FULLY CONVOLUTIONAL AND FEEDFORWARD NETWORKS FOR THE SEMANTIC SEGMENTATION OF REMOTELY SENSED IMAGES

Martina Pastorino [1,2]*, Gabriele Moser* [1]*, Sebastiano B. Serpico* [1]*, and Josiane Zerubia* [2].

[1] University of Genoa, DITEN dept., Genoa, Italy, martina.pastorino@edu.unige.it.
[2] Inria, Université Côte d'Azur, Sophia-Antipolis, France.

ABSTRACT

This paper presents a novel semantic segmentation method of very high resolution remotely sensed images based on fully convolutional networks (FCNs) and feedforward neural networks (FFNNs). The proposed model aims to exploit the intrinsic multiscale information extracted at different convolutional blocks in an FCN by the integration of FFNNs, thus incorporating information at different scales. The purpose is to obtain accurate classification results with realistic data sets characterized by sparse ground truth (GT) data by taking benefit from multiscale and long-range spatial information. The final loss function is computed as a linear combination of the weighted cross-entropy losses of the FFNNs and of the FCN. The modeling of spatial-contextual information is further addressed by the introduction of an additional loss term which allows to integrate spatial information between neighboring pixels. The experimental validation is conducted with the ISPRS 2D Semantic Labeling Challenge data set over the city of Vaihingen, Germany. The results are promising, as the proposed approach obtains higher average classification results than the state-of-the-art techniques considered, especially in the case of scarce, suboptimal GTs.

Index Terms— CNN, FCN, feedforward networks, semantic segmentation, multiresolution satellite images

1. INTRODUCTION

The development of a supervised method for the semantic segmentation–or dense image classification–of remote sensing (RS) images at very high resolution (VHR) is a challenging problem. Current space missions allow to obtain multimodal (e.g., multiresolution, multisensor, multiband, multifrequency) satellite imagery and to reach fine spatial resolutions (up to 30 cm). Aerial imagery often have resolutions of a few centimeters. Methods based on deep learning (DL) are capable to successfully integrate this information and obtain accurate classification results [1, 2, 3]. In particular, models based on FCNs [4] are the state-of-the-art techniques for semantic segmentation tasks. These methods, however, need large training data sets [5], which require high efforts in annotation and most often are not available for RS applications. In this case, the modeling of the information contained at different spatial resolutions (e.g., the spatial details of the finest and the robustness to noise plus outliers of the coarsest), has proven to be effective and guarantees improvements of the segmentation results in spatial precision and accuracy [6].

Indeed, thanks to the operations executed by FCNs [4], characterized by different multiscale processing stages through convolutional and pooling layers, it is possible to derive multiscale information. This is mostly contained in the hidden layers of the FCN and can be retrieved directly via skip connections. Stochastic models, such as probabilistic graphical models (PGMs) (e.g., Markov models, postulated on planar or multilayer graphs), are flexible and powerful models which can extract information from multiscale data for labeling purposes [7, 8].

Several approaches that combine DL and models capable of making structured predictions, such as PGMs, have been studied in order to improve the accuracy of the results of semantic segmentation tasks by capturing the interactions between pixels. Methods that incorporates graphical models into a DL framework such as [9, 10, 11] have also been considered. In [9], convolutional neural networks (CNNs) and conditional random fields (CRFs) are integrated in an end-to-end approach where the mean-field inference for the CRF is approximated with Gaussian pairwise potentials as recurrent neural networks (RNNs). In [10] CRFs and FCNs are combined to model the context exhibited within the middle layers of an FCN and the mean-field inference process of a dense CRF is approximated as a multi-dimensional gated recurrent unit (GRU) layer. In [11] a Markov random field (MRF) is defined by means of a CNN, namely a deep parsing network (DPN). In particular, DPN extends a CNN to model unary terms and additional layers are devised to approximate the mean field (MF) algorithm for pairwise terms.

In a recent approach [12], DL and stochastic models are combined to incorporate multiresolution information present at different layers of the neural networks to address the semantic segmentation of RS images in the case of scarce GT.

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

The code is available at `https://github.com/Ayana-Inria/FCN-FFNET_RS-semantic-segmentation`

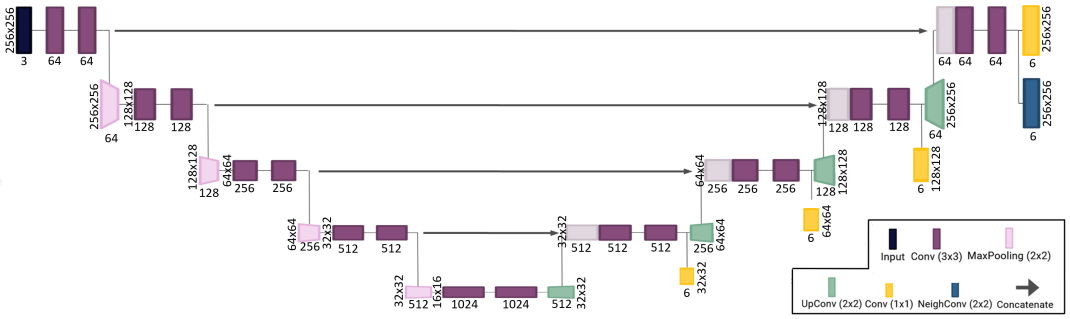


Fig. 1. Overall architecture of the method.

In this paper, our aim is to take advantage of multiscale information extracted by the FCN through the addition of FFNNs at different convolutional blocks of the FCN– corresponding to information at different scales. As compared to the previous approach in [12], the method developed in this paper is based on an architecture capable of integrating multiscale data without the need of ensemble learning techniques, thus making the approach end-to-end. Spatial-contextual information is favored in this framework through a supplementary convolutional layer modeling the interactions between neighboring pixels at the same resolution.

The goal of the proposed method is threefold: firstly, to take advantage of the instrisic multiscale behaviour of FCNs to integrate multiscale information through the addition of FFNNs; secondly, to strengthen the modeling of the spatial information through an additional convolutional layer; finally, to define a loss function capable to take into account both this multiscale and spatial information within an end-to-end neural model (see Fig. 1).

2. METHODOLOGY

FCNs, which are networks with an encoder-decoder architecture, are characterized by an intrinsic quadtree structure: in fact, the pooling and unpooling layers of size 2×2, normally employed within each convolutional block, match the power-of-2 relation typical of the pixel grids S^l $(l = 1, \ldots, L)$ at different L levels of a quadtree [13].

In order to exploit this intrinsic quadtree structure, the idea is to add FFNNs at each convolutional block in the decoder of the FCN, in this case a U-Net [14]. The objective of the FFNNs is to retrieve the multiscale information contained at different scales in the activations of the feature maps of the hidden layers of the FCN. The FFNNs are built through convolutional layers with kernels of size 1×1, in order to work with 2D data and compute pixelwise posterior probabilities.

To integrate further spatial information, a convolutional layer is added after the output layer of the FCN. Zero padding is employed to ensure that every pixel is considered. The size of the kernel defines the number of neighboring pixels that influence a pixel i and, therefore, the spatial-contextual information modeled.

The loss function, expressed in equation (1), is designed as a linear combination of L multiscale weighted cross-entropy loss functions. The losses were computed over the pixelwise softmax [15] of the $L-1$ FFNNs (after a resampling of the GTs at the corresponding resolution) and of the final feature map of the output layer of the FCN. They can be considered as unary terms, since they only depend on the value of each pixel considered individually. The weighting factor introduced in the cross-entropy loss is inversely proportional to the number of samples of each class in the image analysed, in order to take into account the presence of imbalanced training data.

$$\mathcal{L}_u = -\frac{1}{L}\sum_{l=1}^{L}\sum_{k=0}^{M-1}\frac{\hat{P}_{\max}}{\hat{P}_k}\sum_{i\in S^l} t_{ik}\log\left(p_{ik}\right) \tag{1}$$

$$p_{ik} = \frac{\exp\left(z_{ik}\right)}{\sum_{n=0}^{M-1}\exp\left(z_{in}\right)} \tag{2}$$

where M is the number of classes, p_{ik} is the estimated output probability that pixel i belongs to class k, $(z_{i1}, \ldots, z_{ij})$ is the prediction vector, $\hat{P}_k$ is the prior probability of the k-the class, estimated as its relative frequency in the training set ($k = 0, 1, \ldots, M-1$), and $\hat{P}_{\max} = \max_k \hat{P}_k$. $t_{ik} = 1$ only if the sample i belongs to class k, otherwise it is equal to 0. The number of samples in the $l-$th pixel grid S^l is $N_l = W_l \times H_l$, with W_l and H_l being the width and the height of the pixel grid, respectively. Considering N_1 the number of pixels in the smallest grid ($l = 1$), corresponding to the first convolutional block of the decoder, in general, $N_l = N_1 \cdot 2^{2(l-1)}$, given the power-of-2 relation between the pixel grids at consecutive convolutional blocks.

This loss function was integrated with an implicitly pairwise term, shown in equation (3), deriving from the cross-entropy loss function over the pixelwise softmax of the feature map obtained by the additional convolutional layer described above:

$$\mathcal{L}_p = -\sum_{a=1}^{W_L}\sum_{b=1}^{H_L}\sum_{k=0}^{M-1} t_{ik}\log\left(\sum_{j=0}^{D-1}\sum_{q=0}^{D-1} h_{j,q}\cdot p_{(a-j,b-q),k}\right) \tag{3}$$

with $h_{j,q}$ one of the weights of the kernel of size $D \times D$, and $p_{(a,b),k}$ the probability that pixel in position $i = (a, b)$ belongs to class k. The total loss is defined as follows:

$$\mathcal{L} = \mathcal{L}_u + \mathcal{L}_p \tag{4}$$

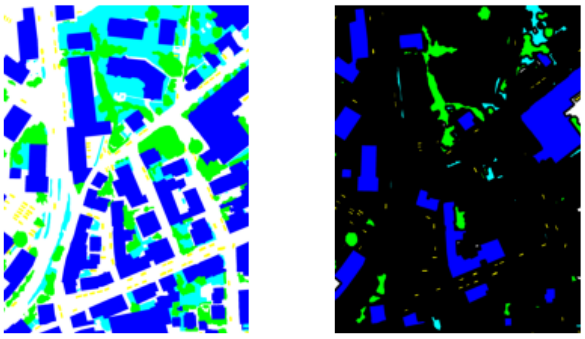
The methodology is briefly summarized in Algorithm 1.

Algorithm 1 FCN + FFNNs and spatial loss

Training of the proposed network with the input VHR data set

1: Unary terms of the loss function (1): linear combination of L weighted cross-entropy losses computed on the softmax layers of the FFNNs and the final output layer of the original FCN.
2: Addition of a convolutional layer integrating spatial information between neighboring pixels.
3: Addition of the pairwise term (3) to the loss function computed with a cross-entropy function on the output of the additional convolutional layer.

Output: final classification map.



(a) (b)

Fig. 2. GT images used to train the proposed architecture: (a) full, (b) sparse. Classes: buildings (blue), impervious (white), vegetation (cyan), trees (green), cars (yellow), unlabeled (black).

3. EXPERIMENTAL VALIDATION

The proposed method was tested with the ISPRS 2D Semantic Labeling Challenge data set containing aerial images of the city of Vaihingen, Germany[1]. It is a data set of VHR images, with a spatial resolution of 9 cm and it contains six classes: impervious surfaces (e.g., roads), buildings, low vegetation, trees, cars, and clutter. The last class is of relatively limited interest, as it comprises all the surfaces not belonging to the previous five and is highly mixed. Furthermore, it accounts for only a small percentage of pixels. Following the example of previous authors [18, 19, 20], the results of the class "clutter" were excluded from the averaged metrics.

The results herein reported were obtained with $L = 4$, i.e., with FFNNs attached to 3 distinct convolutional blocks of the decoder of the FCN; D, the size of the kernel of the additional convolutional layer, is equal to 2, therefore each term in the pairwise loss function contained information of a pixel and three of its neighbors, (other results were omitted for brevity). The equipment used for the experiments is an Alienware Aurora R11 with a RAM of 16 GB and a GPU NVIDIA GeForce RTX 2080 Ti.

Two training conditions were considered (shown in Fig. 2(a)-(b)): (i) the full data set with densely labeled GTs and (ii) a data set of scarce GTs, obtained by removing entire connected components from the original exhaustive GT maps and then applying morphological erosion to the remaining information [12], in order to approximate realistic GTs with isolated patches typically found in RS applications.

The experiments were conducted on both the full images, whose results are reported in Table 1, and subsections of size 1280×1280 of the original test images, to be able to compare the proposed architecture with the previous ones, whose results are collected in [12]. Table 2 shows a quantitative analysis of the results obtained in the latter case.

As it can be seen from the values reported in Tables 1-2, the proposed method exhibits remarkable improvements concerning the average accuracy metrics, especially when the input training data is scarce and approaches the realistic GTs available for RS applications. In particular, when compared to the previous technique [12] and a standard FCN (e.g., U-Net), the proposed approach was capable of reaching higher or similar classwise results and generally higher values for the average metrics. The comparison of the classification maps obtained with U-Net and the proposed method are shown in Fig. 3(b)-(c).

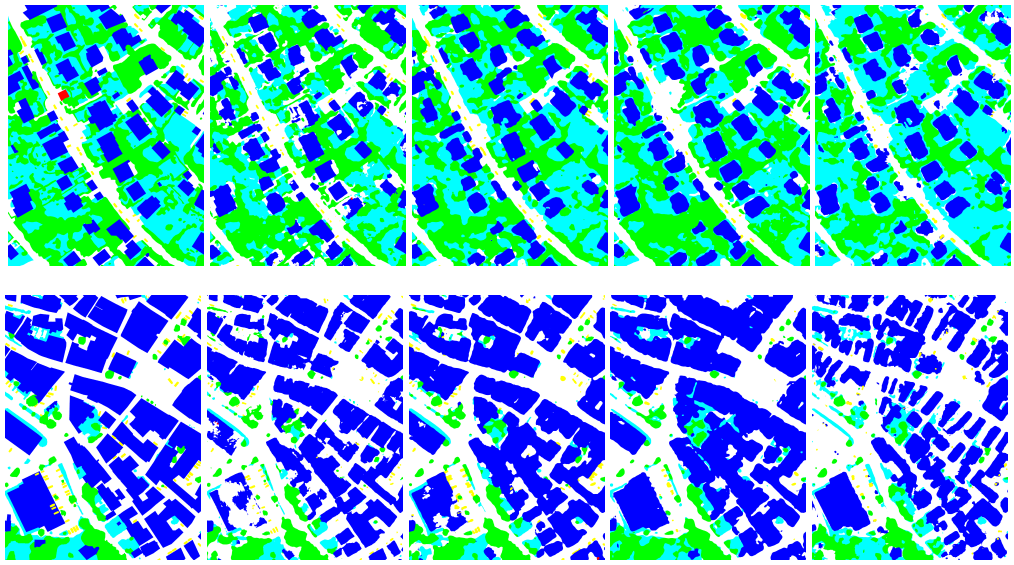
The proposed approach was further compared to a multiscale feature fusion (MFF) technique, namely a light-weight attention network (LWN-Attention[2]) [16], and to HRNet[3] [17], a neural network integrating multiscale information through a set of multiresolution subnetworks connected in parallel. MFF approaches involve multiscale information through the concatenation of feature maps associated with different scales [16]. The qualitative results obtained with these two methods are shown in Fig. 3(d)-(e).

From this comparison it is possible to notice that the proposed approach obtained higher or very similar per-class recalls, and generally higher overall accuracies, recalls and F1 scores. With the sparse GTs, the new approach guaranteed an improvement in overall accuracy, recall, and F1 score of about 5% with respect to the considered state-of-the-art techniques, reaching values up to 86% of accuracy for the full images. The proposed approach also attained improved results for the classification of minority classes in the case of scarce GT data, compared to U-Net, HRNet, and the MFF method.

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

[2] https://github.com/syliudf/LWN-Attention

[3] https://github.com/HRNet/HRNet-Semantic-Segmentation



(a) (b) (c) (d) (e)

Fig. 3. GTs and classification results: (a) GT and classification maps from (b) U-Net, (c) the proposed method, (d) LWN-Attention [16], and (e) HRNet [17]. Classes: buildings (blue), impervious (white), vegetation (cyan), trees (green), cars (yellow).

Table 1. Test-set results with full images. Precision and recall are averaged over the classes.

	Architecture	buildings	impervious	vegetation	trees	cars	overall acc.	recall	precision	F1 score
Full GT	U-Net [14]	0.86	0.91	0.79	0.90	**0.86**	0.89	0.87	0.85	0.86
	HRNet [17]	0.89	0.89	0.50	0.89	0.81	0.79	0.80	0.81	0.81
	Proposed method	**0.96**	0.91	**0.80**	**0.91**	0.81	**0.91**	**0.88**	0.87	**0.88**
	LWN-Attention [16]	**0.96**	**0.93**	0.71	0.89	0.61	0.89	0.82	**0.88**	0.85
Sparse GT	U-Net [14]	0.87	**0.93**	0.64	**0.87**	0.76	0.82	0.81	0.84	0.83
	HRNet [17]	0.84	0.75	**0.82**	0.69	0.49	0.77	0.72	0.79	0.75
	Proposed method	**0.94**	0.78	0.80	**0.87**	**0.90**	**0.86**	**0.86**	**0.85**	**0.86**
	LWN-Attention [16]	**0.94**	0.82	0.65	0.85	0.60	0.81	0.78	0.83	0.80

Table 2. Test-set results with cropped images. Precision and recall are averaged over the classes.

	Architecture	buildings	impervious	vegetation	trees	cars	overall acc.	recall	precision	F1 score
Full GT	U-Net [14]	**0.92**	0.83	0.71	0.92	0.74	0.85	**0.83**	0.84	0.83
	HRNet [17]	0.89	0.81	0.42	0.92	0.63	0.77	0.73	0.76	0.74
	Proposed method	**0.92**	0.85	**0.73**	**0.93**	0.73	**0.86**	**0.83**	**0.85**	**0.84**
	FCN+PGM [12]	0.84	0.81	0.68	0.92	**0.86**	0.81	0.82	0.72	0.77
	LWN-Attention [16]	0.91	**0.97**	0.61	0.88	0.64	0.85	0.80	0.82	0.81
Sparse GT	U-Net [14]	**0.96**	0.65	0.47	**0.89**	0.48	0.76	0.69	**0.81**	0.75
	HRNet [17]	0.84	0.77	0.68	0.80	0.29	0.77	0.68	0.74	0.71
	Proposed method	0.90	**0.79**	**0.70**	**0.89**	0.68	**0.82**	**0.79**	**0.81**	**0.80**
	FCN+PGM [12]	0.94	0.68	0.49	0.86	**0.74**	0.76	0.74	0.75	0.75
	LWN-Attention [16]	0.91	0.75	0.44	0.87	0.51	0.76	0.70	0.74	0.72

4. DISCUSSION AND CONCLUSION

A new end-to-end approach for the semantic segmentation of RS images based on FCNs, FFNNs and a spatial loss function is proposed in this paper. The experimental results demonstrate the effectiveness of the proposed approach for the semantic segmentation of RS images, with remarkable improvements in the overall accuracy and average recall as compared to previous architectures, especially in the case of scarce GT data. This confirms the possibility to obtain classification improvements when integrating the intrinsic multiscale information extracted by FCNs across their layers and by addressing local spatial-contextual information.

Future work may involve the extension of this methodology to the multisensor case, with optical and radar images acquired by different missions and therefore with different spatial resolutions, frequencies, and bands. Moreover, the model may be combined with transfer learning techniques to allow predictions on data sets different–in features and complexity–with respect to the ones used for training. This would allow the use of the proposed methodology on data sets with poorer GT information, for example the ones related to disaster management.

5. REFERENCES

[1] Y. J. E. Gbodjo, O. Montet, D. Ienco, R. Gaetano, and S. Dupuy, "Multisensor land cover classification with sparsely annotated data based on convolutional neural networks and self-distillation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 11485–11499, 2021.

[2] A. Farooq, X. Jia, J. Hu, and J. Zhou, "Transferable convolutional neural network for weed mapping with multisensor imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022.

[3] S. Piramanayagam, E. Saber, W. Schwartzkopf, and F. W. Koehler, "Supervised classification of multisensor remotely sensed images using a deep learning framework," *Remote Sensing*, vol. 10, no. 9, 2018.

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

[5] L. Maggiolo, D. Marcos, G. Moser, S. B. Serpico, and D. Tuia, "A semisupervised CRF model for CNN-based semantic segmentation with sparse ground truth," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

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

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

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

[9] 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 *IEEE International Conference on Computer Vision (ICCV)*, pp. 1529–1537, 2015.

[10] K. Nguyen, C. Fookes, and S. Sridharan, "Context from within: Hierarchical context modeling for semantic segmentation," *Pattern Recognition*, vol. 105, pp. 107358, 2020.

[11] Z. Liu, X. Li, P. Luo, C. C. Loy, and X. Tang, "Deep learning Markov random field for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 8, pp. 1814–1828, 2018.

[12] M. Pastorino, G. Moser, S. B. Serpico, and J. Zerubia, "Semantic segmentation of remote sensing images through fully convolutional neural networks and hierarchical probabilistic graphical models," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022.

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

[14] 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.

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

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

[17] 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.

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

[19] 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.

[20] 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.