Decision Fusion With Multiple Spatial Supports by Conditional Random Fields

Devis Tuia, *Senior Member, IEEE*, Michele Volpi, and Gabriele Moser, *Senior Member, IEEE*

***Abstract*— Classification of remotely sensed images into land cover or land use is highly dependent on geographical information at least at two levels. First, land cover classes are observed in a spatially smooth domain separated by sharp region boundaries. Second, land classes and observation scale are also tightly intertwined: they tend to be consistent within areas of homogeneous appearance, or regions, in the sense that all pixels within a roof should be classified as roof, independently on the spatial support used for the classification. In this paper, we follow these two observations and encode them as priors in an energy minimization framework based on conditional random fields (CRFs), where classification results obtained at pixel and region levels are probabilistically fused. The aim is to enforce the final maps to be consistent not only in their own spatial supports (pixel and region) but also *across* supports, i.e., by getting the predictions on the pixel lattice and on the set of regions to agree. To this end, we define an energy function with three terms: 1) a data term for the individual elements in each support (support-specific nodes); 2) spatial regularization terms in a neighborhood for each of the supports (support-specific edges); and 3) a regularization term between individual pixels and the region containing each of them (intersupports edges). We utilize these priors in a unified energy minimization problem that can be optimized by standard solvers. The proposed 2L$\frac{1}{2}$ CRF model consists of a CRF defined over a bipartite graph, i.e., two interconnected layers within a single graph accounting for interlattice connections. 2L$\frac{1}{2}$ CRF is tested on two very high-resolution data sets involving submetric satellite and subdecimeter aerial data. In all cases, 2L$\frac{1}{2}$ CRF improves the result obtained by the independent base model (either random forests or convolutional neural networks) and by standard CRF models enforcing smoothness in the spatial domain.**

***Index Terms*— Classification, conditional random fields (CRF), convolutional neural networks (CNNs), hierarchical models, region-based analysis, semantic labeling.**

Manuscript received September 21, 2017; revised December 18, 2017; accepted January 17, 2018. Date of publication March 19, 2018; date of current version May 21, 2018. This work was supported by the Swiss National Science Foundation through "Multimodal Machine Learning for Remote Sensing Information Fusion" under Grant 150593. *(Corresponding author: Devis Tuia.)*

D. Tuia was with the MultiModal Remote Sensing Group, University of Zürich, 8006 Zürich, Switzerland. He is now with the GeoInformation Science and Remote Sensing Laboratory, Wageningen University and Research, 6708 PB Wageningen, The Netherlands (e-mail: devis.tuia@wur.nl).

M. Volpi was with the MultiModal Remote Sensing Group, University of Zürich, 8006 Zürich, Switzerland. He is now with the Swiss Data Science Center, ETH Zurich, 8006 Zürich, Switzerland (e-mail: michele.volpi@datascience.ch).

G. Moser is with the Department of Electrical, Electronic, Telecommunication Engineering, and Naval Architecture, University of Genoa, 16145 Genoa, Italy (e-mail: gabriele.moser@unige.it).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TGRS.2018.2797316

I. Introduction

IMAGES with metric to decimetric spatial resolutions are becoming the new standard for very high-resolution (VHR) remote sensing: images acquired by new-generation satellites, as well as by sensors mounted on aircrafts and drones, allow geometrically precise monitoring of the earth's surface and can lead to breakthroughs in agriculture [1], forestry [2], urban characterization [3], and search-and-rescue [4] tasks.

If the promise of VHR images is great, the tasks to be solved also become harder: it is well known that as the spatial resolution increases, so does the complexity of semantic labeling (i.e., tasks aiming at assigning each pixel to a semantic class). This is particularly true because of the increase of the intraclass variance and a parallel decrease of the interclass variance. On the one hand, a single semantic class is generally composed of different materials (e.g., a building class is composed spectrally of several heterogeneous materials found on the roof). On the other hand, different classes share the same materials (e.g., vegetation is found in both forests and meadows). To address such ambiguity, researchers in remote sensing have explored the use of spatial context and structure, for which we will now review the main families of solutions.

A. Object-Based Image Analysis

A first branch of research considers the possibility to find a coarser spatial support for analysis—typically an agglomeration of pixels, a *region*—prior to learning a classifier. Also known as *object-based image analysis* [5], these methods aim at defining a new spatial support, possibly respecting the color gradients of the image, extracting some textural and color features at the region level, and learning a supervised classifier. In computer vision, the task of finding coarser but coherent spatial supports is often referred to as *superpixelization* [6]–[8], where a set of coherent regions are found in the image, most often without semantic meaning nor pretense to enclose a complete object in a single region, but only assuming they contain parts of it and share parts of the borders.

B. Spatial Filters

A second vivid research subfield is concerned with defining spatial *filters*, which are image processing techniques applied to the image bands and aiming at characterizing locally relevant properties of the semantic objects to be labeled. For instance, to encode the property that nearby pixels tend to be of the same class, local convolutions implementing low-pass filters can be used, while morphological filters [9] and

texture filters [10] are often used to encode more complex local relationships. Recently, powerful descriptors from computer vision such as Fisher vectors [11], [12], bag-of-visual-words [13]–[16], or local binary patterns [17] have been proposed to extract spatial information for remote sensing semantic labeling at VHR. But a recurrent problem is to select the right bands to be filtered, the family of filters and their parameters. These are tasks that can prove to be very complex for a nonspecialist and whose failure can lead to very suboptimal models. Recent research addresses this issue by learning the relevant filters in an unconstrained space of filter candidates [18], or to learn further combinations of such filters, therefore, reaching, through the hierarchy of filters, higher level concepts closer to the semantic classes [19]. Volpi and Ferrari [20] build very large feature sets containing all the information which is possibly useful to characterize the classification problem and then use a classifier that is robust to unimportant features (i.e., embedding feature selection).

C. Convolutional Neural Networks

The need to learn the relevant filters and features, instead of predefining them by expert knowledge, focused recent research in *deep learning* as a tool for automatically learning multiscale, nonlinear, semantically tailored, and problem-specific contextual filters jointly with a classifier. *Convolutional neural networks* (CNNs [21]) are the perfect tool for solving such task: they learn unconstrained convolutional filters hierarchically. The filters learned in the first level (the first *convolutional layer*) of the network correspond mostly to local oriented gradients, while those learned at the last convolutional layers have a much wider receptive field (they depend on a wider area of the original images) and, therefore, inform about a larger spatial context, usually semantically more related to the problem to be solved. Since the filters are directly learned by backpropagation, no effort has to be provided for filter engineering, but all effort is moved to the design of the network structure itself (number of layers, type of nonlinearities, spatial pooling, size of filters, etc.). For a comprehensive introduction to CNN building blocks in remote sensing see [22].

Thanks to numerous benchmark data released recently, the development of CNN architectures for the semantic labeling of VHR aerial images has flourished [23]: the first attempts (see [3]) performed inference in a sliding window fashion, therefore, retrieving the complete map pixel by pixel. In this way, one could use a standard classification architecture to map from a patch to a single label (supposed to represent the pixel on which the patch is centered on) and retrieve the whole test map one patch at a time. Besides being inefficient, this also limits the power of the CNN to encode spatial information. A network learned for dense prediction (i.e., outputting the entire map for each pixel in the patch) implicitly encodes spatial relations between the different classes and their features at different scales, while a network predicting a single label and assigning it to the central pixel will consider each patch centered on each pixel independently.

Dense prediction by CNN, for instance inspired by fully convolutional networks [24], hypercolumns [25], or SegNet [26], was proposed in the field of computer vision to upscale the CNN last layers to the original input resolution and *de facto* provided an elegant, fully learnable, solution to the problem of dense semantic labeling. As a consequence, it was quickly adopted in remote sensing: Volpi and Tuia [22], Audebert *et al.* [27], and Marmanis *et al.* [28] propose to use learned deconvolution layers to upsample the activations, while in [29], the activations are simply upsampled by interpolation. Maggiori *et al.* [30] and Marcos *et al.* [31] stack upsampled activations at multiple scales to train other layers performing dense prediction. Finally, Sherrah [32] proposes a network without spatial poolings, i.e., a network that preserves the spatial resolution of the original data throughout all its components.

However, besides the impressive performances reported, most approaches: 1) still showed some residual class inconsistencies and 2) disregarded explicit object structure that could be easily integrated by considering the superpixelization discussed above or some degree of spatial reasoning (e.g., a building rarely occurs in the middle of a river).

D. Structured Prediction

The fourth family of methods addresses this last point by encoding spatial reasoning via interaction potentials between spatial units. Models such as Markov [33]–[35] and conditional random fields [36], [37] (MRFs or CRFs) can be used to jointly account for different kinds of prior information about spatial relationships and local likelihoods. They have been used extensively in remote sensing to encode different kinds of assumptions one has about the data, such as local spatial smoothness [38]–[40], linearity [41], or 3-D arrangements [42]. Owing to the Hammersley–Clifford theorem [33], random field models make it possible to express the posterior joint distribution of the unknown class labels given the observations as a Gibbs distribution. This leads to formulating the maximum *a posteriori* (MAP) criterion as the minimization of an appropriate energy function, modeling conditional relationships across variables. The energy can be defined by a number of factors, generally including local likelihood scores from a (set of) classifier(s) [43] and models of spatial relations across locations and their labels. These relations usually encode the likelihood of a pixel to belong to a class, given the class of its direct neighbors [16], [44], or aim at capturing co-occurrence structures among the classes [20], [45], [46]. In addition, the definition of suitable Markovianity properties on hierarchical graphs, including quad-trees, binary partition trees, or more irregular topologies [47], [48], makes it possible to also formalize multiscale and multiresolution fusion within an MRF/CRF framework. Recent approaches based on random fields have been proposed for the semantic segmentation of data acquired at multiple input resolutions at the same time or in a time series [49], [50]. Finally, higher order random field models exist [51], which can be used to encode higher order relationships between elements forming a suitable group (named clique): in remote sensing, such reasoning was exploited to

detect roads [52] or relationships between different types of classes occurring on different cliques, such as land cover and land use [53]. Even if these models allow to encode much more complex relationships, they also involve large computational load to perform inference.

E. Co-segmentation

Our method also shares similarities with the domain of image co-segmentation. Research in this direction is often framed as a weakly supervised task, where labels for semantic segmentation or object detection are not dense over the pixel lattice but only appear as general "presence" or "absence" attributes for a given class. It is also commonly phrased as an unsupervised step, where objects belonging to the same object class in different images have to be clustered together. co-segmentation aims then at building models able to extract common objects in a series of images under the prior that the same objects are present [54], [55].

In our proposed method, we build on the assumption that the same classes are present at the same locations, rather than at different locations with possibly different viewpoints. We treat the class label for a given spatial coordinate as the random variable to be fit from a set of semantic classes, and we do not additionally fit a binary segmentation masks (foreground background) or bounding boxes while grouping the same classes [54]. This simplification has also been used in [56], where co-segmentation of the multitemporal image pair is driven by the pixelwise difference image.

In this paper, we aim at fusing the advantages of these families of techniques by proposing a model that accommodates different spatial supports (or lattices, connectivity graphs) and encodes spatial reasoning relevant to the problem. We integrate the pixel- and region-based strategies within a multiscale approach, letting two (or more) spatial supports interact and come to a common decision. The likelihood of each granularity level corresponds to an independent classification model on a given spatial support. These responses at all levels are fused probabilistically using a bipartite CRF (i.e., a CRF with two interconnected layers [57]). The model, named 2L$\lightning$CRF (read "two-layered-lightning-CRF"[1]), is sketched in Fig. 1. 2L$\lightning$CRF can use as inputs the output of any classifier providing a distribution over the labels: they can be obtained by using predefined spatial features into standard classifiers or by training CNNs. 2L$\lightning$CRF performs structured prediction accounting simultaneously for a pixel- and a region-based granularity by encoding: 1) spatial structuring via contrast-sensitive pairwise potentials [58] and 2) consistency between the labelings at the pixel and region lattices via an interlayer smoothness assumption. Differently, from higher order models such as the P^{n}Potts [51], our conditional random field avoids the multiscale structure of the graph and the use of auxiliary variables by encoding the multiscale problem as a single (flat) graph with interscale connections, thus allowing to use standard and computationally efficient energy minimization solvers. It is more efficient that the iterative formulation of [59] (which is the base of [53]), since it solves both problems within a single energy minimization step.

[1]The lightning symbol "$\lightning$" is often used in proofs when contradictory statements are involved. Here, it is used on purpose to emphasize that the proposed CRF model is aimed at fusing the two layers, resolving contradictions between their prediction, and "getting pixels and regions" to agree on the final labeling.

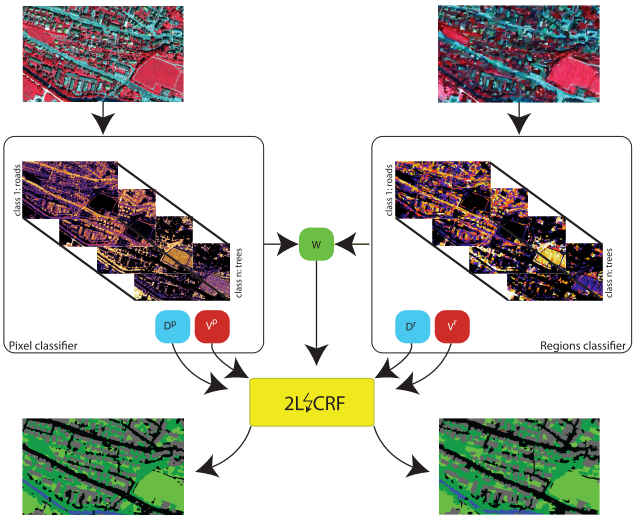

Fig. 1. Proposed 2L$\lightning$CRF: starting from two models trained on different spatial supports, it uses the posterior probability scores D, spatial relations V, and hierarchical information W to find the most likely maps [colors and symbols refer to Equation (6)].

A preliminary version of this paper was previously introduced by the authors in a conference paper [60]. We extend it to this paper, where we provide an in-depth methodological analysis, a study of the impact of CNN unary scores to the system and results on two data sets, Zurich Summer [16] and the challenging 2015 Data Fusion Contest data set [3].

In Section II, we present the formulation of the proposed model, as well as the algorithm used to solve the energy minimization problem. In Section III, we present the data sets and the setup of experiments discussed in Section IV. Section V concludes this paper.

II. Proposed 2L$\lightning$CRF Method

A. Conditional Random Fields for Semantic Labeling

In the framework of semantic labeling, CRFs represent a family of probabilistic models allowing to jointly characterize pixelwise class statistics as well as spatial dependencies between the labeling of different neighboring locations. They are among the most used probabilistic graphical models for remote sensing image analysis.

Let us consider a remote sensing image, from which m features have been extracted. The image presents C thematic classes provided with training samples. Let $\mathcal{I}$ be the regular pixel lattice, and $\boldsymbol{x}_i$ and y_i be the feature vector and the class label of the ith pixel ($i \in \mathcal{I}$; $\boldsymbol{x}_i \in \mathbb{R}^m$; $y_i \in \{1, 2, \ldots, C\}$). The CRF approach considers $\boldsymbol{x}_i$ and y_i as samples from two random fields, i.e., a (generally continuous-valued) random field $\mathcal{X} = \{\boldsymbol{x}_i\}_{i\in\mathcal{I}}$ of feature vectors and a discrete-valued random field $\mathcal{Y} = \{y_i\}_{i\in\mathcal{I}}$ of class labels. Both the random fields are supported on the pixel lattice $\mathcal{I}$, on which a neighborhood

system $\{\partial i\}_{i\in\mathcal{I}}$ is defined [35]. The common choices include the first and second order neighborhood systems, in which ∂i is made of the four pixels adjacent to the ith pixel and the eight pixels surrounding it, respectively [40]. With these notations, the random field $\mathcal{Y}$ of the class labels is said to be a CRF if the following posterior Markovianity property holds:

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

and if the (global, imagewise) posterior distribution $P(\mathcal{Y}|\mathcal{X})$ is strictly positive [37], [40]. Condition (1) means that the labels are spatially Markovian when conditioned to the random field of the feature vectors. This property makes it possible to express the posterior distribution as $P(\mathcal{Y}|\mathcal{X}) \propto \exp[-U(\mathcal{Y}|\mathcal{X})]$, where the energy $U(\mathcal{Y}|\mathcal{X})$ is defined according to the neighborhood system [33], [40]. Minimizing such energy corresponds to the MAP criterion. For a CRF with only pairwise nonzero clique potentials, such energy can be written as [40]

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

where $D_i(y_i|\mathcal{X})$, named *unary* or association potential, is related to the statistics of each individual label given the feature random field. $V_{ij}(y_i, y_j|\mathcal{X})$, named *pairwise* or interaction potential, encodes the spatial relations among the labels of neighboring pixels ($i \in \mathcal{I}, j \in \partial i$). λ is a positive parameter that tunes the tradeoff between the unary and pairwise terms. The possibility to characterize the desired spatial interactions by defining suitable pairwise potentials, tailored to the application considered, and the availability of computationally efficient energy minimization algorithms (see Section II-D) make CRF modeling a powerful and flexible approach to structured prediction in remote sensing image analysis [38], [39].

B. Overview and Methodological Assumptions of 2L$\frac{1}{2}$ CRF

The key idea of the 2L$\frac{1}{2}$ CRF method is to benefit from both pixelwise and region-based image representations by introducing a novel model that connects two CRFs, defined on the pixel lattice and on a segmentation result, to perform structured prediction at both levels simultaneously. Given a VHR image, a segmentation method is first applied to identify a set of homogeneous regions. An arbitrary segmentation technique can be used within 2L$\frac{1}{2}$ CRF, provided that the resulting segments are not particularly coarse.

We denote explicitly with a superscript "p" the quantities introduced in the previous section at the granularity of individual pixels ($\mathcal{I}^p, \partial^p i, \mathcal{X}^p, \mathcal{Y}^p$, etc.). Let $\mathcal{I}^r$ be the set of regions resulting from segmentation, $\boldsymbol{x}_k^r$ be an n-dimensional feature vector extracted from the image data of the kth region, and y_k^r be the class label of the same region ($\boldsymbol{x}_k^r \in \mathbb{R}^n$; $y_k^r \in \{1, 2, \ldots, C\}$; $k \in \mathcal{I}^r$). This is equivalent to introducing a second pair of random fields $\mathcal{X}^r = \{\boldsymbol{x}_k^r\}_{k\in\mathcal{I}^r}$ and $\mathcal{Y}^r = \{y_k^r\}_{k\in\mathcal{I}^r}$ that collect feature vectors and labels at the granularity of regions. The proposed method formalizes the relation between labels and feature vectors at each granularity layer as a CRF, and merges these two CRFs into a unique energy (see Section II-C).

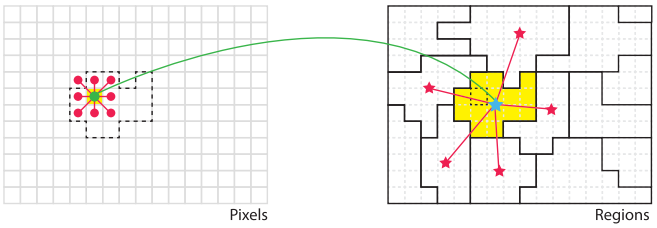

Fig. 2. General idea of the 2L$\frac{1}{2}$ CRF structured prediction model used to connect the pixel lattice and the set of regions. Red links represent the pairwise contrast-sensitive potentials [2) terms in (6) between pixels (represented by circles) or region centroids (represented by stars)]. The green link represents the cross-layer potential relating a region and each pixel included in that region [3) term in (6)]. The locations of the considered pixel and region in the other layer are highlighted with black dashed lines.

From the probabilistic graphical modeling viewpoint, this means using a bipartite graph to combine the two layers and generate a unique labeling at the spatial resolution of the pixel lattice. In the language of data fusion, the method fuses the information associated with pixelwise statistics and with two sources of spatial information, i.e., local neighborhoods and region-based reasoning. From a computational perspective, the resulting energy is also equivalent to a case-specific single-layer model on a planar graph—a property that makes it possible to use time-efficient algorithms to numerically address the energy minimization task (see Section II-D).

More precisely, 2L$\frac{1}{2}$ CRF is based on the following methodological assumptions.

1) In addition to the neighborhood system $\{\partial^p i\}_{i\in\mathcal{I}^p}$ on the pixel lattice, a neighborhood system $\{\partial^r j\}_{j\in\mathcal{I}^r}$ is also defined on the set of regions.
2) The random field $\mathcal{Y}^s$ of the labels at each granularity layer $s \in \{p, r\}$, given the corresponding random field $\mathcal{X}^s$ of the feature vectors, is a CRF with up to pairwise nonzero clique potentials.
3) The following conditional independence assumption holds:

$$f(\mathcal{X}^p, \mathcal{X}^r|\mathcal{Y}^p, \mathcal{Y}^r) = f(\mathcal{X}^p|\mathcal{Y}^p) f(\mathcal{X}^r|\mathcal{Y}^r) \quad (3)$$

where $f(\mathcal{X}^p, \mathcal{X}^r|\mathcal{Y}^p, \mathcal{Y}^r)$ is the joint probability density function (PDF) of all feature vectors conditioned to all labels in both granularity layers, and $f(\mathcal{X}^p|\mathcal{Y}^p)$ and $f(\mathcal{X}^r|\mathcal{Y}^r)$ are the class-conditional PDFs associated with the two layers separately.

In particular, a second order neighborhood is used on the pixel lattice, and two regions are considered to be neighbors (see Assumption 1) if they share some common boundary (see Fig. 2). Equation (3) indicates that the statistics of the features in the two layers are modeled as independent when conditioned to the labels of each layer (see Assumption 3). From a modeling viewpoint, this condition allows the class-conditional statistics within each layer to be characterized separately. This conditional independence assumption ensures mathematical tractability and is often accepted in MRF-based (and more generally Bayesian) approaches to spatial-contextual, multisource, or multitemporal classification [35], [40], [43], [61], [62].

C. Proposed Two-Layer Model

To apply the MAP decision rule over pixels and regions simultaneously, we consider the joint posterior distribution of the combined random field $\mathcal{Y} = (\mathcal{Y}^p, \mathcal{Y}^r)$ of pixel and region labels, given the combined random field $\mathcal{X} = (\mathcal{X}^p, \mathcal{X}^r)$ of all feature vectors on both layers. Based on Assumptions 1–3, we prove that this joint posterior $P(\mathcal{Y}|\mathcal{X})$ can be expressed as a function of: 1) two separate contributions associated with the marginal posteriors $P(\mathcal{Y}^p|\mathcal{X}^p)$ and $P(\mathcal{Y}^r|\mathcal{X}^r)$ of the two granularity layers, which are modeled as CRFs (see Assumption 2) and 2) a cross-layer term that is related to the dependence between $\mathcal{Y}^p$ and $\mathcal{Y}^r$ and provides a prior on the desired relations between region labels and pixel labels.

The Bayes theorem implies that

$$\begin{aligned} P(\mathcal{Y}|\mathcal{X}) &= P(\mathcal{Y}^p, \mathcal{Y}^r|\mathcal{X}^p, \mathcal{X}^r) \\ &\propto f(\mathcal{X}^p, \mathcal{X}^r|\mathcal{Y}^p, \mathcal{Y}^r) \cdot P(\mathcal{Y}^p, \mathcal{Y}^r) \end{aligned} \tag{4}$$

where the proportionality factor depends on the feature vector fields $\mathcal{X}^p$ and $\mathcal{X}^r$, but not on the label fields $\mathcal{Y}^p$ and $\mathcal{Y}^r$, hence, it does not influence MAP decisions. Owing to Assumption 3 and using again the Bayes theorem, we obtain

$$\begin{aligned} P(\mathcal{Y}|\mathcal{X}) &\propto P(\mathcal{Y}^p, \mathcal{Y}^r) \prod_{s\in\{p,r\}} f(\mathcal{X}^s|\mathcal{Y}^s) \\ &\propto P(\mathcal{Y}^p, \mathcal{Y}^r) \prod_{s\in\{p,r\}} \frac{P(\mathcal{Y}^s|\mathcal{X}^s)}{P(\mathcal{Y}^s)} \\ &= \frac{P(\mathcal{Y}^p, \mathcal{Y}^r)}{P(\mathcal{Y}^p)P(\mathcal{Y}^r)} \cdot \prod_{s\in\{p,r\}} P(\mathcal{Y}^s|\mathcal{X}^s) \end{aligned} \tag{5}$$

where the proportionality factor depends again only on $\mathcal{X}^p$ and $\mathcal{X}^r$. Equation (5) provides the desired factorization of the joint posterior.

Accordingly, the MAP rule is expressed as the minimization of the following energy function with respect to $\mathcal{Y}$:

$$\begin{aligned} U(\mathcal{Y}|\mathcal{X}) = &\sum_{s\in\{p,r\}} \sum_{i\in\mathcal{I}_s} \underbrace{D^s\left(y_i^s|\mathcal{X}^s\right)}_{\text{[A]}} \\ &+ \lambda \sum_{s\in\{p,r\}} \sum_{i\in\mathcal{I}_s} \sum_{j\in\partial^s i} \underbrace{V^s\left(y_i^s, y_j^s|\mathcal{X}^s\right)}_{\text{[B]}} \\ &+ \mu \underbrace{W(\mathcal{Y}^p, \mathcal{Y}^r)}_{\text{[C]}} \end{aligned} \tag{6}$$

where λ and μ are positive weight parameters and the three terms highlighted have the following meaning:

[A] This is the unary potential contributing to the CRF model for each single-layer posterior $P(\mathcal{Y}^s|\mathcal{X}^s)$ in (5). As a customary choice in many MRF- and CRF-based methods [16], [40], [45], we compute it as ($i \in \mathcal{I}^s$)

$$D^s\left(y_i^s|\mathcal{X}^s\right) = -\ln \hat{P}^s\left(y_i^s|\boldsymbol{x}_i^s\right) \tag{7}$$

where $\hat{P}^s(y_i^s|\boldsymbol{x}_i^s)$ is an estimate of the elementwise posterior probability of a pixel or region. It can be computed by training, at layer s, a classifier that provides a probabilistic output (e.g., a parametric or nonparametric Bayesian classifier [63], a random forest (RF) [64], a CNN [21], or the postprocessing of the output of a support vector machine [65] by algorithms such as [66]). Note that this model for the unary potential in 2L$\frac{1}{2}$CRF is intrinsically homogeneous, as reflected by the absence of the subscript "i" in (6) and (7), compared to (2) [i.e., $D^s(\cdot)$ instead of $D_i^s(\cdot)$] [40]. This homogeneity in the unary potential is not considered a critical restriction because the region-based analysis incorporates local spatial adaptivity *per se*.

[B] This energy contribution is a pairwise term that favors spatial smoothness at each granularity level (see the red line in Fig. 2). It completes the CRF model for $P(\mathcal{Y}^s|\mathcal{X}^s)$ along with the aforementioned unary term A). $V^s(y_i^s, y_j^s|\mathcal{X}^s)$ is defined as a contrast-sensitive potential that extends the classical Potts MRF model in order to penalize that different classes are predicted for neighboring pixels or regions with similar feature vectors [58]. In 2L$\frac{1}{2}$CRF, contrast sensitivity is modeled using a Gaussian kernel $K(\cdot)$ ($i \in \mathcal{I}^s$, $j \in \partial^s i$)

$$V^s\left(y_i^s, y_j^s|\mathcal{X}^s\right) = \left[1-\delta\left(y_i^s, y_j^s\right)\right] K\left(\boldsymbol{x}_i^s, \boldsymbol{x}_j^s\right) \tag{8}$$

where $\delta(\cdot)$ is the Kronecker symbol [i.e., $\delta(a,b) = 1$ for $a = b$ and $\delta(a,b) = 0$ otherwise]. For instance, if two neighboring pixels have identical feature vectors [hence, $K(\boldsymbol{x}_i^p, \boldsymbol{x}_j^p) = 1$] and are predicted in different classes, then the maximum penalty is applied. On the contrary, if their feature vectors differ substantially [thus $K(\boldsymbol{x}_i^p, \boldsymbol{x}_j^p) \simeq 0$], then assigning the two pixels to different classes is not (or very slightly) penalized. The same comment holds in the case of regions. The previous remark on homogeneity holds in this case (8) as well. Note that the same parameter λ is used in (6) to weigh both $V^p(\cdot)$ and $V^r(\cdot)$. This is consistent with the idea of giving both granularity layers the same relevance in the labeling process. Nevertheless, extending (6) with different weight parameters for the pixel and region granularities is straightforward.

[C] This energy term is a cross-layer pairwise contribution that favors consistency between the labelings at the two granularity levels (see the green line in Fig. 2) and corresponds, up to additive or positive multiplicative constants, to $-\ln\{P(\mathcal{Y}^p, \mathcal{Y}^r)/[P(\mathcal{Y}^p)P(\mathcal{Y}^r)]\}$. This term is related to the joint prior $P(\mathcal{Y}^p, \mathcal{Y}^r)$ and provides a measure of the dependence between the fields $\mathcal{Y}^p$ and $\mathcal{Y}^r$. The rationale of C) is to encode the desired agreement between pixelwise and region-based results. This behavior is favored in the proposed method by using a Potts-like formulation

$$W(\mathcal{Y}^p, \mathcal{Y}^r) = \sum_{i\in\mathcal{I}^p} \left[1-\delta\left(y_i^p, y_{i\uparrow}^r\right)\right] \tag{9}$$

where $i^\uparrow$ indicates the region to which the ith pixel belongs, i.e., y_i^p and $y_{i\uparrow}^r$ are the labels of this pixel at the pixel and region levels, respectively ($i \in \mathcal{I}^p$). Indeed, (9) contributes a penalty for each pixel for which the classes predicted at the two levels differ.

More generally, formulating the decision fusion over multiple supports as the minimization of energy (6) can also

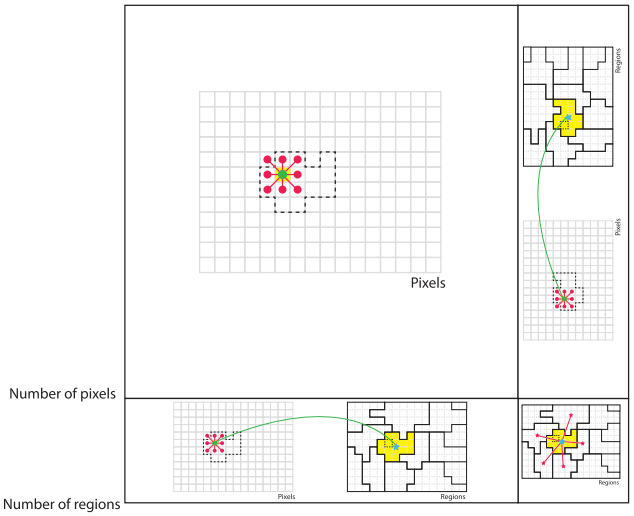

Fig. 3. Intuition behind the graph structure used in the dual layer.

be interpreted as a variational representation, especially in relation to the belief-propagation-type method that is used to address it (see Section II-D) [37], [67].

D. Flattening the Two Layers Into a Single Graph and Minimizing the Energy

The problem of the minimization of CRF energy functions such as (2) is generally a complex combinatorial problem. Nevertheless, in addition to consolidated stochastic optimization approaches such as simulated annealing [33], computationally efficient graph-theoretic algorithms have become very prominent during the past decade [40], [68]. They include graph cut algorithms, which make use of a min-flow/max-cut reformulation of the minimum energy problem [58], and belief propagation-type algorithms, which build on the idea of exchanging messages between neighboring elements to progressively reduce the energy [69]. These techniques have been found successful to minimize energies that are supported on a pixel lattice [as in (2)] or even on more general graphs [68]. On the one hand, in the case of the proposed method, two layers with distinct supports are involved in the minimization task simultaneously, a situation that does not allow applying these methods directly. On the other hand, we show here that 2L$\frac{1}{2}$CRF can also be equivalently reformulated on a unique planar graph, thus making it possible to use the aforementioned energy minimization methods appropriately.

The idea is to combine the two granularity layers into a single undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, in which both the spatial B) and the cross-layer C) pairwise interactions are reorganized. As shown in Fig. 3, the set of nodes is $\mathcal{V} = \mathcal{I}^p \cup \mathcal{I}^r$ (i.e., it includes all pixels and regions), and an edge exists between two nodes u and v, i.e., $(u, v) \in \mathcal{E}$, if and only if one of the following conditions holds ($u, v \in \mathcal{V}$):

1) u and v are neighboring pixels, i.e., $u \in \mathcal{I}^p$, $v \in \partial^p u$;
2) u and v are neighboring regions, i.e., $u \in \mathcal{I}^r$, $v \in \partial^r u$;
3) u is a pixel and v is the region that includes u, i.e., $u \in \mathcal{I}^p$ and $v = u^{\uparrow}$, or vice versa.

The random fields $\mathcal{X}^p, \mathcal{X}^r, \mathcal{Y}^p$, and $\mathcal{Y}^r$ on the pixel lattice and on the set of regions can be rearranged on the graph $\mathcal{G}$ in a straightforward way (see Fig. 3). Denoting with a bar the rearranged quantities (e.g., $\bar{\mathcal{Y}} = \{\bar{y}_v\}_{v \in \mathcal{V}}$, where $\bar{y}_v$ indicates the label of node v and is either y_v^p or y_v^r as $v \in \mathcal{I}^p$ or $v \in \mathcal{I}^r$, respectively), the energy (6) can be rewritten as

$$\bar{U}(\bar{\mathcal{Y}}|\bar{\mathcal{X}}) = \sum_{v \in \mathcal{V}} \bar{D}_v(\bar{y}_v|\bar{\mathcal{X}}) + \sum_{(u,v) \in \mathcal{E}} \bar{V}_{uv}(\bar{y}_u, \bar{y}_v|\bar{\mathcal{X}}) \quad (10)$$

where the unary potential is

$$\bar{D}_v(\bar{y}_v|\bar{\mathcal{X}}) = \begin{cases} D^p\left(y_v^p|\mathcal{X}^p\right), & \text{if } v \in \mathcal{I}^p \\ D^r\left(y_v^r|\mathcal{X}^r\right), & \text{if } v \in \mathcal{I}^r \end{cases} \quad (11)$$

and the pairwise potential is

$$\bar{V}_{uv}(\bar{y}_u, \bar{y}_v|\bar{\mathcal{X}}) = \begin{cases} \lambda[1 - \delta(\bar{y}_u, \bar{y}_v)]K(\bar{\boldsymbol{x}}_u, \bar{\boldsymbol{x}}_v), & \text{if } (u \in \mathcal{I}^p \text{ and } v \in \partial^p u) \\ & \text{or } (u \in \mathcal{I}^r \text{ and } v \in \partial^r u) \\ \mu[1 - \delta(\bar{y}_u, \bar{y}_v)], & \text{if } (u \in \mathcal{I}^p \text{ and } v = u^{\uparrow}) \\ & \text{or } (v \in \mathcal{I}^p \text{ and } u = v^{\uparrow}). \end{cases} \quad (12)$$

Note that, although the unary and pairwise potentials in (7) and (8) are homogeneous, the potentials $\bar{D}_v(\cdot)$ and $\bar{V}_{uv}(\cdot)$ on the flattened planar graph $\mathcal{G}$ are only piecewise homogeneous. To minimize the resulting energy $\bar{U}(\bar{\mathcal{Y}}|\bar{\mathcal{X}})$ with respect to $\bar{\mathcal{Y}}$ on the graph $\mathcal{G}$ (i.e., with respect to $\mathcal{Y}^p$ and $\mathcal{Y}^r$ simultaneously), the sequential tree reweighted message passing algorithm is used in 2L$\frac{1}{2}$CRF. This algorithm integrates the belief propagation approach and the construction of appropriate spanning trees on the considered graph, and makes use of a specific sequential formulation to favor a convergent behavior (details can be found in [69]).

III. Data and Setup

In this section, we introduce the data sets used for the experiments. We also detail the setup of the said experiments.

A. Data

The performance of the proposed 2L$\frac{1}{2}$CRF is tested using two very high-resolution data sets:

1) *Zurich Summer Data Set [20]:* This data set is a collection of 20 images from a single large QuickBird [near infrared response (NIR)-RGB bands] acquisition of 2002. The images picture different neighborhoods of the city of Zürich, Switzerland. Each image is pansharpened to 0.6-m resolution and labeled in eight classes. The labeling is not dense, meaning that some pixels are either unassigned or belonging to unseen classes. False color infrared images and the ground truth for five tiles can be seen in the first two columns in Fig. 6. The data set can be freely downloaded at https://sites.google.com/site/michelevolpiresearch/data/zurich-dataset.
2) *Zeebruges, or the Data Fusion Contest 2015 Data Set (grss_dfc_2015) [3]:* The Image Analysis and Data Fusion Technical Committee of the IEEE Geoscience

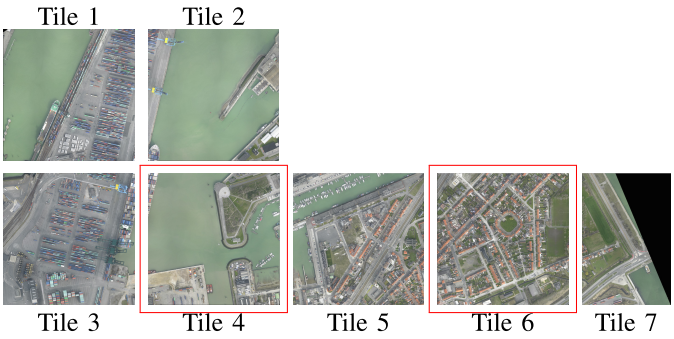

Fig. 4. Seven tiles of the Data Fusion Contest 2015 data set (RGB) (from [3]). Test tiles are highlighted in red. The data can be freely downloaded from http://www.grss-ieee.org/community/technical-committees/data-fusion/2015-ieee-grss-data-fusion-contest/.

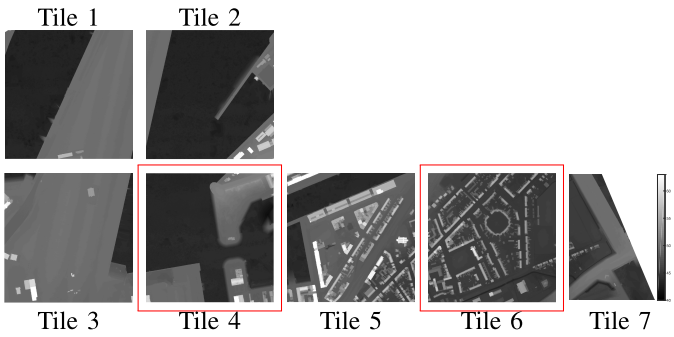

Fig. 5. Seven tiles of the DSM issued from the LiDAR point cloud of the Data Fusion Contest 2015 (from [3]). Test tiles are highlighted in red. The data can be freely downloaded from http://www.grss-ieee.org/community/technical-committees/data-fusion/2015-ieee-grss-data-fusion-contest/.

and Remote Sensing Society (GRSS) organized in 2015 a data processing competition which aimed at 5-cm resolution land cover mapping. Both an RGB aerial image (5-cm spatial resolution) and a dense lidar point cloud (65 pts/m^2) were acquired over the harbor of Zeebruges, Belgium. The data are organized on seven 10 000 $\times$ 10 000 pixels tiles. We used a downgraded version on 5000 $\times$ 5000 pixels for training the models and upsampled the final prediction using interpolation (the final numerical scores were not significantly affected, but computational efficiency was dramatically improved). All the tiles have been densely annotated in 8 land classes, including land use (buildings, roads) and small objects (vehicles, boats) classes by the authors of [70]. The seven RGB tiles, as well as the LiDAR normalized DSM are illustrated in Figs. 4 and 5. The data can be obtained from the IEEE GRSS Data and Algorithm Standard Evaluation Website (DASE) http://dase.ticinumaerospace.com/. From DASE, users can download the seven tiles and labels for five tiles. To assess models on the two remaining tiles, we uploaded the classification maps on the DASE server, which automatically computes confusion matrices and accuracy scores. For more information about the data please refer to [3] and https://www.grss-ieee.org/community/technical-committees/data-fusion/2015-ieee-grss-data-fusion-contest/.

B. Experimental Setup

The experiments presented in the following section aim at showcasing the effectiveness of the proposed 2L$\lightning$CRF in different settings using different classifiers. In all the experiments, we compare the results obtained using models aware of a single support (i.e., only pixel- or region-based mapping units) with those obtained by models exploiting structure in the spatial domain of the support (in all cases, the CRF of [51]) and with the results of the proposed model, which lets the two supports interact via the two-layers CRF structure. Below we present the setups used for the two data sets.

1) Zurich Summer: We consider two base classifiers: RFs [64] and CNNs [21] with independent patch-response.

- *RF*: In the case of RFs, we train two separate models, one at the pixel and another at the region level [60]. At the pixel level, we use spectral features (R-G-B-NIR raw values, normalized to standard scores), the normalized difference vegetation and water indices (NDVI and NDWI), and contextual features (average over 3 $\times$ 3 and 5 $\times$ 5 local windows, computed on the R and NIR bands). Therefore, we have a 10-dimensional input space. Regarding the region level, we first extract regions using the Felzenszwalb and Huttenlocher graph-based algorithm [6]. We used as region descriptors the min, max, average and standard deviation values of the pixel intensities included in each region for the R, G, B, NIR, NDVI, and NDWI channels. The input space of the region classifier is therefore 24-dimensional. To train the models, we derived a region ground truth from the available pixelwise training map by assigning the majority class found within each region.
- *CNN:* In the case of CNNs, we opted for a classical image classification architecture, i.e., an architecture that predicts a single label per patch (in our case, of size 65 $\times$ 65 pixels) and predicts the final map using a sliding window approach, as in [3]. The reasoning behind it is that, since the ground truth is not dense, spatial structures become less explicit to be learned, and complex optimization schemes (as the one in [16]), which are not trivial to stack on top of a dense prediction CNN, should be involved to learn them properly. Specifically, we use four convolutional blocks, the first two with 5 $\times$ 5 filters (expressed as two consecutive 3 $\times$ 3 filters, as proposed in [71]) and the other two with plain 3 $\times$ 3 filters. Each filter is followed by a ReLU nonlinearity and by a spatial pooling halving the spatial extent of the activations. The fourth block is followed by a fully connected layer outputting a 256-dimensional vector per patch. This vector is then used to predict the class conditional probabilities with a softmax classifier. A dropout [72] rate of 50% is used on the fully connected layer to reduce the risk of overfitting the training data. Given the computational efforts that would have been involved, we did not train a specific region-based model, but rather averaged and renormalized the class probabilities per region and used them as scores for the region. The results were not affected by using the averaged CNN pixelwise class probabilities instead

TABLE I

ZURICH SUMMER DATA SET: NUMERICAL RESULTS ON THE FIVE TEST TILES. EACH PAIR OF COLUMNS ILLUSTRATES THE CHANGE OF PERFORMANCE BETWEEN THE UNARIES (RF) AND THE SPATIAL MODEL USING THEM AS A BASE PROBABILISTIC INPUT IN THE UNARY POTENTIAL. THE CRF IS BASED ON [58]

	Ov. Accuracy (OA)						Kappa (κ)						Av. Accuracy (AA)					
Tile	Pixels			Regions			Pixels			Regions			Pixels			Regions		
#	RF	CRF	2L$\nmid$CRF	RF	CRF	2L$\nmid$CRF	RF	CRF	2L$\nmid$CRF	RF	CRF	2L$\nmid$CRF	RF	CRF	2L$\nmid$CRF	RF	CRF	2L$\nmid$CRF
16	80.3	81.8	85.1	82.6	84.8	87.5	0.72	0.75	0.79	0.76	0.79	0.82	60.0	60.0	58.4	63.7	63.3	58.9
17	77.7	79.2	83.4	82.9	83.3	86.3	0.71	0.73	0.78	0.77	0.78	0.82	59.8	60.6	62.8	69.3	68.3	64.5
18	69.6	71.8	78.1	71.2	73.4	79.9	0.57	0.6	0.68	0.58	0.61	0.70	62.2	63.4	66.9	61.2	62.5	65.9
19	67.4	68.6	69.5	65.5	66.8	69.4	0.57	0.59	0.60	0.55	0.57	0.60	60.0	60.5	60.5	68.7	70.3	62.2
20	76.2	77.7	82.9	78.0	79.2	84.4	0.69	0.71	0.78	0.72	0.73	0.80	65.6	67.0	71.1	67.2	68.3	72.3
Avg.†	75.4	76.9	80.9 (+4.0)	77.3	78.8	82.7 (+3.9)	0.69	0.71	0.76 (+0.05)	0.71	0.73	0.78 (+0.05)	66.0	67.0	69.5 (+2.5)	67.1	67.8	70.5 (+2.7)
Ov.*	74.2	75.8	79.8 (+4.0)	76.0	77.5	81.5 (+4.0)	0.65	0.67	0.73 (+0.06)	0.68	0.70	0.75 (+0.05)	61.5	62.3	63.9 (+1.6)	66.0	66.6	64.8 (–1.8)

† metric over all test samples.
* average over the metric of the 5 test tiles.

of training a specific model for the region lattice (be it a CNN or the RF model presented above).

We use images ID 1–15 (out of the 20 images composing the data set) to train the models and images 16–20 to test its generalization. Accuracy figures are provided per image and averaged over the test set, for both the pixel and region levels. For the case of RF, we trained the pixel-level classifier with 0.01% of the available labeled pixels, with a selection stratified by class (corresponding to 122 658 pixels pooled from all the training images). At the region level, the second RF classifier used all the available 34 020 labeled regions. The increase of training set size at pixel level with respect to [60] is due to the intention of providing a better numerical comparison against the CNN model, for which 10 000 patches are used for training, randomly selected, and resampled at every 10 training epochs. Since we run 150 epochs of the model, the CNN actually saw roughly 150 000 patches during the whole training, whose diversity was also increased by data augmentation, i.e., random flips and rotations at each 10 epochs resampling. The CNN was trained by stochastic gradient descent with momentum (0.9), with a weight decay of 0.001 and a constant learning rate of 10^{-3}.

2) Zeebruges: In the case of Zeebruges, we opted for the CNN case only, as Random Forest did not perform nearly as accurate. In [3], two main observations were made: CNN outperformed other nonlinear methods (such as support vector machines) and a fully trained model provided the most accurate solution on this data set. We also opted for a dense prediction architecture: for this data set, we have a dense ground truth [70], and therefore, spatial relationships can be explicitly learned. Following these two reasons, we used the CNN in [31], whose architecture follows the concept of hypercolumns [25], but with added equivariance to rotation. The CNN is based on a set of convolutional blocks with spatial pooling and all the activations at each level are upsampled to the original image resolution, stacked and then fed into a second block employing this multiscale descriptions to map to class likelihoods. In our case, the convolutional blocks are RotEqNet layers enforcing rotation equivariance of the final prediction (see [73] for details). There are six such blocks, learning $[14, 14, 21, 28, 28, 28]$ filters of size 7×7 each, respectively. After concatenation of the activations, the CNN has three 1×1 convolution layers with 350, 350, and 6 (the number of classes) neurons, and a softmax normalization. The model was trained for 34 iterations with decreasing weight decay (from 10^{-1} to $4 \cdot 10^{-3}$) and learning rates (from 10^{-2} to $4 \cdot 10^{-4}$).

Finally, for both data set hyperparameters μ and λ were set by cross-validation and the σ parameter of the kernel in the contrast sensitive term (8) was set as half the Euclidean features distance between samples.

IV. EXPERIMENTAL RESULTS

In this section, we present the results obtained on the Zurich Summer and Zeebruges data sets.

A. Zurich Summer Data Set

For the Zurich Summer data set, we first discuss the results obtained by the approach considering RFs with contextual features and then compare it against the results obtained by the CNN base classifier.

1) Random Forests: Table I shows the numeric scores obtained using RF. The results are in line with those reported in [60]. The slight increases in accuracies are due to the larger training set used to train the pixel-based classifier and are consistent across the experiments. When using RF, the 2L$\nmid$CRF approach allows to increase consistently the performances of the base independent models with average increases of accuracy of around 4%. An exception is found when considering the mean of the average accuracy (AA) over the five tiles ("Ov." in the table): in this case, the 1.8% decrease in AA observed is due to poor performance in tile #16 (−5%). The poor performance is explained by the disappearance of the class "bare soil" from the prediction, which decreases strongly the AA. When pooling confusion matrices over the five tiles ("Avg." in the table) this effect disappears, as bare soil is correctly predicted in the other tiles.

2) Random Forests Versus CNNs: Table II presents the numerical results of the same data set, this time obtained by a CNN predicting single labels per patch. The first striking observation is that all the figures of merit have a sharp improvement of about 10–15% in overall accuracy (OA), 10–18 κ points or 8%–10% in AA. Certainly, the RF results could have been improved by a more in-depth research on feature engineering of the input space. However, considering that the CNN learns end-to-end, we only devoted efforts in finding the appropriate architecture training well. This is in line with observations in several other recent papers.

TABLE II

ZURICH SUMMER DATA SET: NUMERICAL RESULTS ON THE FIVE TEST TILES. EACH PAIR OF COLUMNS ILLUSTRATES THE CHANGE OF PERFORMANCE BETWEEN THE UNARIES (CNN) AND THE SPATIAL MODEL USING THEM AS A BASE PROBABILISTIC INPUT IN THE UNARY POTENTIAL. THE CRF IS BASED ON [58]

	Ov. Accuracy (OA)						Kappa (κ)						Av. Accuracy (AA)					
Tile	Pixels			Regions			Pixels			Regions			Pixels			Regions		
#	CNN	CRF	2L$\frac{1}{2}$CRF	CNN	CRF	2L$\frac{1}{2}$CRF	CNN	CRF	2L$\frac{1}{2}$CRF	CNN	CRF	2L$\frac{1}{2}$CRF	CNN	CRF	2L$\frac{1}{2}$CRF	CNN	CRF	2L$\frac{1}{2}$CRF
16	90.7	90.7	91.0	87.0	87.0	92.3	0.87	0.87	0.87	0.82	0.82	0.89	73.2	73.2	73.2	69.1	69.0	78.1
17	90.2	90.2	90.4	88.5	88.5	93.0	0.87	0.87	0.87	0.85	0.85	0.91	69.5	69.4	69.5	70.0	70.0	71.9
18	91.1	91.1	91.2	90.7	90.7	92.7	0.87	0.87	0.87	0.86	0.86	0.89	86.9	86.9	87.1	85.3	85.3	88.0
19	90.4	90.4	90.4	87.7	87.7	91.4	0.87	0.87	0.87	0.84	0.84	0.89	91.1	91.1	91.2	87.2	87.2	92.9
20	89.6	89.7	89.9	88.0	88.0	91.1	0.86	0.86	0.87	0.84	0.84	0.89	76.0	76.0	76.2	74.1	74.1	77.3
Avg.†	90.3	90.3	90.5 (+0.2)	88.1	88.1	92.0 (+3.9)	0.88	0.88	0.88	0.85	0.85	0.90 (+0.05)	77.7	77.7	77.8 (+0.1)	76.5	76.5	79.3 (+2.8)
Ov.*	90.4	90.4	90.6 (+0.2)	88.3	88.4	92.1 (+3.7)	0.87	0.87	0.87	0.84	0.84	0.89 (+0.05)	79.3	79.4	79.4	77.1	77.1	81.6 (+4.5)

† metric over all test samples.
* average over the metric of the 5 test tiles.

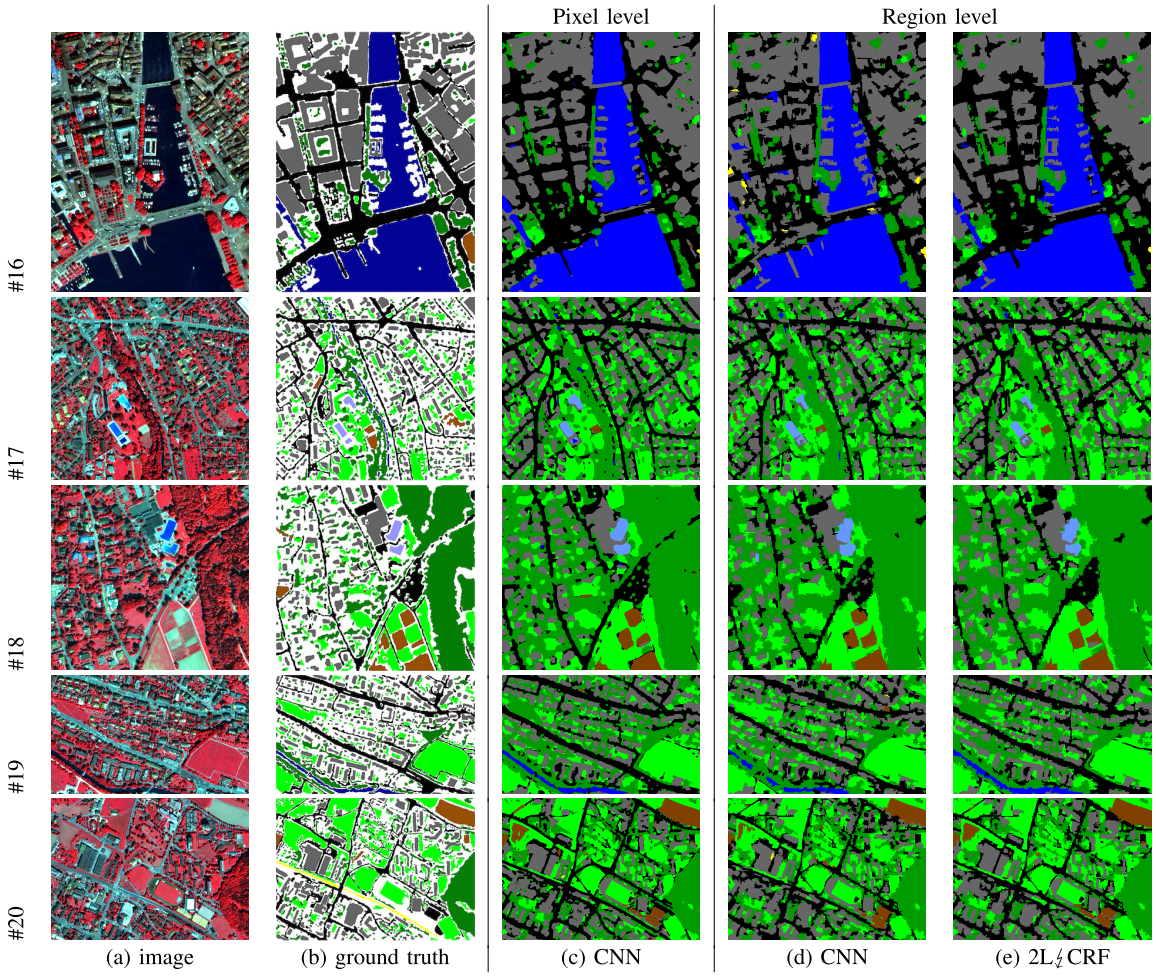

Fig. 6. Zurich Summer data set: results on the five test images. (a) Original image. (b) Ground truth. (c) CNN, pixel based. (d) CNN, region level. (e) 2L$\frac{1}{2}$CRF, region level. For the accuracies of the single maps, please refer to Table II (color legend: residential, street, trees, meadows, railway, water, swimming pools, bare soil).

3) CNNs: Considering the role of the proposed 2L$\frac{1}{2}$CRF when using unary potentials from a CNN, we observe a general improvement of the accuracy figures, although it is less striking than in the previous case of RFs, in particular on pixels results. The reason lies in the high accuracy of the initial result, for which the unary scores are sharp and spatially consistent (see the third column in Fig. 6). This can be seen in the difference between the CNN results and those of the CRF based on [58], which considers only spatial interactions for each lattice separately. The 2L$\frac{1}{2}$CRF model only needs to correct for inconsistent labelings among the two layers, which results in small increases in the accuracies. On the contrary, the high accuracy and consistency of the unary scores allows to correct for several misclassifications at the region level, for which the pooling of the unary scores resulted in erroneous region scores, in particular in the case of regions with high entropy of the posterior probability at the pixel level. For instance, the road network is poorly reconstructed in tile

TABLE III

NUMERICAL RESULTS ON THE ZEEBRUGES DATA SET (GRSS_DFC_2015). EACH PAIR OF COLUMNS ILLUSTRATES THE CHANGE OF PERFORMANCE BETWEEN THE UNARIES (CNN) AND THE SPATIAL MODEL USING THEM AS A BASE IN THE UNARY POTENTIAL. THE CRF IS BASED ON [58]

	Ov. Accuracy (OA)						Kappa (κ)						Av. Accuracy (AA)					
Tile	Pixels			Regions			Pixels			Regions			Pixels			Regions		
#	CNN	CRF	2L$\nmid$CRF	CNN	CRF	2L$\nmid$CRF	CNN	CRF	2L$\nmid$CRF	CNN	CRF	2L$\nmid$CRF	CNN	CRF	2L$\nmid$CRF	CNN	CRF	2L$\nmid$CRF
Tile #4	87.3	87.4	87.7	87.5	87.7	87.7	0.80	0.80	0.80	0.81	0.81	0.81	79.2	79.3	79.8	80.6	80.7	80.7
Tile #6	77.8	78.0	78.9	79.3	79.6	80.0	0.68	0.68	0.69	0.70	0.70	0.71	65.2	65.2	65.4	65.9	65.9	66.2
Avg.†	82.6	82.7	83.3 (+0.6)	83.3	83.5	83.7 (+0.2)	0.77	0.77	0.78 (+0.01)	0.78	0.79	0.79	75.2	75.3	75.6 (+0.3)	76.4	76.4	76.5 (+0.1)
Ov.*	82.6	82.7	83.3 (+0.6)	83.4	83.7	83.8 (+0.1)	0.74	0.74	0.75 (+0.01)	0.75	0.76	0.76	72.2	72.2	72.6 (+0.4)	73.3	73.3	73.4 (+0.1)

† metric over all test samples.
* average over the metric of the 5 test tiles.

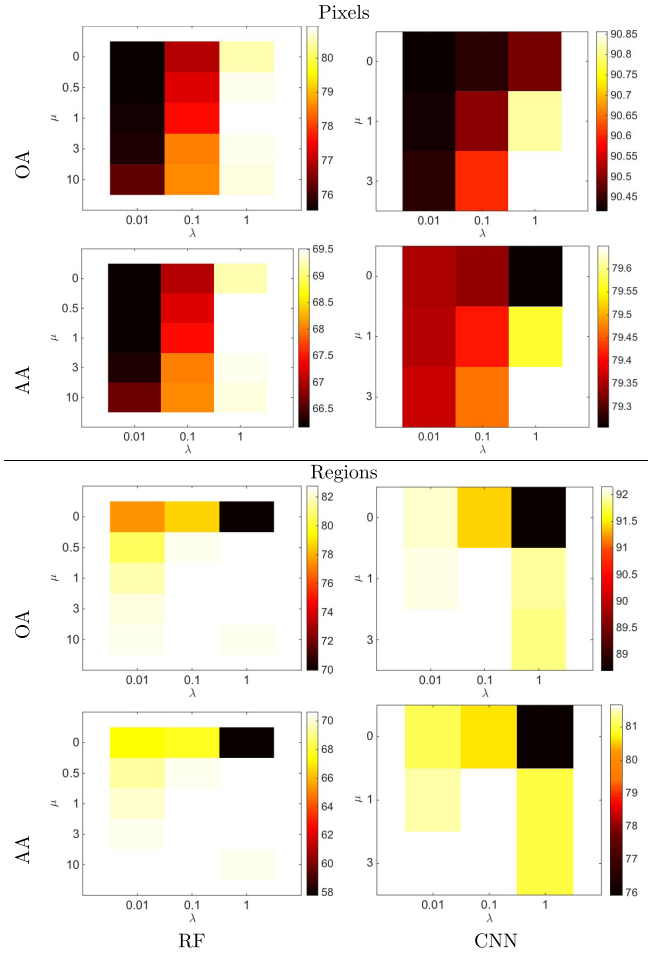

Fig. 7. Sensibility analysis for the μ and λ parameters in (6) and the Zurich Summer data set.

#16 in Fig. 6, as regions composing it are often incorrectly classified as buildings. On the contrary, they are assigned to the roads class by 2L$\nmid$CRF, since the evidence from the pixel representation is strong enough to influence the response at the region level, thus leading to a map that is correct (by the pixel response) and sharp (by the region topology). Another example is in tile #19, where the river is correctly recovered by 2L$\nmid$CRF, while the original region results were providing a mix between water, trees, and roads.

4) Parameters Analysis: Fig. 7 reports the sensitivity analysis for the μ and λ parameters, involved in (6). As a reminder, μ control the strength of the relations between the spatial supports and λ the strength of the local spatial smoothing within each spatial support. In all experiments, the model not making any use of intersupport connections ($\mu = 0$, which would roughly correspond to optimizing two separate CRFs) does not provide the best results, thus showing the interest of a joint model. This is particularly visible at the region level, where the evidence from the pixel level is able to successfully revert many errors, in particular for the RF classifier, which often provides ambiguous class-likelihoods. Also, using contrast sensitivity seems particularly important at the pixel level, while less at the region level: this is expected, since given the high resolution of the data, local inconsistencies at the pixel level are more frequent that noisy predictions at the region level.

B. Zeebruges Data Set

In the case of the Zeebruges data set, we retrieve unary scores directly from the model of [31]. This model achieves results that compare favorably to the state-of-the-art methods on this data set. It still suffers from some prediction noise leading to small artifacts, as it can be seen in the maps in Fig. 8 for the entire tiles and in Fig. 9 for details. This is certainly due to the increased spatial resolution (and therefore complexity) of the problem, since we are working, compared to the Zurich Summer data set, on a 10 times higher spatial resolution and without an infrared channel.

Numerical results are reported in Table III: as a first observation, the standard CRF approach seems to improve the results only marginally, which can be explained by two factors: first, and as for the previous data set, the CNN unaries are very sharp and confident (even when misclassifying), making the change of a label very unlikely. Secondly, the use of CRFs with up to pairwise nonzero clique potentials rather than more sophisticated higher order CRFs (which was necessary to keep the problem computationally feasible) made label swaps improbable, since the pixel classifier was not much affected by salt and pepper noise, but rather by larger spatial artifacts that could not be corrected by looking only at the direct pixel neighborhood. This is the reason why, for this data set, 2L$\nmid$CRF greatly increases performance at the pixel level: by letting the pixel lattice be influenced by the region structure, where the scores are pooled per region, the pixel CNN becomes aware of larger neighborhood structures and can therefore correct for larger confident misclassifications at pixel level (see the zoomed-in-views in Fig. 9 for some examples). Such beneficial effect could have been maybe reinforced if the region level CNN were a classifier trained specifically to predict land cover at the region level (and

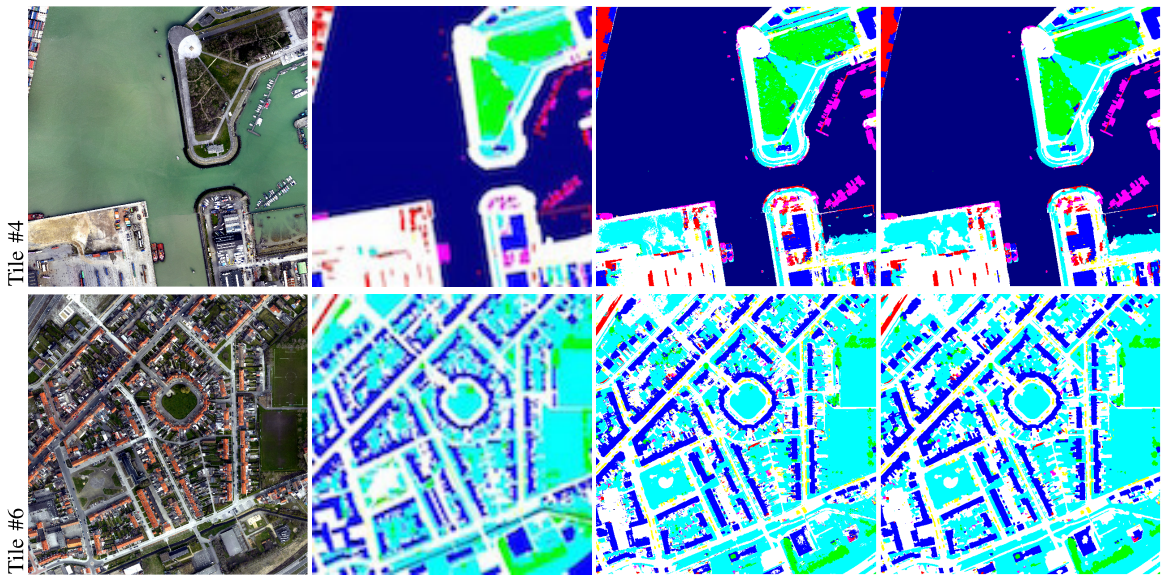

Fig. 8. Zeebruges data set (`grss_dfc_2015`): results on the two test tiles. (a) Original image. (b) Ground truth. (c) CNN, pixel based. (d) 2L$\frac{1}{2}$CRF, pixel level. For the accuracies of the single maps, please refer to Table III [color legend: impervious (white colored), water, clutter, low vegetation, buildings, trees, boats, cars]. The ground-truth images are blurred as in [3], since they are undisclosed.

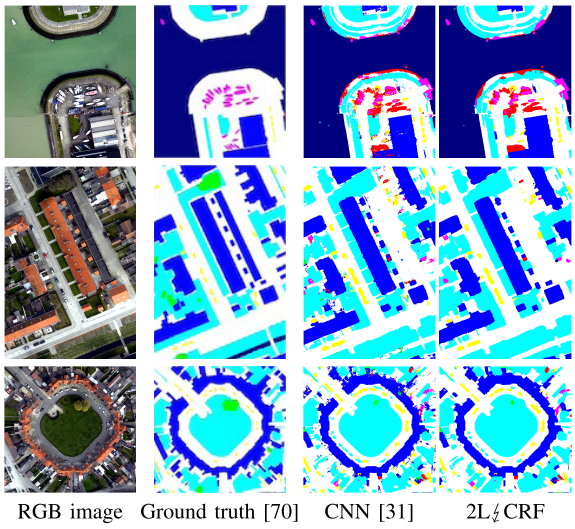

Fig. 9. Zeebruges data set (`grss_dfc_2015`): zoomed results on details of the two test tiles of Fig. 8. (a) Original image. (b) Ground truth. (c) CNN, pixel based. (d) 2L$\frac{1}{2}$CRF, pixel level. For the accuracies of the single maps, please refer to Table III [color legend: impervious (white colored), water, clutter, low vegetation, buildings, trees, boats, cars]. The ground-truth images are blurred as in [3], since they are undisclosed.

not a set of region-pooled scores over the pixel unaries), but this would have implied training a separate CNN model for the region scale. The consequent significant increase in memory and computational resources would diminish—in our opinion—the interest of the combined approach.

V. Conclusion

In this paper, we proposed a probabilistic discriminative graphical model relying on a CRFs formulation for the fusion of land-cover and land-use classification results from VHR remote sensing images. The system is able to find agreement between probabilistic decisions with multiple spatial supports. We explored the fusion of pixel- and region-based spatial supports within an energy minimization framework, where each individual classification result is tributary of: 1) the posterior distribution of the land cover classes at the single instance level (the pixel or the region); 2) the spatial smoothness of the predictions (i.e., the consistency of the prediction among the spatial neighbors); and 3) the smoothness across supports (i.e., the consistency of the predictions between a region and the pixels composing the region itself). These three goals are addressed jointly within a CRF model with connections across layers corresponding to different spatial support representations. It is also proven that the proposed two-layer model can be flattened into a single CRF, for which energy minimization can be addressed efficiently with standard energy minimization solvers.

Applications to two VHR benchmark data sets showed the potential of the approach that, using common models (we considered RFs and CNNs), can improve the final maps consistently and joining the spatial detail of the pixel support with the geometrical object accuracy of the region support. In the future, we would like to test the 2L$\frac{1}{2}$CRF model for the fusion of multitemporal data, multiple segmentations (>2), or for applications involving different classes to be predicted for different spatial supports. Furthermore, although the impact of the weight parameters of the proposed CRF model on the resulting performance was limited, it would also be interesting to automatically optimize their values, for example using log-likelihood-like (e.g., through pseudo-likelihood approximations or the expectation-maximization algorithm) or mean-square-error concepts [35], [74], [75].

Acknowledgment

The authors would like to thank the Belgian Royal Military Academy, for acquiring and providing the Zeebruges data used in this paper, ONERA—The French Aerospace Lab, for providing the corresponding ground-truth data [70], and

the IEEE GRSS Image Analysis and Data Fusion Technical Committee.

REFERENCES

[1] J. A. J. Berni, P. J. Zarco-Tejada, L. Suarez, and E. Fereres, "Thermal and narrowband multispectral remote sensing for vegetation monitoring from an unmanned aerial vehicle," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 3, pp. 722–738, Mar. 2009.

[2] R. Zahawi, J. Dandois, K. Holl, D. Nadwodny, J. Reid, and E. Ellis, "Using lightweight unmanned aerial vehicles to monitor tropical forest recovery," *Biol. Conservation*, vol. 186, pp. 287–295, Jun. 2015.

[3] 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.

[4] T. Tomic *et al.*, "Toward a fully autonomous UAV: Research platform for indoor and outdoor urban search and rescue," *IEEE Robot. Autom. Mag.*, vol. 19, no. 3, pp. 46–56, Sep. 2012.

[5] T. Blaschke, "Object based image analysis for remote sensing," *ISPRS J. Photogramm. Remote Sens.*, vol. 65, no. 1, pp. 2–16, Jan. 2010.

[6] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *Int. J. Comput. Vis.*, vol. 59, no. 2, pp. 167–181, 2004.

[7] M.-Y. Liu, O. Tuzel, S. Ramalingam, and R. Chellappa, "Entropy rate superpixel segmentation," in *Proc. CVPR*, Jun. 2011, pp. 2097–2104.

[8] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.

[9] M. Dalla Mura, J. A. Benediktsson, B. Waske, and L. Bruzzone, "Morphological attribute profiles for the analysis of very high resolution images," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 10, pp. 3747–3762, Oct. 2010.

[10] F. Pacifici, M. Chini, and W. J. Emery, "A neural network approach using multi-scale textural metrics from very high-resolution panchromatic imagery for urban land-use classification," *Remote Sens. Environ.*, vol. 113, no. 6, pp. 1276–1292, 2009.

[11] L. Huang, C. Chen, W. Li, and Q. Du, "Remote sensing image scene classification using multi-scale completed local binary patterns and Fisher vectors," *Remote Sens.*, vol. 8, no. 6, p. 483, 2016.

[12] U. Budak, U. Halıcı, A. cSengür, M. Karabatak, and Y. Xiao, "Efficient airport detection using line segment detector and Fisher vector representation," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 8, pp. 1079–1083, Aug. 2016.

[13] Q. Zhu, Y. Zhong, B. Zhao, G.-S. Xia, and L. Zhang, "Bag-of-visual-words scene classifier with local and global features for high spatial resolution remote sensing imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 6, pp. 747–751, Jun. 2016.

[14] J. Tu, H. Sui, W. Feng, K. Sun, and L. Hua, "Detection of damaged rooftop areas from high-resolution aerial images based on visual bag-of-words model," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 12, pp. 1817–1821, Dec. 2016.

[15] N. Rey, M. Volpi, S. Joost, and D. Tuia, "Detecting animals in African Savanna with UAVs and the crowds," *Remote Sens. Environ.*, vol. 200, pp. 341–351, Oct. 2017.

[16] M. Volpi and V. Ferrari, "Structured prediction for urban scene semantic segmentation with geographic context," in *Proc. JURSE*, Lausanne, Switzerland., 2015, pp. 1–4.

[17] D. Konstantinidis, T. Stathaki, V. Argyriou, and N. Grammalidis, "Building detection using enhanced HOG–LBP features and region refinement processes," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 3, pp. 888–905, Mar. 2017.

[18] D. Tuia, M. Volpi, M. Dalla Mura, A. Rakotomamonjy, and R. Flamary, "Automatic feature learning for spatio-spectral image classification with sparse SVM," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 10, pp. 6062–6074, Oct. 2014.

[19] D. Tuia, N. Courty, and R. Flamary, "Multiclass feature learning for hyperspectral image classification: Sparse and hierarchical solutions," *ISPRS J. Photogramm. Remote Sens.*, vol. 105, pp. 272–285, Jul. 2015.

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

[21] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[22] M. Volpi and D. Tuia, "Dense semantic labeling of subdecimeter resolution images with convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 881–893, Feb. 2017.

[23] X. Zhu *et al.*, "Deep learning in remote sensing: A review," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, 2017.

[24] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. CVPR Workshops*, 2015, pp. 3431–3440.

[25] B. Hariharan, P. A. Arbeláez, R. Girshick, and J. Malik, "Hypercolumns for object segmentation and fine-grained localization," in *Proc. CVPR*, 2015, pp. 447–456.

[26] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[27] N. Audebert, B. Le Saux, and S. Lefèvre, "Semantic segmentation of earth observation data using multimodal and multi-scale deep networks," in *Proc. ACCV*, 2016, pp. 180–196.

[28] D. Marmanis, K. Schindler, J. D. Wegner, S. Galliani, M. Datcu, and U. Stilla. (2016). "Classification with an edge: Improving semantic image segmentation with boundary detection." [Online]. Available: https://arxiv.org/abs/1612.01337

[29] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Convolutional neural networks for large-scale remote-sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 645–657, Feb. 2017.

[30] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez. (2016). "High-resolution semantic labeling with convolutional neural networks." [Online]. Available: https://arxiv.org/abs/1611.01962

[31] D. Marcos, M. Volpi, B. Kellenberger, and D. Tuia, "Land cover mapping at very high resolution with rotation equivariant CNNs: Towards small yet accurate models," *ISPRS J. Int. Soc. Photogramm. Remote Sens.*, to be published.

[32] J. Sherrah. (2016). "Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery." [Online]. Available: https://arxiv.org/abs/1606.02585

[33] 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.

[34] J. Besag, "Statistical analysis of dirty pictures," *J. Appl. Statist.*, vol. 20, nos. 5–6, pp. 63–87, 1993.

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

[36] J. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. ICML*, 2001, pp. 282–289.

[37] C. Sutton and A. McCallum, "An introduction to conditional random fields," *Found. Trends Mach. Learn.*, vol. 4, no. 4, pp. 267–373, 2011.

[38] G. Moser, S. B. Serpico, and J. A. Benediktsson, "Land-cover mapping by Markov modeling of spatial–contextual information in very-high-resolution remote sensing images," *Proc. IEEE*, vol. 101, no. 3, pp. 631–651, Mar. 2013.

[39] K. Schindler, "An overview and comparison of smooth labeling methods for land-cover classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 11, pp. 4534–4545, Nov. 2012.

[40] S. Z. Li, *Markov Random Field Modeling in Image Analysis*. London, U.K.: Springer, 2009.

[41] P. C. Smits and S. G. Dellepiane, "Synthetic aperture radar image segmentation by a detail preserving Markov random field approach," *IEEE Trans. Geosci. Remote Sens.*, vol. 35, no. 4, pp. 844–857, Jul. 1997.

[42] C. Luo and G. Sohn, "Scene-layout compatible conditional random field for classifying terrestrial laser point clouds," *ISPRS Ann. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 2, no. 3, pp. 79–86, 2014.

[43] A. H. S. Solberg, T. Taxt, and A. K. Jain, "A Markov random field model for classification of multisource satellite imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 34, no. 1, pp. 100–113, Jan. 1996.

[44] D. Marcos, R. Hamid, and D. Tuia, "Geospatial correspondences for multimodal registration," in *Proc. CVPR*, Las Vegas, NV, USA, 2016, pp. 5091–5100.

[45] T. Hoberg, F. Rottensteiner, R. Q. Feitosa, and C. Heipke, "Conditional random fields for multitemporal and multiscale classification of optical satellite imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 2, pp. 659–673, Feb. 2015.

[46] J. D. Wegner, S. Branson, D. Hall, K. Schindler, and P. Perona, "Cataloging public objects using aerial and street-level images—Urban trees," in *Proc. CVPR*, Las Vegas, NV, USA, 2016, pp. 6014–6023.
[47] A. S. Willsky, "Multiresolution Markov models for signal and image processing," *Proc. IEEE*, vol. 90, no. 8, pp. 1396–1458, Aug. 2002.
[48] G. Scarpa, R. Gaetano, M. Haindl, and J. Zerubia, "Hierarchical multiple Markov chain model for unsupervised texture segmentation," *IEEE Trans. Image Process.*, vol. 18, no. 8, pp. 1830–1843, Aug. 2009.
[49] G. Moser, A. De Giorgi, and S. B. Serpico, "Multiresolution supervised classification of panchromatic and multispectral images by Markov random fields and graph cuts," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 9, pp. 5054–5070, Sep. 2016.
[50] I. Hedhli, G. Moser, S. B. Serpico, and J. Zerubia, "A new cascade model for the hierarchical joint classification of multitemporal and multiresolution remote sensing data," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 11, pp. 6333–6348, Nov. 2016.
[51] P. Kohli, L. Ladicky, and P. H. Torr, "Robust higher order potentials for enforcing label consistency," in *Proc. CVPR*, 2008, pp. 1–8.
[52] J. D. Wegner, J. A. Montoya-Zegarra, and K. Schindler, "Road networks as collections of minimum cost paths," *ISPRS J. Photogramm. Remote Sens.*, vol. 108, pp. 128–137, Oct. 2015.
[53] L. Albert, F. Rottensteiner, and C. Heipke, "A higher order conditional random field model for simultaneous classification of land cover and land use," *ISPRS J. Photogramm. Remote Sens.*, vol. 130, pp. 63–80, Aug. 2017.
[54] C. Rother, V. Kolmogorov, T. Minka, and A. Blake, "Cosegmentation of image Pairs by histogram matching—Incorporating a global constraint into MRFs," in *Proc. CVPR*, 2006, pp. 993–1000.
[55] A. Joulin, F. Bach, and J. Ponce, "Discriminative clustering for image co-segmentation," in *Proc. CVPR*, 2010, pp. 1943–1950.
[56] P. Xiao, M. Yuan, X. Zhang, X. Feng, and Y. Guo, "Cosegmentation for object-based building change detection from high-resolution remotely sensed images," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 3, pp. 1587–1603, Mar. 2017.
[57] Y. Huangy, M. S. Brown, and D. Xu, "A framework for reducing ink-bleed in old documents," in *Proc. CVPR*, Anchorage, AK, USA, 2008, pp. 1–7.
[58] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 11, pp. 1222–1239, Nov. 2001.
[59] G. Roig, X. Boix, H. Ben Shitrit, and P. Fua, "Conditional random fields for multi-camera object detection," in *Proc. ICCV*, 2011, pp. 563–570.
[60] D. Tuia, M. Volpi, and G. Moser, "Getting pixels and regions to agree with conditional random fields," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Beijing, China, Jul. 2016, pp. 3290–3293.
[61] F. Melgani and S. B. Serpico, "A Markov random field approach to spatio-temporal contextual image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 11, pp. 2478–2487, Nov. 2003.
[62] P. H. Swain, "Bayesian classification in a time-varying environment," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-8, no. 12, pp. 879–883, Dec. 1978.
[63] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.
[64] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[65] V. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer, 2000.
[66] T.-F. Wu, C.-J. Lin, and R. C. Weng, "Probability estimates for multi-class classification by pairwise coupling," *J. Mach. Learn. Res.*, vol. 5, pp. 975–1005, Dec. 2004.
[67] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families, and variational inference," *Found. Trends Mach. Learn.*, vol. 1, nos. 1–2, pp. 1–305, 2008.
[68] R. Szeliski *et al.*, "A comparative study of energy minimization methods for Markov random fields," in *Proc. ECCV*, 2006, pp. 16–29.
[69] V. Kolmogorov, "Convergent tree-reweighted message passing for energy minimization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 10, pp. 1568–1583, Oct. 2006.
[70] A. Lagrange *et al.*, "Benchmarking classification of earth-observation data: From learning explicit features to convolutional networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2015, pp. 4173–4176.
[71] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2015.
[72] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.
[73] D. Marcos, M. Volpi, N. Komodakis, and D. Tuia, "Rotation equivariant vector field networks," in *Proc. ICCV*, Venice, Italy, 2017. [Online]. Available: http://openaccess.thecvf.com/content_ICCV_2017/papers/Marcos_Rotation_Equivariant_Vector_ICCV_2017_paper.pdf
[74] J. Gimenez, A. C. Frery, and A. G. Flesia, "When data do not bring information: A case study in Markov random fields estimation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 1, pp. 195–203, Jan. 2015.
[75] S. B. Serpico and G. Moser, "Weight parameter optimization by the Ho–Kashyap algorithm in MRF models for supervised image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 12, pp. 3695–3705, Dec. 2006.

Devis Tuia (S'07–M'09–SM'15) received the Ph.D. degree in environmental sciences from the University of Lausanne, Lausanne, Switzerland, in 2009.

He was a Post-Doctoral Researcher with the University of Valéncia, Valencia, Spain, the University of Colorado, Boulder, CO, USA, and the École Polytechnique Fédérale de Lausanne, Lausanne. From 2014 to 2017, he was an Assistant Professor with the University of Zürich, Zürich, Switzerland. He is currently an Associate Professor with the GeoInformation Science and Remote Sensing Laboratory, Wageningen University, Wageningen, The Netherlands. His research interests include algorithms for information extraction and data fusion of geospatial data (including remote sensing) using machine learning and computer vision.

Michele Volpi received the Ph.D. degree from the University of Lausanne, Lausanne, Switzerland, in 2013. From 2014 to 2016, he was a Visiting Post-Doctoral Researcher with the CALVIN Group, School of Informatics, University of Edinburgh, Edinburgh, Scotland, and then with the Multimodal Remote Sensing Group, University of Zürich, Zürich, Switzerland. Since 2016, he has been the Co-Chair of the ISPRS Working Group II/6 "Large scale machine learning for geospatial data analysis." He is currently with the Swiss Data Science Center, ETH Zürich, Zürich, where he is involved in developing and applying machine learning to a wide variety of topics. His research interests include the interface of remote sensing, machine and deep learning, and computer vision for the extraction of information from aerial, satellite, and geospatial data.

Gabriele Moser (S'03–M'05–SM'14) received the Ph.D. degree from the University of Genoa (Unige), Genoa, Italy, in 2005. Since 2013, he has been the Head of the Remote Sensing for Environment and Sustainability Laboratory, Unige, where he is an Associate Professor in Telecommunications. He has collaborated with the Image Processing and Pattern Recognition for Remote Sensing Laboratory, Unige, since 2001. His research interests include pattern recognition and image processing methodologies for remote sensing and energy applications.

Dr. Moser was the Chair of the IEEE GEOSCIENCE AND REMOTE SENSING SOCIETY (GRSS) Image Analysis and Data Fusion Technical Committee from 2013 to 2017. He is an Area Editor of PATTERN RECOGNITION LETTERS, and IEEE GEOSCIENCE AND REMOTE SENSING LETTERS. In 2015, he was a Guest Editor in IEEE GEOSCIENCE AND REMOTE SENSING MAGAZINE. He was a Publication Co-Chair of the 2015 IGARSS conference, a Program Co-Chair of the 2015/2017 IEEE GRSS EarthVision workshops held in the CVPR conference.