MLMT-CNN for Object Detection and Segmentation in Multi-layer and Multi-spectral Images

Majedaldein Almahasneh, Adeline Paiement, Xianghua Xie, Jean Aboudarham

Noname manuscript No.
(will be inserted by the editor)

MLMT-CNN for Object Detection and Segmentation in Multi-layer and Multi-spectral Images

Majedaldein Almahasneh · Adeline Paiement · Xianghua Xie · Jean Aboudarham ·

Received: date / Accepted: date

Abstract Precisely localising solar Active Regions (AR) from multi-spectral images is a challenging but important task in understanding solar activity and its influence on space weather. A main challenge comes from each modality capturing a different location of the 3D objects, as opposed to typical multi-spectral imaging scenarios where all image bands observe the same scene. Thus, we refer to this special multi-spectral scenario as *multi-layer*. We present a multi-task deep learning framework that exploits the dependencies between image bands to produce 3D AR localisation (segmentation and detection) where different image bands (and physical locations) have their own set of results. Furthermore, to address the difficulty of producing dense AR annotations for training supervised machine learning (ML) algorithms, we adapt a training strategy based on weak labels (i.e. bounding boxes) in a recursive manner. We compare our detection and segmentation stages against baseline approaches for solar image analysis (multi-channel coronal hole detection, SPOCA for ARs) and state-of-the-art deep learning methods (Faster RCNN, U-Net). Additionally, both detection and segmentation stages are quantitatively validated on artificially created data of similar spatial configurations made from annotated multi-modal magnetic resonance images. Our framework achieves an average of 0.72 IoU (segmentation) and 0.90 F1 score (detection) across all modalities, comparing to the best performing baseline methods with scores of 0.53 and 0.58, respectively, on the artificial dataset, and 0.84 F1 score in the AR detection task comparing to baseline of 0.82 F1 score. Our segmentation results are qualitatively validated by an expert on real ARs.

Keywords Image segmentation · object detection · deep learning · weakly supervised learning · multi-spectral images · solar image analysis · solar active regions

M. Almahasneh
Department of Computer Science, Swansea University, Swansea, UK
E-mail: 809508@swansea.ac.uk

A. Paiement
Université de Toulon, Aix Marseille Univ, CNRS, LIS, Marseille, France
E-mail: adeline.paiement@univ-tln.fr

X. Xie
Department of Computer Science, Swansea University, Swansea, UK
E-mail: x.xie@swansea.ac.uk

J. Aboudarham
Observatoire de Paris/PSL, Paris, France
E-mail: Jean.Aboudarham@obspm.fr

1 Introduction

Solar features (e.g. active regions (ARs)) detection and segmentation are essential in studying solar weather and behaviours. This analysis can be carried out by remotely monitoring the solar atmosphere continuously on multiple wavelengths, e.g. as shown in Figs. 1 and 2, captured from different ground- and space-based sensors.

However, unlike traditional multi-spectral scenarios such as Earth imaging from space, e.g. [1, 2, 3, 4, 5, 6, 7], where multiple imaging bands reveal different aspects (e.g. composition) of a same scene, in solar physics, different bands capture the solar atmosphere at different temperatures, which correspond to different altitudes [8].

Indeed, the solar atmosphere consists of various atoms, each of which emits light of a certain wavelength

when they reach a specific temperature, in a context of strong temperature gradient across the solar atmosphere. Therefore, different wavelengths show different 2D layers of the 3D objects (e.g. ARs) that span the solar atmosphere. We refer to this scenario as *multi-layer analysis*. For this reason, handling the multi-spectral (and multi-layer) nature of the problem is not straightforward. Moreover, the variety in shapes, fuzzy boundaries, and differing brightness of ARs also make their precise localisation complex.

Very few solutions were presented to the AR localisation problem. Most of these methods exploited single image bands only, e.g. [9, 8]. Authors justified this by the fact that each band provides information from a different solar altitude, they show how areas of ARs differ from band to band [8]. We , however, argue that inter-dependencies exist between bands, which can be exploited for increased robustness.

The SPOCA method [10] used clustering to extract (pixel-wise) ARs and coronal holes from SOHO/EIT 171 Å and 195 Å combined images, assuming that they should yield identical detection. This approximation may result in a poor analysis of at least one of these bands. SPOCA's detection is based on Fuzzy C-means and Possibilistic C-means [11], followed by post-processing with morphological operations. The use of fuzzy logic in SPOCA addresses the uncertainty in defining AR boundaries [10]. The quality of results was subjectively evaluated on 112 observations. SPOCA is now used in the HFC online catalogue.

Generally, these methods are mainly based on clustering and morphological operations, thus are pre- and post-processing dependant, which makes them difficult to adapt to new image domains and hyperparameter-dependant.

In this work, we investigate the possibilities offered by deep learning (DL) methods and exploit more bands than previous methods, for richer information on the solar atmosphere. In the past two decades, object detection has evolved dramatically, from hand-crafted features based detection (e.g. Haar [12], and HOG [13]) to deep neural networks (DNN) such as YOLO [14], SSD [15], R-FCN [16], Cornernet [17], or Faster RCNN [18]. Generally, DL based detectors rely on convolutional neural networks (CNN) to analyse images.

These may be split into two categories, 1) two stage detection, in which images are analysed in two steps, region proposal (generate a set of suspicious locations) and a final classification stage, and 2) one stage detection, where a DNN learns to regress object locations and classes in a single step. In general, two stage detectors (e.g. Faster RCNN) can achieve higher accuracy over single stage detectors [19, 20]. However, such methods aim at analysing 2D images or dense 3D volumes, and are therefore not suited to directly handle the sparse 3D nature of the solar imaging data. Hence, we design a specialised DL framework that can accommodate for different DL architectures as a backbone. We demonstrate this by applying our framework to different backbones (Faster RCNN and U-Net) and tasks (object detection and segmentation)

Multi-spectral images are commonly treated in a similar fashion to RGB images, by stacking different bands into multi-channel images, [6, 5, 21, 3, 22]. These methods are designed under the assumption that the different image bands capture different aspects of the same scene, which makes it ill-suited for our multi-layer case, where spatial positioning indeed differs from band to band. Another common approach is to aggregate information from different bands at different levels (e.g. feature level and image level) [23, 24, 7, 2, 1, 22, 25]. This feature fusion strategy demonstrates the potential for DNNs to improve localisation by exploiting the multi-spectral aspect of the data. Some works found that feature level fusion assists CNNs in producing a more consistent detection than using image level fusion for pedestrian detection from RGB and thermal images [2]. Contrary, image fusion worked best when segmenting soft tissue sarcomas in multi-modal medical images [22]. This suggests that there is no universal best fusion strategy. Thus, we investigate different types of fusion and different stages to apply fusion. Another feature fusion strategy was used to segment coronal holes from SDO's 7 EUV bands and line-of-site magnetogram in [26]. The method relies on training a CNN, using weak labels, to segment coronal holes from a single band, followed by fine-tuning the learned CNN over the other bands consecutively. The feature maps of each specialised CNN are used in combination as input to a final segmentation CNN, resulting in a unique final prediction. This unique localisation result for all multi-spectral images is a common limitation to all cited works for our multi-layer scenario, which we address in this study with a multi-task network.

In this work, We introduce a novel MultiLayer MultiTask CNN (MLMT-CNN), a multi-tasking DNN framework, as a robust solution for the solar AR localisation problem (i.e. detection and segmentation) that takes into consideration the multi-layer aspect of the data and the 3-dimensional spatial dependencies between image bands. In a preliminary work [27], we demonstrated its potential of analysing multiple layers simultaneously for AR detection in the form of bounding box. In this paper, we extend on this work, applying the MLMT-CNN framework to new tasks (segmenta-

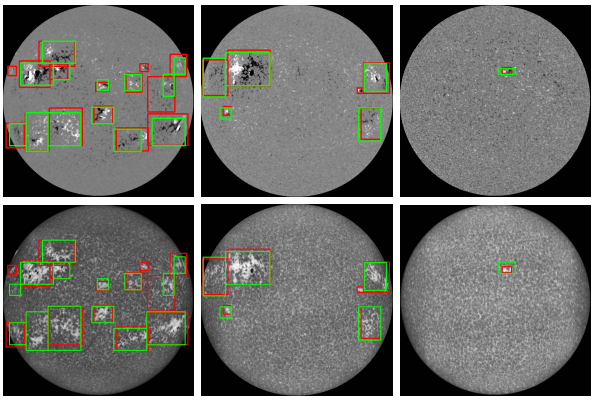

Fig. 1 Ground-truth (red) and MLMT-CNN's (green) detection of ARs at three levels of solar activity (left to right: high, medium, low) in randomly selected images from (top to bottom) SOHO/MDI Magnetogram and PM/SH 3934 Å.

tion) and to new datasets of different types, using new DNN backbones.

The 3D nature of our multi-spectral and multi-layer imaging scenario, which differs from other multi-spectral cases such as Earth observations, requires a new benchmark. Therefore, we introduce two annotated datasets comprised of images of the solar atmosphere from both ground and space-based sensors. They cover evenly all phases of solar activity, which follows an 11-year cycle. To the best of our knowledge, no localisation ground-truth was previously available for such data. A labelling tool was hence designed to cope with its temporal, multi-spectral, and multi-layer nature and will be also released. The solar data with bounding box labels were first presented in our preliminary work [27]. Here, we further extend the datasets with additional weak segmentation labels.

Furthermore, we propose a training approach that accounts to the different objectives of the individual MLMT components using their correspondent losses, in contrast to the classical training in which all components are deemed to reach an optimal solution simultaneously according to their overall loss.

Our contributions may be summarised as:

1. We present a paradigm to handle multi-spectral solar images that show several layers of a 3D object that span the solar atmosphere (i.e. multi-layer).
2. We demonstrate the effectiveness of our approach in MLMT, a multi-task DL framework for solar AR localisation. Localisation includes both detection in the form of bounding boxes, and pixel-wise segmentation. We further explore and demonstrate the potential of our proposed paradigm by implementing it with different state-of-the-art CNN backbones, as well as handling different data types and arbitrary number of bands.
3. We propose a training strategy for MLMT that optimises the DNN weights more effectively for each objective than the classical training strategy.
4. To address the difficulty of producing accurate and detailed annotations for AR segmentation, we propose a recursive training approach based on weak labels (i.e. bounding boxes).
5. We introduce two balanced and annotated datasets of multi-layer images of the solar atmosphere for AR detection, from both ground- and space-based data.
6. We release a multi-spectral and multi-layer image annotation tool that facilitates bounding box labelling using temporal and spectral information.
7. We further validate our approach on an artificially created dataset of multi-modal medical images of similar spatial configurations to the multi-layer solar images.

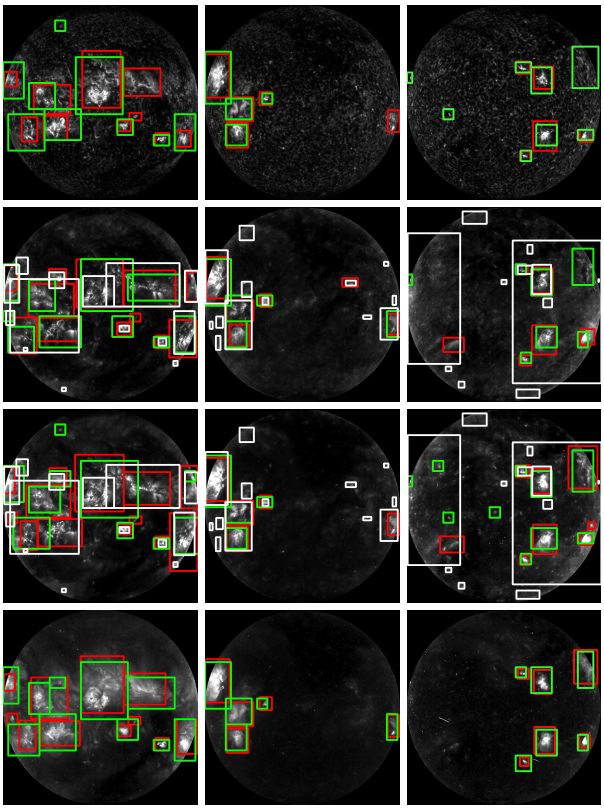

Fig. 2 Ground-truth (red) and MLMT-CNN (green) and SPOCA's (white) detection of ARs at three levels of solar activity (left to right: high, medium, low) in randomly selected images from (top to bottom) SOHO/EIT 304 Å, 171 Å, 195 Å, and 284 Å.

2 Methodology

Our framework exploits several time-matched multi-layer images in parallel, to predict separate, although related, localisation results for each image. These time-matched observations are possibly acquired by different instruments or at different orientations of the same instrument. As such they are spatially aligned prior to analysis. Our localisation involves two stages: detection, in the form of bounding box around an object and its classification of object type, followed by a segmentation stage to produce a pixel-wise classification map enclosed in the predicted bounding box.

For both stages, we deploy a new multi-layer and multi-task DL framework that analyses information from neighbouring layers (i.e. image bands). The network learns band-specific features, these features are then fused at multiple levels in the network, inducing the network to learn correlations between the different bands. Finally, the resulting embeddings are jointly analysed, exploiting information from neighbouring layers to produce their separate but related results.

This framework is general and may be used with various DNN backbones. We experiment with Faster RCNN and U-Net backbones, for detection and segmentation respectively, demonstrating the benefits of our joint analysis scheme in learning the inter-dependencies between the different image bands in both stages. Our framework may be easily adopted to serve other applications, as demonstrated with BraTS-prime and Cloud-38-prime, cf. Section 3.

In this section, we introduce the main concepts of the MLMT-CNN framework in Section 2.1, the backbone networks used in our framework in Section 2.2, and the details of our two detection and segmentation stages in Sections 2.3 and 2.4, respectively.

2.1 MultiLayer-MultiTask (MLMT) framework

While some existing works were developed for analysing multi-spectral images, to our best knowledge, the problem of detecting objects over multi-layer imagery, which is a sparse 3D multi-spectral case in which different bands show different scenes (i.e. layers), was not yet addressed. We introduce a new multi-layer and multi-task framework (MLMT) to tackle this scenario. The intuition behind our framework manifests in three key principles:

1. Extracting features from different image bands individually using parallel feature extraction branches. This allows the network to learn independent features from each band, according to their specific modality.
2. Aggregating the learned features from the different branches using some appropriate fusion operator. This assists the network to jointly analyse the extracted features from different bands and thus learn interdependencies between the image bands. In this work, we test fusion by addition and concatenation, at different feature levels (i.e. early and late fusion).
3. Generating a set of results per image band, based on a multi-task loss, allowing the detection of different sections or layers of 3D objects within the different bands.

Points 1 and 3 are motivated by the nature of the multi-layer data, where different bands capture different locations in a 3D scene, each providing some unique information. Our multi-tasking framework aims at obtaining specialised results for each image band, in contrast to most existing works where focus is on producing an independent prediction to all image bands. This is crucial since the localisation information may differ from one band to another in cases of multi-layer images (e.g. solar data). Yet, all bands are correlated, which motivates point 2. Our framework exploits the inter-dependencies between the different bands by its joint analysis strategy, increasing the robustness of its performance in individual bands.

Furthermore, our framework emulates how experts manually detect ARs, where a suspected region's correlation with other bands is evaluated prior to its final classification. This demonstrates the usefulness and importance of accounting for (spatially and temporally) neighbouring slices in robustly detecting ARs.

Moreover, this framework is very modular and flexible. It can accommodate any number of available image bands (i.e. layers) and perform different tasks (e.g. detection and segmentation). Additionally, since different scenarios may require different fusion strategies (as suggested by existing works), the modularity of our framework allows it to be easily adapted to different cases. We demonstrate this by applying our framework to different applications in Section 3 (solar ARs, BraTS-prime, and Cloud-38-prime datasets), where we investigate the best type and level of feature fusion (e.g. addition and concatenation, early and late).

2.2 Backbone networks

The modular design of our framework allow it to adopt different backbone architectures. Indeed, the 3 key principles are applicable to different backbones, as they are not architecture dependent. We demonstrate this in this section and discuss different backbone networks for different tasks in which we adopt our framework to.

2.2.1 Detection backbone: Faster RCNN

For detection, we adopt the Faster RCNN architecture as the backbone. Faster RCNN is a DL-based detector that may be trained to detect and classify a number of objects from a (usually RGB) image. It consists of three main parts: 1) convolutional layers extract features from the input image, as in any CNN. From these features, 2) a region proposal network (RPN) proposes locations that might contain objects, and 3) a detection network predicts the object class of each proposed locations. We apply our framework to the three stages detection strategy of Faster RCNN, thus generalising it to jointly analysing multiple images that span different locations (or layers) of a same 3D scene.

Comparing to other state-of-the-art architectures (e.g. YOLO and SSD), the multi-stage design of Faster RCNN allows aggregating information from different bands at different levels, namely low level (i.e. feature extraction stage) and high level information (i.e. region proposals). Additionally, Faster RCNN has scored the highest accuracy in [20].

2.2.2 Segmentation backbone: U-Net

We experiment with U-Net as the backbone of our segmentation stage. Nevertheless, other competing networks can also be used, and we also experimented with FCN8 [28] in early tests. U-Net [29] is a fully convolutional network that consists of 3 main parts: 1) contraction path, 2) bottleneck, and 3) expansion path.

In our segmentation stage, we apply our MLMT framework to the building blocks from U-Net to demonstrate the benefits of the joint analysis in segmenting ARs. MLMT takes advantage of U-Net's skip connections that allow combining features from different semantic levels within the same band. This maximises the learned information within individual bands while combining this information with feature fusion at the U-Net's bottleneck stage. Thus, information from different bands are combined for classification, while preserving the spatial information of individual images.

2.3 MLMT-CNN: Detection stage

Our detection DNN is presented in Fig. 3. It takes the pre-processed multi-layer image as input. A CNN (ResNet50 or VGG16 in our experiments) is first used as a feature extraction network. Parallel branches (subnetworks) produce a feature map per image band, following the late (or feature map) fusion strategy. Since individual bands provide different information, this allows the subnetworks' filters to be optimised for their

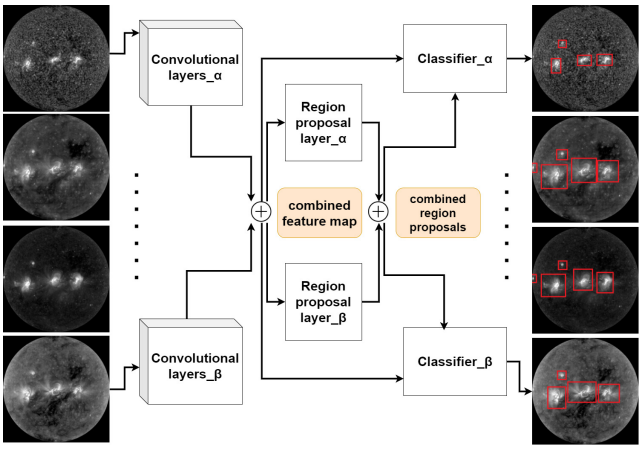

Fig. 3 MLMT for detection using the Faster RCNN backbone. The '+' sign denotes concatenation of the feature maps, or of the lists of region proposals.

input bands individually. The feature maps are then concatenated across the bands, assisting the network to learn correlations between them.

The combined feature map is jointly analysed by one parallel module per image band that performs Faster RCNN's RPN. The RPN stage uses three aspect ratios ([1:1], [1:2], [2:1]) and four sizes of anchor (32, 64, 128, and 256 pixel width). We found empirically that these match well the typical size and shape of ARs. One specialised RPN per image band is trained.

At training, for each band, the correspondent region proposals along with the combined feature map are used by a detection module to perform the final prediction for the band. However, at testing time, the band-specialised detector modules use the region proposals from all bands. This combination of region proposals helps finding potential AR locations (i.e. region proposals) in bands where they are more difficult to identify. This aids the network to learn the correlation between the different bands more dynamically, benefiting from information from different bands simultaneously while having band-specialised region proposal and detection models.

It is worth noting that during training, the RPN proposals for a band are filtered (i.e. labelled as positive or negative) with respect to their overlap with the band's own ground-truth. Hence, combining them in the training time would mean implicitly inheriting the ground-truth of a band to another, in contradiction with the band-specific ground-truth used for training the detector module. Indeed, different bands show distinct cuts of a 3D object in which each cut must have its own ground-truth. Combining ground-truths of different bands at training time may hinder the learning of both the RPN and detector modules. Therefore, region

proposals are only combined at testing time to ensure a better learning of the final detection modules.

Using the combined feature map aids the network to learn the relationship between the image bands, in both region proposal and classification stages, hence providing a more robust prediction in line with the nature of the data. This prediction is still band-specialised thanks to the different ground-truths being used for each band at training time. We demonstrate in Section 3 that this is particularly helpful in cases where an AR is difficult to detect in a single band.

We train our MLMT framework using all input bands and branches according to a combined loss function:

$$L = \sum_{b} \left(\frac{1}{N_{cls}} \sum_{i} L_{cls}(p_{b_i}, p^*_{b_i}) + \lambda \frac{1}{N_{reg}} \sum_{i} p^*_{b_i} L_{reg}(t_{b_i}, t^*_{b_i}) \right) \quad (1)$$

where b and i refer to the image band and the index of the bounding box being processed, respectively. The terms L_{cls} and L_{reg} are the bounding-box classification loss and the bounding-box regression loss defined in [18]. N_{cls} and N_{reg} represent the size of the mini batch being processed and the number of anchors, respectively. λ balances the classification and the regression losses (we set λ to 10 as suggested in [18]). p and $p^\star$ are the predicted anchor's class probability and its actual label, respectively. Lastly, t and $t^\star$ represent the predicted bounding box coordinates and the ground-truth coordinates, respectively. It is worth noting that our proposed framework is not limited to using Faster RCNN's loss and may be trained with using other task-suitable loss functions.

During training, the weights of each stage (i.e. feature extraction, region proposal, and detection) are stored independently whenever the related Faster RCNN loss decreases. At testing time, the best performing set of weights is retrieved per stage. We refer to this practice as 'Multi-Objective Optimisation' (MOO). The improved performance that we observe in Section 3 may be explained by each stage having a different objective to optimise, which may be reached at different times.

In this paper, we experiment with a 2, 3, and 4-band pipeline. However, the approach may generalise straightforwardly to n bands and new imaging modalities.

2.4 MLMT-CNN: Segmentation stage

Our segmentation framework is presented in Fig. 4. It consists of 3 parts: 1- feature extraction, 2- feature fusion, and 3- mask reconstruction. The network takes as input the AR detections (patches) produced by the detection stage. Each detection is cropped from all image bands, and resized into 224x224 pixel before entering the segmentation network.

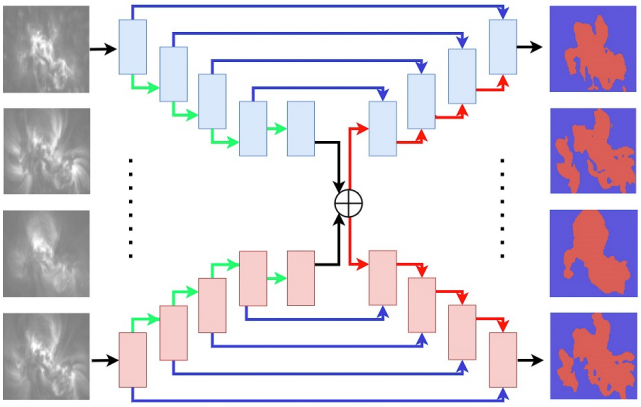

Fig. 4 MLMT for segmentation using the U-Net backbone. The '+' sign denotes fusion of the feature maps. Coloured boxes are convolutional blocks for each branch (band) respectively. Green and red arrows denote max pooling and up sampling operations, respectively. Blue arrows are skip connections, applied to the appropriate channel of the joint feature map for each branch.

The feature extraction part consists of parallel U-Net contracting paths (one per band), each specialised to extract a feature map from its band individually. The resulting feature maps are then combined in the latent space (i.e. late fusion). It is worth noting that different feature fusion operations may be used. In this work, we experiment with addition and concatenation. The combined feature map is passed to the mask reconstruction part where parallel U-Net expensive paths (a specialised path per band) perform the final prediction. Skip connections are utilised between each band's contracting path and its correspondent expensive path to preserve fine details learned in early layers of that band (blue arrows in Fig. 4).

To overcome the lack of dense AR annotation, we use weak labels to train our segmentation network along with a recursive training approach. In the first round of iterations, weak annotations are used to guide the training. Once the network converges, the training is repeated from random weights using the new labels predicted by the model from the previous round. This process is repeated until validation loss stops decreasing, or starts to increase. The idea is inspired by [30, 31, 32, 33], where authors demonstrate that iteratively training segmentation CNNs with weak labels can achieve results close to fully supervised.

Our weak label was carefully designed to provide a conservative representation of ARs, favouring a high precision over recall, to accelerate the first training

round (as detailed in 3). Recursive training allows the network to learn a more generalised representation by supervising itself through the recursion process, while limiting the bias that may be introduced by the initial weak label. This is in line with the discovery that sampling as little as 4% of the pixels to compute the training loss enables CNNs to achieve a close performance to fully supervised, caused by the strong correlation within the training data of a pixel-level task [34]. The results of our recursive approach were validated by a solar physics expert, and will be further discussed in Section 3.

Moreover, the solar data suffers from a class imbalance by nature, since most of the solar disk is covered by quite sun (solar background). The use of AR crops (patches from previous detection) helps in reducing this imbalance significantly, yet it does not solve the matter completely. Hence, we train our model using a weighted categorical cross entropy loss that combines information from all image bands as follows:

$$L(y,\hat{y}) = -\sum_{b=0}\sum_{c=0}\omega_c \sum_{i=0} y_{icb} \ * \ log\ (\hat{y}_{icb}) \tag{2}$$

where y and $\hat{y}$ are the actual and the predicted classes, respectively, ω_c is the weight of the c^{th} class, and i and b denote the pixel and the band being processed, respectively. We use the values 2, 1, and 2 as the weights for the three AR, solar background (quite sun), and image background classes, respectively. These weights were found to be best performing by experimenting with different values based on prior computed class ratios. Adding the weighting term to the combined loss prevents any bias that might be caused by the dominating solar background class.

3 Experiments

All experiments were implemented using Tensorflow with an NVIDIA GeForce GTX 1080 Ti GPU. The detection and segmentation stages were trained for 3000, and 250 epochs (∼4 and ∼0.41 days), respectively, using Adam optimiser [35] with learning rates of 2e-5 and 4e-3, respectively.

3.1 Data

3.1.1 Labelled AR datasets

We work with images from SOHO spacecraft and Paris-Meudon (PM) observatory. Multi-layer solar images comprise of measurements at different ultraviolet and X-ray wavelengths (denoted as *bands*) and centred on the emission wavelengths of ionised atoms of interest. Since these ionised atoms exist at given temperatures, they allow imaging different altitude regions of the solar atmosphere, following its temperature gradient. ARs are areas of strong magnetic field. Therefore, the multi-spectral and multi-layer images may be complemented by magnetograms that inform on the intensity and polarity of the magnetic field. With current technologies, magnetograms are mainly available for the photosphere. The images of this study were acquired in the 171 Å, 195 Å, 284 Å, and 304 Å bands (SOHO/EIT imager), 3934 Å band (PM Spectroheliograph (PM/SH) imager), and the magnetogram images (SOHO/MDI imager) as illustrated in Figs. 1 and 2. These correspond to observing the photosphere (magnetogram), chromosphere (3934 Å), chromosphere and base of the transition region (304 Å), transition region (171 Å and 195 Å), and corona (284 Å). Solar observations are acquired frequently to study the evolution of solar features and events over time.

Our work requires ground-truth annotations of ARs in the form of bounding boxes (detection) and pixel-wise masks (segmentation). To the best of our knowledge, no such annotated dataset is currently publicly available. Therefore, we publish the Lower Atmosphere Dataset (LAD) and Upper Atmosphere Dataset (UAD). Both datasets include bounding box annotations produced with a new multi-spectral labelling tool. which displays, side by side, images from an auxiliary modality and from a sequence of 3 previous and 3 subsequent time steps. ARs have a high spatial coherence in 3934 Å and magnetogram images due to the physical proximity of the two imaged regions, hence they share the same bounding boxes. The UAD additionally includes weak segmentation labels produced by thresholding and morphological operations so as to label only pixels that have an evident activity, i.e. being the brightest regions in the solar disk. This is motivated by the discovery in [34] that training data of a pixel-level task has a strong between-sample correlation, and that randomly sampling as little as 4% of the pixels to train a CNN can achieve about the same performance as full supervision. Both datasets are augmented using north-south mirroring, east-west mirroring, and a combination of the two. All annotations were validated by a solar physics expert.

We split the datasets into training and testing sets in the following proportions. For LAD, we use 213 images (1380 bounding box) for training, and 53 images (406 bounding box) for testing. For UAD, we use 283 images for training, and 40 images for testing. This amounts to 2205, 1919, 2341, and 2016 training bounding boxes in the 304 Å, 171 Å, 195 Å, and 284 Å bands

respectively, and 287, 262, 330, and 263 testing bounding boxes. Furthermore, in order to compare against the localisation of SPOCA, we consider a subset of the UAD testing set for which SPOCA detection results are available in HFC: the SPOCA subset. It consists of 26 testing images (181, 168, 213, and 166 bounding boxes in the 304 Å, 171 Å, 195 Å and 284 Å images respectively).

3.1.2 Weak-BraTS-prime

To further demonstrate the benefits of our joint analysis based approach, we create a synthetic dataset from the BraTS multi-modal dataset [36] of similar spatial configurations to the solar imaging bands. BraTS consists of full 3D MR image volumes of brain in 4 modalities (T1GD, T1, T2, and Flair) and 3 classes: enhancing tumour (ET), necrotic and non-enhancing tumour core (NCR/NET), and peritumoural edema (ED). We create the synthetic dataset by selecting one 2D slice of each image modality separated by (spatial) gaps of size g. This emulates the solar images scenario where each band shows ARs in a different solar altitude. We experiment with g being either 1, 2, or 3 voxels, to show the influence of the image modalities having different levels of spatial correlation on the segmentation. For each modality, we use a total of 11,533 and 190 training and testing images, respectively.

3.1.3 Weak-Cloud-38

We further evaluate our recursive training approach on a third weakly labelled dataset derived from the Cloud-38 [4] multi-modal (4 bands) dataset. This dataset has resemblance to our solar images in that there are a variety of cloud shapes, sizes and densities, albeit the multi-layer (3D) aspect is missing. It consists of 2,502 (2,382 training and 120 testing) images per band. We augment the training set using similar transformations to solar images.

3.2 Detection stage

A detection is considered a true positive if its intersection with a ground-truth box is greater or equal to 50% of either the predicted or ground-truth area. NMS is used in all experiments to discard any redundant detections.

All tested deep learning architectures were initialised with a pre-trained CNN with ImageNet weights (similar transfer learning strategy has been found useful in, for instance, depth estimation [37]). Its worth noting that the components of each detection branch (feature extraction network, RPN, and detection network) adopt a similar hyper-parameter configuration to that suggested in Faster RCNN [18].

A single-channel solar image was repeated along the depth axis resulting in a 3-channel image matching the pre-trained CNN's input depth.

HFC's SPOCA detections were obtained from 171 Å and 195 Å images only, combined as two channels of an RGB image, and SPOCA produces a single detection for both bands. We compare this detection against the ground-truth detections of each of the bands, individually. SPOCA may only combine image bands that are located close to each other in the solar atmosphere and for which it makes sense to produce a common set of detection results. Thus, HFC's SPOCA results are only available for bands of the transition region (171 Å) and low corona (195 Å), and no images from the chromosphere (304 Å) or the high corona (284 Å) were used. However, to prove the robustness and versatility of our detector, we also experiment with a combination of chromosphere, transition region, and corona bands on the SPOCA subset in addition to the whole UAD.

Table 1 Baseline detection performance of the single image band detectors.

Dataset	Image band	Precision	Recall	F1
LAD	3934 Å	0.93	0.82	0.87
	Magn.	0.89	0.78	0.83
UAD	304 Å	0.73	0.83	0.78
	171 Å	0.84	0.89	0.86
	195 Å	0.81	0.75	0.78
	284 Å	0.86	0.82	0.84
SPOCA	304 Å	0.72	0.82	0.77
	171 Å	0.87	0.87	0.87
	195 Å	0.82	0.73	0.77
	284 Å	0.86	0.82	0.84

3.2.1 Independent detection on single image bands

We first compare detection results produced by Faster RCNN over individual image bands (Table 1). This serves as baseline to assess our proposed framework. Different DL-based feature extraction networks are tested (ResNet50 and VGG), and we present here results of the best performing, namely ResNet50.

When comparing the detection results per image band, we notice that 304 Å images are repeatedly amongst the most difficult to analyse in UAD, having the lowest F1-scores in all tests. On the other hand, 171 Å shows the highest results of all UAD bands, followed by 284 Å and 195 Å, respectively. This may be explained by ARs having a denser or less ambiguous appearance in 171 Å, 195 Å, and 284 Å image bands than in 304 Å since they are higher in the corona. A similar observation can be made in the LAD dataset when comparing the Magnetogram results to PM/SH

Table 2 Detection performance of the MLMT-CNN detectors. For each band, the highest scores are highlighted in bold.

Detector	Fusion	Dataset	Bands	Prec.	Recall	F1
MLMT-CNN (ResNet50 – MOO)	Early – concat.	LAD	3934 Å	0.96	**0.82**	**0.89**
			Magn.	0.95	0.82	0.88
	Late – concat.		3934 Å	**0.97**	**0.82**	**0.89**
			Magn.	**0.96**	**0.85**	**0.90**
	Late – addition		3934 Å	0.95	**0.82**	0.88
			Magn.	0.94	0.80	0.87
MLMT-CNN (ResNet50)	Late – concat.	UAD	171 Å	0.92	0.77	0.84
			284 Å	0.90	0.81	0.85
			171 Å	0.82	0.85	0.83
			195 Å	0.86	0.72	0.78
			195 Å	0.88	0.67	0.77
			284 Å	0.84	0.78	0.81
			304 Å	0.82	0.79	**0.80**
			195 Å	0.87	0.75	0.80
MLMT-CNN (ResNet50 – MOO)	Late – concat.	UAD	171 Å	0.90	0.83	**0.87**
			284 Å	**0.93**	0.80	**0.86**
		SPOCA	171 Å	**0.89**	0.83	**0.86**
			284 Å	**0.92**	**0.80**	**0.86**
		UAD	171 Å	0.86	0.77	0.82
			195 Å	0.89	0.75	0.81
		SPOCA	171 Å	0.83	0.77	0.80
			195 Å	0.86	0.73	0.79
		UAD	195 Å	0.88	0.68	0.77
			284 Å	0.84	0.78	0.81
		SPOCA	195 Å	**0.87**	0.67	0.75
			284 Å	0.81	0.78	0.80
		UAD	304 Å	0.82	0.78	**0.80**
			195 Å	0.88	**0.78**	**0.83**
		SPOCA	304 Å	**0.79**	**0.78**	**0.79**
			195 Å	0.85	0.77	**0.81**
		UAD	304 Å	0.78	0.74	0.76
			171 Å	0.76	0.76	0.76
			284 Å	0.79	0.78	0.78
		UAD	304 Å	**0.93**	0.69	0.79
			171 Å	**0.94**	0.66	0.78
			195 Å	**0.91**	0.72	0.80
			284 Å	**0.93**	0.66	0.77
SPOCA	Early – concat.	SPOCA	171 Å	0.54	**0.93**	0.68
			195 Å	0.58	**0.82**	0.68
[26] using Faster RCNN (ResNet50)	Sequential fine-tuning	UAD	304 Å	0.73	**0.83**	0.78
			171 Å	0.80	**0.90**	0.84
			195 Å	0.83	0.72	0.77
			284 Å	0.86	0.80	0.83

Table 3 F1-scores of single image band based detectors against MLMT-CNN with different fusion strategies over BraTS-prime (with 1 slice gap). All detectors are based on ResNet50. For each band, the highest scores are highlighted in bold.

Bands	Faster RCNN	MLMLT-CNN (Early - addition)	MLMLT-CNN (Early - concat.)	MLMLT-CNN (Late - concat.)
T1Gd	0.73	0.74	0.83	**0.89**
T1	0.54	0.78	0.89	**0.91**
T2	0.56	0.76	0.86	**0.89**
Flair	0.48	0.75	0.86	**0.91**

3934 Å, where Magnetograms observe a lower altitude than PM/SH 3934 Å. This demonstrates that the these bands are not equal in how difficult they may be analysed, even though they were acquired at the same time with same size and resolution. These observations suggest that detecting ARs using information provided by a single band may be an under-constrained problem.

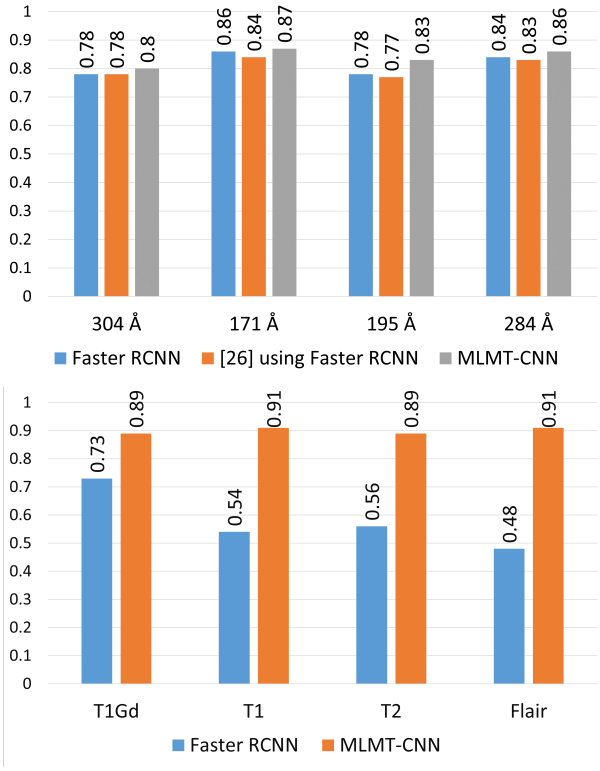

Fig. 5 Comparison of the detection results over UAD (top) and BraTS-prime (bottom) datasets. Each group of bars represents an imaging modality. Different colors represent different methods.

3.2.2 Joint detection on multiple image bands

We now present the results of our framework when detecting ARs over the UAD bands jointly. We experiment with different types of feature fusion and different combinations of bands. We compare against the state-of-the-art AR detector HFC's SPOCA [10]. We further compare against a sequential fine-tuning method derived from [26] through adapting the first stage of their approach to Faster RCNN by sequentially fine tuning it over the neighbouring image bands. We evaluate this approach on UAD. Moreover, we compare against Faster RCNN on single bands to demonstrate the benefit of jointly processing the image bands, taking into account their inter-dependencies for more robust individual detections.

In our first experiment, we compare early fusion (pixel level concatenation) against late fusion (feature level concatenation or addition), on the LAD dataset. Overall, the three approaches show an enhanced performance in contrast to single band based detection. However, we find that late fusion with concatenation shows higher performance than early fusion, having 0.90 F1-score versus 0.88 for magnetograms, while both scored 0.89 over 3934 Å. We further test late fusion using el-

Table 4 Performance of single image segmentation over BraTS-prime. For each class, the highest scores are highlighted in bold.

Architecture	Supervision	Bands	IoU score per class			Mean
			NCR/NET	ED	ET	IoU
FCN8	Fully supervised	T1Gd	0.54	0.43	0.70	0.56
		T1	0.08	0.33	0.0	0.14
		T2	0.49	0.48	0.23	0.40
		Flair	0.43	0.51	0.19	0.38
U-Net	Fully supervised	T1Gd	**0.69**	0.52	**0.80**	**0.67**
		T1	0.56	0.50	0.19	**0.42**
		T2	0.63	0.56	0.36	**0.52**
		Flair	0.50	**0.59**	0.29	**0.46**
U-Net	Weakly supervised	T1Gd	0.66	0.33	0.53	0.51
		T1	0.58	0.39	0.0	0.32
		T2	0.58	0.43	0.1	0.37
		Flair	0.44	0.49	0.0	0.31

Table 5 Segmentation performance of MLMT-CNN (U-Net) with full supervision over BraTS-prime for different numbers of modalities and feature fusions. For each class, the highest scores are highlighted in bold.

Architecture	Fusion	Slice gap	Bands	IoU score per class			Mean
				NCR/NET	ED	ET	IoU
MLMT-CNN (U-Net)	Early - concat.	1	T1	0.60	0.56	0.41	0.52
			T2	0.59	0.59	0.39	0.52
			T1	0.62	0.59	0.35	0.52
			T2	0.63	0.61	0.36	0.53
			Flair	0.63	0.63	0.39	0.55
			T1Gd	**0.75**	**0.66**	0.78	**0.73**
			T1	**0.75**	**0.69**	**0.76**	**0.73**
			T2	**0.74**	**0.71**	**0.70**	**0.72**
			Flair	**0.73**	**0.68**	0.62	**0.68**
	Early - addition	1	T1Gd	0.74	0.64	0.78	0.72
			T1	0.73	0.67	0.74	0.71
			T2	0.69	0.66	0.65	0.67
			Flair	0.68	0.67	0.61	0.65
	Late - concat.	1	T1Gd	0.71	0.63	0.81	0.72
			T1	0.73	0.65	0.74	0.71
			T2	0.71	0.68	**0.70**	0.70
			Flair	0.69	0.67	**0.64**	0.67
	Late - addition	1	T1Gd	0.70	0.60	0.81	0.70
			T1	0.71	0.63	0.73	0.69
			T2	0.67	0.66	0.67	0.67
			Flair	0.68	0.66	0.60	0.65
	Early - concat.	2	T1Gd	0.68	0.55	0.76	0.66
			T1	0.67	0.61	0.66	0.65
			T2	0.64	0.65	0.55	0.61
			Flair	0.57	0.62	0.46	0.55
		3	T1Gd	0.63	0.51	0.73	0.62
			T1	0.63	0.60	0.62	0.62
			T2	0.57	0.64	0.43	0.55
			Flair	0.59	0.61	0.41	0.54
[26] using U-Net	Sequential fine-tuning	1	T1Gd	0.71	0.55	**0.82**	0.69
			T1	0.56	0.50	0.19	0.42
			T2	0.65	0.58	0.26	0.50
			Flair	0.57	0.61	0.33	0.50

Table 6 Evaluation of weakly supervised MLMT-CNN (U-Net) on BraTS-prime. For each class, the highest scores are highlighted in bold.

# train. stages	Bands	IoU score per class			Mean
		NCR/NET	ED	ET	IoU
1	T1Gd	0.67	0.40	0.38	0.48
	T1	0.66	0.41	**0.40**	0.49
	T2	0.62	0.45	0.39	0.49
	Flair	0.64	0.46	0.38	0.49
2	T1Gd	**0.69**	0.43	**0.40**	**0.51**
	T1	**0.69**	0.41	**0.40**	**0.50**
	T2	0.66	0.45	0.38	**0.50**
	Flair	0.67	**0.47**	0.38	**0.51**
3	T1Gd	0.67	0.40	0.37	0.48
	T1	0.67	0.40	0.37	0.48
	T2	0.64	0.42	0.36	0.47
	Flair	0.64	0.45	0.34	0.48

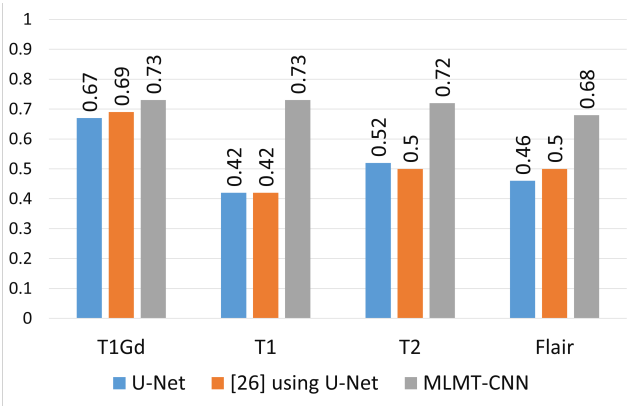

Fig. 6 Comparison of the segmentation results over BraTS-prime dataset. Each group of bars represents an imaging modality. Different colors represent different methods.

Table 7 Comparison of full and weak supervision for MLMT-CNN (U-Net) over weak-Cloud-38. For each band, the highest scores of the weakly-supervised models are highlighted in bold.

Super-vision	# train. stages	IoU score per band				Mean
		Red	Green	Blue	NIR	IoU
Fully	NA	0.95	0.95	0.95	0.95	0.95
Weakly	1	0.78	**0.80**	**0.83**	**0.83**	**0.81**
	2	**0.79**	**0.80**	**0.83**	**0.83**	**0.81**
	3	0.78	**0.81**	0.82	**0.83**	**0.81**

ement wise addition and observe a decrease of 1% and 3% in the F1-score over 3934 Å and Magnetogram, respectively. Late fusion is thus adopted for all following experiments.

We also evaluate the benefit of our MOO strategy using our 2-band based architecture on the UAD dataset. As seen in Table 2, this approach generally improves the F1-scores in most bands comparing to the non-MOO architectures. This behaviour may indicate that the two feature extraction stages were indeed more effectively optimised for their different tasks at different epochs. Thus we use this MOO approach for all other experiments.

On the UAD dataset, with various combinations of 2 bands, we notice a general improvement over single band detections. In addition, the performance varies in correspondence to the bands being used. Combining bands that are difficult to analyse (304 Å or 195 Å that have lowest F1-scores in the single band analyses) with easier bands (171 Å and 284 Å) unsurprisingly enhances their respective performance. More interestingly, combining the difficult 304 Å and 195 Å bands together also improve on their individual performance. Similarly, when combining bands that are easier to analyse (171 Å and 284 Å), performances are also improved over their individual analyses. Following these settings, our 2-band based approach was able to record higher

or similar F1-scores in contrast to the best performing single-band detector. This supports our hypothesis that joint detection may provide an increased robustness through learning the inter-dependencies between the image bands. Moreover, the most dramatic improvement in F1-scores across both LAD and UAD datasets is for the 3934 Å images when magnetograms are added to the analysis. This is in line with the current understanding of AR having strong magnetic signatures.

Generally, in the UAD dataset, we find that using a combination of 2 bands produces the best F1 scores in comparison to using 3 or 4 bands in the analysis, see Table 2. This may be caused by the fact that optimising the network for multiple tasks (2, 3, or 4 detection tasks) simultaneously increases the complexity of the problem. While the network successfully learned to produce better detections in the case of 2 bands, it was difficult to find a generalised yet optimal model for 3 or 4 bands at the same time. Thus, for 4 bands, the model obtains the best precision but at the expense of a poor recall.

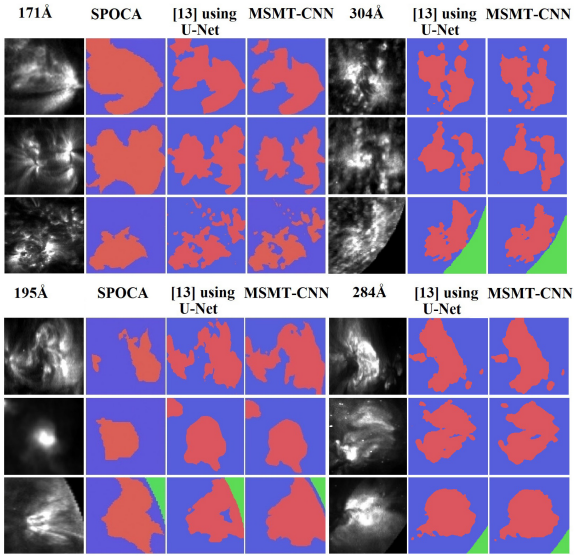

Fig. 7 AR segmentation comparison between our presented method, SPOCA, and sequentially fine-tuned DNNs similar to [26], over the SPOCA subset. Red is AR, blue denotes the quite Sun background, and green is outside of the solar disk.

On the SPOCA subset, over the bands 171 Å and 195 Å for which it is designed, the SPOCA method obtains the poorest performance of all multi-band and single-band experiments. It is worth noting that this method relies on manually tuned parameters according to the developers' own definition and interpretation of AR boundaries, which may differ from the ones we used when annotating the dataset. While supervised DL-based methods could integrate this definition during training, SPOCA could not perform such adaptation. This may have had a negative impact on its scores. Furthermore, visual inspection shows a poor performance for SPOCA on low solar activity images, see Fig. 2. This may be due to the use of clustering in SPOCA, since in low activity periods the number of AR pixels (if any) is significantly smaller than solar background pixels, which makes it difficult to identify clusters.

Moreover, the sequential fine tuning approach similar to [26] shows a close performance to single band detection using Faster RCNN with an identical precision, recall and F1-score over the band 304 Å and a slight decrease over the other 3 bands, See Table 2 and Fig. 5. This may be due to the fact that its transfer learning does not incorporate the bands' inter-dependencies when analysing the different bands. Moreover, the method was designed in [26] to produce a single prediction for the different bands, this differs from our usage where we predict a different set of detections per band.

We further evaluate our detection approach with different fusions, over the 4 bands of BraTS-prime dataset, and compare it against single band based detection. All fusion strategies significantly outperform single band detectors, with late concatenation fusion being the highest, showing an average F1-score increase of 39% across all modalities. See Table 3 and Fig. 5. This confirms our hypothesis that exploiting inter-dependencies between the image bands by the joint analysis may provide a superior performance in contrast to single band based detection.

3.3 Segmentation stage

Our AR segmentation results were all qualitatively assessed and validated by a solar physics expert. We also visually compare the results against SPOCA and a sequentially fine-tuned U-Net model (similar to the first stage of [26]).

Additionally, to quantitatively demonstrate the benefit of the joint analysis, and due to the lack of manual AR pixel-wise ground-truth, we evaluate our approach using the BraTS-prime synthetic dataset. Weak-Cloud-38 may not be used for this purpose because of its different bands capturing the same scene, rather than different layers of a 3D object. It is worth noting that we do not aim to achieve state-of-the-art performance in tumour segmentation, but rather to confirm the benefit of the joint analysis in scenarios similar to our solar case, where different modalities show different cuts of a 3D object. Since ground-truth is available for this dataset, we follow the classical fully-supervised training procedure. Furthermore, we use Weak-BraTS-prime

and Weak-Cloud-38 to evaluate our iterative training strategy from weak labels against full supervision.

Its worth noting that the segmentation subnetworks adopt the same layers configuration of their correspondent blocks in U-Net [29].

3.3.1 Independent segmentation on single image band

We first compare segmentation results produced by U-Net and FCN8 over the AR and BraTS-prime (Table 4) individual image bands, analysed independently, to evaluate different DL-based segmentation networks. These results also serve as baseline to assess our joint analysis based approach in Section 3.3.2.

We notice that U-Net produces higher IoU values over all bands for BraTS-prime, as well as smoother AR boundaries, compared to FCN8. This is expected since U-Net utilises skip connections to help retrieving fine details in the mask reconstruction process. Therefore, we use the building blocks of U-Net in our joint segmentation framework.

When comparing the results of U-Net over different modalities, we notice that the T1-Gd modality gets the highest IoU score for the ET class. A similar trend can be seen when comparing the results of the NCR/NET class over different modalities. On the other hand, we find that Flair gets the highest IoU for the ED class comparing to the other modalities. This contrast in the IoU scores is in line with the understanding that different modalities provide different information.

3.3.2 Joint segmentation on multiple image bands

Similar to our detection experiment, we assess our framework using different combinations of image bands and different types of feature fusion to evaluate their influence on the segmentation performance.

Quantitative results First, we present our BraTS-prime segmentation on combined bands using our joint analysis approach (Table 5). We note that all combinations improve on the single-band results, with the best improvement coming from combining all four modalities. All following BraTS-prime experiments use a four-band architecture.

We compared four fusion strategies, namely fusing features after one block of convolution only (early) and at the end of convolutions (late), using addition and concatenation. We find that early fusion with concatenation shows higher results. This differs from our observation in the AR detection experiment, hence confirming that the fusion strategy needs to be adapted to the analysis scenario. Accordingly, we continue using early fusion with concatenation for all BraTS-prime segmentation experiments.

As expected, there is a negative correlation between the IoU score and the width of slice gap, where the overall increase in the IoU was the highest for smaller gaps and higher levels of spatial correlation (gap of 1 pixel). This observation, together with the improved results from combining bands, suggest that jointly analysing related multi-modal images in scenarios similar to our solar case may indeed aid the network in learning the inter-dependencies between the different modalities.

We compare against sequentially fine-tuned U-Net models similar to the first stage of [26] in Table 5 and Fig. 6. They achieved comparable IoU scores to those produced by U-Net on single bands. Hence, they do not benefit from the combination of modalities as our framework does.

Additionally, as a mean to assess our iterative training steps, we use weak-BraTS-prime and weak-Cloud-38 to evaluate this strategy against manual annotations, and compare it to the classical training approach.

When evaluating the recursively trained model using weak-BraTS-prime dataset against the fully supervised model on BraTS-prime manual annotations, we notice an increase in the IoU scores after one step of recursion (i.e. 2 stages of training, first using the weak labels, then using the previous predictions as labels), achieving 71% of the fully supervised performance (Table 6). Moreover, this iterative training process achieves 85% of the fully supervised approach over the Weak-Cloud-38 dataset, with the best performance also being after one round of recursion, with an increase of 1% over the Red band (Table 7). These observations indicate that our recursive training strategy is beneficial in cases where manual annotations are not available, such as solar ARs.

In contrast to the single band based segmentation of weak-BraTS-prime (last 4 rows of Table 4), we also note that performance still benefits from the joint analysis even when trained – classically or recursively – with weak labels (Table 6).

Qualitative results Lastly, we compare visually our segmentation results on the SPOCA subset, using our proposed architecture, against SPOCA and sequentially fine-tuned DNNs similar to [26] (without their final stage of fusing the CNNs' individual predictions) (Fig. 7). The results show that our framework generally finds more detailed AR shapes than SPOCA, while at the same time being more robust to fainter regions of ARs.

Additionally, we compare our AR segmentation results to SPOCA by finding the IoU between the predictions produced by the two approaches over the SPOCA subset. This may be used to indicate the agreement between the two methods. We find that both 171 Å and 195 Å achieve a higher agreement of 44% and 46%, respectively, in contrast to 304 Å and 284 Å scoring 33% and 41%, respectively. This is expected since SPOCA was designed to segment ARs in 171 Å and 195 Å. Overall, the similarity between our predictions and SPOCA's is relatively low. However, as discussed in Section 3.2.2, SPOCA was manually tuned by the developers according to their own interpretation of AR boundaries which may be different from our interpretation when annotating the dataset. Hence, care must be taken when interpreting the results.

Comparison against sequentially fine-tuned CNNs in the spirit of [26] is fairer, since the DNNs were trained on our data. Segmentation of the sequentially fine-tuned CNNs appears to be of similar quality to ours, although shapes of an AR between neighbouring bands evolve more smoothly with our method. This is an advantage of accounting for the 3D geometry of ARs in performing the 2D segmentation.

4 Conclusion

We presented a multi-layer and multi-tasking framework to tackle the 3D solar AR detection and segmentation problem from multi-spectral images that observe different layers of the 3D solar atmosphere. MLMT-CNN analyses multiple bands jointly to produce consistent localisation. It is a flexible framework that may use different CNN backbones, and may be generalised to any number and modalities of images. We find that by fusing information from different image bands at different feature levels, CNNs were able to localise objects more robustly and more consistently across layers. Additionally, our study suggests that different imaging scenarios may require different types of feature fusion strategies. We also show that the number of bands used in the analysis might affect the performance and must be optimised to each case. Furthermore, we demonstrate that CNNs may show a satisfactory localisation performance when iteratively trained from weak annotations. MLMT-CNN showed competitive results against both baseline and state-of-the-art detection and segmentation methods. Future research could investigate the information importance of different image bands and its influence on task learning in both multi-spectral and multi-layer scenarios.

References

1. S. Hwang et al. Multispectral pedestrian detection: Benchmark dataset and baselines. In CVPR, 2015.
2. J. Wagner et al. Multispectral pedestrian detection using deep fusion convolutional neural networks. In ESANN, 2016.
3. S. Mohajerani and P. Saeedi. Cloud-Net: An end-to-end cloud detection algorithm for Landsat 8 imagery. In IGARSS, 2019.
4. S. Mohajerani, T. A. Krammer, and P. Saeedi. A cloud detection algorithm for remote sensing images using fully convolutional neural networks. In IEEE MMSP, 2018.
5. T. Ishii et al. Detection by classification of buildings in multispectral satellite imagery. In ICPR, 2016.
6. O. A. B. Penatti, K. Nogueira, and J. A. dos Santos. Do deep features generalize from everyday objects to remote sensing and aerial scenes domains? In CVPR, 2015.
7. K. Takumi et al. Multispectral object detection for autonomous vehicles. In Thematic Workshops, 2017.
8. K. Revathy, S. Lekshmi, and S. R. P. Nayar. Fractal-based fuzzy technique for detection of active regions from solar images. Solar Phys., 2005.
9. A. Benkhalil et al. Active region detection and verification with the solar feature catalogue. Solar Phy., 2006.
10. C. Verbeeck et al. The SPoCA-suite: Software for extraction, characterization, and tracking of active regions and coronal holes on EUV images. A&A, 2013.
11. R. Krishnapuram and J. Keller. The possibilistic C-means algorithm: Insights and recommendations. IEEE TFS, 1996.
12. Paul Viola and Michael J Jones. Robust real-time face detection. IJCV, 2004.
13. N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In CVPR, 2005.
14. J. Redmon et al. You only look once: Unified, real-time object detection. In CVPR, 2016.
15. W. Liu et al. SSD: Single shot multibox detector. In ECCV, 2016.
16. J. Dai et al. R-fcn: Object detection via region-based fully convolutional networks. In Advances in Neural Information Processing Systems. Curran Associates, Inc., 2016.
17. H. Law and J. Deng. Cornernet: Detecting objects as paired keypoints. In ECCV, 2018.
18. S. Ren and J. others. Faster R-CNN: Towards real-time object detection with region proposal networks. In NIPS, 2015.

19. P. Soviany and R. Ionescu. Optimizing the trade-off between single-stage and two-stage deep object detectors using image difficulty prediction. In SYNASC. IEEE, 2018.
20. J. Huang et al. Speed/accuracy trade-offs for modern convolutional object detectors. In CVPR, 2017.
21. M. Gani et al. Multispectral object detection with deep learning, 2021.
22. Z. Guo et al. Deep learning-based image segmentation on multimodal medical imaging. IEEE TRPMS, 2019.
23. K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. arXiv preprint arXiv:1406.2199, 2014.
24. A. Eitel et al. Multimodal deep learning for robust rgb-d object recognition. In IEEE IROS, 2015.
25. X. Song et al. A multispectral feature fusion network for robust pedestrian detection. Alexandria Enginering Journal, 2021.
26. R. Jarolim et al. Multi-channel coronal hole detection with a convolutional neural network. In ML-Helio, 2019.
27. M. Almahasneh et al. Active region detection in multi-spectral solar images. In ICPRAM, 2021.
28. E. Shelhamer, J. Long, and T. Darrell. Fully convolutional networks for semantic segmentation. IEEE TPAMI, 2017.
29. O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. 2015.
30. A. Khoreva et al. Weakly supervised semantic labelling and instance segmentation. 2016.
31. Q. Li, A. Arnab, and P. HS. Torr. Weakly-and semi-supervised panoptic segmentation. In ECCV, 2018.
32. J. Dai, K. He, and J. Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In ICCV, 2015.
33. S. Wang et al. Weakly supervised deep learning for segmentation of remote sensing imagery. Remote Sensing, 2020.
34. A. Bansal et al. Pixelnet: Representation of the pixels, by the pixels, and for the pixels. arXiv preprint arXiv:1702.06506, 2017.
35. Kingma D. P. and Ba J. Adam: A method for stochastic optimization. In 3rd International Conference on Learning Representations, ICLR 2015, 2015.
36. B. H. Menze et al. The multimodal brain tumor image segmentation benchmark (BRATS). IEEE T-MI, 2015.
37. B. Crabbe et al. Skeleton-free body pose estimation from depth images for movement analysis. In ICCVW, 2015.