Multiresolution Fusion and Classification of Hyperspectral and Panchromatic Remote Sensing Images

Martina Pastorino
DITEN Dept.
University of Genoa, Italy
martina.pastorino@edu.unige.it

Gabriele Moser
DITEN Dept.
University of Genoa, Italy
gabriele.moser@unige.it

Sebastiano B. Serpico
DITEN Dept.
University of Genoa, Italy
sebastiano.serpico@unige.it

Josiane Zerubia
Inria, Université Côte d'Azur, France
josiane.zerubia@inria.fr

Abstract

This paper proposes a supervised method for the joint classification and fusion of multiresolution panchromatic and hyperspectral data based on the combination of probabilistic graphical models (PGMs) and deep learning methods. The idea is to exploit the spatial and spectral information contained in panchromatic and hyperspectral images at different resolutions with the aim to generate a classification map at the spatial resolution of the panchromatic channel, while exploiting the richness of the spectral information provided by the hyperspectral channels. The proposed technique is based on deep learning, with FCN-type architectures, and PGMs, through the definition of a conditional random field (CRF) model approximating the behavior of the ideal fully connected CRF in a computationally tractable manner. The neural architecture aims to integrate hyperspectral and panchromatic data at the corresponding spatial resolution and generate posterior probability estimates, while the CRF incorporates information associated with not only local but also long-distance spatio-spectral relationships. The algorithm has been experimentally validated with PRISMA data from the Italian Space Agency with promising results.

1. Introduction

Hyperspectral image classification is a highly active and investigated area of remote sensing, for which effective techniques provide very accurate results. In particular, several approaches for supervised classification of hyperspectral data based on the analysis of spatial and spectral information have been proposed [1, 6, 8, 14, 16, 23, 24, 30, 38, 40, 41] based on different image processing techniques. The methods in [6, 14] develop architectures combining Markov random field (MRF) and dictionary learning or support vector machine (SVM), respectively, while [23] combines SVM and deep learning techniques. The architectures presented in [8, 16, 24, 30, 38, 40, 41] are also based on deep learning models, specifically [16, 40] involving the combination of neural networks and CRFs, [8, 41] graph convolutional network (GCN) models, [38] graph transformer networks and [31, 33, 34, 36, 37] convolutional neural networks integrating the analysis of spectral information to the more traditional analysis of spatial information guaranteed by two-dimensional convolutional layers.

Other approaches for the supervised classification of hyperspectral images include the use of preliminary classification methods, such as superpixels [2, 10, 29, 39], in some cases to overcome the scarcity of information available in input [39] typical of applications related to remote sensing.

The proposed method addresses the complex problem of the hyperspectral-panchromatic multiresolution supervised classification with the aim of generating a classification map at the resolution of the panchromatic channel, but exploiting the richness of the spectral information provided by the hyperspectral channels. This multiresolution problem is very promising from the point of view of the extraction of thematic information, precisely thanks to the opportunity to benefit from both spatial and spectral information captured by the two different sensors.

At the same time, unlike the above-mentioned classification of single-resolution hyperspectral data, panchromatic hyperspectral multiresolution classification is a very new problem, substantially not addressed in the literature so far. Panchromatic-multispectral classification techniques have been developed in [28] and [22] by means of MRF-based approaches, but with reference only to sensors characterized by a limited number of spectral channels and without

taking into account the potential and challenges of the hyperspectral data. Other panchromatic-multispectral classification methods present in the literature are based on feature fusion, deep learning models [17] and pyramid representations [20]. The latter have also been used in the conjunctive RGB-hyperspectral classification [35].

Hyperspectral pansharpening methods [18, 26] address the problem of multi-resolution fusion of hyperspectral and panchromatic channels but with a different objective, oriented to generate not a classification result but a simulation of the hyperspectral data on the pixel grid of the panchromatic channel. These techniques do not take into account the relationship between the hyperspectral and panchromatic measures and the spatial field of thematic class labels (e.g., land cover or use), do not use training samples for the classes, and determine, on the basis of concepts of signal processing and without a goal of supervised classification, a transformed image that is not optimised in order to correctly discriminate the classes.

The proposed multiresolution fusion technique for classification is based on methodological contributions related to deep learning and probabilistic graphical models (structured output learning). Great effectiveness in satellite image classification has been demonstrated by deep neural networks, typically based on the interconnection of multiple CNN subnetworks. The latter are characterized by an intrinsically multiresolution structure, thanks to the presence of pooling layers that perform spatial subsampling processes, and convolutional layers, whose spatial windows determine the extraction of levels of representation at distinct spatial scales (representation learning) [9]. This structure typical of two-dimensional CNNs on pixel networks then naturally extends to datasets associated with topologies of different size, as well as the three-dimensional datacube associated with a hyperspectral image. Such extensions make use of 3D (3D-CNN) or 1D convolutions along the channel set (equivalent to the wavelength axis; 1D-CNN) [25].

In this context, the proposed method is based, firstly, on CNN-type architectures (including also fully convolutional networks, FCN [19]) dedicated to the estimation, on the panchromatic network, of the posterior probabilities of thematic classes, given both hyperspectral and panchromatic observations. To this end, the neural architecture aims to integrate the hyperspectral and panchromatic data at the corresponding spatial resolution levels and generate estimates of the posterior probabilities. To effectively train such deep networks, a very large number of precisely annotated training samples is usually needed. This condition is easily verified in the application to benchmark data made available for international scientific contests (see, for example, [5]), but it often turns out to be a critical restriction in real-world applications, in which the available training sets are usually composed of a relatively limited number of samples or spatially disjoint regions assigned to the classes [21]. More detailed information on the spatial structures in the scene (objects and regions) is therefore often absent from the training set. This negatively influences the performance of a neural learning model trained on such data and constitutes an operational limitation. The technique proposed here also aims to address, in the context of multiresolution panchromatic-hyperspectral classification, this criticality related to the requirements in terms of quantity and quality of the training set. To this end, the proposed method integrates the approach based on deep neural networks with a probabilistic graphical model, namely a conditional random field (CRF). The technique extends the previous method in [21], which focused on the classification of three-channel optical aerial images with a spatial resolution of a few centimeters, to the case of the classification of multiresolution panchromatic-hyperspectral satellite images. The key idea consists in mitigating the impact of the training set insufficiencies by incorporating, with an innovative CRF model, information associated with not only local but also long-range spatial-spectral relations. It is assumed that the neural network is trained with the available training data, albeit limited in quantity and/or quality. Ideally, an effective approach to compensate for the impact of such a training set on accuracy would be to use a fully connected CRF, i.e. a PGM that models the interactions between all pairs of pixels in the considered grid — a model whose computational complexity would be intractable. A traditional MRF or CRF model characterizes relationships on a local basis, i.e. in terms of a system of neighborhoods [11, 32].

The CRF model proposed here approximates in a computationally tractable way the behavior of the ideal fully connected CRF, introducing a graph that includes not only the local relations defined by the neighborhood scheme but also relations between a set of additional virtual nodes aimed at representing long-distance dependencies. Such virtual nodes are defined on the basis of a clustering result of the computed activations of the neural network in the set of all its hidden layers. On the composite graph, which extends to the pixel grid and the resulting set of virtual nodes, a CRF is defined that characterizes the interactions (i) between neighboring pixels, (ii) between virtual nodes interconnected with a full mesh, and (iii) between pixels and virtual nodes interconnected with a dedicated neighborhood scheme. The application of state-of-the-art graph-based minimum energy algorithms, involving graph cut [12], ensures convergence to solutions characterized by strong optimality conditions. Methodologically, the proposed technique allows to capture and integrate into a probabilistic decision process the information extracted from the network at various spatial scales, taking advantage of both the great potential of deep neural networks in learning from high-dimensional datasets information at different levels of

abstraction and the ability of probabilistic graphical models to optimize classification results, explicitly formalizing the dependencies associated with the input data.

2. Methodology

2.1. Overview of the proposed methodology

The proposed methodology integrates multiresolution and multispectral panchromatic and hyperspectral image data for semantic segmentation by leveraging a novel spatial-spectral neural network architecture, multiresolution fusion, clustering, and a cluster fully connected CRF (Cl-FC-CRF, see Fig. 1). The approach focuses on the complementary strengths of high spatial resolution from panchromatic data and rich spectral resolution from hyperspectral data.

The problem involves image data of the same scene acquired at multiple resolutions and with different spectral information. In particular, the focus is on panchromatic, $X_{PAN} \in \mathbb{R}^{1\times H\times W}$ and hyperspectral, $X_{HYS} \in \mathbb{R}^{C_{HYS}\times \frac{H}{r}\times \frac{W}{r}}$ acquisitions, with r the spatial resolution ratio.

The multiresolution fusion is firstly addressed by the proposed neural network, which employs a two-branch encoder. One branch processes the high spatial resolution data through convolutional and max-pooling layers, while the other processes high spectral resolution data with 3D pointwise convolutions and spectral max pooling, compressing spectral information into spatially meaningful features. Features from both branches are fused in a bottleneck layer, which aligns their resolutions for subsequent decoding.

Afterwards, the panchromatic data is concatenated with a reduced set of the features extracted by the hyperspectral modeling branch of the neural network and of the features extracted by the decoder. These features are upsampled to match the panchromatic resolution, resulting in a unified tensor that preserves both spatial and spectral information. Clustering is applied to this tensor using k-means, where clusters represent spatio-spectral similarity. This step ensures connections between similar pixels across the image, independently of their spatial proximity.

To further refine segmentation, a CRF approximating full connectivity is employed. The CRF energy function incorporates unary potentials derived from network predictions and pairwise potentials to enforce spatial smoothness and label consistency. These potentials connect pixels locally, clusters globally, and pixels to their nearest clusters, capturing hierarchical relationships.

Overall, the methodology combines multiresolution data fusion through neural network-based feature extraction, clustering, and CRF-based long-range spatial dependencies modeling, resulting in an effective framework for semantic segmentation of panchromatic and hyperspectral data.

2.2. Spatial-spectral neural architecture

The employed neural network architecture belongs to the family of fully convolutional networks (FCNs) [19]. In this case, the architecture is designed for a semantic segmentation task that integrates two different multiresolution data: panchromatic and hyperspectral channels.

The encoder is divided into two separate branches which process panchromatic and hyperspectral information separately. The first branch processes the high-resolution panchromatic data X_{PAN}. The input passes through a series of convolution and downsampling max pooling layers. The second branch processes the coarser spatial resolution hyperspectral data X_{HYS} applying 3D pointwise convolutions followed by max pooling along the spectral dimension. We emphasize the role of 3D convolutions, which are aimed at extracting a meaningful representation from the hyperspectral channels, to benefit from its rich spectral content. In particular, this processing operation compresses the spectral information while retaining spatial features, effectively condensing multi-channel spectral data.

The features extracted from both channels are fused in a bottleneck layer, which learns the combined representations of spatial and spectral data. In order to have a match between the two original spatial resolutions, the sequence of pooling layers have an overall size which is a multiple of the original resolution ratio, $\nu \cdot r$ with $\nu \in \mathbb{N}$. The bottleneck consists of two convolutional blocks with batch normalizations and rectified linear unit (ReLU) activations.

The decoder reconstructs the segmentation map by progressively upscaling the features. Upsampling is performed using either bilinear interpolation or transposed convolutions, followed by convolutional layers which refine the features. The final segmentation map is generated at the spatial resolution of the panchromatic image. The network includes dropout layers for regularization, preventing overfitting.

2.3. Multiresolution fusion and clustering

The aforementioned neural architecture leverages the multiresolution information at its native resolution, producing feature maps at multiple resolutions which depend either on the original fine resolution panchromatic channel of the image, on the coarse resolution hyperspectral channels, or on their combination (in the architecture decoder). In the proposed method, a multiresolution tensor is constructed starting from the original panchromatic channel with the addition of upsampled results of a feature reduction (through principal component analysis, PCA) over the network activations computed (i) over the hyperspectral channels of the input image and (ii) over the combination of the processed panchromatic and hyperspectral channels.

$$x_i = x_{PAN,i} \oplus f_{HYS,i} \oplus f_{DEC,i} \tag{1}$$

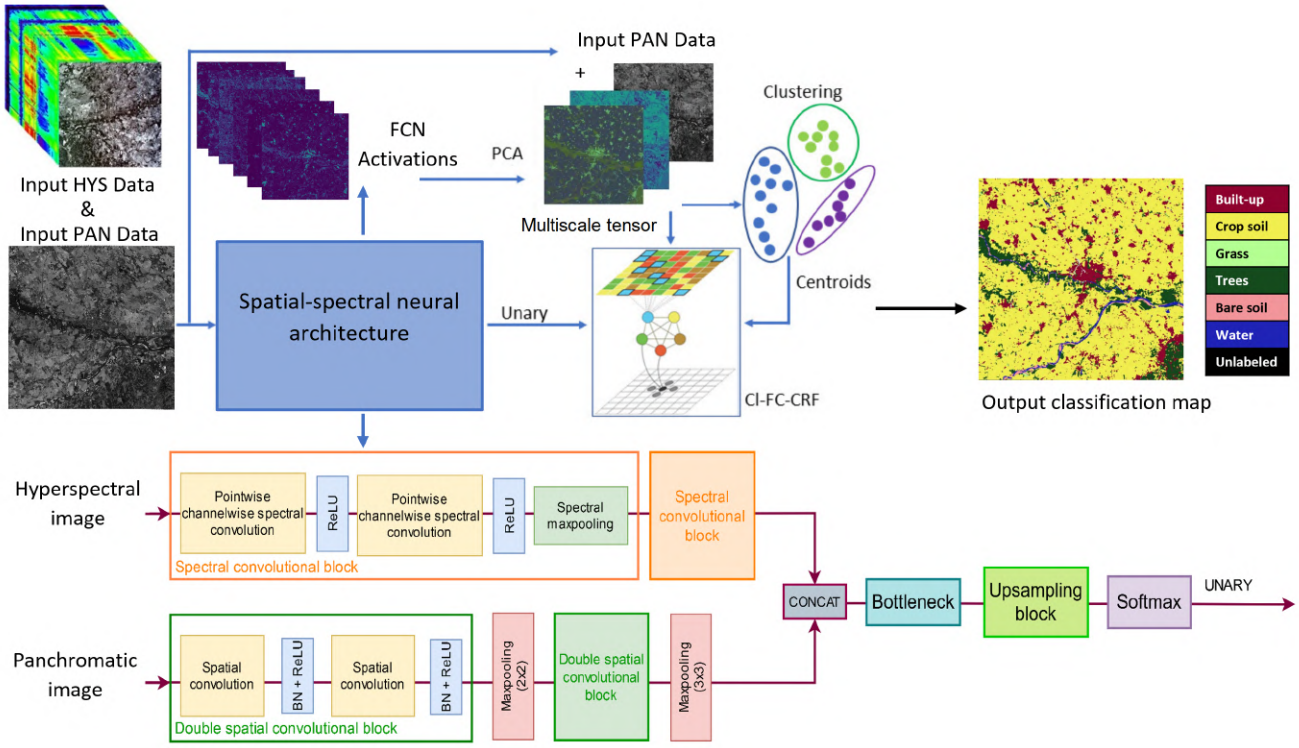

Figure 1. Overall architecture of the proposed methodology for multiresolution panchromatic-hyperspectral image fusion and classification. The CRF connects each pixel (black circle) with its neighbors (gray circles) in the image and with the h most similar clusters (colored circles), which are fully connected (middle).

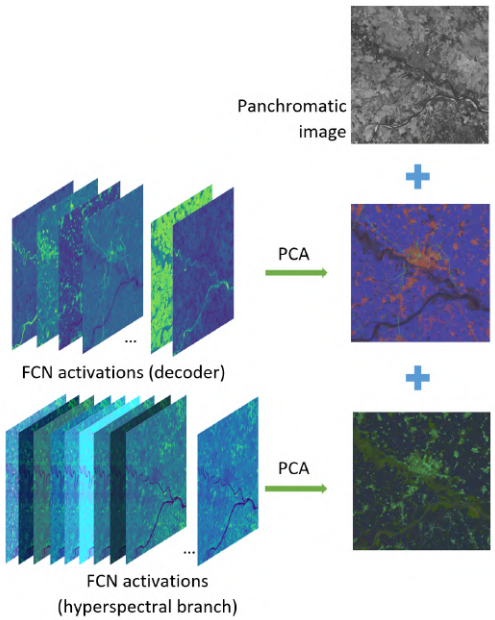

Figure 2. Multiscale tensor containing panchromatic and hyperspectral data.

with x_i the overall feature vector of pixel $i \in \mathcal{I}$, $x_{PAN,i}$, $f_{HYS,i}$, and $f_{DEC,i}$ are respectively the panchromatic acquisition, the feature vector of the second channel-wise convolutional layer of the neural architecture and of the first convolutional block of the decoder (with spatial resolution 2 times coarser than the original panchromatic image), in the same location i, $\oplus$ is the concatenation operator, and $\mathcal{I}$ is the pixel lattice at the panchromatic scale. The rationale is to model the spatial and spectral information contained in the panchromatic and hyperspectral data. Since the intermediate activations of the neural network applied on the hyperspectral data have a significant number of channels – while still being less than the number of original channels of the hyperspectral image – feature dimensionality reduction is applied. In particular, PCA is applied to the feature maps at coarse resolution, and only a number p of principal components is kept as in [21]. The result of the PCA is upsampled at the panchromatic spatial resolution through bilinear interpolation. Likewise for the feature maps derived from the decoder.

k-means is run on the obtained multiscale tensor (see Fig. 2). On the one hand, we do so to benefit, within the clustering stage, from the spatio-spectral information contained in the multisensory acquisitions. On the other hand,

the k-means partition joins similar pixels – a similarity influenced by both the panchromatic and the hyperspectral channels – all over the image in the same cluster, thus allowing connections through points at any distance on the image itself. Consider $\mathcal{C}$ and Ω the set of clusters and labels, respectively. The cluster feature vector $x_c \in \mathbb{R}^n$ is defined as the centroid of cluster $c \in \mathcal{C}$:

$$x_c = \frac{1}{|c|} \sum_{i \in c} x_i \tag{2}$$

with $|c|$ the numbers of pixels in cluster c, and the label y_c is given by the maximum of the posterior probability computed by the neural network averaged over cluster c.

2.4. Cluster fully connected CRF (Cl-FC-CRF)

The developed CRF model is an extension, aimed at the multiresolution fusion and classification of panchromatic and hyperspectral images, of the methodology proposed in [21] for the case of single-resolution multispectral imagery. Let us consider a CRF model with up to pairwose nonzero clique potentials, i.e., models considering at most the interactions between pairs of pixels. In this case, the energy is expressed as follows [15]:

$$\mathcal{U}(\mathcal{Y}|\mathcal{X}) = \sum_{i \in \mathcal{I}} D_i(y_i|\mathcal{X}) + \sum_{\substack{j \in \partial i \\ i \in \mathcal{I}}} V_{ij}(y_i, y_j|\mathcal{X}), \tag{3}$$

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

The proposed CRF model approximates a fully connected behavior through a computationally feasible solution, which is determined by the clustering partition. Given the pixel lattice $\mathcal{I}$, the feature vector $x_i \in \mathbb{R}^d$ of pixel $i \in \mathcal{I}$ is composed of the multiscale tensor described in the previous section, and $y_i \in \Omega$ is the associated label. The random field of observations and labels are $\mathcal{X} = \{x_i, x_c\}_{i \in \mathcal{I}, c \in \mathcal{C}}$ and $\mathcal{Y} = \{y_i, y_c\}_{i \in \mathcal{I}, c \in \mathcal{C}}$, respectively. The total energy of the proposed CRF is [21]:

$$\begin{aligned}\mathcal{U}(\mathcal{Y}|\mathcal{X}) = &\sum_{i \in \mathcal{I}} D_i(y_i|x_i) + \lambda_{\mathcal{I}\mathcal{I}} \sum_{\substack{j \in \partial i \\ i \in \mathcal{I}}} V_{ij}(y_i, y_j|x_i, x_j) \\ &+ \gamma \sum_{c \in \mathcal{C}} D_c(y_c|x_c) + \lambda_{\mathcal{C}\mathcal{C}} \sum_{\substack{c,d \in \mathcal{C} \\ c \neq d}} V_{cd}(y_c, y_d|x_c, x_d) \\ &+ \lambda_{\mathcal{I}\mathcal{C}} \sum_{\substack{i \in \mathcal{I} \\ c \in \bar{\partial} i}} V_{ic}(y_i, y_c|x_i, x_c),\end{aligned} \tag{4}$$

D_i and D_c are unary potentials for the pixel and cluster layers, and are computed as the log-posterior pixelwise probability predicted by the neural network (the softmax) and its weighted version for each cluster and class, respectively. γ is a weight to balance the two terms.

The other terms represent the pairwise energy contributions favoring spatial smoothness. In particular, in the proposed model, pixels are connected locally using a first-order neighborhood system (represented by ∂i), while clusters are fully connected, and a pixel is connected to the clusters corresponding to the h nearest neighbors (h-NN) among the centroids (represented by $\bar{\partial} i$), including the cluster the pixel belongs to. V_{ij} enforces spatial smoothness between neighboring pixels, V_{cd} prompts similar clusters to be assigned to the same class, and V_{ic} is the pixel-cluster pairwise potential. The λ-terms are weights to tune the contribution of each term. The pairwise potentials are defined by a contrast-sensitive Potts potential [4] to favor consistency in the labeling while simultaneously weighting on the similarity among the corresponding features.

The energy function is minimized through the $\alpha - \beta$ swap graph cut method [7] which decomposes a multiclass inference problem in a sequence of binary ones. The converges to a local minimum with strong optimality properties [3, 4, 13].

3. Experimental Results

3.1. Dataset and experimental setup

The proposed technique was tested on two datasets acquired over Lombardy and Emilia Romagna, Italy, in mainly urban areas. Both the datasets consist of three PRISMA images collected in April, 2021 for Lombardy and July, 2022 for Emilia Romagna. The PRISMA acquisitions include a panchromatic image, with a 5-m spatial resolution, and a hyperspectral image with 234 spectral bands and 30-m spatial resolution.

The ground truth was derived with regional land use archives: the 2021 DUSAF[1] for the Lombardy area and a 2020 regional archive for the urban areas of Emilia Romagna[2]. The mapping classes aim to discriminate different vegetated areas (cultivated fields, low and high vegetation), water bodies and built-up areas (buildings and roads).

The obtained dataset was split into two disjoint subsets for training and validation of the proposed multisensor fusion architecture. In particular, the Lombardy dataset in-

[1] https://www.regione.lombardia.it/wps/portal/istituzionale/HP/DettaglioServizio/servizi-e-informazioni/Enti-e-Operatori/Territorio/sistema-informativo-territoriale-sit/uso-suolo-dusaf/uso-suolo-dusaf

[2] https://geoportale.regione.emilia-romagna.it/download/dati-e-prodotti-cartografici-preconfezionati/pianificazione-e-catasto/uso-del-suolo/2020-coperture-vettoriali-uso-del-suolo-di-dettaglio-edizione-2023/cartella-dei-dati

cludes three fully overlapping images of the urban area of Lombardy, while the Emilia Romagna dataset includes three partially overlapping images that show several urban centers. Consequently, for the first dataset the three images were split into small subimages corresponding to training and test sets, while for the second, two complete images were used for training and one for testing.

The unary potentials are obtained through an FCN classifier providing probability scores at the pixel level and the feature maps over the hyperspectral channels. Three different architectures were employed for experimental validation, all of the kind presented in Section 2.2: a standard U-Net [27] with a pointwise channelwise spectral convolution (U-Net_λ), a lightweight U-Net with fewer convolutional blocks and a pointwise channelwise spectral convolution (LU-Net_λ), and its version with two channelwise convolutions (LU-Net_$\lambda\lambda$). The FCNs are trained on an RTX2080Ti GPU. The overall amount of trainable parameters is shown in Tables 1-2. These backbones were selected for their robustness, stability, and encoder-decoder architecture, which allows to obtain outputs with the same size of the input [19, 27].

The proposed method involves several hyperparameters, which were experimentally set through trial and error. In particular, the weights in the energy function $\lambda_{\mathcal{II}}$, $\lambda_{\mathcal{CC}}$, $\lambda_{\mathcal{IC}}$, and γ were set to $2, 1, 1, \frac{|\mathcal{I}|}{k}$, respectively, with $|\mathcal{I}|$ the number of pixels in the image patch and k the number of cluster centroids. The standard deviation σ of the Gaussian kernel in the contrast sensitive Potts pairwise potential is the median Euclidean distance between all considered pairs of feature vectors. The number of principal components p is experimentally set to 3, according to the behavior of the corresponding eigenvalues. PCA is performed on the feature maps output of the hyperspectral modeling branch of the network before the upsample.

After computing the multiscale tensor, with the output of the network, the clustering is run on a subset of pixels, to maintain a low computational complexity. The number of clusters k is chosen empirically as 256, and a sensitivity analysis was performed (see Table 3). The CRF energy minimization is implemented by subdividing the image into patches of 600×600 pixels. This allows the proposed methodology to be adapted to larger scale images, with limited increase in computational burden. The number h of nearest centroids is set to 4 to capture long-range while keeping a rather low computational cost.

3.2. Results and discussion

The results obtained through the experimental validation of the method on the test set of the Lombardy dataset are shown in Fig. 3 (*first row*) and in Table 1. The accuracies are reported in terms of recall for each class, overall accuracy, and class-averaged recall, precision, and F1 score.

As can be seen from the results in Table 1, the proposed neural architectures effectively distinguish the six land cover classes identified in the ground truth data. Specifically, all three architectures achieve accurate results not only for the classification of majority classes such as built-up areas ("built-up") and vegetated areas ("crop soil" and "trees"), but also for the minority classes "bare soil" and "water."

The results are slightly less accurate for the classification of low-vegetation, non-cultivated areas ("grass"), primarily due to the overlap in feature space between the "trees" and "grass" classes. The "trees" class, being significantly more prevalent, is also characterized by some false positives.

Regarding the average accuracies, the U-Net_λ is the neural architecture that achieves the best results, thanks to its higher number of parameters (which also entails a greater computational time and high sensitivity to the amount of available training samples). In general, the average values of recall, precision, and F1 score are around 80%, reaching up to 90% for the U-Net_λ.

As shown in Fig. 3, the results obtained by the method closely reflect the ground truth, with no significant oversmoothing or evident false alarms. Notably, the proposed method demonstrates its ability to generate classification maps where the spatial boundaries between classes are well-defined and effectively identifies minority classes such as "bare soil," "water," and "grass."

As also highlighted in Table 1, the inclusion of the CRF and the consequent modeling of spatial-spectral relationships, both local and long-range, improves the accuracy of results across all the considered classes, leading to enhancements in average accuracies. The U-Net_λ remains the architecture capable of achieving the most accurate results, with an overall accuracy of 96% and an F1 score of 92%.

For the Emilia Romagna dataset, the results are shown in Fig. 3 (*second row*) and in Table 2. As with the previous dataset, the accuracies are reported in terms of recall for each class, overall accuracy, mean recall, precision, and F1 score.

In this case, the results are generally less accurate. This is primarily because the ground truth for this region is derived from a land-use archive with a lower spatial resolution compared to DUSAF (the archive used as ground truth for the Lombardy area). Consequently, the minority classes identified have even fewer available samples, leading to less accurate classification results compared to the previous dataset (with a maximum accuracy of 30%).

Another important point to note is that while for the Lombardy dataset both the PRISMA acquisitions and the DUSAF map are from 2021, for the Emilia Romagna dataset the PRISMA acquisition is of 2022 and the regional archive dates back to 2020. As a result, the nominal extent of the mapping classes in the ground truth may not perfectly

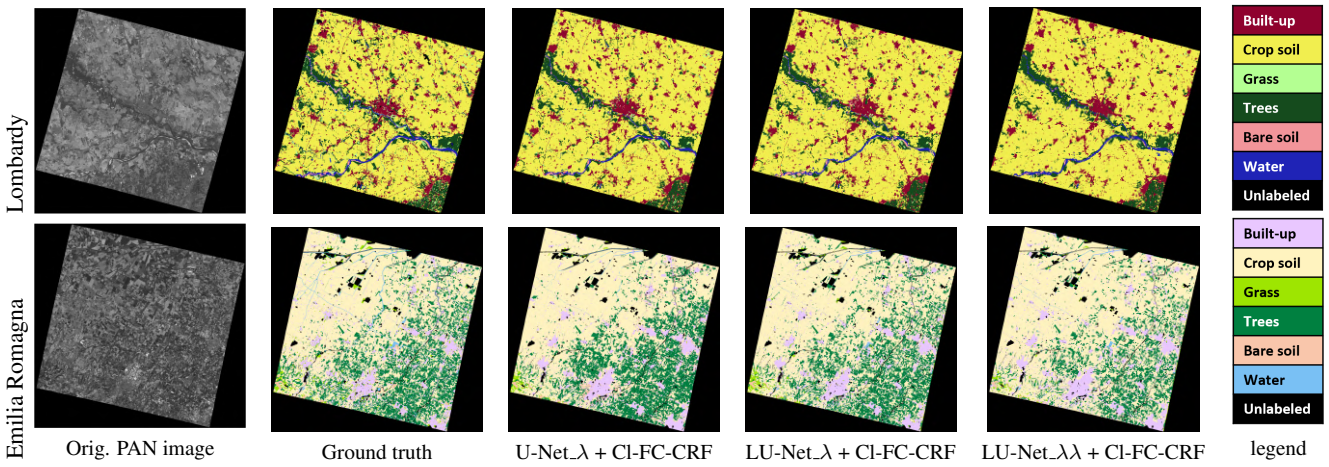

Figure 3. Test ground truths and classification maps for the test tiles in the Lombardy and Emilia Romagna datasets.

Table 1. Test-set results for the Lombardy dataset. Per-class values are recalls. Overall accuracy (OA), recall, precision, and F1 score are averaged over the classes.

Architecture	built-up	crop soil	tall veg.	grass	bare soil	water	OA	recall	prec.	F1 score	# train param.	train time
U-Net_λ	92.75	98.28	91.34	74.88	90.75	91.68	95.69	89.91	91.63	90.77	$17{\cdot}10^6$	2h
LU-Net_λ	88.72	97.10	83.77	43.72	87.02	76.71	91.99	79.51	84.73	82.03	$23{\cdot}10^4$	2h
LU-Net_$\lambda\lambda$	86.35	97.73	82.30	32.99	83.67	76.73	91.54	76.63	85.17	80.68	$38{\cdot}10^4$	2h
U-Net_λ + Cl-FC-CRF	**94.77**	**99.00**	**93.01**	**75.32**	**92.98**	**91.76**	**96.69**	**91.14**	**94.63**	**92.85**	$17{\cdot}10^6$	(2+6)h
LU-Net_λ + Cl-FC-CRF	89.81	98.29	85.94	43.72	88.85	78.37	93.26	80.83	86.51	83.57	$23{\cdot}10^4$	(2+6)h
LU-Net_$\lambda\lambda$ + Cl-FC-CRF	88.00	98.76	84.38	33.01	83.69	79.06	92.76	77.81	93.08	84.77	$38{\cdot}10^4$	(2+6)h

align with their actual extent in the PRISMA acquisitions, particularly for classes subject to temporal changes due to climate change or seasonal variations (e.g., the extent of low vegetation and water bodies).

Fig. 3 displays the classification maps obtained using the proposed method with the three different neural architectures. For this dataset, as well, the method accurately reflect the ground truth for the majority classes. The LU-Net_λ and LU-Net_$\lambda\lambda$ architectures demonstrate better discrimination of the "water" class but tend to underestimate the "bare soil" and "trees" classes, with the latter being particularly pronounced for LU-Net_λ. On the other hand, U-Net_$\lambda\lambda$ tends to overestimate the "trees" class, almost entirely suppressing the "grass" and "water" classes.

Given the lower accuracies of the underlying network, multiple experiments were conducted by varying the CRF parameters. The experiments reported in Table 3 for LU-Net_$\lambda\lambda$ focus on the number of k-means clusters k, which ranged in powers of 2 from 32 to 256. As highlighted in the table, the most accurate results are obtained when $k = 64$. The "trees," "grass," and "bare soil" classes exhibit better discrimination, albeit slightly at the expense of the "water" and "built-up" classes. Specifically, it can be observed that lower values of k correspond to higher precision, indicating a lower presence of false positives. The classification maps suggest the effectiveness of the proposed method in distinguishing the considered land cover classes, despite the aforementioned challenges. Notably, the maps do not exhibit any spatial oversmoothing, and the boundaries between classes are well-defined. However, the discrimination of minority classes remains insufficient.

Similar to the Lombardy dataset, the inclusion of the CRF, with its ability to model both local and long-range spatial-spectral relationships, resulted in improvements in the accuracy of all considered classes and in the average accuracies. The LU-Net_$\lambda\lambda$ achieved the most accurate results, with an overall accuracy of 89% and an F1 score of 65%.

4. Conclusion

This paper introduced a multiresolution fusion method for the joint classification – or semantic segmentation – of panchromatic and hyperspectral images. In particular, the proposed methodology is based on FCNs and a cluster level fully connected CRF to model the spatial and spectral information provided by the multiresolution and hyperspectral input data.

The methods were applied to the PRISMA product of the Italian Space Agency, in a case study of land cover mapping in mainly urban zones over Northern and Central Italy. The results of the experiments demonstrate the effectiveness of the proposed method in classifying panchromatic-

Table 2. Test-set results for the Emilia Romagna dataset. Per-class values are recalls. Overall accuracy (OA), recall, precision, and F1 score are averaged over the classes.

Architecture	built-up	crop soil	tall veg.	grass	bare soil	water	OA	recall	prec.	F1 score	# train param.	train time
U-Net_λ	81.30	81.97	78.07	8.37	26.51	27.53	77.54	50.63	67.91	58.00	$17{\cdot}10^6$	2h
LU-Net_λ	77.96	**97.09**	39.79	5.70	12.77	22.73	78.19	42.68	69.95	53.00	$23{\cdot}10^4$	2h
LU-Net_$\lambda\lambda$	86.44	91.82	62.34	15.17	20.62	**30.05**	81.16	51.08	68.03	58.34	$38{\cdot}10^4$	2h
U-Net_λ + Cl-FC-CRF	81.48	89.01	**83.06**	15.62	**34.89**	27.54	83.15	**55.27**	69.91	**61.74**	$17{\cdot}10^6$	(2+6)h
LU-Net_λ + Cl-FC-CRF	**87.19**	94.24	65.18	**23.95**	22.36	30.01	83.55	53.83	69.07	60.50	$23{\cdot}10^4$	(2+6)h
LU-Net_$\lambda\lambda$ + Cl-FC-CRF	87.03	94.89	65.07	22.57	22.46	**30.05**	**83.86**	53.68	**70.43**	60.92	$38{\cdot}10^4$	(2+6)h

Table 3. Test-set results for the Emilia Romagna dataset varying the number k of clusters.

Architecture	built-up	crop soil	tall veg.	grass	bare soil	water	OA	recall	precision	F1 score
LU-Net_$\lambda\lambda$ + Cl-FC-CRF ($k = 32$)	86.84	96.67	65.65	18.21	22.79	**30.05**	84.95	53.36	**74.28**	62.11
LU-Net_$\lambda\lambda$ + Cl-FC-CRF ($k = 64$)	80.70	**97.49**	**85.87**	**27.84**	**31.15**	22.74	**88.99**	**57.63**	73.62	**64.66**
LU-Net_$\lambda\lambda$ + Cl-FC-CRF ($k = 128$)	86.92	95.79	65.39	20.56	22.61	30.05	84.42	53.55	72.37	61.56
LU-Net_$\lambda\lambda$ + Cl-FC-CRF ($k = 256$)	**87.03**	94.89	65.07	22.57	22.46	30.05	83.86	53.68	70.43	60.92

hyperspectral imagery, with particular emphasis on its ability to model both spatial and spectral relationships. The inclusion of the fully connected CRFs further improves classification accuracy by refining spatial boundaries and incorporating long-range dependencies, particularly for minority classes.

While the proposed architectures, especially the U-Net_λ and LU-Net_$\lambda\lambda$ formulations, achieve high overall accuracy and F1 scores, challenges remain in discriminating minority classes in datasets with limited training samples or outdated ground truth data. These findings underline the importance of aligning data sources and employing models that can handle temporal inconsistencies and class imbalances.

Future work will focus on improving the robustness of the framework against such inconsistencies, integrating more advanced contextual models, and exploring semi-supervised techniques to address sample scarcity.

5. Acknowledgments

This work was partially supported by the Italian Space Agency (ASI) within the Framework of the Project under Grant PRISMA-Learn-ASI no. 2022-12-U.0. PRISMA Products, ©ASI, delivered under a license to use by ASI.

References

[1] Muhammad Ahmad, Sidrah Shabbir, Swalpa Kumar Roy, Danfeng Hong, Xin Wu, Jing Yao, Adil Mehmood Khan, Manuel Mazzara, Salvatore Distefano, and Jocelyn Chanussot. Hyperspectral image classification—traditional to deep models: A survey for future prospects. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:968–999, 2022. 1

[2] Jing Bai, Wei Shi, Zhu Xiao, Amelia C. Regan, Talal Ahmed Ali Ali, Yongdong Zhu, Rui Zhang, and Licheng Jiao. Hyperspectral image classification based on superpixel feature subdivision and adaptive graph structure. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–15, 2022. 1

[3] Yuri Boykov and Vladimir Kolmogorov. An experimental comparison of min-cut/max- flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1124–1137, 2004. 5

[4] Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, 2001. 5

[5] Manuel Campos-Taberner, Adriana Romero-Soriano, Carlo Gatta, Gustau Camps-Valls, Adrien Lagrange, Bertrand Le Saux, Anne Beaupère, Alexandre Boulch, Adrien Chan-Hon-Tong, Stephane Herbin, Hicham Randrianarivo, Marin Ferecatu, Michal Shimoni, Gabriele Moser, and Devis Tuia. Processing of extremely high-resolution LiDAR and RGB data: Outcome of the 2015 IEEE GRSS data fusion contest–part A: 2-D contest. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 9(12):5547–5559, 2016. 2

[6] Xianghai Cao, Xiaozhen Wang, Da Wang, Jing Zhao, and Licheng Jiao. Spectral–spatial hyperspectral image classification using cascaded Markov random fields. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(12):4861–4872, 2019. 1

[7] Brian Fulkerson, Andrea Vedaldi, and Stefano Soatto. Class segmentation and object localization with superpixel neighborhoods. In *2009 IEEE 12th International Conference on Computer Vision*, pages 670–677, 2009. 5

[8] Zhi Gong, Lei Tong, Jun Zhou, Bin Qian, Lijuan Duan, and Chuangbai Xiao. Superpixel spectral–spatial feature fusion graph convolution network for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2022. 1

[9] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. USA: MIT Press, Boston, Massachussetts, 2016. 2

[10] Sen Jia, Xianglong Deng, Meng Xu, Jun Zhou, and Xiuping Jia. Superpixel-level weighted label propagation for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 58(7):5077–5091, 2020. 1

[11] Zoltan Kato and Josiane Zerubia. Markov random fields in image segmentation. *Found. Trends Signal Process.*, 5(1-2):1–155, 2012. 2

[12] Vladimir Kolmogorov and Carsten Rother. Minimizing nonsubmodular functions with graph cuts-a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(7):1274–1279, 2007. 2

[13] Vladimir Kolmogorov and Ramin Zabin. What energy functions can be minimized via graph cuts? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):147–159, 2004. 5

[14] Elham Kordi Ghasrodashti, Mohammad Sadegh Helfroush, and Habibollah Danyali. Sparse-based classification of hyperspectral images using extended hidden Markov random fields. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(11):4101–4112, 2018. 1

[15] Stan Z. Li. *Markov random field modeling in image analysis*. Springer, 3rd edition, 2009. 5

[16] Yi Liang, Xin Zhao, Alan J. X. Guo, and Fei Zhu. Hyperspectral image classification with deep metric learning and conditional random field. *IEEE Geoscience and Remote Sensing Letters*, 17(6):1042–1046, 2020. 1

[17] Sicong Liu, Hui Zhao, Qian Du, Lorenzo Bruzzone, Alim Samat, and Xiaohua Tong. Novel cross-resolution feature-level fusion for joint classification of multispectral and panchromatic remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022. 2

[18] Laetitia Loncan, Luis B. de Almeida, Jose M. Bioucas-Dias, Xavier Briottet, Jocelyn Chanussot, Nicolas Dobigeon, Sophie Fabre, Wenzhi Liao, Giorgio A. Licciardi, Miguel Simões, Jean-Yves Tourneret, Miguel Angel Veganzones, Gemine Vivone, Qi Wei, and Naoto Yokoya. Hyperspectral pansharpening: A review. *IEEE Geoscience and Remote Sensing Magazine*, 3(3):27–46, 2015. 2

[19] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 3431–3440, 2015. 2, 3, 6

[20] Wenping Ma, Jianchao Shen, Hao Zhu, Jun Zhang, Jiliang Zhao, Biao Hou, and Licheng Jiao. A novel adaptive hybrid fusion network for multiresolution remote sensing images classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–17, 2022. 2

[21] Luca Maggiolo, Diego Marcos, Gabriele Moser, Sebastiano B. Serpico, and Devis Tuia. A semisupervised CRF model for CNN-based semantic segmentation with sparse ground truth. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–15, 2022. 2, 4, 5

[22] Gabriele Moser, Andrea De Giorgi, and Sebastiano Bruno Serpico. Multiresolution supervised classification of panchromatic and multispectral images by Markov random fields and graph cuts. *IEEE Transactions on Geoscience and Remote Sensing*, 54(9):5054–5070, 2016. 1

[23] Onuwa Okwuashi and Christopher E. Ndehedehe. Deep support vector machine for hyperspectral image classification. *Pattern Recognition*, 103:107298, 2020. 1

[24] Bin Pan, Xia Xu, Zhenwei Shi, Ning Zhang, Huanlin Luo, and Xianchao Lan. Dssnet: A simple dilated semantic segmentation network for hyperspectral imagery classification. *IEEE Geoscience and Remote Sensing Letters*, 17(11):1968–1972, 2020. 1

[25] Charlotte Pelletier, Geoffrey I. Webb, and François Petitjean. Temporal convolutional neural network for the classification of satellite image time series. *Remote Sensing*, 11(5), 2019. 2

[26] Jiahui Qu, Yanzi Shi, Weiying Xie, Yunsong Li, Xianyun Wu, and Qian Du. MSSL: Hyperspectral and panchromatic images fusion via multiresolution spatial–spectral feature learning networks. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022. 2

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Med. Image Comput. Comput. Ass. Interv.*, volume 9351 of *LNCS*, pages 234–241, 2015. Springer. 6

[28] Geir Storvik, Roger Fjortoft, and Anne H. Schistad Solberg. A Bayesian approach to classification of multiresolution remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 43(3):539–547, 2005. 1

[29] Subhashree Subudhi, Ram Narayan Patro, Pradyut Kumar Biswal, and Fabio Dell'Acqua. A survey on superpixel segmentation as a preprocessing step in hyperspectral image analysis. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:5015–5035, 2021. 1

[30] Hao Sun, Xiangtao Zheng, and Xiaoqiang Lu. A supervised segmentation network for hyperspectral image classification. *IEEE Transactions on Image Processing*, 30:2810–2825, 2021. 1

[31] Yifan Sun, Bing Liu, Xuchu Yu, Anzhu Yu, Zhixiang Xue, and Kuiliang Gao. Resolution reconstruction classification: fully octave convolution network with pyramid attention mechanism for hyperspectral image classification. *International Journal of Remote Sensing*, 43(6):2076–2105, 2022. 1

[32] Charles Sutton and Andrew McCallum. An introduction to conditional random fields. *Found. Trends Mach. Learn.*, 4(4):267–373, 2011. 2

[33] Bing Tu, Wangquan He, Qianming Li, Yishu Peng, and Siyuan Chen. Fully convolutional network-based nonlocal-dependent learning for hyperspectral image classification. *IEEE Transactions on Instrumentation and Measurement*, 71:1–14, 2022. 1

[34] Xue Wang, Kun Tan, Peijun Du, Chen Pan, and Jianwei Ding. A unified multiscale learning framework for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–19, 2022. 1

[35] Qingsong Xu, Xin Yuan, Chaojun Ouyang, and Yue Zeng. Attention-based pyramid network for segmentation and classification of high-resolution and hyperspectral remote sensing images. *Remote Sensing*, 12(21), 2020. 2

[36] Dabing Yu, Qingwu Li, Xiaolin Wang, Chang Xu, and Yaqin Zhou. A cross-level spectral–spatial joint encode learning framework for imbalanced hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–17, 2022. 1

[37] Hongqi Zhang, Xudong Sun, Yuan Zhu, Fengqiang Xu, and Xianping Fu. A global-local spectral weight network based on attention for hyperspectral band selection. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022. 1

[38] Xiaofeng Zhao, Jiahui Niu, Chuntong Liu, Yao Ding, and Danfeng Hong. Hyperspectral image classification based on graph transformer network and graph attention mechanism. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022. 1

[39] Chengyong Zheng, Ningning Wang, and Jing Cui. Hyperspectral image classification with small training sample size using superpixel-guided training sample enlargement. *IEEE Transactions on Geoscience and Remote Sensing*, 57(10):7307–7316, 2019. 1

[40] Zilong Zhong, Jonathan Li, David A. Clausi, and Alexander Wong. Generative adversarial networks and conditional random fields for hyperspectral image classification. *IEEE Transactions on Cybernetics*, 50(7):3318–3329, 2020. 1

[41] Wenxiang Zhu, Chunhui Zhao, Shou Feng, and Boao Qin. Multiscale short and long range graph convolutional network for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–15, 2022. 1