i.e.,

$$\begin{aligned}\ln P^{(\mathsf{crf})}\{y_i = \omega_k|\mathbf{y}_{\partial i}, \mathcal{X}\} &= -\mathcal{U}_i(\omega_k|\mathbf{y}_{\partial i}, \mathcal{X}) + \phi_i'(\mathcal{X}) \\ &= \phi_i'(\mathcal{X}) + h_0(\omega_k) f_i(\omega_k|\mathcal{X}) \\ &\quad + \sum_{j\in\partial i} h_{i-j}(\omega_k) f_j(\omega_k|\mathcal{X})\delta(\omega_k, y_j)\end{aligned} \tag{22}$$

where $\phi_i'(\mathcal{X})$ indicates again a term that depends on the pixel location and the observations but is constant with respect to the class label.

If $y_j = \omega_k$ for all $j \in \partial i$, i.e., if the neighborhood of pixel i shares the same class label ω_k, then $\delta(\omega_k, y_j) = 1$ for all $j \in \partial i$. Therefore, (20) and (22) coincide up to an additive constant

$$\ln \hat{P}_{ik}(\mathcal{X}) = \ln P^{(\mathsf{crf})}\{y_i = \omega_k|\mathbf{y}_{\partial i}, \mathcal{X}\} + \phi_i''(\mathcal{X}) \tag{23}$$

where $\phi_i''(\mathcal{X})$ is a further term independent on ω_k. Due to the obvious sum-to-1 constraint with respect to k, (23) implies (18), thus proving the theorem.

Acknowledgment

The authors would like to thank the authors of the codes of HRNet [58], LWN-Attention [59], and dc-Swin [60], available on GitHub. The Vaihingen and Potsdam datasets were provided by the German Society for Photogrammetry, Remote Sensing and Geoinformation (DGPF). The authors would also like to thank the Belgian Royal Military Academy, for acquiring and providing the Zeebruges dataset, ONERA—The French Aerospace Lab, for providing the corresponding ground-truth data, and the IEEE GRSS Image Analysis and Data Fusion Technical Committee. University of Genoa and Université Côte d'Azur are part of the Ulysseus Alliance (European University). https://ulysseus.eu/

[4]The superscript "(crf)" emphasizes that the distribution on the right-hand side is modeled by the CRF established by (15)–(17).

REFERENCES

[1] G. Csurka, R. Volpi, and B. Chidlovskii, "Semantic image segmentation: Two decades of research," *Found. Trends Comput. Graph. Vis.*, vol. 14, nos. 1–2, pp. 1–162, 2022.

[2] R. Qin and T. Liu, "A review of landcover classification with very-high resolution remotely sensed optical images—Analysis unit, model scalability and transferability," *Remote Sens.*, vol. 14, no. 3, p. 646, Jan. 2022.

[3] M. Kampffmeyer, A.-B. Salberg, and R. Jenssen, "Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2016, pp. 680–688.

[4] L. Wang et al., "UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 190, pp. 196–214, Aug. 2022.

[5] Y. Xu, Z. Xie, Y. Feng, and Z. Chen, "Road extraction from high-resolution remote sensing imagery using deep learning," *Remote Sens.*, vol. 10, no. 9, p. 1461, Sep. 2018.

[6] M. Pastorino, F. Gallo, A. Di Febbraro, G. Moser, N. Sacco, and S. B. Serpico, "Multimodal fusion of mobility demand data and remote sensing imagery for urban land-use and land-cover mapping," *Remote Sens.*, vol. 14, no. 14, p. 3370, Jul. 2022.

[7] B. Benjdira, Y. Bazi, A. Koubaa, and K. Ouni, "Unsupervised domain adaptation using generative adversarial networks for semantic segmentation of aerial images," *Remote Sens.*, vol. 11, no. 11, p. 1369, Jun. 2019.

[8] F. Lateef and Y. Ruichek, "Survey on semantic segmentation using deep learning techniques," *Neurocomputing*, vol. 338, pp. 321–348, Apr. 2019.

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

[10] L. Maggiolo, D. Marcos, G. Moser, and D. Tuia, "Improving maps from CNNs trained with sparse, scribbled ground truths using fully connected CRFs," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Valencia, Spain, Jul. 2018, pp. 2099–2102.

[11] 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 J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 2387–2402, 2021.

[12] Y. Li, T. Shi, Y. Zhang, 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 J. Photogramm. Remote Sens.*, vol. 175, pp. 20–33, May 2021.

[13] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-6, no. 6, pp. 721–741, Nov. 1984.

[14] R. C. Dubes and A. K. Jain, "Random field models in image analysis," *J. Appl. Stat.*, vol. 16, no. 2, pp. 131–164, 1989.

[15] D. Zheng, X. Zhang, K. Ma, and C. Bao, "Learn from unpaired data for image restoration: A variational Bayes approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 5, pp. 5889–5903, May 2023.

[16] J. Shen, P.-C. Su, S. S. Cheung, and J. Zhao, "Virtual mirror rendering with stationary RGB-D cameras and stored 3-D background," *IEEE Trans. Image Process.*, vol. 22, no. 9, pp. 3433–3448, Sep. 2013.

[17] M. Pastorino, G. Moser, S. B. Serpico, and J. Zerubia, "Semantic segmentation of remote-sensing images through fully convolutional neural networks and hierarchical probabilistic graphical models," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5407116.

[18] D. Liu, G. Hua, and T. Chen, "A hierarchical visual model for video object summarization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 12, pp. 2178–2190, Dec. 2010.

[19] J. van de Ven, F. Ramos, and G. D. Tipaldi, "An integrated probabilistic model for scan-matching, moving object detection and motion estimation," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2010, pp. 887–894.

[20] Y. Wang et al., "A low rank promoting prior for unsupervised contrastive learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 2667–2681, Mar. 2023.

[21] N. Friedman, D. Geiger, G. Provan, P. Langley, and P. Smyth, "Bayesian network classifiers," *Mach. Learn.*, vol. 29, pp. 131–163, May 1997.

[22] P. A. Devijver, "Hidden Markov mesh random field models in image analysis," *J. Appl. Statist.*, vol. 20, nos. 5–6, pp. 187–227, Jan. 1993.

[23] S. Li, *Markov Random Field Modeling in Image Analysis*. Cham, Switzerland: Springer, 2009.

[24] J.-M. Laferte, P. Perez, and F. Heitz, "Discrete Markov image modeling and inference on the quadtree," *IEEE Trans. Image Process.*, vol. 9, no. 3, pp. 390–404, Mar. 2000.

[25] A. Blake, P. Kohli, and C. Rother, *Markov Random Fields for Vision and Image Processing*. Cambridge, MA, USA: MIT Press, 2011.

[26] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. 18th Int. Conf. Mach. Learn. (ICML)*. San Francisco, CA, USA: Morgan Kaufmann, 2001, pp. 282–289.

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

[28] M. Pastorino et al., "Multisensor and multiresolution remote sensing image classification through a causal hierarchical Markov framework and decision tree ensembles," *Remote Sens.*, vol. 13, no. 5, p. 849, Feb. 2021.

[29] Y. Yang, X. Tang, Y.-M. Cheung, X. Zhang, and L. Jiao, "SAGN: Semantic-aware graph network for remote sensing scene classification," *IEEE Trans. Image Process.*, vol. 32, pp. 1011–1025, 2023.

[30] L.-C. Chen, A. Schwing, A. Yuille, and R. Urtasun, "Learning deep structured models," in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 37, F. Bach and D. Blei, Eds., Lille, France: PMLR, 2015, pp. 1785–1794.

[31] T. Do and T. Artieres, "Neural conditional random fields," in *Proc. Int. Conf. Artif. Intell. Stat. (AISTATS)*, vol. 9, Y. W. Teh and M. Titterington, Eds., Sardinia, Italy: PMLR, 2010, pp. 177–184.

[32] X. Lan and K. Barner, "From MRFS to CNNS: A novel image restoration method," in *Proc. 52nd Annu. Conf. Inf. Sci. Syst. (CISS)*, Mar. 2018, pp. 1–5.

[33] H. Xiong and N. Ruozzi, "General purpose MRF learning with neural network potentials," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Jul. 2020, pp. 2769–2776.

[34] Y. Abouqora, O. Herouane, L. Moumoun, and T. Gadi, "A hybrid CNN-CRF inference models for 3D mesh segmentation," in *Proc. 6th IEEE Congr. Inf. Sci. Technol. (CiSt)*, Jun. 2020, pp. 296–301.

[35] J. Xiang, G. Xu, C. Ma, and J. Hou, "End-to-end learning deep CRF models for multi-object tracking deep CRF models," *IEEE Trans. IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 1, pp. 275–288, Jan. 2021.

[36] S. Zheng et al., "Conditional random fields as recurrent neural networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1529–1537.

[37] A. Arnab et al., "Conditional random fields meet deep neural networks for semantic segmentation: Combining probabilistic graphical models with deep learning for structured prediction," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 37–52, Jan. 2018.

[38] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[39] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected CRFs with Gaussian edge potentials," in *Proc. 24th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, 2011, pp. 109–117.

[40] G. Lin, C. Shen, A. van den Hengel, and I. Reid, "Efficient piecewise training of deep structured models for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3194–3203.

[41] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[42] F. Liu, G. Lin, and C. Shen, "Discriminative training of deep fully connected continuous CRFs with task-specific loss," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2127–2136, May 2017.

[43] K. Fu, I. Y. Gu, and J. Yang, "Saliency detection by fully learning a continuous conditional random field," *IEEE Trans. Multimedia*, vol. 19, no. 7, pp. 1531–1544, Jul. 2017.

[44] F. I. Alam, J. Zhou, A. W. Liew, and X. Jia, "CRF learning with CNN features for hyperspectral image segmentation," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2016, pp. 6890–6893.

[45] F. I. Alam, J. Zhou, A. W. Liew, X. Jia, J. Chanussot, and Y. Gao, "Conditional random field and deep feature learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 3, pp. 1612–1628, Mar. 2019.

[46] D. Tuia, M. Volpi, and G. Moser, "Decision fusion with multiple spatial supports by conditional random fields," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 6, pp. 3277–3289, Jun. 2018.

[47] C. Sutton, "An introduction to conditional random fields," *Found. Trends Mach. Learn.*, vol. 4, no. 4, pp. 267–373, 2012.

[48] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, vol. 9351. Cham, Switzerland: Springer, 2015, pp. 234–241.

[49] Z. Kato and J. Zerubia, "Markov random fields in image segmentation," *Found. Trends Signal Process.*, vol. 5, nos. 1–2, pp. 1–155, 2012.

[50] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques* (Adaptive Computation and Machine Learning). Cambridge, MA, USA: MIT Press, 2009.

[51] M. Campos-Taberner et al., "Processing of extremely high-resolution LiDAR and RGB data: Outcome of the 2015 IEEE GRSS data fusion contest—Part A: 2-D contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 12, pp. 5547–5559, Dec. 2016.

[52] A.-V. Vo et al., "Processing of extremely high resolution LiDAR and RGB data: Outcome of the 2015 IEEE GRSS data fusion contest—Part B: 3-D contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 12, pp. 5560–5575, Dec. 2016.

[53] N. Audebert, B. Le Saux, and S. Lefèvre, "Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks," *ISPRS J. Photogramm. Remote Sens.*, vol. 140, pp. 20–32, Jun. 2018.

[54] Q. Liu, M. Kampffmeyer, R. Jenssen, and A.-B. Salberg, "Dense dilated convolutions' merging network for land cover classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 9, pp. 6309–6320, Sep. 2020.

[55] 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 Geosci. Remote Sens. Lett.*, vol. 18, no. 12, pp. 2172–2176, Dec. 2021.

[56] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2014, pp. 1–15.

[57] S. Wu, G. Wang, P. Tang, F. Chen, and L. Shi, "Convolution with even-sized kernels and symmetric padding," in *Proc. Int. Conf. Neural Inf. Process. Syst. (NeurIPS)*. Red Hook, NY, USA: Curran Associates, 2019, no. 108, pp. 1194–1205.

[58] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 5686–5696.

[59] S. Liu, C. He, H. Bai, Y. Zhang, and J. Cheng, "Light-weight attention semantic segmentation network for high-resolution remote sensing images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Sep. 2020, pp. 2595–2598.

[60] L. Wang, R. Li, C. Duan, C. Zhang, X. Meng, and S. Fang, "A novel transformer based semantic segmentation scheme for fine-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[61] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, Jun. 1947.

[62] G. Foody, "Thematic map comparison: Evaluating the statistical significance of differences in classification accuracy," *Photogramm. Eng. Remote Sens.*, vol. 70, pp. 627–633, May 2004.

[63] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Comput.*, vol. 10, no. 7, pp. 1895–1923, Oct. 1998.

[64] A. Arnab, S. Jayasumana, S. Zheng, and P. H. S. Torr, "Higher order conditional random fields in deep neural networks," in *Computer Vision—ECCV 2016*. Cham, Switzerland: Springer, 2016, pp. 524–540.

[65] J. D. Wegner, J. A. Montoya-Zegarra, and K. Schindler, "A higher-order CRF model for road network extraction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1698–1705.

[66] Q. Wang, J. Gao, and X. Li, "Weakly supervised adversarial domain adaptation for semantic segmentation in urban scenes," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4376–4386, Sep. 2019.

[67] H. Xu, M. Yang, L. Deng, Y. Qian, and C. Wang, "Neutral cross-entropy loss based unsupervised domain adaptation for semantic segmentation," *IEEE Trans. Image Process.*, vol. 30, pp. 4516–4525, 2021.

[68] D. Tuia, C. Persello, and L. Bruzzone, "Domain adaptation for the classification of remote sensing data: An overview of recent advances," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 41–57, Jun. 2016.

[69] G. Matasci, M. Volpi, M. Kanevski, L. Bruzzone, and D. Tuia, "Semisupervised transfer component analysis for domain adaptation in remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 7, pp. 3550–3564, Jul. 2015.

[70] O. Tasar, A. Giros, Y. Tarabalka, P. Alliez, and S. Clerc, "DAugNet: Unsupervised, multisource, multitarget, and life-long domain adaptation for semantic segmentation of satellite images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 2, pp. 1067–1081, Feb. 2021.

[71] 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," *Proc. IEEE*, vol. 100, no. 10, pp. 2946–2970, Oct. 2012.

[72] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5967–5976.

[73] L. Maggiolo, D. Solarna, G. Moser, and S. B. Serpico, "Registration of multisensor images through a conditional generative adversarial network and a correlation-type similarity measure," *Remote Sens.*, vol. 14, no. 12, p. 2811, Jun. 2022.

[74] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2242–2251.

[75] L. T. Luppino et al., "Deep image translation with an affinity-based change prior for unsupervised multimodal change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4700422.

[76] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006.

[77] A. Plaza et al., "Recent advances in techniques for hyperspectral image processing," *Remote Sens. Environ.*, vol. 113, pp. S110–S122, Aug. 2009.

[78] L. Mou, P. Ghamisi, and X. X. Zhu, "Deep recurrent neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3639–3655, Jul. 2017.

[79] D. Solarna, L. Maggiolo, G. Moser, and S. B. Serpico, "A tiling-based strategy for large-scale multisensor optical-SAR image registration," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2022, pp. 127–130.

[80] M. Pastorino, G. Moser, S. B. Serpico, and J. Zerubia, "Multimission, multifrequency, and multiresolution SAR image classification through hierarchical Markov models and convolutional networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024.

Martina Pastorino (Member, IEEE) received the B.Sc. degree in electronic engineering and information technologies from the University of Genoa, Genoa, Italy, in 2018, the double M.Sc. degree in telecommunication engineering from the University of Genoa, and IMT Atlantique, Brest, France, in 2020, and the Ph.D. degree in science and technology for electronic and telecommunication engineering from the University of Genoa, and Inria d'Université Côte d'Azur, Sophia-Antipolis, France, in 2023.

She is currently a Post-Doctoral Research Fellow with the University of Genoa. Her research interests include combination of stochastic models and deep learning techniques for remote sensing image analysis.

Dr. Pastorino received the Best Student Paper Award at IGARSS 2021 and the Prix d'excellence d'Université Côte d'Azur in 2021.

Gabriele Moser (Fellow, IEEE) 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, Genoa, Italy, in 2001 and 2005, respectively.

Since 2001, he has been cooperated with the Image Processing and Pattern Recognition for Remote Sensing Laboratory, University of Genoa. Since 2013, he has been the Head of the Remote Sensing for Environment and Sustainability Laboratory, Savona Campus of the University of Genoa. From January 2004 to March 2004, he was a Visiting Student with the Institut National de Recherche en Informatique et en Automatique (INRIA), Sophia Antipolis, France, where he has been an External Collaborator with the Ayin and Ayana Laboratories since 2012. In 2016, he spent a period as a Visiting Professor with the Institut National Polytechnique de Toulouse, Toulouse, France. Since 2019, he has been the Head of the M.Sc. Program of Engineering for Natural Risk Management with the University of Genoa.

From 2021 to 2023, he was a member of the National Evaluation Committee for the National Scientific Qualification ("Abilitazione Scientifica Nazionale") as a Full and an Associate Professors in the Telecommunications field in Italy. He is a Full Professor of telecommunications with the University of Genoa. His research interests include pattern recognition and image processing methodologies for remote sensing and energy applications.

Dr. Moser received the Best Paper Award at the 2010 IEEE Workshop on Hyperspectral Image and Signal Processing (WHISPERS) and the Interactive Symposium Paper Award at IGARSS 2016. He was a co-author of the paper awarded with the 2021 IEEE GRSS Mikio Takagi Student Prize at IGARSS 2021. He served as the Chair for the IEEE GRSS Image Analysis and Data Fusion Technical Committee (IADF TC) from 2013 to 2015, and the IADF TC Co-Chair from 2015 to 2017. He was the Publication Co-Chair of the 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), the Technical Program Co-Chair of the IEEE GRSS EarthVision Workshop at the 2015 IEEE/CVF Computer Vision and Pattern Recognition conference (CVPR), and the Co-Organizer of the Second Edition of EarthVision at CVPR 2017. He has been an Associate Editor of 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 Coeditor of the September 2015 special issue of *IEEE Geoscience and Remote Sensing Magazine*.

Sebastiano B. Serpico (Life Fellow, IEEE) received the Laurea (M.S.) degree in electronic engineering and the Ph.D. degree in telecommunications from the University of Genoa, Genoa, Italy, in 1982.

He is a Full Professor of telecommunications with the Polytechnic School, University of Genoa. He is the Coordinator of the Research Group on Signal Processing and Recognition Methods and Systems with the Department of Electrical, Electronic, Telecommunications Engineering, and Naval Architecture (DITEN), University of Genoa. 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, and Italian Ministry of Education and Research. He has authored (or co-authored) of about 300 scientific articles published in journals and conference proceedings. His research interests include pattern recognition for remote sensing image analysis.

Dr. Serpico 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 was the Chairperson of the Institute of Advanced Studies in Information and Communication Technologies (ISICT) from 2003 to 2019. He is the Chairperson of the Advanced Education School, University of Genoa, in 2022. From 1998 to 2002, he was the Chairperson of the SPIE/EUROPTO series of conferences on *Signal and Image Processing for Remote Sensing*. He was an Associate Editor of the international journal IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING (TGRS) (2001–2022). 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 PROCEEDINGS OF THE IEEE on "Remote Sensing of Natural Disasters." He was the Co-Chair of the IEEE International Geoscience and Remote Sensing Symposium in 2015 (Milan, Italy).

Josiane Zerubia (Fellow, IEEE) received the M.Sc. degree from the Department of Electrical Engineering, ENSIEG, Grenoble, France, in 1981, the Doctor of Engineering, the Ph.D., and the "Habilitation" degrees from the University of Nice, Nice, France, in 1986, 1988, and 1994, respectively, and the Doctor Honoris Causa degree from the University of Szeged, Szeged, Hungary, in 2020.

She has been a Permanent Research Scientist with INRIA, Sophia Antipolis, France, since 1989 and the Director of Research since July 1995 (DR Exceptional Class since 2023; DR first Class from 2002 to 2022). She was the Head of the PASTIS Remote Sensing Laboratory (INRIA) from 1995 to 1997 and 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 the Head of the 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 the Head of the 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 a Professor (PR1) with SUPAERO (ISAE), Toulouse, France, from 1999 to 2020. Her main research interest is in image processing using probabilistic models. She also works on parameter estimation, statistical learning, optimization techniques, and artificial intelligence.

Dr. Zerubia is a fellow of EURASIP in 2019 and IAPR in 2020, and IEEE SP Society Distinguished Lecturer 2016–2017. She was a member of the IEEE IMDSP TC (SP Society) from 1997 to 2003, of the IEEE BISP TC (SP Society) from 2004 to 2012 and of the IVMSP TC (SP Society) from 2008 to 2013. She was a member of *IEEE Signal Processing Magazine* Senior Editorial Board from September 2018 to January 2022. She was Member-at-Large of the Awards Board of the IEEE SP Society from 2020 to 2022. Finally, she was a member of the Best Paper Award Committee for EURASIP JIVP in 2021 and also member of the IAPR Fellow committee in 2021 and 2022. She received the Excellence Award from University of Nice (now UCA) for three times in 2020, 2019, and 2016. She has been an Editorial Board Member of the Foundation and Trends in Signal Processing since 2007 and of the IEEE WISP Committee since 2024. She was the General Co-Chair of the EarthVision workshop at IEEE CVPR 2015 (Boston, USA) and a Organizing Committee Member of the IEEE-EURASIP EUSIPCO 2015 (Nice, France), of the EarthVision Workshop (Co-Chair) at IEEE CVPR 2017 (Honolulu, USA), and GRETSI 2017 Symposium (Juan-les-Pins, France). She was a Scientific Advisor and a Co-Organizer of ISPRS 2020 (virtual), 2021 (virtual) and 2022 congress (Nice, France) and the Technical Co-Chair of IEEE-EURASIP EUSIPCO 2021 (virtual, Dublin, Ireland). She was an Associate Editor of IEEE TRANSACTION ON IMAGE PROCESSING from 1998 to 2002, an Area Editor of IEEE TRANSACTION ON IMAGE PROCESSING from 2003 to 2006, a Guest Coeditor of a special issue of IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE in 2003, an Editorial Board Member of the *International Journal of Computer Vision* from 2004 to March 2013, and the Member-at-Large of the Board of Governors of the IEEE SP Society from 2002 to 2004. She was an Editorial Board Member of the French Society for Photogrammetry and Remote Sensing (SFPT) from 1998 to 2020, and the Member-at-Large of the Board of Governors of the SFPT from 2014 to 2020. She is currently part of the organizing committee for IEEE ICASSP'28 proposal (Washington, DC, USA). See http://www-sop.inria.fr/members/Josiane.Zerubia/index-eng.html for more detail.

Open Access funding provided by 'Università degli Studi di Genova' within the CRUI CARE Agreement