Probabilistic Fusion Framework Combining CNNs and Graphical Models for Multiresolution Satellite and UAV Image Classification

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

Probabilistic Fusion Framework Combining CNNs and Graphical Models for Multiresolution Satellite and UAV Image Classification

Martina Pastorino[1,2], Gabriele Moser[1], Fabien Guerra[3], 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
[3] INRAE, RECOVER, Aix-Marseille University, Aix-en-Provence, France

Abstract. Image classification – or semantic segmentation – from input multiresolution imagery is a demanding task. In particular, when dealing with images of the same scene collected at the same time by very different acquisition systems, for example multispectral sensors onboard satellites and unmanned aerial vehicles (UAVs), the difference between the involved spatial resolutions can be very large and multiresolution information fusion is particularly challenging. This work proposes two novel multiresolution fusion approaches, based on deep convolutional networks, Bayesian modeling, and probabilistic graphical models, addressing the challenging case of input imagery with very diverse spatial resolutions. The first method aims to fuse the multimodal multiresolution imagery via a posterior probability decision fusion framework, after computing posteriors on the multiresolution data separately through deep neural networks or decision tree ensembles. The optimization of the parameters of the model is fully automated by also developing an approximate formulation of the expectation maximization (EM) algorithm. The second method aims to perform the fusion of the multimodal multiresolution information through a pyramidal tree structure, where the imagery can be inserted, modeled, and analyzed at its native resolutions. The application is to the semantic segmentation of areas affected by wildfires for burnt area mapping and management. The experimental validation is conducted with UAV and satellite data of the area of Marseille, France. The code is available at `https://github.com/Ayana-Inria/BAS_UAV_satellite_fusion`.

Keywords: graphical models · deep learning · probabilistic fusion · multiresolution imagery · semantic segmentation · wildfires · UAVs.

1 Introduction

In the framework of pattern recognition, a semantic segmentation problem, whose goal is to assign a class label to each individual pixel in an image, can

be formalized as a supervised image-classification problem [33]. Within semantic segmentation tasks, the use of multimodal data has been shown to favor accuracy and spatial precision of the classification results [10]. From a computer vision perspective, the development of processing methods that can benefit from multimodal information (e.g., synoptic and detailed views from multiresolution data, different band information from multisensor imagery) and take advantage of the complementary information therein contained presents huge potentials.

Thanks to the advent of deep learning, the performances of semantic segmentation algorithms have significantly improved. However, there are still some challenges. For example, when dealing with remote sensing images, one of the main issues is the variability of features within the same category in the image, leading to confusion in segmentation. Moreover, the availability of training data is a key requirement for deep learning architectures, not always feasible for computer vision applications related to remote sensing. To address these challenges, one way is to leverage contextual and multiscale information for accurate segmentation [29, 30].

For example, focusing on land-cover mapping applications, on the one hand there are satellite imaging sensors, which provide an efficient and large-scale coverage of the Earth surface, thanks to their wide range and short revisit time. Optical satellite imagery with spatial resolution as fine as 10 m is made available by space missions with open data policies (e.g., the ESA Copernicus program). However, optical satellite sensors are sensitive to weather conditions and Sun illumination. On the other hand, in recent years, unmanned aerial vehicles (UAV) – or drones – have also sparked a lot of interest thanks to their high flexibility, low-cost, and ability to cover wide areas during the day or night [43]. UAV monitoring is undertaken at low-to-medium altitudes, thus effectively avoiding the cloud interference, and allowing for very high spatial resolution up to few centimeters. However, the imagery captured by UAVs is typically characterized by a small area coverage, irregular contours, susceptibility to forest cover, making land-cover mapping from UAV imagery a challenging task [1].

The joint availability of satellite and UAV acquisitions of the same geographical zones, with their complementary features, presents a huge potential for semantic segmentation applications and, simultaneously, a big challenge for the development of a method capable to fully take advantage of this multimodal information. The resulting multiresolution fusion task is quite extreme, and currently under-exploited, since the resolution ratio between the input image sources is of the order of the hundreds – a situation that is normally not addressed by traditional multiresolution schemes [6, 13, 29–32, 35–39, 42].

In this paper, two approaches based on deep learning, Bayesian fusion, and probabilistic graphical fusion are proposed for the semantic segmentation of multiresolution imagery with a huge ratio between resolutions. The focus is on binary classification problems, which have many applications in natural disaster management, such as the detection of areas affected by floods, wildfires, or earthquakes [24], the mapping of urban areas and human settlements [8], of snow covers [22], and cloud masking [23]. The first method proposes a pixelwise prob-

abilistic fusion of the multiresolution information after computing the posterior probabilities with separate classifiers – neural networks and decision tree ensembles – on the multimodal images separately. The parameters of the method are automatically optimized by developing a case-specific formulation of the EM algorithm, based on a pseudo-likelihood-type approximation. The second considers multiresolution fusion in a pyramidal tree graph topology through the marginal posterior mode (MPM) criterion, an extension to the case of great spatial resolution ratio of the approach proposed in [29, 30] for multimodal and multiresolution images.

The main novel contributions of this paper are twofold: (i) the development of two semantic segmentation methods for input multiresolution imagery with great mismatch in spatial resolution; (ii) the combination, within the two novel methods, of deep learning, stochastic modeling, decision fusion, and an EM-based automatic parameter optimization.

2 Related Work

Here, we briefly review the literature on semantic segmentation from input multiresolution imagery. Models for multimodal data, in particular multiscale and multiresolution methods, are gaining importance in order to face the requirements of several applications, for example remote sensing [10] and medical image processing [32, 37]. The idea is to jointly use multiple images associated with distinct spatial resolutions to benefit from their complementary perspectives.

Wavelet-based methods [25] are often employed to perform multiresolution image processing. In [13] an image segmentation method for human face detection based on multiresolution wavelet transforms and watershed segmentation algorithm is presented. In [6] a wavelet-based multiresolution pyramid applied to multitemporal or multisensor satellite data is combined with a stochastic gradient based on two similarity measures, correlation and mutual information. In [42] several wavelet pyramids aimed at performing invariant feature extraction and accelerating image fusion through multiple spatial resolutions are evaluated.

Deep learning methods are state-of-the-art techniques for computer vision tasks [26]. Fully convolutional networks (FCNs) structurally involve several multiscale processing stages, through their encoder-decoder architecture and their convolutional and pooling layers. In [37], a semantic segmentation model for histopathology whole-slide images, which combines multiresolution context and details via multiple branches of encoder-decoder neural networks, is proposed. A multi-scale representation learning network integrating CNNs and Transformers was proposed in [20] to exploit multi-scale local detailed feature and global contextual information for the segmentation of lesions in lung CT images. The joint potential of CNNs and Transformers for the analysis of local and multiscale information was explored in [35], as well, for the semantic segmentation of urban remote sensing images.

The literature on multiresolution fusion in remote sensing is vast and dates back a few decades [33], with approaches rooted in several methodological areas, such as statistical pattern recognition [15, 31], neural networks [17, 39], decision fusion [40], kernel-based approaches [36], and Markov random fields [3, 38].

Furthermore, with the advent and diffusion of UAV platforms, images with extremely high spatial resolution have become available at a relatively low cost [19]. UAVs are often equipped with simple, lightweight sensors, such as RGB cameras [36] that capture small portions of land. In [36] classification of a high spatial resolution RGB image and a lower spatial resolution hyperspectral image of the same scene is addressed. Contextual information is obtained from the RGB image through color attribute profiles, and spectral information is extracted from the hyperspectral image; a composite decision fusion strategy exploiting kernel-based techniques is proposed.

3 Methodology

The aim of the proposed techniques is to perform the fusion of multiresolution imagery – with big mismatch in spatial resolution – for binary semantic segmentation tasks without the need of resampling techniques. The two proposed approaches integrate stochastic modeling, decision fusion, deep learning, ensemble learning, and the EM algorithm. The overall diagrams of the proposed approaches are shown in Fig. 1.

In this framework, neural networks and decision tree ensembles act as non-parametric estimators of posterior probability, thus allowing multimodal data fusion. Specifically, fully convolutional networks (FCN) [21] are employed to estimate the posterior probabilities on the image with the finest spatial resolution (i.e., the UAV acquisition), and random forest (RF) [4] on the image with the coarsest spatial resolution (i.e., the resolution of the satellite acquisition). Indeed, the ratio between the two spatial resolutions is very high. Therefore, even though the pixel lattice of the UAV image can be quite large, the corresponding satellite image is expected to be composed of relatively few pixels, hence generally unfit for deep learning methods. That is the rationale of the use of a decision tree ensemble to predict pixelwise posteriors on the pixel grid of the satellite image.

In general, the proposed approaches can be combined with an arbitrary FCN model. In particular, U-Net [34] is used as the reference model on the UAV lattice, since it is widely employed and has been found to be effective in applications to remote sensing imagery. Likewise, for the ensemble learning technique, RF was selected for its well-known computational efficiency and flexibility to model heterogeneous data.

After the computation of the pixelwise posterior probabilities of the multiresolution image pixel lattices by the FCN and RF, the first proposed method (see Section 3.1 and Fig. 1(a)) performs a pixelwise probabilistic fusion to obtain the final classification results exploiting the information carried by the UAV and the satellite imagery. A formulation of the EM algorithm allows to automatically estimate the transition probabilities that determine the chance of having a certain label at the finer spatial resolution given the label at the coarser spatial resolution.

For the second method (see Section 3.3 and Fig. 1(b)), on the other hand, the pixelwise posterior probabilities computed on the multiresolution image lattices at the native resolutions are fused through a hierarchical probabilistic graphi-

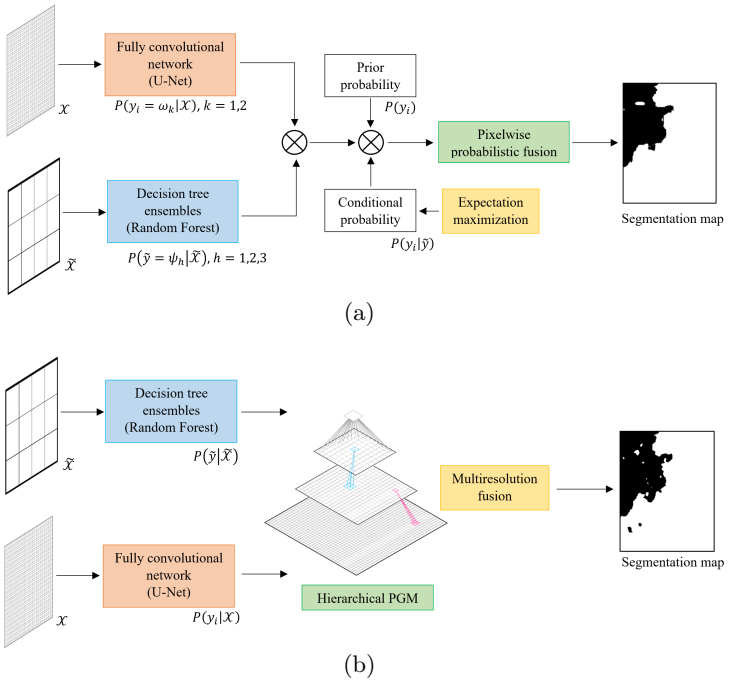


Fig. 1: Architecture of the two proposed methods: (a) pixelwise probabilistic fusion and (b) multiresolution fusion through hierarchical PGM.

cal framework based on a hierarchical Markov random field, which models the multiresolution transition probabilities over a quadtree structure.

3.1 Pixelwise probabilistic fusion

The first proposed method introduces a pixelwise probabilistic decision fusion framework to combine the information contained at different resolutions [2, 5]. With the assumption that the two images are well registered, let us consider a patch of size $D \times D$ of the image at the finer resolution, with size corresponding exactly to one pixel in the lattice associated with the image at the coarser spatial resolution. Accordingly, D represents the resolution ratio associated with the input multiresolution dataset. The idea of the proposed method is to separately extract the thematic information contained in the two acquisitions collected at very different spatial resolutions and with generally different spectral bands, and perform a posterior probability pixelwise decision fusion.

As the ultimate task is to perform supervised binary image classification, the method requires a training map at both considered spatial resolutions. We assume that a training (ground truth, GT) map for two classes ω_1 and ω_2 is available for the acquisition at the finer spatial resolution. It is necessary to also define classes and their training information on the coarser lattice. Focusing on the aforementioned $D \times D$ patch, this is determined through the following rules:

1. If all $D \times D$ finer-resolution pixels are training samples for ω_1, then the corresponding coarser-resolution pixel is a training sample for class ψ_1;
2. If all $D \times D$ finer-resolution pixels are training samples for ω_2, then the corresponding coarser-resolution pixel is a training sample for class ψ_2;
3. Else, the coarser resolution pixel is a training sample for class ψ_3.

Accordingly, the two resolution levels correspond to distinct sets of classes: $\Omega = \{\omega_1, \omega_2\}$ on the finer resolution lattice and $\tilde{\Omega} = \{\psi_1, \psi_2, \psi_3\}$ on the coarser resolution grid. Semantically, ψ_1 and ω_1 represent the same land-cover class, but observed at the two very diverse resolutions – and the same comment holds about ψ_2 and ω_2 as well. On the contrary, ψ_3 represents a "mixed" class on the coarser-resolution lattice. The presence of this class is consistent with the fact that this pixel in the satellite image is necessarily a mixed pixel, corresponding to a ground area that is covered by partly ω_1 and partly ω_2.

Let $x_i \in \mathbb{R}^n$ and $y_i \in \Omega$ be the feature vector and the class label, respectively, of the ith pixel of the $D \times D$ patch in the finer-resolution image, and let $\tilde{x} \in \mathbb{R}^m$ and $\tilde{y} \in \tilde{\Omega}$ be the feature vector and the class label, respectively, of the corresponding coarser-resolution pixel. We collect all finer-resolution feature vectors x_i within the patch in a tensor $X \in \mathbb{R}^{D \times D \times n}$. The first proposed method is formalized as follows in terms of a decision fusion approach from suitable input posteriors. Specifically, the posterior distribution of y_i, given all available input observations at both resolutions, i.e., given both X and $\tilde{x}$, can be expressed as:

$$P(y_i|X, \tilde{x}) = \sum_{\tilde{y} \in \tilde{\Omega}} P(y_i, \tilde{y}|X, \tilde{x}). \tag{1}$$

Applying the Bayes theorem:

$$\sum_{\tilde{y} \in \tilde{\Omega}} P(y_i, \tilde{y}|X, \tilde{x}) = \sum_{\tilde{y} \in \tilde{\Omega}} p(X, \tilde{x}|y_i, \tilde{y}) \frac{P(y_i, \tilde{y})}{p(X, \tilde{x})} \propto \sum_{\tilde{y} \in \tilde{\Omega}} p(X, \tilde{x}|y_i, \tilde{y}) P(y_i, \tilde{y}), \tag{2}$$

where $P(y_i, \tilde{y})$ is the pixelwise joint probability of the labels of the images at the two resolutions. The proportionality constant in (2) depends only on the features and not on the labels, hence it does not affect the decision. In the first proposed approach, we state the following conditional independence assumption:

$$p(X, \tilde{x}|y_i, \tilde{y}) = p(X|y_i) p(\tilde{x}|\tilde{y}). \tag{3}$$

Similar conditional independence assumptions are widely accepted in the development of Bayesian and Markovian approaches (e.g., in [12, 16, 18]). Under this assumption and considering again the Bayes theorem, plugging (3) into (2) implies:

$$P(y_i|X, \tilde{x}) \propto \sum_{\tilde{y} \in \tilde{\Omega}} P(y_i|X) \frac{p(X)}{P(y_i)} P(\tilde{y}|\tilde{x}) \frac{p(\tilde{x})}{P(\tilde{y})} P(y_i, \tilde{y}) \propto \sum_{\tilde{y} \in \tilde{\Omega}} P(y_i|X) P(\tilde{y}|\tilde{x}) \frac{P(y_i|\tilde{y})}{P(y_i)}, \tag{4}$$

where $P(y_i|X)$ is the posterior probability of the acquisition at finer spatial resolution conditioned on all feature vectors in the $D \times D$ patch, $P(\tilde{y}|\tilde{x})$ is the posterior probability computed for the acquisition at coarser spatial resolution on the individual pixel corresponding to the patch, $P(y_i)$ is the prior probability at the finer resolution, and $P(y_i|\tilde{y})$ is the conditional probability of the labels at the finer resolution given those at the coarser resolution.

Given the size of the input multiscale data, as mentioned in the previous section, $P(y_i|X)$ is estimated as the prediction $\hat{P}^{(\mathsf{fcn})}(y_i|X)$ at the output of the softmax layer of the FCN and $P(\tilde{y}|\tilde{x})$ is predicted by the RF classifier in a pixelwise manner as $\hat{P}^{(\mathsf{rf})}(\tilde{y}|\tilde{x})$.

Concerning the conditional probability $P(y_i|\tilde{y})$, first, stationarity is assumed. Specifically, for each pair (ω_k, ψ_h) of classes at the two resolutions, the joint probability $P\{y_i = \omega_k, \tilde{y} = \psi_h\}$ $(k = 1, 2; h = 1, 2, 3)$ is assumed independent on the pixel location i. Therefore, the conditional probability $P\{y_i = \omega_k|\tilde{y} = \psi_h\}$ is independent of the location as well. Denoting $\theta_{k,h} = P\{y_i = \omega_k, \tilde{y} = \psi_h\}$, the joint probability matrix $\Theta = [\theta_{k,h}] \in \mathbb{R}^{2\times 3}$ collects the parameters of the proposed method. Θ is estimated through an approximate formulation of the EM algorithm [7, 27], as explained in the next section.

3.2 EM-based estimation of the transition probabilities

EM is a well-known iterative method to address maximum-likelihood parameter estimation when the observations can be viewed as incomplete data [7, 27]. EM has been proved to converge (under mild assumptions) to a stationary point of the log-likelihood function [7, 41], although it does not converge, in general, to a global maximum point.

Let S be the coarser-resolution lattice. If $j \in S$ is a coarser-resolution pixel, $\tilde{x}_j, \tilde{y}_j$, and X_j are its feature vector, its label, and the corresponding finer-resolution tensor, respectively. The related $D \times D$ subset of the finer-resolution lattice is denoted as $\mathcal{D}_j$. Let $\mathcal{X}$ be the tensor collecting all feature vectors x_i at the finer resolution and $\mathcal{Y}$ be the matrix collecting all corresponding labels y_i $(\forall i \in \mathcal{D}_j, \forall j \in S)$. Similarly, let $\tilde{\mathcal{X}}$ and $\tilde{\mathcal{Y}}$ be the tensor of all feature vectors $\tilde{x}_j$ and the matrix of all label $\tilde{y}_j$ at the coarser resolution, respectively $(\forall j \in S)$. EM iteratively maximizes the following function with respect to the matrix Θ of the parameters [7, 41]:

$$Q(\Theta|\Theta^t) = \mathbb{E}\left\{\ln p(\mathcal{X}, \tilde{\mathcal{X}}, \mathcal{Y}, \tilde{\mathcal{Y}}|\Theta) \,\middle|\, \mathcal{X}, \tilde{\mathcal{X}}, \Theta^t\right\}, \tag{5}$$

where the superscript t is the iteration index $(t = 0, 1, 2, \ldots)$ and $p(\mathcal{X}, \tilde{\mathcal{X}}, \mathcal{Y}, \tilde{\mathcal{Y}}|\Theta)$ is the joint distribution of all feature vectors and labels, in which the dependence on Θ is explicitly emphasized. Equivalently:

$$Q(\Theta|\Theta^t) = \mathbb{E}\left\{\ln p(\mathcal{X}, \tilde{\mathcal{X}}|\mathcal{Y}, \tilde{\mathcal{Y}}) + \ln P(\mathcal{Y}, \tilde{\mathcal{Y}}|\Theta) \,\middle|\, \mathcal{X}, \tilde{\mathcal{X}}, \Theta^t\right\}. \tag{6}$$

Here, Θ determines the joint distribution $P(\mathcal{Y}, \tilde{\mathcal{Y}}|\Theta)$ of all labels, whereas the probability density function $p(\mathcal{X}, \tilde{\mathcal{X}}|\mathcal{Y}, \tilde{\mathcal{Y}})$ of all observations, given all labels, is not parameterized on Θ and does not depend on it.

Specifically, in the proposed method, the function $Q(\Theta|\Theta^t)$ is replaced by an approximate formulation, in which we accept the following conditions:

1. The joint label distribution can be factored out as:

$$P(\mathcal{Y}, \tilde{\mathcal{Y}}|\Theta) = \prod_{j \in S} \prod_{i \in \mathcal{D}_j} P(y_i, \tilde{y}_j|\Theta); \tag{7}$$

2. For each coarser-resolution pixel $j \in S$, the label $\tilde{y}_j$ and all labels y_i of the related finer-resolution pixels $i \in \mathcal{D}_j$ are independent on the observations associated with all other coarser-resolution samples $\tilde{x}_s, s \neq j$ and all the related finer-resolution samples $x_r, r \in \mathcal{D}_s$.

We recall that approximate formulations, based for instance on pseudo-likelihood or mean-field concepts, have been widely used in the application of EM-type algorithms to favor analytical feasibility or computational efficiency [14, 28, 44]. Here, conditions 1 and 2 are used precisely for this purpose, in the estimation of the parameters Θ. However, it is worth noting that such approximation is not involved at all in the training or prediction of the FCN and the RF classifiers.

Plugging (7) into (6), dropping the terms of (6) that do not depend on Θ, and applying condition 2 lead to the following approximate formulation:

$$\begin{aligned} \bar{Q}(\Theta|\Theta^t) &= \sum_{j \in S} \sum_{i \in \mathcal{D}_j} \mathbb{E}\left\{ \ln P(y_i, \tilde{y}_j|\Theta) \left| \mathcal{X}, \tilde{\mathcal{X}}, \Theta^t \right. \right\} \\ &= \sum_{j \in S} \sum_{i \in \mathcal{D}_j} \mathbb{E}\left\{ \ln P(y_i, \tilde{y}_j|\Theta) \left| X_j, \tilde{x}_j, \Theta^t \right. \right\}. \end{aligned} \tag{8}$$

Since $\theta_{k,h} = P\{y_i = \omega_k, \tilde{y}_j = \psi_h\}$ $(i \in \mathcal{D}_j)$, we can write explicitly:

$$\bar{Q}(\Theta|\Theta^t) = \sum_{j \in S} \sum_{i \in \mathcal{D}_j} \sum_{k=1}^{2} \sum_{h=1}^{3} \alpha_{i,k,h}^t \ln \theta_{k,h}, \tag{9}$$

where $\alpha_{i,k,h}^t = P\{y_i = \omega_k, \tilde{y}_j = \psi_h | X_j, \tilde{x}_j, \Theta^t\}$ $(i \in \mathcal{D}_j; j \in S; k = 1, 2; h = 1, 2, 3; t = 0, 1, 2, \ldots)$ and where:

$$\sum_{k=1}^{2} \sum_{h=1}^{3} \theta_{k,h} = 1. \tag{10}$$

The updated parameter matrix Θ^{t+1} is obtained by maximizing the function $\bar{Q}(\Theta|\Theta^t)$ in (9) with respect to Θ under the constraint in (10). Solving the maximization through the Lagrangian multipliers, we obtain, after straightforward algebraic calculations $(k = 1, 2; h = 1, 2, 3; t = 0, 1, 2, \ldots)$:

$$\theta_{k,h}^{t+1} = \frac{1}{|S|D^2} \sum_{j \in S} \sum_{i \in \mathcal{D}_j} \alpha_{i,k,h}^t, \tag{11}$$

where $|S|$ is the total number of pixels in the coarser-resolution lattice (i.e., the cardinality of S). Owing to the conditional-independence assumption in (3), one can also prove that ($i \in \mathcal{D}_j; j \in S; t = 0, 1, 2, \ldots$):

$$\alpha_{i,k,h}^t = A_i^t \frac{\theta_{k,h}^t P\{y_i = \omega_k | X_j, \Theta^t\} P\{\tilde{y}_j = \psi_h | \tilde{x}_j, \Theta^t\}}{\left(\sum_{\ell=1}^2 \theta_{\ell,h}^t\right)\left(\sum_{\ell=1}^3 \theta_{k,\ell}^t\right)}, \tag{12}$$

where A_i^t is a normalization constant that ensures that $\sum_{k=1}^2 \sum_{h=1}^3 \alpha_{i,k,h}^t = 1$. In the proposed method, we evaluate $\alpha_{i,k,h}^t$ by estimating the posteriors in the numerator of (12) as in the previous section, i.e., through their predictions $\hat{P}^{(\mathsf{fcn})}(y_i|X_s)$ and $\hat{P}^{(\mathsf{rf})}(\tilde{y}_s|\tilde{x}_s)$ computed by the FCN on the finer-resolution lattice and by RF on the coarser-resolution one, respectively:

$$\alpha_{i,k,h}^t = A_i^t \frac{\theta_{k,h}^t \hat{P}^{(\mathsf{fcn})}\{y_i = \omega_k | X_j\} \hat{P}^{(\mathsf{rf})}\{\tilde{y}_j = \psi_h | \tilde{x}_j\}}{\left(\sum_{\ell=1}^2 \theta_{\ell,h}^t\right)\left(\sum_{\ell=1}^3 \theta_{k,\ell}^t\right)}, \tag{13}$$

The approximate EM algorithm, integrated in the proposed method for the estimation of the joint pixelwise probabilities Θ of the labels at the two resolutions, is initialized with a uniform distribution Θ^0 (i.e., $\theta_{k,h}^0 = 1/6$ for $k = 1, 2$ and $h = 1, 2, 3$). Then, it iteratively alternates (13) and (11) until convergence.

3.3 Multiresolution fusion through hierarchical probabilistic graphical model

The second proposed method aims to perform the fusion of the multimodal multiresolution information through a pyramidal tree structure, where the imagery can be inserted, modeled, and analyzed at its native resolution (see Fig. 2).

In this case, the root level (level 0) of the tree contains the coarse-resolution image and the leaf level (level L) contains the fine-resolution image. Accordingly, each root pixel corresponds to $D \times D$ leaf pixels. Starting from the leaf level, intermediate levels $(L-1), \ldots, 2, 1$ are constructed as in a traditional quadtree, by progressively halving the spatial resolution, and by associating the intermediate activations of the neural network at the corresponding resolution. Then, the root, i.e., level 0 of the tree, is linked directly to level 1. Differently from a conventional quadtree, where each consecutive level has a power-of-two relationship with the previous one, here, many more connections are present between the root and level 1. In particular, each pixel at the root corresponds to a patch of $(2^{1-L}D) \times (2^{1-L}D)$ pixels on level 1.

A hierarchical probabilistic graphical model (PGM) is defined over this pyramidal tree. As compared to PGMs on hierarchical quadtrees, this partially irregular topology affects the formulation of the inference criterion and the top-down and bottom-up flow of information across the levels.

Specifically, let S^ℓ be the pixel lattice of level ℓ of the tree ($\ell = 0, 1, \ldots, L$)[4]. We focus again, like in Section 3.1, on a single individual coarse-resolution pixel

[4] In Section 3.2, the pixel lattice of the input coarser-resolution image was indicated S. Here, it is denoted S^0 to distinguish it from the other lattices in the tree.

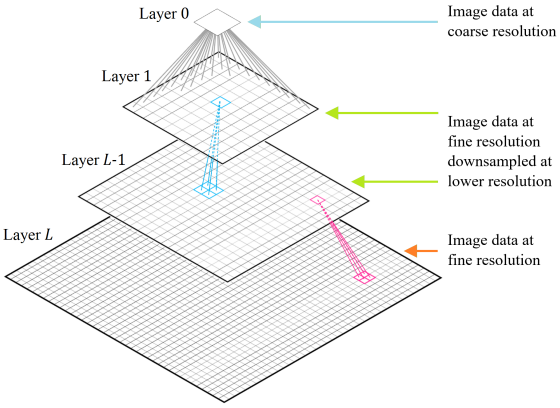


Fig. 2: Architecture of the MPM information fusion based on a quadtree.

(i.e., a single root pixel) and on the corresponding $D \times D$ patch at the fine resolution (i.e., at the leaves), and we use the same notations $\tilde{x}, \tilde{y}, x_i, y_i$, and X for the observations and labels at the root and at the leaves ($i \in S^L$). Here, we extend the notation x_i and y_i to the feature vector and the label of a pixel $i \in S^\ell$ also in an intermediate level of the tree ($\ell = 1, 2, \ldots, L-1$).

The inference is performed through the marginal posterior mode (MPM) criterion [12]. Similar to [30], under suitable conditional independence assumptions MPM can be formulated on the proposed pyramidal tree as follows (the proof is omitted for brevity). Firstly, a top-down pass to compute the prior probability of the class label, starting from the root to the leaves is performed. From the root to level 1, this implies:

$$P(y_i) = \sum_{\tilde{y} \in \tilde{\Omega}} P(y_i|\tilde{y}) P(\tilde{y}) \quad \forall i \in S^1. \tag{14}$$

Then, from level 1 to the leaves:

$$P(y_i) = \sum_{y_i^- \in \Omega} P(y_i|y_i^-) P(y_i^-) \quad \forall i \in S^\ell,\ \ell = 2, 3, \ldots, L, \tag{15}$$

where $i^- \in S^{\ell-1}$ denotes the parent of a pixel $i \in S^\ell$ not on the root ($\ell > 0$).

Secondly, a bottom-up pass is performed from the leaves to the root to compute the distribution of the label y_i of each pixel i, given all observations of the descendants of i in the tree (collected in a vector x_i^d) [30]:

$$P(y_i|x_i^d) \propto P(y_i|x_i) \prod_{r \in i^+} \sum_{y_r \in \Omega} \frac{P(y_r|x_r^d) P(y_r|y_i)}{P(y_r)}, \tag{16}$$

$$P(y_i|y_i^c, x_i^d) \propto \frac{P(y_i|x_i^d) P(y_i|y_{i^-}) P(y_{i^-})}{P(y_i)^{n_i}} \quad \forall i \in S^\ell,\ \ell = L-1, L-2, \ldots, 0$$

where $i^+ \subset S^{\ell+1}$ is the set of the children of a pixel $i \in S^\ell$ not on the leaves ($\ell < L$), y_i^c collects the labels of all pixels connected to i in the tree, and n_i is the number of such pixels. Finally, a second top-down pass is performed to compute $P(y_i|X, \tilde{x})$ on all pixels i in the tree [30]:

$$P(y_i|X, \tilde{x}) = \sum_{y_i^c \in \Omega^{n_i}} P(y_i^c|y_i, x_i^d) P(\tilde{y}|X, \tilde{x}) \quad \forall i \in S^1, \tag{17}$$

$$P(y_i|X, \tilde{x}) = \sum_{y_i^c \in \Omega^{n_i}} P(y_i^c|y_i, x_i^d) P(y_{i^-}|X, \tilde{x}) \quad \forall i \in S^\ell,\ \ell = 2, 3, \ldots, L. \tag{18}$$

Accordingly, a pixel $i \in S^L$ is assigned the label that maximizes $P(y_i|X, \tilde{x})$.

More details can be found in [30]. On the leaf level the predictions $\hat{P}^{(\mathsf{fcn})}(y_i|X)$ of the FCN on the finer-resolution image are used and incorporated in the PGM through (16). On the root, the predictions $\hat{P}^{(\mathsf{rf})}(\tilde{y}|\tilde{x})$ from RF on the coarser-resolution image are used. On the intermediate levels, the pixelwise posteriors are computed through a softmax over the intermediate activations of the FCN, after a pass through a convolutional layer whose number of filters is equal to $|\Omega|$. Accordingly, in the proposed approach, the hierarchical PGM on the pyramidal tree addresses multiresolution fusion, merging the predictions from the deep neural and ensemble components.

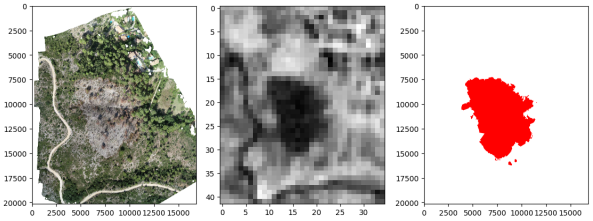
4 Experimental results

The proposed methods were tested on a multiresolution dataset for burnt area mapping in case of wildfires. The dataset consists of an RGB image acquired by an UAV with a spatial resolution of about 2 cm and the NIR channels of a Sentinel-2 image with a spatial resolution of 10 m (Fig. 3(a)-(b)). In particular, the UAV image has size of 16904×20324 pixels. Given the relationship between resolutions, $D = 480$. To maintain a reasonable number of levels and, simultaneously, model multiscale information, the drone imagery was resized to 4 cm and 8 cm of resolution (i.e., $L = 3$). Hence, each Sentinel-2 pixel is the parent of the 14400 pixels of the layer at 8 cm of resolution (with size 120×120 pixels).

The study area is La Destrousse, Provence-Alpes-Côte d'Azur, France. The drone image was acquired by INRAE (Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement) Provence-Alpes-Côte d'Azur research centre, Aix-en-Provence, shortly after the fire of 11 July 2018. The first available Sentinel-2 image of the same zone is dated 14 July 2018.

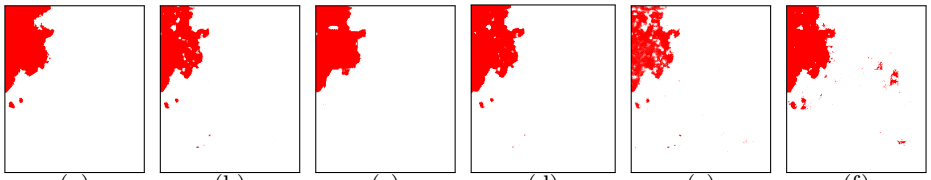
The GT boundaries of the burnt area, provided by the experts, were found with the canopy height model (CHM), measuring the height of trees, buildings, and other structures above the ground topography [11] (see Fig. 3(c)). This dataset was properly split in separate zones for training and testing the two proposed methods (see Fig. 4(a)).

To our knowledge, the proposed approaches are the first ones combining multiresolution UAV and satellite images at their – very different – native resolutions, for the mapping of burnt areas, therefore comparisons with state-of-the-art methods developed for this specific task were not possible. Nevertheless, the results of the proposed approaches were compared with those of the baseline U-Net,



(a) (b) (c)

Fig. 3: **Input images and GT:** (a) drone image at 2 cm resolution, (b) Sentinel-2 image at 10 m resolution (the normalized difference vegetation index, NDVI, is displayed), (c) the GT with the same resolution as of the drone image.



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

Fig. 4: **GT and classification results on the UAV image:** (a) GT used for testing (crop of Fig. 3(c)), and the classification results from (b) U-Net trained on UAV data, (c) the first and (d) the second proposed methods, (e) deep learning multiresolution fusion used for comparison, and (f) DBINet [9] trained on UAV data. Class legend: burnt (red) and non-burnt (white).

trained on the drone image at fine resolution, with those of RF trained on the satellite data at coarse resolution, and with a deep learning multiresolution fusion architecture where the satellite data at coarse resolution are included in the first convolutional layer as a bias scalar term given by the spectral information of the pixel $\tilde{x}$ multiplied by a learnable weighting vector.

The method was also compared with a recent state-of-the-art approach for burnt area segmentation combining CNNs and transformers, DBINet [9]. Since the methodology does not involve a multiresolution input, it was trained either with the UAV very-high-resolution data or with the satellite imagery.

The quantitative results obtained by the proposed methods and the two approaches used for comparison are reported in Table 1 in terms of false and missed alarm rates, and overall error rate with respect to the GT test tile. The classification maps are shown in Fig. 4(b)-(f). On the one hand, the baseline U-Net directly applied to the UAV image is quite effective in the discrimination of burnt and non-burnt areas, as suggested by the results shown in Fig. 4(b) with an overall error rate equal to 1.61%. The same can be said for DBINet [9], which attains similar results in terms of overall error rate, 1.66%. However, despite the output classification maps following the silhouette of the original GT map, there are several missed alarms inside the burnt zone for U-Net. On the other hand, the proposed pixelwise probabilistic fusion combining a few centimeters

Table 1: Classification accuracies of the proposed methods and of the comparison techniques.

Architecture	False alarm rate	Missed alarm rate	Overall error rate
U-Net on UAV [34]	0.28	10.81	1.61
RF on Sentinel-2 [4]	0.27	20.34	3.17
DL multires. fusion	**0.19**	27.08	3.51
First proposed method	0.48	8.55	1.47
Second proposed method	**0.19**	7.43	**1.17**
DBINet on UAV [9]	1.01	**6.29**	1.66
DBINet on Sentinel-2 [9]	0.69	12.45	2.15

very high resolution RGB drone image and a multispectral satellite image with ten-meter resolution shows more accurate results in terms of overall error rate, alas with a small loss in terms of false alarm rate with respect to U-Net, and a small loss in terms of missed alarm rate with respect to DBINet. On the contrary, DBINet presents the highest false alarm rate, 1.01%, thus several false positives, consistently with low missed alarm rate (see Fig. 4(f)). The classification map of the first proposed method (Fig. 4(c)) is more visually smooth and accurate than the result of U-Net and DBINet, thanks to the integration of the multispectral Sentinel-2 data through the proposed approach.

The second proposed method, the multiresolution fusion through the hierarchical PGM, outputs the classification map after considering all the information of the observation of the descendant pixels and the labels of all the connected pixels. Thanks to this multiresolution multispectral information fusion, the method attains the best performances for the experimental validation with the UAV and satellite images processed at their native resolutions, in terms of all the accuracy metrics considered, except missed alarm rate. The overall error rate is slightly higher than 1% and the false alarm rate is about 0.2%. As compared to DBINet, it attains a slightly higher missed alarm rate, yet maintaining low values for both false positives and false negatives. The classification map shown in Fig. 4(d) confirms the potential of this proposed model, as it is visually smooth and accurate, especially in comparison with the original GT, outperforming not only the baseline but also the previous fusion method.

The results of the deep learning multiresolution fusion (see Table 1 and Fig. 4(e)) suggest its potential in mapping burnt areas, reaching the lowest false alarm rate of 0.19%, same as the second proposed method. However, the classification map and the performances in terms of missed alarm rate and overall error rate are poorer than those obtained by the two proposed techniques and U-Net.

The performances of the two methodologies trained on the satellite imagery, RF and DBINet (on Sentinel-2) appear to be suboptimal with respect to the ones obtained by the methodologies trained on the UAV imagery (U-Net and DBINet on UAV), or on the fusion of the two multiresolution inputs (the two proposed methods and the deep learning multiresolution fusion). This can be explained by the lower number of training samples and the coarser spatial resolution of the satellite imagery.

In general, the results in terms of missed alarm rate are worse than those of false alarm rate, due to the imbalance of the classes in the dataset, where "non-burnt" is clearly a majority class, thus prompting this behavior.

5 Conclusion

This paper introduced two probabilistic fusion methods for the joint use of multiresolution imagery with a big mismatch in spatial resolution in the framework of semantic segmentation tasks. In particular, the focus was on RGB images collected by UAV and multispectral satellite data, thus bringing to a resolution ratio between the input image sources of the order of the hundreds.

The methods were applied to a case study of wildfire burnt zones semantic segmentation and experimentally validated with a real dataset consisting of drone and Sentinel-2 image data collected over the South of France. The experiments show the effectiveness of the two proposed methods for the detection and mapping of zones affected by fires. The two developed techniques obtain accurate classification results and maps, in particular for the approach fusing multiresolution information through an irregular quadtree topology, a hierarchical PGM, and an FCN. This confirms the potential of the combination of FCN architectures with PGMs on appropriate graphs.

Perspectives for future developments will involve the integration of the proposed methodologies with transfer learning techniques to test it with image data acquired by different sensors, thus characterized by different features, and associated with different geographical areas. Furthermore, it would be interesting to apply the method to different case studies related to other applications involving multiresolution input imagery with great mismatch in spatial, and possibly spectral, resolution.

References

1. Alvarez-Vanhard, E., Corpetti, T., Houet, T.: UAV & satellite synergies for optical remote sensing applications: A literature review. Sci. Remote Sens. **3**, 100019 (2021)
2. Benediktsson, J., Kanellopoulos, I.: Classification of multisource and hyperspectral data based on decision fusion. IEEE Trans. Geosci. Remote. Sens. **37**(3), 1367–1377 (1999)
3. Bouman, C., Liu, B.: Multiple resolution segmentation of textured images. IEEE Trans. Pattern Anal. Mach. Intell. **13**(2), 99–113 (1991)
4. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
5. Briem, G., Benediktsson, J., Sveinsson, J.: Multiple classifiers applied to multisource remote sensing data. IEEE Trans. Geosci. Remote. Sens. **40**(10), 2291–2299 (2002)
6. Cole-Rhodes, A., Johnson, K., LeMoigne, J., Zavorin, I.: Multiresolution registration of remote sensing imagery by optimization of mutual information using a stochastic gradient. IEEE Trans. Image Process. **12**, 1495–1511 (2003)
7. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Series B (Stat. Methodol.) **39**(1), 1–38 (1977)
8. Esch, T., Heldens, W., Hirner, A., Keil, M., Marconcini, M., Roth, A., Zeidler, J., Dech, S., Strano, E.: Breaking new ground in mapping human settlements from space – the global urban footprint. ISPRS J. Photogramm. Remote Sens. **134**, 30–42 (2017)

9. Fang, W., Fu, Y., Sheng, V.S.: Dual backbone interaction network for burned area segmentation in optical remote sensing images. IEEE Geoscience and Remote Sensing Letters **21**, 1–5 (2024)
10. Gómez-Chova, L., Tuia, D., Moser, G., Camps-Valls, G.: Multimodal classification of remote sensing images: A review and future directions. Proc. of the IEEE **103**(9), 1560–1584 (2015)
11. Hyyppä, J., Hyyppä, H., Leckie, D., Gougeon, F., Yu, X., Maltamo, M.: Review of methods of small-footprint airborne laser scanning for extracting forest inventory data in boreal forests. Int. J. Remote Sens. **29**(5), 1339–1366 (2008)
12. Kato, Z., Zerubia, J.: Markov random fields in image segmentation. Found. Trends Signal Process. **5**(1-2), 1–155 (2012)
13. Kim, J.B., Kim, H.J.: Multiresolution-based watersheds for efficient image segmentation. Pattern Recognit. Lett. **24**(1), 473–488 (2003)
14. Kuhn, E., Matias, C., Rebafka, T.: Properties of the stochastic approximation EM algorithm with mini-batch sampling. Stat. Comput. **30**, 1725–1739 (2020)
15. Laferté, J.M., Heitz, F., Perez, P.: A multiresolution EM algorithm for unsupervised image classification. In: Int. Conf. Pattern Recognit. (ICPR). vol. 2, pp. 849–853 vol.2 (1996)
16. Laferté, J.M., Pérez, P., Heitz, F.: Discrete Markov image modeling and inference on the quadtree. IEEE Trans. Image Process. **9**(3), 390–404 (2000)
17. Laine, A., Fan, J.: Texture classification by wavelet packet signatures. IEEE Trans. Pattern Anal. Mach. Intell. **15**(11), 1186–1191 (1993)
18. Li, S.Z.: Markov random field modeling in image analysis. Springer, 3rd edn. (2009)
19. Liu, H., Li, W., Jia, W., Sun, H., Zhang, M., Song, L., Gui, Y.: Clusterformer for pine tree disease identification based on UAV remote sensing image segmentation. IEEE Trans. Geosci. Remote Sens. **62**, 1–15 (2024)
20. Liu, S., Cai, T., Tang, X., Wang, C.: MRL-Net: Multi-scale representation learning network for COVID-19 lung CT image segmentation. IEEE J. Biomed. Health Inform. **27**(9), 4317–4328 (2023)
21. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 3431–3440, 2015
22. Luan, W., Zhang, X., Xiao, P., Wang, H., Chen, S.: Binary and fractional MODIS snow cover mapping boosted by machine learning and big Landsat data. IEEE Trans. Geosci. Remote Sens. **60**, 1–14 (2022)
23. Luotamo, M., Metsämäki, S., Klami, A.: Multiscale cloud detection in remote sensing images using a dual convolutional neural network. IEEE Trans. Geosci. Remote Sens. **59**(6), 4972–4983 (2021)
24. Luppino, L.T., Hansen, M.A., Kampffmeyer, M., Bianchi, F.M., Moser, G., Jenssen, R., Anfinsen, S.N.: Code-aligned autoencoders for unsupervised change detection in multimodal remote sensing images. IEEE Trans. Neural Netw. Learn. Syst. **35**(1), 60–72 (2024)
25. Mallat, S.: A wavelet tour of signal processing – The sparse way. Academic Press, 3rd edn. (2009)
26. Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., Terzopoulos, D.: Image segmentation using deep learning: A survey. IEEE Trans. Pattern Anal. Mach. Intell. **44**(7), 3523–3542 (2022)
27. Moon, T.: The expectation-maximization algorithm. IEEE Signal Process. Mag. **13**(6), 47–60 (1996)

28. Moser, G., Serpico, S.B.: Unsupervised change detection from multichannel SAR data by Markovian data fusion. IEEE Trans. Geosci. Remote Sens. **47**(7), 2114–2128 (2009)
29. Pastorino, M., Montaldo, A., Fronda, L., Hedhli, I., Moser, G., Serpico, S.B., Zerubia, J.: Multisensor and multiresolution remote sensing image classification through a causal hierarchical Markov framework and decision tree ensembles. Remote Sens. **13**(5), 849 (2021)
30. Pastorino, M., Moser, G., Serpico, S.B., Zerubia, J.: Semantic segmentation of remote-sensing images through fully convolutional neural networks and hierarchical probabilistic graphical models. IEEE Trans. Geosci. Remote Sens. **60**(5407116), 1–16 (2022)
31. Pyun, K., Lim, J., Won, C.S., Gray, R.M.: Image segmentation using hidden Markov Gauss mixture models. IEEE Trans. Image Process. **16**(7), 1902–1911 (2007)
32. Rezaee, M., van der Zwet, P., Lelieveldt, B., van der Geest, R., Reiber, J.: A multiresolution image segmentation technique based on pyramidal segmentation and fuzzy clustering. IEEE Trans. Image Process. **9**(7), 1238–1248 (2000)
33. Richards, J.A.: Remote sensing digital image analysis: An introduction. Springer, 5th edn. (2013)
34. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Med. Image Comput. Comput. Assist. Interv. LNCS, vol. 9351, pp. 234–241. Springer (2015)
35. Song, P., Li, J., An, Z., Fan, H., Fan, L.: CTMFNet: CNN and Transformer multiscale fusion network of remote sensing urban scene imagery. IEEE Trans. Geosci. Remote Sens. **61**, 1–14 (2023)
36. Thoonen, G., Mahmood, Z., Peeters, S., Scheunders, P.: Multisource classification of color and hyperspectral images using color attribute profiles and composite decision fusion. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **5**(2), 510–521 (2012)
37. van Rijthoven, M., Balkenhol, M., Siliņa, K., van der Laak, J., Ciompi, F.: Hooknet: Multi-resolution convolutional neural networks for semantic segmentation in histopathology whole-slide images. Med. Image Anal. **68**, 101890 (2021)
38. Wang, L., Liu, J.: Texture classification using multiresolution Markov random field models. Pattern Recognit. Lett. **20**(2), 171–182 (1999)
39. Wang, L., Zhang, C., Li, R., Duan, C., Meng, X., Atkinson, P.M.: Scale-aware neural network for semantic segmentation of multi-resolution remote sensing images. Remote Sens. **13**(24) (2021)
40. Waske, B., Benediktsson, J.A.: Fusion of support vector machines for classification of multisensor data. IEEE Trans. Geosci. Remote Sens. **45**(12), 3858–3866 (2007)
41. Wu, C.F.J.: On the convergence properties of the EM algorithm. Annal. Stat. **11**(1), 95 – 103 (1983)
42. Zavorin, I., Moigne, J.: Use of multiresolution wavelet feature pyramids for automatic registration of multisensor imagery. IEEE Trans. Image Process. **14**, 770–82 (2005)
43. Zheng, G., Jiang, Z., Zhang, H., Yao, X.: Deep semantic segmentation of unmanned aerial vehicle remote sensing images based on fully convolutional neural network. Front. Earth Sci. **11** (2023)
44. Zhou, F., Kong, Q., Deng, Z., Kan, J., Zhang, Y., Feng, C., Zhu, J.: Efficient inference for dynamic flexible interactions of neural populations. J. Mach. Learn. Res. **23**(211), 1–49 (2022)