Learning CRF potentials through fully convolutional networks for satellite image semantic segmentation 1 st Martina Pastorino

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

Learning CRF potentials through fully convolutional networks for satellite image semantic segmentation

1st Martina Pastorino
DITEN dept.
University of Genoa, Italy
Inria, Université Côte d'Azur
Sophia-Antipolis, France
email: martina.pastorino@edu.unige.it

2nd Gabriele Moser
DITEN dept.
University of Genoa, Italy
email: gabriele.moser@unige.it

3rd Sebastiano B. Serpico
DITEN dept.
University of Genoa, Italy
email: sebastiano.serpico@unige.it

4th Josiane Zerubia
Inria, Université Côte d'Azur
Sophia-Antipolis, France
email: josiane.zerubia@inria.fr

***Abstract*—This paper introduces a method to automatically learn the unary and pairwise potentials of a conditional random field (CRF) from the input data in a non-parametric fashion, within the framework of the semantic segmentation of remote sensing images. The proposed model is based on fully convolutional networks (FCNs) and fully connected neural networks (FCNNs) to extensively exploit the semantic and spatial information contained in the input data and in the intermediate layers of an FCN. The idea of the model is twofold: first to learn the statistics of a CRF via a convolutional layer, whose kernel defines the clique of interest, and, second, to favor the interpretability of the intermediate layers as posterior probabilities through the FCNNs. The method was tested with the ISPRS 2D Semantic Labeling Challenge Vaihingen dataset, after modifying the ground truths to approximate the ones found in realistic remote sensing applications, characterized by scarce and spatially non-exhaustive annotations. The results confirm the effectiveness of the proposed technique for the semantic segmentation of satellite images.**

***Index Terms*—semantic segmentation, satellite images, CNN, FCN, CRF**

I. Introduction

Semantic segmentation, also known as pixel-level image classification, is the computer vision task of clustering together and labeling parts of an image that belong to the same semantic class. It plays a major role in several applications [1], for instance land cover mapping [2] and urban management [3], [4]. In the context of semantic segmentation, various techniques rooted in the methodological areas of stochastic models and deep learning (DL) have been proposed.

On the one hand, DL models achieve state-of-the-art results in image classification tasks, especially with the fully convolutional networks (FCNs) [5], capable to combine semantic information at different resolutions through skip connections and to yield classification results with arbitrary size [5], [6]. Yet, these models have heavy requirements in terms of quality and quantity of the input data used for training, which most often cannot be met in real-world remote sensing applications. On the other hand, probabilistic graphical models (PGMs) based on random fields, such as Markov random fields (MRFs) [7], and conditional random fields (CRFs) [8], are powerful tools for 2D image analysis tasks. They are capable to express dependencies between random variables over a multidimensional space through a graph representation, thus being convenient to model spatial and multiresolution information.

Several techniques combining PGMs and DL for semantic segmentation have been proposed [9] in order to obtain more accurate classification results in various scenarios. For example, the method in [10] presents the advantage of using the output of an FCN as the unary potentials of a dense CRF model with Gaussian pairwise potentials [11]. However, in this case the approach is not end-to-end, the CRF is simply used as a post-processing technique and its parameters are set through cross-validation [10]. Approaches allowing to train the CRF have been developed for semantic segmentation problems, for example through a piecewise CNN-based training, where two different networks compute the unary and pairwise potentials of the CRF [12]; or embedding the CRF in memory networks, such as with recurrent neural networks (RNNs) [13], [14]. In particular, in [13] the mean-field inference of a dense CRF with Gaussian pairwise potentials [11] is incorporated in a DL model as an RNN, enabling the joint end-to-end training of both the neural network and the CRF parameters by backpropagation. Since the mean-field inference is iterative, it is modeled across its time-steps to form an RNN [14]

The aim of this paper is to present a technique to automatically learn the potentials of a CRF up to the second order through an FCN. This allows the method to be completely non-parametric and directly learn from the input data. The multiscale information extracted by the FCN is explicitly modeled through the addition of fully connected neural networks (FCNNs) at different scales, favoring the interpretability of the hidden layers of the FCN as posterior probabilities.

The goal of the proposed method is twofold: (i) to define an end-to-end neural architecture based on FCNs to learn the

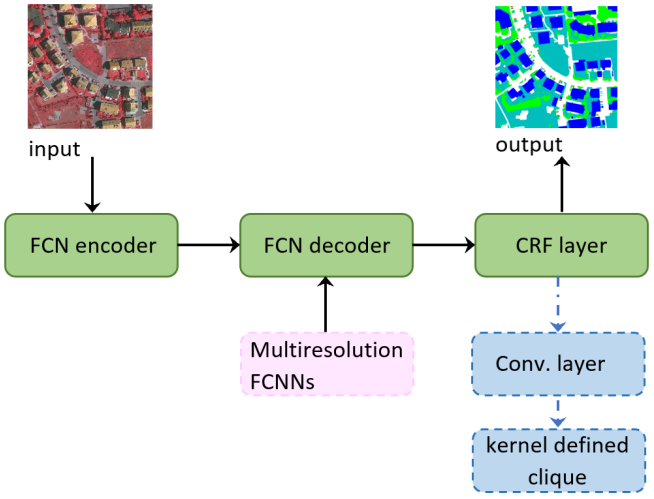


Fig. 1. Overall architecture of the proposed method.

potentials of a CRF model up to the second order and (ii) to take advantage of the multiscale information contained in very high resolution (VHR) images [15], thanks to the intrisic multiscale behaviour of FCNs through the addition of FCNNs, all framed in the application of semantic segmentation of remote sensing images.

II. METHODOLOGY

The proposed method aims to automatically learn the potentials of a categorical-valued CRF model with up to second order non-zero potentials, leveraging the modeling capabilities of DL architectures to directly learn semantic relationships from the input data. The model consists of an FCN [5] integrated with FCNNs at different convolutional blocks to manipulate the multiscale information intrinsically extracted by the intermediate convolutional layers, enforcing the interpretability of the network and its intermediate activations as posterior probabilities. Each FCNN is modeled as a convolutional layer with a kernel of size 1×1. The overall architecture is summarized in Fig. 1.

Briefly recalling the basics of CRF models [16], let's consider an image and the associated pixel lattice S. Let $\Omega = \{\omega_1, \omega_2, \ldots, \omega_M\}$ be the set of M semantic classes to which every pixel $s \in S$ can belong to. Calling $\mathcal{X} = \{x_s\}_{s \in S}$ and $\mathcal{Y} = \{y_s\}_{s \in S}$ the random fields of the observations and of the labels, respectively, $\mathcal{Y}$ is a CRF if, first:

$$P(y_s|y_q, q \neq s, \mathcal{X}) = P(y_s|y_q, q \in \partial s, \mathcal{X}) \tag{1}$$

where ∂s is a neighborhood of pixel s, and second, $P(\mathcal{Y}|\mathcal{X})$, the global posterior distribution, is strictly positive [7]. For CRF models considering up to the pairwise (second order) potentials, the energy function is defined as:

$$\mathcal{U}(\mathcal{Y}|\mathcal{X}) = \sum_{s \in S} D_s(y_s|\mathcal{X}) + \sum_{\substack{q \in \partial s \\ s \in S}} V_{sq}(y_s, y_q|\mathcal{X}) \tag{2}$$

with D_s and V_{sq} are the unary and the pairwise potentials, respectively. They define the statistics of the labels of each pixel s and the spatial relations among neighboring pixels s and q in a clique, given the random field of observations. In particular, this paper refers to CRFs where the pairwise potentials are of the kind $V_{sq}(y_s, y_q|\mathcal{X}) = E_{sq}(y_s, y_q|\mathcal{X})\delta(y_s, y_q)$ (where δ is the Kronecker impulse), shifting the focus on models where neighboring pixels likely share the same label, hence enforcing spatial consistency.

According to the Hammersley-Clifford theorem, the energy function of a CRF is related to the global posterior distribution [8]. In particular, focusing on the local posterior distribution of pixel s, we can write [7], [16]:

$$\mathcal{U}_s(y_s|\mathcal{X}) \propto -\ln P(y_s|\mathcal{X}) \tag{3}$$

where the local posterior energy $\mathcal{U}_s(y_s|\mathcal{X})$ includes all terms of (2) that regard pixel s. In the proposed approach, the CRF is modeled through an FCN by making use of an additional convolutional layer, whose kernel size defines the clique [14], [16], therefore influences the span of spatial information integrated by the CRF. For example, a 3×3 kernel can define a first or a second order neighborhood system (in the first case, some of the kernel weights are set to zero to consider only the 4 adjacent pixels to the central one). The number of inputs and outputs of this layer is equivalent to the number of classes M. Accordingly, the estimated pixelwise posterior probability that pixel s belongs to class k, deriving from the softmax operation, is expressed as:

$$-\ln \hat{P}(y_s = \omega_k|\mathcal{X}) = -\sum_{q \in \partial s \cup \{s\}} h^k_{q-s} f_q(\omega_k|\mathcal{X}) \tag{4}$$

with $k = 1, 2, \ldots, M$. In (4), $\hat{P}$ is used to acknowledge that this distribution is estimated through the network, and we recall that Markovian neighborhoods do not include their central pixel (i.e., $s \notin \partial s$). h is the kernel of the last convolutional layer and f are the feature maps output of the previous layer, whose output generally depends on the input image $\mathcal{X}$. It is possible to prove that this expression relates to the form of the unary and pairwise potentials of a CRF. Specifically, the corresponding potentials are defined by:

$$\begin{cases} D_s(\omega_k|\mathcal{X}) = -h^k_0 f_s(\omega_k|\mathcal{X}) \\ V_{sq}(\omega_k, \omega_m|\mathcal{X}) = -h^k_{q-s} f_q(\omega_k|\mathcal{X})\delta(\omega_k, \omega_m) \end{cases} \tag{5}$$

with the unary potential only depending on the central value of the kernel h_0 and the pairwise potential containing the Kronecker delta term, whose smoothing properties are desired in order to favor spatially homogeneous regions typical of natural images [17]. Accordingly, the energy of the CRF is entirely learnt by the DL architecture applied on the input

data, and is defined as:

$$\mathcal{U}(\mathcal{Y}|\mathcal{X}) = -\sum_{s}\Bigg[h_0^k f_s(y_s|\mathcal{X}) + \\ + \sum_{q\in\partial s} h_{q-s}^k f_q(y_q|\mathcal{X})\delta(y_q, y_s)\Bigg] \quad (6)$$

The loss function of the proposed method is a linear combination of the cross-entropy losses of the multiscale terms – associated to the FCNNs – and of the output of the final convolutional layer, in order to take into account both the multiscale and the spatial information.

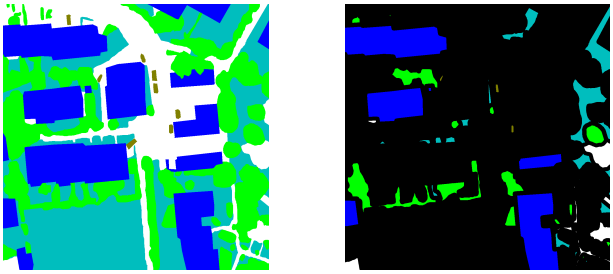
III. EXPERIMENTAL VALIDATION

The proposed method was experimentally validated with the ISPRS 2D Semantic Labeling Challenge dataset consisting of optical aerial images of the city of Vaihingen, Germany[1]. The images are at VHR, with a spatial resolution of 9 cm, and are characterized by six land-cover classes: buildings, impervious surfaces (e.g., roads), low vegetation, trees, cars, and clutter. The last class only appears in few training tiles and does not have a strong semantic meaning, since it includes all of the surface covers that do not belong to the other well-defined classes. Hence, it was removed from the experimentation (similar to previous works [22], [23]).

The original ground truth maps available in the dataset were modified to approximate scarce input information typically found in real-world remote sensing applications and characterized by sparse patches. The new ground truth maps were obtained removing entire connected components and applying morphological erosion to the remaining map (see Fig. 2(b)). The FCNNs were connected to the three deconvolutional blocks of the FCN, thus modeling three different spatial resolutions other than the one of the original input image. The overall loss function of the neural network takes into account this information after a proper resampling of the ground truth.

The convolutional layer dedicated to the learning of the potentials of the CRF is responsible for the further modeling of the spatial information. The size of its kernel defines the neighborhood system over which the contextual information is analyzed, i.e., the clique on which the CRF is built. In the following experiments, a 3×3 kernel is employed, thus defining a second-order neighborhood system (an eight pixel neighborhood). The training learning rate is fixed to 0.01 with a decay rate of 0.0005, and the optimizer employed is the Adam algorithm [24]. The experiments were conducted with a GPU NVIDIA GeForce RTX 2080 Ti.

The proposed method was compared to other approaches relating DL and CRFs in the framework of semantic segmentation, such as the method in [10], which employs a dense CRF [11] as a post-processing technique, whose unary potentials are computed by the DL model; and the one in [13] defining an end-to-end method were the CRF is modeled as an RNN. In this paper, all these techniques were formulated to use the U-Net as a backbone, in order to be consistent across the various experiments. Hence, the results of U-Net are also presented as baseline, together with those of SegNet [19], another FCN frequently used for segmentation problems. Further comparisons were performed with HRNet [21], a method explicitly integrating multiresolution information through multiresolution subnetworks connected in parallel, and DeepLabV3+ [20], an FCN with atrous spatial pyramid pooling (ASPP), able to encode multiscale contextual information by applying atrous convolutions with different rates.



(a) (b)

Fig. 2. Original (a) and scarce (b) ground truth.

The results reported in Table I suggest the effectiveness of the proposed methodology, which takes into account both multiresolution information, through the FCNNs, and spatial-contextual information, through the CRF modeled by the additional convolutional layer. They are expressed in terms of classwise accuracies and averaged metrics (overall accuracy, recall, precision, and F1 score). The approach presented in this paper manages to obtain remarkable improvements with respect to the techniques used for comparison for most of the averaged metrics, as well as the highest accuracy values for the classes "buildings", "trees", and "cars". This last class represents the smallest objects in the considered dataset. Consistently, it is the most affected by the scarcity of the input ground truth information. The improvement in classification accuracy for this class further confirms the effectiveness of the proposed model integrating information between neighboring pixels both inter-scale and intra-scale.

The classification maps reported in Fig. 3 further support these comments. In general, the classification results obtained by the three methods employing a CRF model (see Fig. 3(b)-(d)), either defined end-to-end with a neural architecture or used as post-processing, appear to be visually smoother, confirming the advantages of modeling spatial-contextual information. More in details, the maps obtained by the proposed method (see Fig. 3(b)) have sharper edges and more homogeneous zones, particularly noticeable for the class "buildings", where the impact of the scarcity of the input data is better attenuated than in the other techniques (see Fig. 3(c)-(e)).

In general, all the methods involving a CRF model obtain comparable results for all the majority classes, "buildings" and "impervious surfaces", with some having better classification performances for the first (U-Net coupled with the dense CRF

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

TABLE I
TEST-SET RESULTS ON THE ISPRS 2D SEMANTIC LABELING CHALLENGE VAIHINGEN DATASET WITH SCARCE INPUT INFORMATION.

Architecture	**buildings** (recall)	**impervious** (recall)	**low-veg** (recall)	**trees** (recall)	**cars** (recall)	**overall accuracy**	**average recall**	**average precision**	**average F1 score**
Baseline FCN									
U-Net [18]	0.87	**0.93**	0.64	**0.87**	0.76	0.82	0.81	0.84	0.82
SegNet [19]	**0.93**	0.82	**0.84**	0.76	0.80	**0.85**	0.83	0.84	0.83
FCN with ASPP									
DeepLabV3+ [20]	**0.93**	0.78	0.67	0.80	0.33	0.79	0.65	0.75	0.70
CRF as post-processing									
DL + dense CRF [11]	**0.93**	0.80	0.73	0.85	0.60	0.82	0.78	0.84	0.81
End-to-end CRF									
CRF-RNN [13]	0.86	0.92	0.63	**0.87**	0.70	0.81	0.80	0.84	0.82
Proposed method	**0.93**	0.82	0.79	**0.87**	**0.93**	**0.85**	**0.87**	**0.86**	**0.86**
Multiresolution									
HRNet [21]	0.84	0.75	0.82	0.69	0.49	0.77	0.72	0.79	0.75

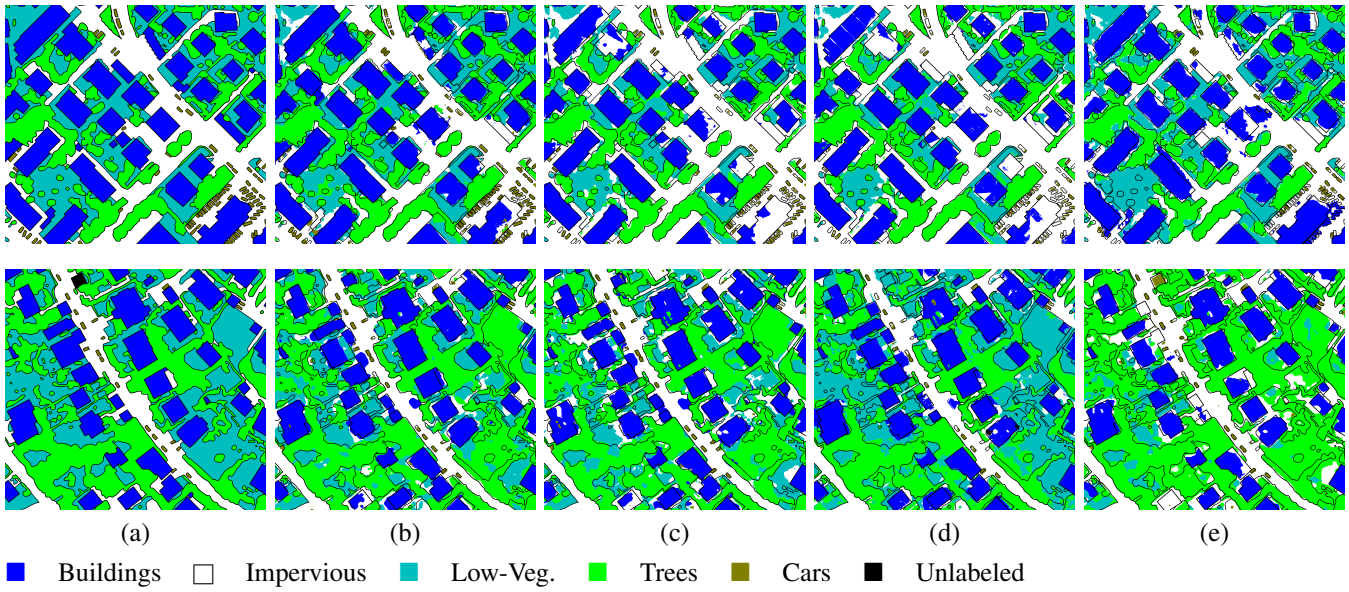


Fig. 3. GTs and classification results: (a) GT and classification maps from (b) the proposed method, (c) CRF-RNN [13] (d) DL + dense CRF [11] as post-processing, and (e) HRNet [21].

post-processing, the proposed method, but also SegNet and U-Net) and some for the latter (U-Net, CRF-RNN). The methods explicitly modeling spatial-contextual information also allow a better discrimination of the two vegetated classes in the dataset, "low vegetation" and "trees". Conversely, the approach modeling multiresolution information is characterized by the second highest classwise accuracy value in terms of "low vegetation", to the detriment of the class "trees", further confirmed by the appearance of the associated classification maps (see Fig. 3(e)) and by the lower values of precision and recall with respect to the other analyzed approaches.

The SegNet [19] architecture obtains the most accurate results for the class "low vegetation" and, together with the proposed method, the highest values for the recall on "buildings" and the overall accuracy. Its classification maps appear to be quite smooth and coherent with the GT information, as reported in Fig. 4(e). However, as compared to the proposed technique, it attains lower values for the average precision, recall, and F1 score, due to the fact that its classification performances are poorer for the minority classes of the dataset ("trees" and "cars"). This confirms once more the potential of modeling explicitly multiresolution information.

Contrarily, DeepLabV3+ [20], with its encoder-decoder architecture, spatial pyramid pooling, and atrous convolutions, tends to obtain generally less accurate results in the case of very scarce ground truth information, as reported by Table I. Nonetheless, consistently with the proposed method, SegNet and the technique with the CRF as post-processing, reaches accurate results for the discrimination of the class "buildings".

The methodology using the dense CRF described in [11] as post-processing, which was here applied to the posteriors extracted by the U-Net to reproduce the method proposed in [10] (employing a VGG16 instead of a U-Net), remarkably recovers semantic information for the class "buildings". How-

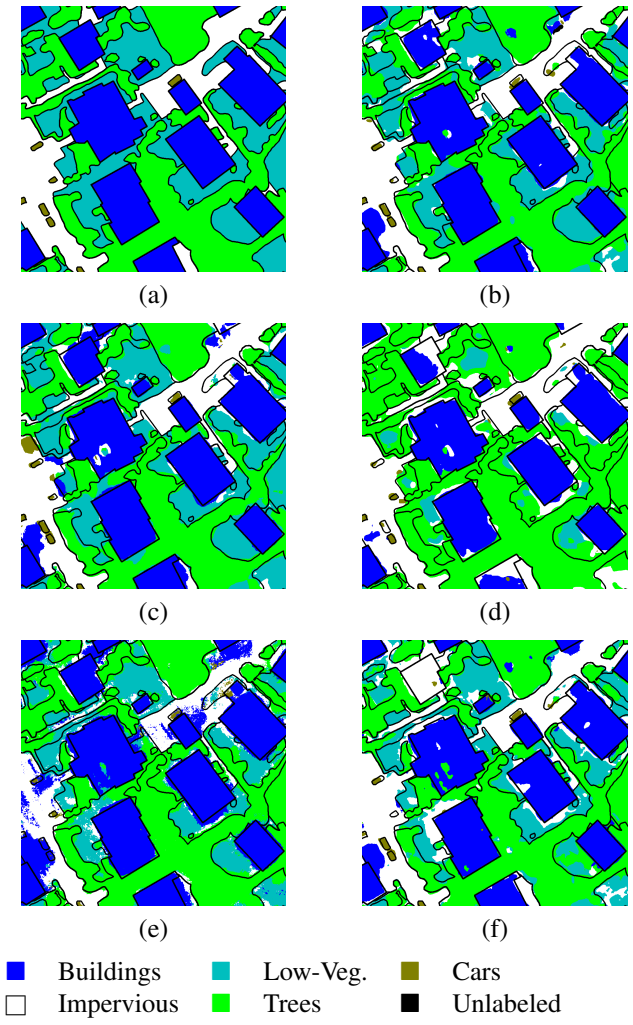


Fig. 4. Zoom-in of the GT and the classification maps: (a) GT and classification maps from (b) the proposed method, (c) U-Net [18], (d) HRNet [21], (e) SegNet [19], and (f) CRF-RNN [13].

ever, it looses in terms of "impervious surfaces", the class with the highest relative frequency in the dataset, thus justifying the slight loss in average recall and, consequently, F1 score.

More details on the classification maps obtained by the different techniques are shown in Fig. 4, that presents a zoom-in of an area of the original segmented image. It is possible to notice how the proposed technique recovers some information lost, in particular for the class "buildings" (see Fig. 4(b)), while limiting the false alarms over other class covers (see for example Fig. 4(d)-(f)). The classification of the edges of the buildings is presented clearly in Fig. 5, where the results of the proposed method are compared to those of U-Net, HRNet and CRF-RNN. The proposed technique shows more defined and sharper contours, while avoiding trespassing the limits of the buildings of the original GT, as mostly happens, on the contrary, for the results of HRNet (see Fig. 5(c)).

The introduction of multiscale FCNNs at different convolutional blocks of the FCN and their multiscale loss terms favors the interpretability of the intermediate layers of the FCN as

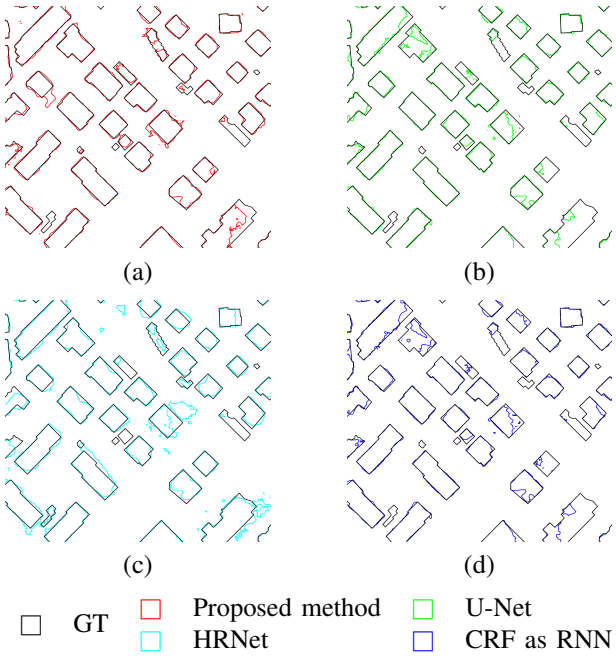


Fig. 5. Building contours extracted by the various techniques on the scarce dataset: (a) proposed method, (b) U-Net [18], (c) HRNet [21], and (d) CRF-RNN [13].

posterior probabilities, as it can be seen in Fig. 6. This figure presents the feature and classification maps extracted at 40 m of resolution – at the first FCNNs starting from the first convolutional block of the decoder of the FCN – and at 5 m – at the output of the FCN–, the original resolution. The addition of the FCNNs, aiming at integrating and modeling multiscale information, actually allows to interpret the feature maps as posterior probabilities even at a coarser scale (see the top row of Fig. 6).

This experimental validation confirms the potential of jointly leveraging both multiresolution and spatial information, which favors higher classification accuracy.

IV. DISCUSSION AND CONCLUSION

A new end-to-end approach to automatically learn the potentials of a CRF model through a DL architecture based on FCNs and FCNNs for dense image classification tasks was presented in this paper. The method leverages the semantic and spatial modeling capabilities of neural networks to learn directly from the input data in a non-parametric fashion within the framework the semantic segmentation of remote sensing images. Multiscale information is modeled through the FCNNs in order to favor accurate class discrimination by the integration of details extracted at different spatial resolutions.

The experimental results in the case of scarce ground truth data, a common scenario in remote sensing applications, demonstrate the effectiveness of the proposed approach. This approach generally attains classwise classification accuracies comparable to or higher than the ones of other state-of-the-art techniques, and better performances in terms of all the averaged accuracy metrics. The presented methodology, integrating

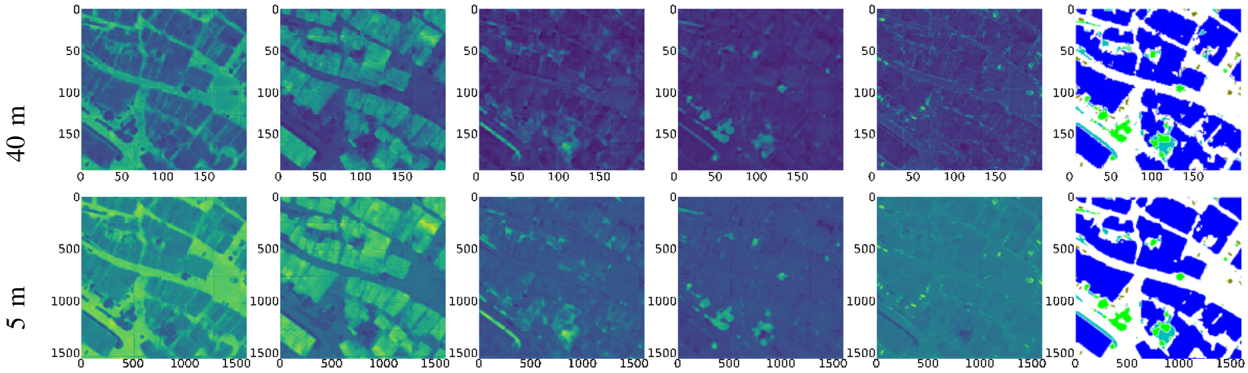


Fig. 6. Classwise posterior probabilities and relative classification maps: first FCNN of the intermediate layers (spatial resolution 40 m *top*); and output layer of the FCN (original spatial resolution 5 m *bottom*). From left to right: impervious surfaces, buildings, low vegetation, trees, and cars.

multiresolution information, thanks to the FCNNs, and spatial-contextual information, related to the CRF model, is capable to produce smoother, more homogeneous classification maps with sharper edges.

Future work may involve the introduction of domain adaptation techniques [25] in order to apply this supervised technique to datasets dealing with natural disasters [26], typically characterized by very scarce ground truth information, if any at all.

REFERENCES

[1] G. Csurka, R. Volpi, and B. Chidlovskii, "Semantic image segmentation: Two decades of research," *Foundations and Trends in Computer Graphics and Vision*, vol. 14, no. 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 Sensing*, vol. 14, no. 3, 2022.

[3] M. Volpi and V. Ferrari, "Semantic segmentation of urban scenes by learning local class interactions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2015, pp. 1–9.

[4] L. Wang, R. Li, C. Zhang, S. Fang, C. Duan, X. Meng, and P. M. Atkinson, "UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 190, pp. 196–214, 2022.

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

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

[7] S.Z. Li, *Markov random field modeling in image analysis*, Springer, 3rd edition, 2009.

[8] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the Eighteenth International Conference on Machine Learning (ICML)*, San Francisco, CA, USA, 2001, p. 282–289, Morgan Kaufmann Publishers Inc.

[9] A. Arnab, S. Zheng, S. Jayasumana, B. Romera-Paredes, M. Larsson, A. Kirillov, B. Savchynskyy, C. Rother, F. Kahl, and P.H.S. Torr, "Conditional random fields meet deep neural networks for semantic segmentation: Combining probabilistic graphical models with deep learning for structured prediction," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 37–52, 2018.

[10] 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 Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[11] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected CRFs with Gaussian edge potentials," in *Proceedings of Neural Information Processing Systems*, pp. 109–117, 2011.

[12] G. Lin, C. Shen, A. van den Hengel, and I. Reid, "Efficient piecewise training of deep structured models for semantic segmentation," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3194–3203.

[13] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr, "Conditional random fields as recurrent neural networks," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1529–1537.

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

[15] Y. Cai, L. Fan, and Y. Fang, "SBSS: Stacking-based semantic segmentation framework for very high-resolution remote sensing image," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–14, 2023.

[16] C. Sutton, A. McCallum, and F. Pereira, "An introduction to conditional random fields," *Foundations and Trends in Machine Learning*, vol. 4, no. 4, pp. 267–373, 2011.

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

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

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

[20] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Computer Vision – ECCV 2018*, pp. 833–851, 2018.

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

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

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

[24] Diederik Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 2014.

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

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