International Journal of Applied Earth Observation and Geoinformation 124 (2023) 103499

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

journal homepage: www.elsevier.com/locate/jag

A dual-branch weakly supervised learning based network for accurate mapping of woody vegetation from remote sensing images

Youwei Cheng [a,1], Shaocheng Lan [a,1], Xijian Fan [a,*], Tardi Tjahjadi [b], Shichao Jin [c], Lin Cao [a,d]

[a] College of Information Science and Technology, Nanjing Forestry University, Nanjing, 210037, Jiangsu, China
[b] School of Engineering, The University of Warwick, Coventry, CV4 7AL, West Midlands, UK
[c] Academic for Advanced Interdisciplinary Studies, Nanjing Agriculture University, Nanjing, 210095, Jiangsu, China
[d] College of Forestry, Nanjing Forestry University, Nanjing, 210037, Jiangsu, China

ARTICLE INFO

Keywords:
Vegetation remote sensing
Weakly supervised learning
Semantic segmentation
Environment monitoring

ABSTRACT

Mapping woody vegetation from aerial images is an important task bluein environment monitoring and management. A few studies have shown that semantic segmentation methods involving deep learning achieve significantly better performance in mapping than methods involving field-based measurement and handcrafted features. However, current deep networks used for mapping vegetation require labour-intensive pixel-level annotations. Thus, this paper proposes the use of image-level annotations and a weakly supervised semantic segmentation (WSSS) network for mapping woody vegetation based on Unmanned Aerial Vehicle (UAV) imagery. The network comprises a Localization Branch (LB) and an Attention Relocation Branch (ARB). The LB is trained in stage 1 of the mapping to identify regions with the most discriminative vegetation, while the ARB is introduced to better mine semantic information, which enhances the ability of the class activation maps (CAMs) to represent useful information. The ARB inherits the weights from the LB in stage 2 and uses a Multi-layer Attention Refocus Structure (MARS) into the network to expand the receptive field to enable the model to process global features. Thus, same-category regions that are located farther apart are better captured. Finally, the region focused by the dual branches are integrated to more accurately cover the areas to be segmented. Using UAV imagery datasets, namely UOPNOA and MiniFrance, along with quantitative metrics and qualitative results, the network demonstrates performance better than existing state-of-the-art related methods. The effectiveness and generalization of each module of the network are validated by ablation experiments. The code for implementing the network will be accessible on https://github.com/Mr-catc/DWSLNet.

1. Introduction

Woody vegetation constitutes a vital component of global ecosystems, playing a pivotal role in facilitating the circulation of energy and nutrients within the ecosystems, enhancing water quality, and contributing to the regulation of floods and land erosion (Chapin et al., 2011). However, the combined impacts of climate change and human activities have adversely affected woody vegetation, thereby undermining the overall stability of the ecosystems. Consequently, accurate and timely monitoring of woody vegetation is crucial for promoting social sustainability and maintaining ecological balance.

Vegetation mapping is an essential means of monitoring woody vegetation (Chen et al., 2023; Wu et al., 2022). The traditional way for mapping woody vegetation involves field-based measurements, which tends to be time-consuming and labour-intensive, particularly when applied over large areas. Remote sensing (RS) technology, especially satellite imagery, has proven to be an effective tool for addressing the challenges encountered by traditional vegetation mapping methods (Shen et al., 2021; Shafeian et al., 2021; Xiao et al., 2021). More recently, new technologies such as the combination of Unmanned Aerial Vehicle (UAV) and deep learning (Li et al., 2022), have shown promising value for accurately mapping woody vegetation. Fig. 1 shows example images from two datasets acquired by UAV, namely UOPONA (Pedrayes et al., 2021) and MiniFrance (Castillo Navarro et al., 2020), that we investigated for such mapping.

The recent advancement of deep learning methods has facilitated substantial progress in the research and application of vegetation mapping. For instance, Trenčanová et al. (2022) introduced an approach by utilizing convolutional neural network based framework (U-Net) for forest fire management, planning, and prevention to automatically detect shrub coverage in high-resolution images captured by

* Corresponding author.
E-mail addresses: xijian.fan@njfu.edu.cn (X. Fan), T.Tjahjadi@warwick.ac.uk (T. Tjahjadi), ginkgocao@gmail.com (L. Cao).
[1] These authors contribute equally to these work.

https://doi.org/10.1016/j.jag.2023.103499
Received 18 June 2023; Received in revised form 16 August 2023; Accepted 18 September 2023
Available online 29 September 2023

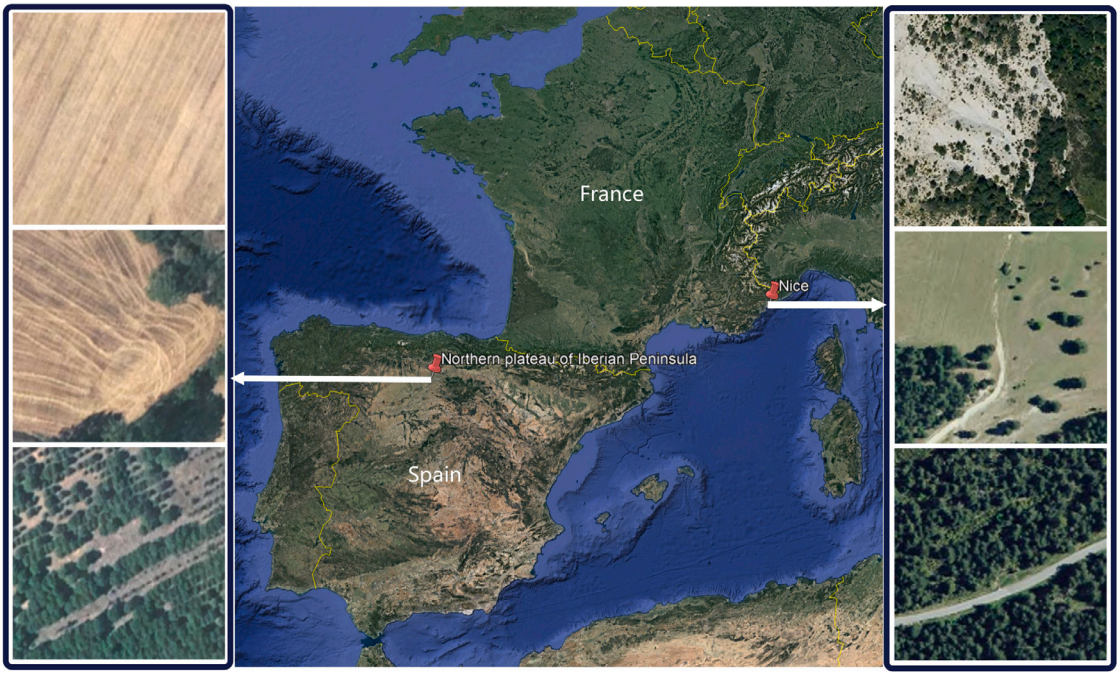


Fig. 1. Example images from UOPONA and MiniFrance datasets.

an UAV. Wang et al. (2021) investigated optimal methods for early detection and classification of invasive eastern red cedar using UAV imagery and semantic segmentation algorithms in multi-species forests. However, these methods depend on strongly supervised learning which necessitates the presence of pixel-level annotations during the training. The creation of such data involves significant human effort, making the creation of such a dataset extremely expensive.

To address this issue, a weakly supervised semantic segmentation (WSSS) approach employing labels at the image level is proposed in our study, which is specifically designed for woody vegetation. This approach only requires image-level labels, which is easier to be obtained than pixel-level labels. Fig. 2 shows the basic stages, where class activation maps (CAMs) (Zhou et al., 2016) only requires the original image and image-level labels to generate pseudo-labels. CAMs are capable of highlighting the feature regions that are sensitive for classification tasks, and obtaining the initial class-specific areas. The method is then extended to produce pseudo labels (Ahn et al., 2019), which can be used in fully supervised semantic segmentation. Overall, there are two challenges in designing a weakly supervised method for mapping woody vegetation.

The first challenge is due to woody vegetation in RS images having high intra-class heterogeneity and low inter-class heterogeneity. For example, Fig. 3(a) shows sparse gaps without canopy cover may exist within forest areas, while areas with canopy cover may occur within shrub areas. Furthermore, as demonstrated in Fig. 3(b), distinguishing between foreground and background in some images is challenging. To overcome this, we incorporate Multi-layer Attention Refocus Structure (MARS) into our network, which enhances the ability of the network to extract contextual information and reassign weights to channels and spaces by incorporating position information into channel attention. This significantly improves the classification performance of the network for distinguishing foreground and background with similar features and different woody vegetation classes, as well as the ability of the network to accurately respond to different features in CAMs.

The second challenge is due to the features of woody vegetation RS images being heavily repetitive and having a multi-region distribution. However, the CAM generated by traditional classification networks can only focus on the most discriminative region, which significantly reduces the accuracy of weakly supervised methods. Thus, we propose a novel dual-branch network which incorporates MARS. The two branches of this network are the Localization Branch (LB) and Attention Relocation Branch (ARB). The LB is similar to a traditional classification network, emphasizing the most significant feature regions. A training strategy involving two stages is implemented. First, the LB is trained, and then the ARB is trained by adding MARS based on the inherited weights of the LB. By modelling long-range dependencies more effectively, the MRAS enables the ARB to focus on crucial regions for segmentation that may be overlooked by the LB. Additionally, we add the CAM generated by the LB to the training of the ARB as an extra guidance, enabling the ARB branch to respond more accurately to the boundaries of the region to be segmented. Finally, we integrate the CAM generated by both the LB and ARB to more accurately cover the region to be segmented.

Our contributions in this paper are as follows:

- We have pioneered the application of WSSS for mapping woody vegetation using only labels at the image level, and addressed the problem associated with the level of heterogeneity in woody vegetation, which causes to a certain extent the difficulty of segmentation using image-level annotations.
- We present a novel dual-branch integrated framework. The LB is initially trained to identify crucial target regions, followed by the ARB inheriting the weights of the LB for a distinct perspective. Ultimately, the integrated CAMs from the dual-branch significantly improve the precision of the segmentation.
- We propose ARB, which relocates the target area by fusing MARS, to facilitate the responsiveness of the network to large-scale feature-repeated areas, and accurately locate multiple segmented regions. Thus, enhancing the ability of the model to effectively segment the target.
- The effectiveness of the network is verified via experiments on UOPNOA and the MiniFrance datasets. We compare our method with other WSSS networks through qualitative and quantitative evaluations of their performance. The results show a substantial enhancement in the segmentation performance of our method over other networks.

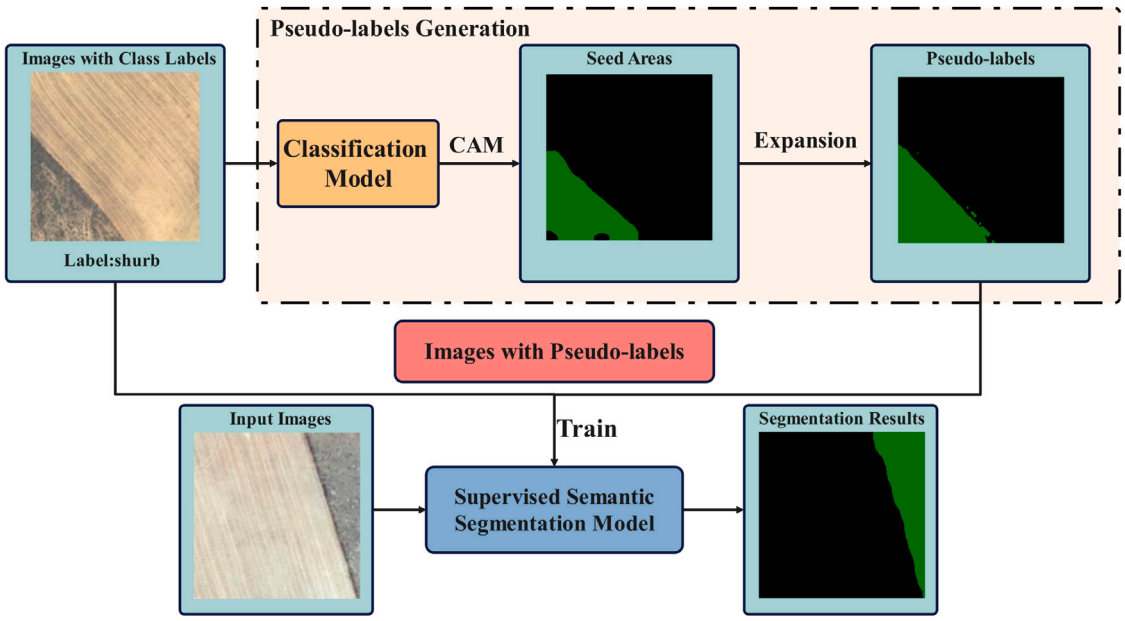


Fig. 2. Weakly supervised semantic segmentation pipeline.

(a) (b)

Fig. 3. Example images of the first challenge: (a) high intra-class heterogeneity; and (b) low inter-class heterogeneity.

2. Related work

2.1. Weakly supervised semantic segmentation

It is prohibitively expensive to generate pixel-level annotations. To address this challenge, weakly supervised methods that require minimal human intervention have been developed to make them more practical and accessible. These methods aim to strike a balance between data limitations and maintaining the accuracy of semantic segmentation tasks. They can be classified according to the type of annotations used: scribble (Lin et al., 2016), image-level (Ahn and Kwak, 2018), bounding box (Dai et al., 2015), and point (Bearman et al., 2016). It is essential to use the least human intervention for image-level annotations for the method to be accessible. Hitherto, most studies related to image-level annotations involve leveraging CAMs produced by classification networks. From the CAMs, high-quality pseudo-labels are generated and subsequently employed to train full supervised semantic segmentation networks. The research directions in this field can be broadly classified into two categories: generating more accurate initial seed regions and refining pseudo labels to obtain greater precision.

On the one hand, certain methods focus on expanding CAMs that only attend to salient regions. For instance, adversarial erasing (Wei et al., 2017) trains a network on more challenging regions by selectively removing the most easily identifiable areas. Multiple attention maps are integrated at different stages of the training of the OAA method (Jiang et al., 2019). The quality of segmentation by SEAM (Wang et al., 2020) is enhanced by reducing the disparity between the affine transformed and the original image result. On the other hand, some networks refine pseudo labels to obtain more accurate ones. AffinityNet (Ahn and Kwak, 2018) learns the semantic similarity between neighbouring coordinates and achieves semantic propagation through a random walk strategy. IRNet (Ahn et al., 2019) generates transformation matrices from boundary activation mapping. Furthermore, some approaches employ additional supervision. Approaches (Jiang et al., 2022; Lee et al., 2021b) utilize saliency maps as extra supervision, providing relevant foreground and background information to the model. CLIMS (Xie et al., 2022) introduces natural language supervision by employing contrastive language-image pre-training models to obtain more accurate CAM.

From the perspective of network architecture, existing weakly supervised methods mostly adopt a single-branch or double-branch framework. Single-branch networks add modules after traditional classification networks, which can easily lead to the loss of some original information. Double-branch networks add special structures to one branch and train two mutually influential branches in parallel. However, due to the indistinct features of woody vegetation in RS images, the branch with the added special structure may easily introduce significant errors during training, affecting the training of the other branch. In contrast, our method innovatively employs a double-branch phased training approach. The LB branch focuses accurately on the

most salient features of the segmentation area but struggles to cover the entire area to be segmented. Therefore, we introduce the ARB branch, which inherits the weights from the LB branch and incorporates MARS, enabling it to pay attention to a broader range of areas to be segmented. Finally, we integrate the CAMs generated by the two branches through weighted aggregation, resulting in high-quality CAMs.

2.2. Mapping woody vegetation in remote sensing imagery

Mapping woody vegetation cover represents a significant domain within RS imagery. Owing to complex terrains and diverse vegetation types, mapping woody vegetation has consistently faced numerous challenges. In traditional processing of woody vegetation RS images, manual feature extraction methods are prevalent, such as those based on spectral features (De Petris et al., 2019; Fisher et al., 2017). Due to the advancement in machine learning, methods that combine manual feature extraction with machine learning techniques have been widely adopted (Purwanto et al., 2022; Zhou et al., 2021; Nasiri et al., 2023). Additionally, the fusion of multi-source RS data (multispectral, high-resolution, LiDAR, etc.) has emerged as a significant research trend (Yang et al., 2021; Rüetschi et al., 2021; Schindler et al., 2021). However, traditional machine learning algorithms typically classify images based on shallow features, providing scope for performance enhancement.

The recent development of deep learning, benefiting from its automated feature extraction capabilities and robust generalization performance, has motivated the direct application of deep learning algorithms to woody vegetation RS imagery in order to obtain the corresponding mapping of woody vegetation cover. These studies generally target visible light band woody vegetation RS images and convert woody vegetation cover mapping tasks into semantic segmentation. Numerous methods have achieved promising results in the RS domain by adapting semantic segmentation models which were initially intended for natural scene images, e.g., U-Net (Ronneberger et al., 2015) and FCN (Long et al., 2015) have achieved high performance. Methods using U-Net-based semantic segmentation (Flood et al., 2019; Alzu'bi and Alsmadi, 2022; Waldeland et al., 2022) for woody vegetation in various regions have yielded favourable results. Research investigating the application of FCN architecture in woody vegetation mapping has also been conducted (La Rosa et al., 2021). However, these aforementioned methods depend on fully supervised learning which necessitates an enormous quantity of pixel-level annotations. The majority of existing woody vegetation RS images lacks such pixel-level annotations, leading to substantial costs when obtaining labelled data. To tackle the high labelling costs associated with fully supervised approaches, methods that require less annotation effort have increasingly garnered interest. For instance, Schmitt et al. (2020) attempted to train semantic segmentation models using lower-resolution annotation data to generate higher-precision results, achieving relatively good outcomes in global forest cover mapping. Additionally, Puthumanaillam and Verma (2023) proposed a few-shot semantic segmentation method for forest semantic segmentation, which somewhat reduced labelling costs. Overall, efforts to minimize labelling costs for mapping woody vegetation cover are still in their infancy. Concurrently, reducing labelling costs may diminish effective information during model training, lowering model accuracy. Therefore, considerable effort is required from researchers to minimize labelling costs while maintaining high model accuracy.

3. Methodology

First, we introduce the conventional method of generating CAMs. This is followed by the basic architecture of the dual-branch integration network incorporating MARS. The specific details of the MARS are then discussed. Finally, we provide the loss functions for the two-stage training process.

3.1. Preliminary (CAMs generation)

CAMs display specific attention regions for a given class within the input image $\mathcal{I}$. A multi-label classification network is used to encode the features of all classes. Since there is a corresponding image-level label y for each $\mathcal{I}$, the total number of classes is c. In training the network, the output of the last convolutional layer of the network is the feature $F(\mathcal{I})$ with N channels. Subsequently, the last fully connected layer of the network is replaced by a global average pooling layer, transforming the feature into a vector V. Finally, the vector of size N is converted into predicted soft labels through a 1×1 convolution with N input channels and c output channels. The specific process is

$$V = GAP(F(\mathcal{I})), \tag{1}$$

$$S_c = w_c^T \cdot V, \tag{2}$$

where $GAP(F(\mathcal{I}))$ performs global average pooling operation on the feature map, w_c^T are channel weights for class c, and S_c denotes the final predicted score for class c. The CAMs $M_c(I)$ for class c is generated using

$$M_c(\mathcal{I}) = w_c^T \cdot F(\mathcal{I}), \tag{3}$$

where $F(\mathcal{I})$ denotes the feature map before being input to GAP, and w_c^T denotes the weights learned by the network for class c. Fig. 4 illustrates the process of CAM.

However, as pointed out in most related works, while CAMs are able to locate the most discriminative regions, they are unable to detect challenging areas that are crucial for semantic segmentation tasks. This is because the network is trained for classification, which causes the generated CAMs to excessively focus on the most discriminative regions. As a result, the efficacy of weakly supervised semantic segmentation is greatly reduced. In this paper we address the issue by using a novel dual-branch integration network that incorporates MARS. By adding an attention-based relocation branch, the network can discover other target areas and achieve better segmentation results through weighted integration.

3.2. Network architecture

There are two branches, LB and ARB, in the network as shown in Fig. 5. Similar to previous works, the LB employs the classification loss function for optimization and generates Localization CAMs, $M_{LB}(\mathcal{I})$. Since the LB targets classification tasks, the most distinct features for classification are often activated during training. Consequently, the resulting CAMs highlight the most distinguishing areas of the object of interest.

Additionally, the MARS incorporated in ARB redistributes channel and spatial weights, thereby serving an "expansion" and "complementary" function with respect to the CAM attention areas of the LB. This approach addresses the task gap issue encountered in previous works that utilized single-branch classification CAMs for segmentation tasks and provides additional cues beneficial for semantic segmentation. More specifically, the ARB is trained by integrating MARS into the pre-trained LB network. A detailed introduction to MARS will be presented in the subsequent section. Owing to the larger receptive field offered by the attention mechanism, the ARB can better focus on context and capture more feature information globally. Furthermore, the MARS operates across multiple stages of the ARB, facilitating the identification of semantically similar information and the response to a more extensive range of features in the CAMs.

Since woody vegetation in RS images provides high intra-class heterogeneity and low inter-class heterogeneity, the two branches of the network provide distinct perspectives. This helps the network to focus on the various regions to be segmented, effectively mitigating the issue of single-branch CAMs only concentrating on the most salient regions. This is akin to the "multi-view" theory (Allen-Zhu and Li, 2020),

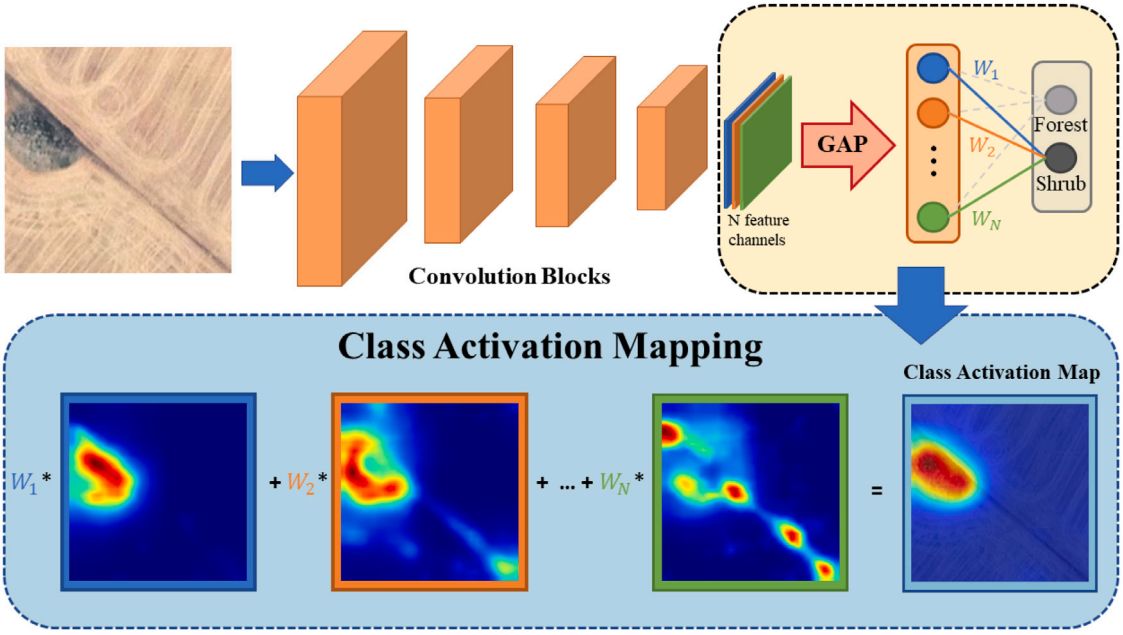


Fig. 4. The CAM process involves the weighting of each feature channel, resulting in the final CAM that reveals the network's region of interest.

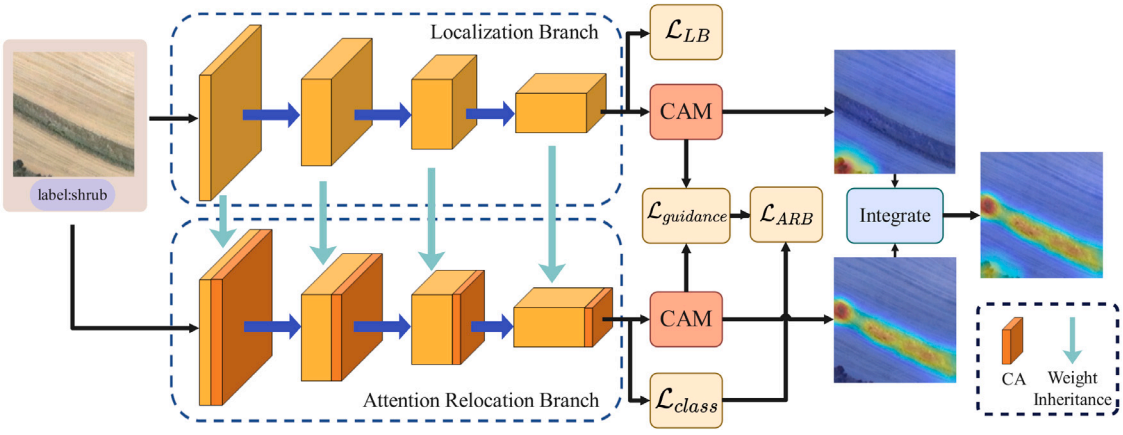


Fig. 5. The network architecture.

where similarly structured but separately trained models learn different sub-views of the features in the dataset, and the model integration effectively enhances accuracy. Finally, the integration of CAMs between the two branches can be represented as

$$M_{all}(I) = \beta M_{LB}(I) + (1-\beta) M_{ARB}(I), \tag{4}$$

where $M_{all}(I)$ represents the integrated CAMs, β denotes the weighting coefficient, $M_{LB}(I)$ represents the CAM output from the LB, and $M_{ARB}(I)$ represents the CAM output from the ARB.

Moreover, we observed that using $M_{LB}(I)$ as guidance during the training process of the ARB improves accuracy to a certain extent. Although there is currently no quantitative analysis on the impact of attention mechanism on the response region of CAMs, we believe this might be related to the imprecise response of the attention mechanism to boundary regions in classification networks. While the inclusion of the attention mechanism helps the model to understand the contextual semantic information, the larger receptive field also results in inaccurate responses of the generated CAMs in boundary regions. By incorporating guidance from the CAM generated by the LB, the CAM produced by the ARB is better constrained within the target segmentation area, thereby improving the performance of the network. The following pseudo code shows the training and generation process of our network.

3.3. Multi-layer attention refocus structure

Within our network, the ARB incorporates MARS to enable it to inherit the LB weights while reconstructing spatial and channel information across multiple stages of the network, offering distinct attention perspectives compared to LB and facilitating target region relocalization. This MARS-based relocalization approach is particularly effective for addressing the issue of a single branch focusing solely on the most prominent feature areas in woody vegetation RS images with extensive feature repetition or multiple separated regions. Moreover, the multi-layered structure of MARS enhances network robustness and generalizability compared to single-layer attention module, circumventing performance instability caused by single-pass attention mechanisms. Furthermore, MARS contributes to more pronounced target area responses in the CAM generated by ARB, ultimately improving the accuracy of pseudo label segmentation determined by CAM. We employ Coordinate Attention (CA) which was proposed by Hou et al.

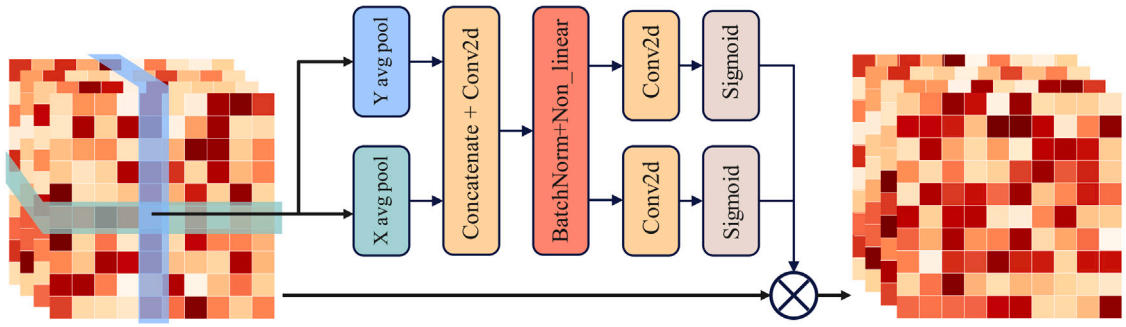


Fig. 6. Individual attention module in MARS enables synchronized embedding of channel and spatial information through coordinate pooling in both X and Y directions.

Algorithm 1: Pseudocode of DMAMNet

Input: n woody vegetation RS images with image-level labels $[D_1, \ldots, D_n]$
Output: CAMs $[M_1, \ldots, M_n]$
Training Phase 1:
Input labelled images and minimize the loss function $\mathcal{L}_{LB}$ to train LB.
Training Phase 2:
Integrate LB weights into ARB and incorporate MARS.
Input labelled images and minimize the loss function $\mathcal{L}_{ARB}$ to train ARB.
Generate CAMs:
for $i \leftarrow 1$ **to** n **do**
  Input test images D_i into LB to get predicted CAM M_{LB};
  $M_{LB} \leftarrow LB(D_i)$;
  Input test images D_i into the ARB to get predicted CAM M_{ARB};
  $M_{ARB} \leftarrow ARB(D_i)$;
  Fuse M_{LB} and M_{ARB} to generate the final CAM M_{all};
  $M_{all} \leftarrow \beta M_{LB} + (1-\beta) M_{ARB}$;
return CAMs $[M_1, \ldots, M_n]$;

(2021) as the individual module within MARS due to its lightweight efficacy. As a streamlined Channel and Spatial Attention mechanism, CA simultaneously redistributes spatial and channel weights, preventing long-distance dependency information loss caused by distribution modelling. Fig. 6 shows the basic structure of the individual attention module in MARS.

The CA involves two sequential phases: embedding coordinate information and generating coordinate attention. It captures long-range dependencies and channel correlations with accurate positional information, which aids the network in accurately localizing objects of interest. Typically, channel attention employs global average pooling to encode spatial information, resulting in the loss of substantial positional information. Coordinate information embedding, however, preserves positional information by utilizing one-dimensional average pooling in both vertical and horizontal directions. Specifically, it encodes each channel separately in the vertical and horizontal directions by employing two different convolution kernels, either $(h, 1)$ or $(1, w)$, i.e.,

$$z_c^h(h) = \frac{1}{W} \sum_{0 \le i < W} x_c(h, i), \tag{5}$$

$$z_c^w(w) = \frac{1}{H} \sum_{0 \le j < H} x_c(j, w), \tag{6}$$

where $x_c(h, w)$ represents the value at channel c, height h, and width w of the input X, while $z_c^h(h)$ denotes the average pooling result at height h for channel c, and $z_c^w(w)$ represents the average pooling result at width w for channel c.

Function F_1 is a shared 1×1 convolution transformation employed to concatenate the outputs of the above two pooling layers, i.e.,

$$f = \delta(F_1([z^h; z^w])), \tag{7}$$

where the concatenation operation along the spatial dimensions is denoted by $[\cdot;\cdot]$, δ denotes the non-linear activation function, and the combined encoding result obtained from both horizontal and vertical directions is denoted by f.

The resulting tensor is represented by two attention enhanced vectors, each with an identical number of channels, corresponding to the vertical and horizontal directions, respectively, i.e.,

$$f^h, f^w = Split(f). \tag{8}$$

Next, F_h and F_w are two 1×1 convolution transformations that are exploited to convert the channel numbers of f^h and f^w to match the channel count of the input tensor. This process can be represented as

$$s^h = \sigma(F_h(f^h)), \tag{9}$$

$$s^w = \sigma(F_w(f^w)), \tag{10}$$

where σ denotes the sigmoid activation function. Finally, the output of CA at location (i, j) for input X, i.e., $Y(i, j)$, is given by

$$Y(i, j) = X(i, j) \cdot s^h(i) \cdot s^w(j). \tag{11}$$

Thus, unlike conventional channel attention modules that only process the redistribution of different channel weights, CA also encodes spatial information. The pooling results generated in both directions allow each pixel to reflect whether the object of interest is present in the corresponding row and column. By employing MARS, the network has a larger receptive field without losing positional information, enabling more accurate modelling of long-range dependencies on a global scale. In addition, the inclusion of MARS allows the ARB to generate different attention perspectives from the LB, discovering some hard-to-identify target areas. Therefore, after integrating the dual branches, the CAM response regions become more accurate, leading to improved segmentation results.

3.4. Loss function

The predicted result Y is obtained at the beginning stage, which denotes the probability of predictions for all categories. It is optimized using $\mathcal{L}_{class}$, the multi-label soft margin loss given by

$$\mathcal{L}_{class} = -\frac{1}{N} \sum_{i=1}^{N} (\widetilde{Y}_i \log\left(\frac{1}{1+e^{-Y_i}}\right) + \left(1-\widetilde{Y}_i\right) \log\left(\frac{e^{-Y_i}}{1+e^{-Y_i}}\right)), \tag{12}$$

where N represents the categories count, $\widetilde{Y}_i$ represents the predicted label for category i, and Y_i represents the true label for category i.

After the first stage LB is trained, we begin training ARB. The training loss for the second stage ARB incorporates both $\mathcal{L}_{class}$ and the

guidance from the CAM generated by the LB. The guidance loss can be represented as

$$\mathcal{L}_{guidance} = \|M_{LB} - M_{ARB}\|_1. \tag{13}$$

This is equivalent to adding semantic regularization to the training of ARB, and our experiments show that this improves accuracy to some extent. Ultimately, the loss function consists of the classification network's multi-label soft margin loss and guidance loss, i.e.,

$$\mathcal{L}_{ARB} = \mathcal{L}_{class} + \lambda \mathcal{L}_{guidance}. \tag{14}$$

4. Experiments

4.1. Datasets

Our method is applied to two UAV-acquired RS datasets, namely UOPNOA and MiniFrance. Of these, the UOPNOA dataset has images from the northern highlands of the Iberian Peninsula in Spain and consists of 33,699 images of size 256 × 256. The MiniFrance dataset contains high-resolution RGB images of 16 cities in different regions of France. Since only some of data are labelled, we chose the dataset acquired in Nice. We cropped each original 10 000 × 10 000 pixels to a size of 256 × 256 pixels. For both datasets, 1100 images are randomly selected from the images containing forests and shrubs, and grouping them into three sets in the ratio of 7:2:1, respectively for training, validation and test.

4.2. Evaluation metrics

We chose the following evaluation metrics to evaluate the efficiency of our network and for performance comparison:

$$\text{mIoU} = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i + FP_i + FN_i}, \tag{15}$$

$$\text{mAccuracy} = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i}, \tag{16}$$

$$Precision_i = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i + FP_i}, \tag{17}$$

$$Recall_i = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i + FN_i}, \tag{18}$$

$$\text{mF1-score} = \frac{1}{N}\sum_{i=1}^{N} 2 \times \frac{Precision_i \times Recall_i}{Precision_i + Recall_i}, \tag{19}$$

where N (equals 3) denotes the number of categories in the datasets we used, and TP, TN, FP and FN respectively denote the true positive, true negative, false positive and false negative.

4.3. Implementation

We used Pytorch to implement our experiments on NVIDIA's 3090 GPU with a memory size of 24G. During data preprocessing, we employed data augmentation techniques of horizontal flipping and random scaling, followed by random cropping the original images to 515 × 512 images as input for the network. For both datasets, we adopted the Resnet50 architecture (He et al., 2016) as the foundational framework and incorporated the pretrained weights used in ImageNet (Russakovsky et al., 2015). In the seed generation period, we set both the training epochs and the batch size to 16. For UOPNOA and MiniFrance, the learning rates are respectively set to 0.02 and 0.023, and SGD is used as the optimizer, incorporating a weight decay of 0.0001. Subsequently, we refined the generated CAMs using the random walk algorithm (IRNet) to produce pseudo labels. Lastly, we trained fully supervised semantic segmentation network SegNeXt (Guo et al., 2022) using the generated pseudo labels and pretrained weights from ImageNet.

Table 1
Quality of CAM and pseudo labels (Pseudo) generated by various networks.

Methods	UOPNOA		MiniFrance	
	CAM	Pseudo	CAM	Pseudo
AffinityNet	55.77	64.23	51.16	56.10
IRNet	54.05	61.11	55.16	57.86
SEAM	54.89	46.61	36.81	27.66
AdvCAM	50.55	57.15	50.52	53.85
OC-CSE	55.83	57.75	54.95	56.14
SIPE	56.65	57.37	48.54	47.87
VWE	49.31	56.05	45.7	48.03
MCTformer	44.75	47.66	45.17	46.95
Ours	**58.18**	**65.93**	**59.22**	**64.24**

4.4. Comparison with state-of-the-methods

Our method was compared with the following eight weakly supervised methods: AffinityNet (Ahn and Kwak, 2018), IRNet (Ahn et al., 2019), SEAM (Wang et al., 2020), AdvCAM (Lee et al., 2021a), OC-CSE (Kweon et al., 2021), SIPE (Chen et al., 2022), VWE (Ru et al., 2021), and MCTformer (Xu et al., 2022).

4.4.1. Comparison on CAM and pseudo labels

By generating more exact CAMs, our method improves the accuracy of the pseudo-labels. To demonstrate this, Table 1 compares the quality of the CAM seeds and pseudo labels generated by different networks in terms of mIoU and shows the performance of our network is better than other networks of WSSS in both stages. Specifically, our network achieves mIoU scores of 59.22% and 64.24% for CAM seeds and pseudo labels, respectively on MiniFrance, and 58.18% and 65.93% respectively on UOPNOA dataset.

Furthermore, in order to better demonstrate the advantages of our network, we visualized the generated CAMs in Fig. 7. The CAMs for the first and fifth input images demonstrate the superiority of our network in multi-region focus. The CAMs for the second and third input images showcase the advantages of our approach in identifying small-scale regions, while the CAMs for the fourth and fifth input images illustrate the capability of our network for comprehensive coverage of large regions. In summary, our method excels in generating CAM of higher quality compared to other approaches.

In addition, we also conducted an analysis of the parameters and inference speed at the CAMs generation stage, as shown in the Table 2. It is observed that our network maintains relatively low parameter count and fast inference speed while achieving optimal performance.

4.4.2. Comparison of semantic segmentation results

We conducted experiments on SegNext using the generated pseudo labels. Table 3 shows the performance of various networks on both datasets. The evaluation metric mIoU is commonly used for semantic segmentation, where the highest mIoU achieved by our network indicates that it segments the target area more accurately than other weakly supervised networks. High mF1-score and high average accuracy also reflect the higher performance of our network. Specifically, our network achieved mIoU, mF1-score, and mAccuracy of 69.27%, 81.3%, and 81.46%, respectively on UOPONA, and 56.82%, 71.88%, and 72.51%, respectively on MiniFrance. Compared to other weakly supervised networks, our network achieved the best performance.

We also conducted comparative experiments on three classic fully supervised networks. In these experiments, we maintained consistency in the samples with weakly supervised networks, differing only in the labels used. The results indicate that our network even outperforms FCN, which is a well-known early work in fully supervised semantic segmentation. The performance of our network is comparable to UNet, a highly effective classic fully supervised semantic segmentation network. While our network falls short of the performance of the classic network Deeplabv3, the mIoU on the two datasets also

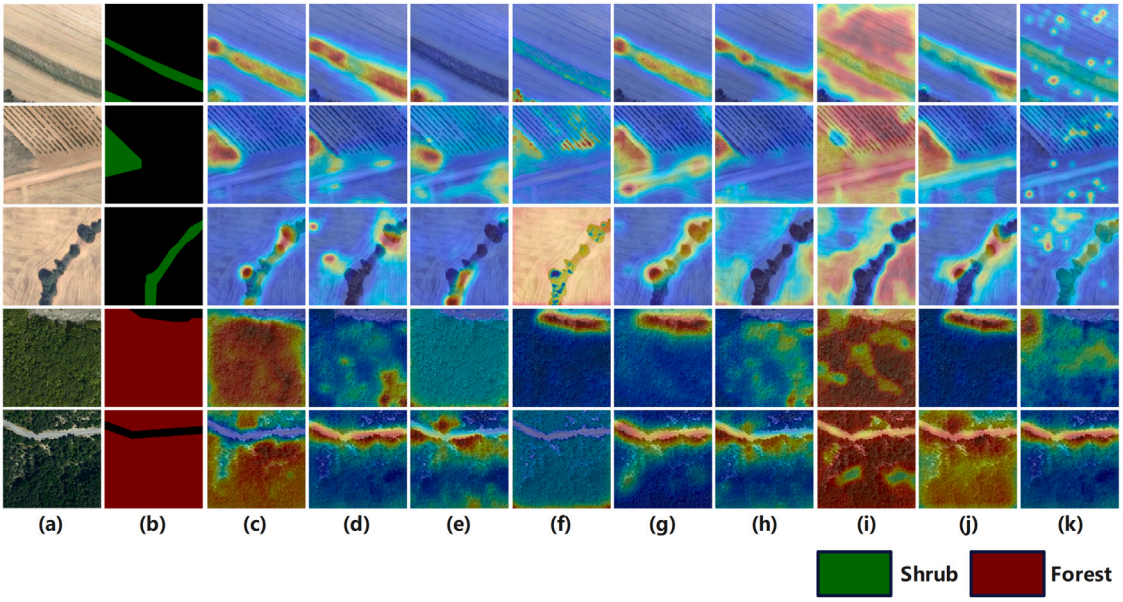


Fig. 7. CAMs generated by different networks. (a) Input image, (b) Ground truth, (c) Our network, (d) AffinityNet, (e) IRNet, (f) SEAM, (g) AdvCAM, (h) OC-CSE, (i) SIPE, (j) VWE, and (k) MCTformer.

Table 2
Parameters(M) and Frames Per Second (FPS) at the CAMs generation stage for different networks.

Methods	Affinity	IRNet	SEAM	AdvCAM	OC-CSE	SIPE	VWE	MCTformer	Ours
Parameters	105.08	70.69	105.33	70.37	105.13	72.46	129.92	21.67	75.80
FPS	2.98	15.5	2.77	0.2	9.6	6.3	14.34	6.51	6.21

Table 3
Performances of various networks on UOPNOA and MiniFrance. The final results were obtained by training SegNeXt with pseudo labels generated by each network and testing on the test set. The supervision types (Sup.) indicate: $\mathcal{I}$-Image-level label and $\mathcal{F}$-segmentation label.

Methods	Sup.	UOPNOA			MiniFrance		
		mIoU (%)	mF1-score (%)	mAcc. (%)	mIoU (%)	mF1-score (%)	mAcc. (%)
AffinityNet	$\mathcal{I}$	66.35	78.92	79.06	53.41	68.7	70.34
IRNet	$\mathcal{I}$	67.54	79.66	80.03	54.28	69.85	70.95
SEAM	$\mathcal{I}$	40.87	57.33	57.75	27.8	41.58	47.49
AdvCAM	$\mathcal{I}$	61.89	74.3	76.46	42.61	56.29	63.24
OC-CSE	$\mathcal{I}$	58.77	72.86	72.67	47.55	63.82	65.28
SIPE	$\mathcal{I}$	65.78	78.19	79.0	37.4	48.75	59.51
VWE	$\mathcal{I}$	61.13	74.83	74.87	44.53	59.85	64.26
MCTformer	$\mathcal{I}$	47.76	58.15	66.25	36.22	46.89	55.42
FCN	$\mathcal{F}$	64.44	78.32	78.18	52.08	68.18	70.19
UNet	$\mathcal{F}$	70.49	82.85	82.99	57.11	73.74	72.33
Deeplabv3	$\mathcal{F}$	76.45	86.59	86.5	63.92	77.83	78.2
Ours	$\mathcal{I}$	69.27	81.3	81.46	56.82	71.88	72.51

achieves 90.6% and 88.9% performance of Deeplabv3, which is highly inspiring. In comparison to fully supervised networks, our network achieves competitive performance while utilizing extremely low-cost annotations.

Fig. 8 shows the P-R curves and TPR-FPR curves of different networks. The figure shows that for both datasets, our network attained the highest accuracy.

4.4.3. Qualitative comparison

The comparison of segmentation results for different methods on the two datasets is illustrated in Fig. 9. Firstly, our network demonstrates its capability to accurately identify small target regions, as illustrated by the segmentation results of the sixth and seventh input images. Moreover, our network also exhibits accurate identification of large target regions and multi-target areas, as demonstrated by the results of the fifth and second input images. Secondly, our network is able to address the issues associated with the heterogeneity that exist in vegetation RS images, thus enabling it to perform accurate overall segmentation of images, as illustrated by the results of the third and fourth input images. Thirdly, our network has a greater advantage in handling boundary details, accurately segmenting target regions with complex boundaries, thereby capable of processing more complex scene graphs, as illustrated by the result of the first input image. In general, our network exhibits significant advantages over other networks.

4.4.4. Ablation experiments

The effectiveness of individual modules. In order to validate the soundness of the proposed network, each component of the network architecture was incrementally integrated. We conducted experiments on UOPONA and MiniFrance with consistent hyperparameters across all trials. Initially, we deployed a baseline model that utilized a singular ResNet50 branch pretrained on ImageNet weights. For assessing the effectiveness of an attention module, we performed different experiments

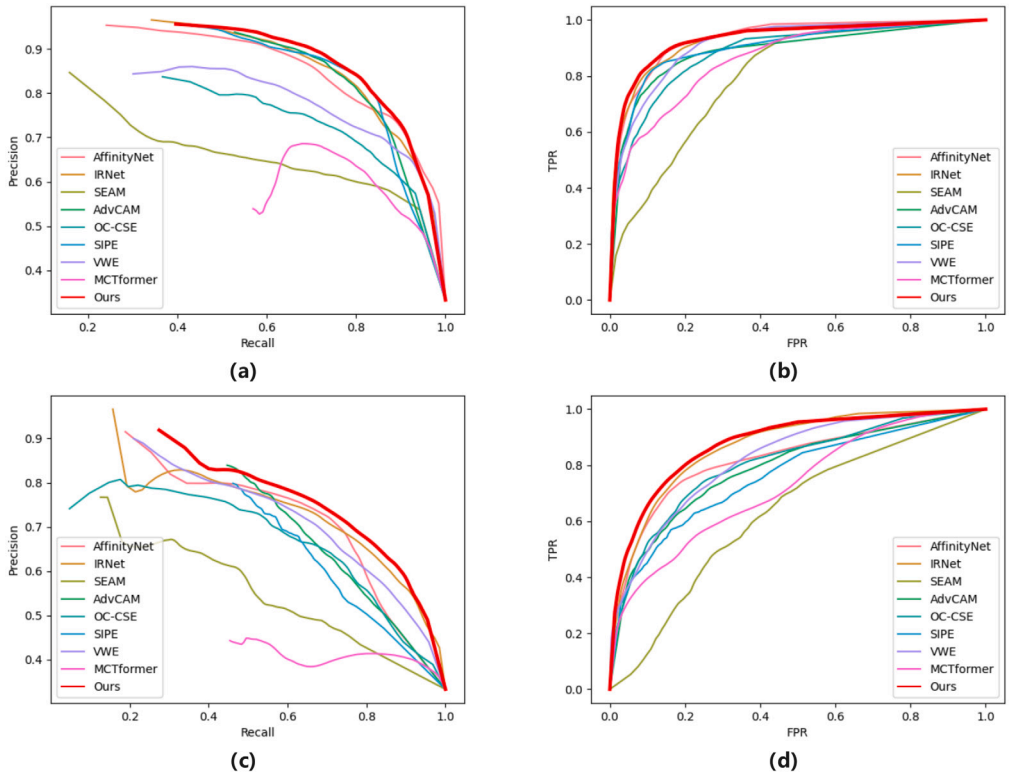


Fig. 8. Quantitative curves of various methods: (a) P-R curves on UOPONA; (b) TPR-FPR curves on UOPONA; (c) P-R curves on MiniFrance; and (d) TPR-FPR curves on MiniFrance.

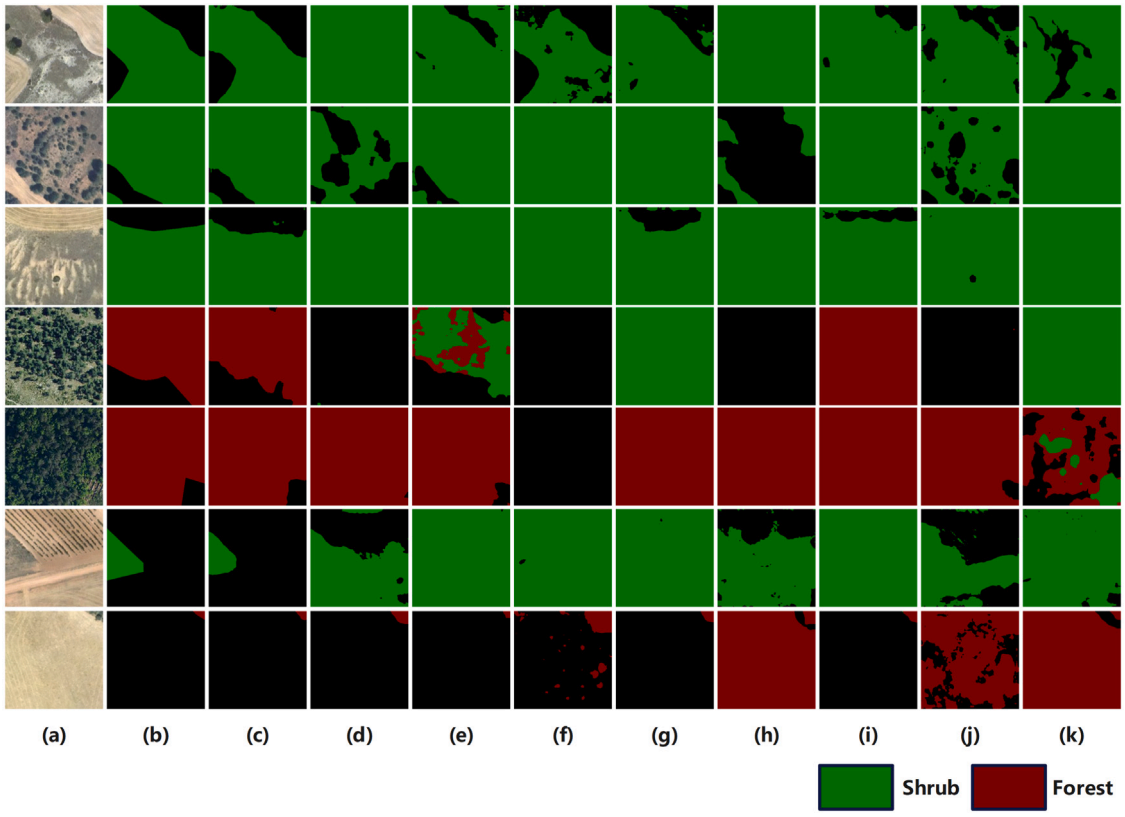


Fig. 9. Segmentation results using different networks. Green represents shrub areas and red represents forest areas. (a) Input image, (b) Ground truth, (c) Our network, (d) AffinityNet, (e) IRNet, (f) SEAM, (g) AdvCAM, (h) OC-CSE, (i) SIPE, (j) VWE, and (k) MCTformer. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

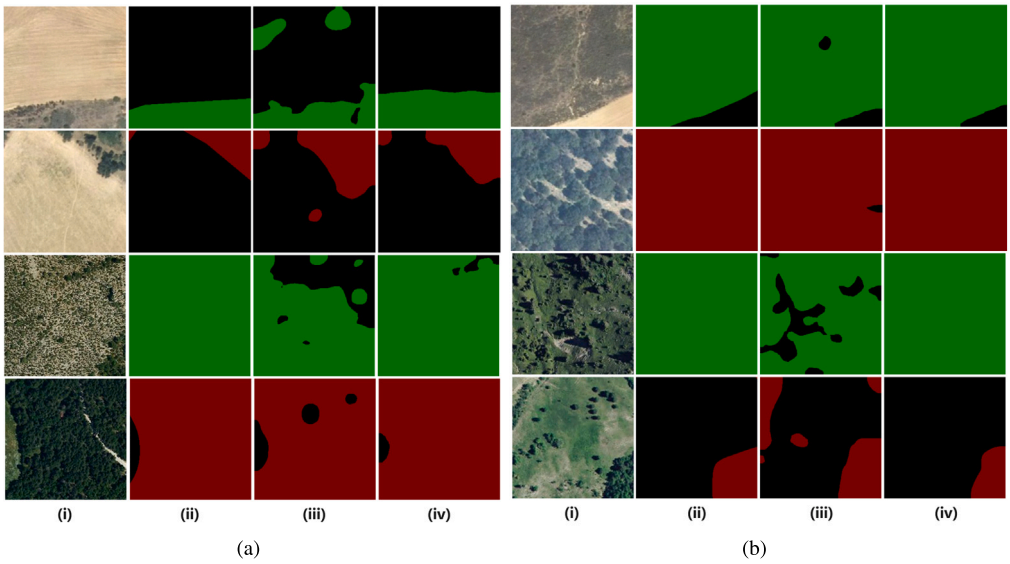


Fig. 10. Visualization of the ablation experiments. (a) Comparison of visualization for baseline and dual branch integration: (i) Input image; (ii) Ground truth; (iii) Baseline; and (iv) Baseline+ARB. (b) Visualization comparison before and after adding guidance loss: (i) Input image; (ii) Ground truth; (iii) Without adding guidance loss; and (iv) With guidance loss added.

Table 4
Results of ablation experiments on UOPONA.

Base	A1	A2	A3	A4	Aall	Loss	mIoU (%)	mAccuracy (%)	mF1-score (%)
√							54.05	70.22	68.77
√	√						56.46	72.02	71.15
√		√					56.02	72.26	70.38
√			√				56.3	72.48	70.67
√				√			56.88	72.55	71.41
√					√		57.53	73.13	71.97
√					√	√	**58.18**	**73.68**	**72.53**

Table 5
Results of ablation experiments on MiniFrance.

Base	A1	A2	A3	A4	Aall	Loss	mIoU (%)	mAccuracy (%)	mF1-score (%)
√							55.16	70.17	69.82
√	√						57.58	73.24	71.17
√		√					57.74	73.85	71.08
√			√				57.37	73.92	70.44
√				√			57.56	73.58	70.93
√					√		58.35	74.17	71.70
√					√	√	**59.22**	**74.30**	**73.09**

on individual module integrated at diverse positions to demonstrate the advantages of MARS. Specifically, "Base" denotes the scenario with only a localization branch, "AX" indicates the addition of an attention module after the Xth layer of ARB, "Aall" represents the situation where an attention module is added after every layer of the relocation branch, and "Loss" denotes the employment of guidance loss. Tables 4 and 5 respectively show the results of the experiments conducted on UOPONA and MiniFrance. The mIoU scores of the baseline model are 54.05% and 55.16% for UOPONA and MiniFrance, respectively. Although the single attention module enhances the performance of the network, the accuracy is still lower than the result achieved with the addition of MARS. Finally, the efficacy of the network was further enhanced by including a guidance loss. Ultimately, our network achieved 58.18% and 59.22% mIoU for UOPONA and MiniFrance, respectively.

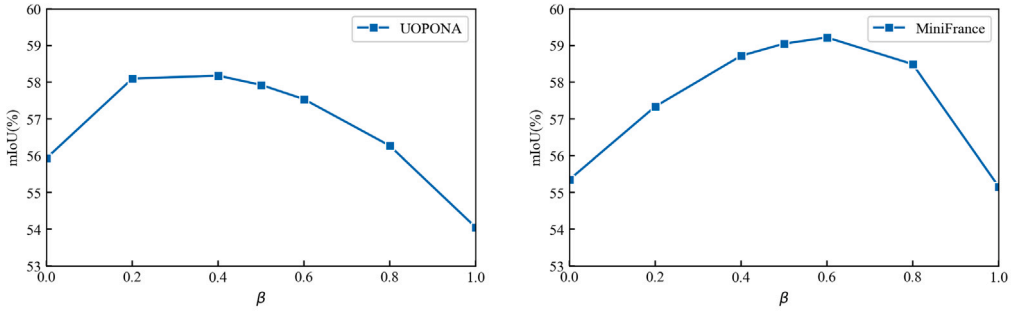
We visualized the results to better demonstrate the effectiveness of each designed component by conducting ablation studies, as shown in Fig. 10. Fig. 10(a) shows that in comparison with the baseline network, the dual-branch structure of our network which incorporates MARS inhibits background noise, captures more precise features, and produces more accurate CAMs. In addition, Fig. 10(b) illustrates that incorporating guidance loss avoids the identification of incorrect regions as well as highlights the identification of correct regions, which effectively improves the correctness of the overall segmentation results.

Effectiveness of integration. To investigate the integration coefficients, we conducted experiments on the two datasets, evaluating various coefficients. Fig. 11 presents the results of our experimentation, indicating that the maximum metric scores are achieved when the β coefficient is set to 0.4 for UOPNOA and 0.6 for MiniFrance. Notably, when the coefficient increases or decreases, the performance metric decreases as well. When the coefficient is 1, it is equivalent to using only the first branch, whereas a coefficient of 0 is equivalent to using only the second branch, both of which are far less effective than the two-branch fusion approach. Our integration strategy, therefore, enhances the network performance.

We also visualized the generated CAMs and pseudo labels in Fig. 12 to better illustrate the effectiveness of our network. The figure shows that our first branch focuses on the most discriminative regions, while the second branch is repositioned to capture other important features. In addition, consistent with what was mentioned earlier, the ARB will have a complementary (e.g., columns 1 and 2 in Fig. 12) or expansive (e.g., columns 8 and 9 in Fig. 12) effect on the LB. Finally, weighting and fusing these features can lead to higher quality pseudo labels.

5. Conclusion

The paper proposes a novel dual-branch integration network that includes a LB and an ARB. The LB uses a ResNet50 classification network to accurately identify some important areas, while the ARB incorporates MARS based on the localization branch to effectively extract important contextual semantic information and compensate for the target areas that the LB cannot identify. The experimental findings demonstrate that the utilization of our network in woody vegetation RS image semantic segmentation effectively addresses the issues associated with the heterogeneity of woody vegetation in RS images. Our network also offers a solution to address challenging integrity segmentation of



(a) UOPNOA (b) MiniFrance

Fig. 11. mIoU(%) for two datasets with different integration coefficients.

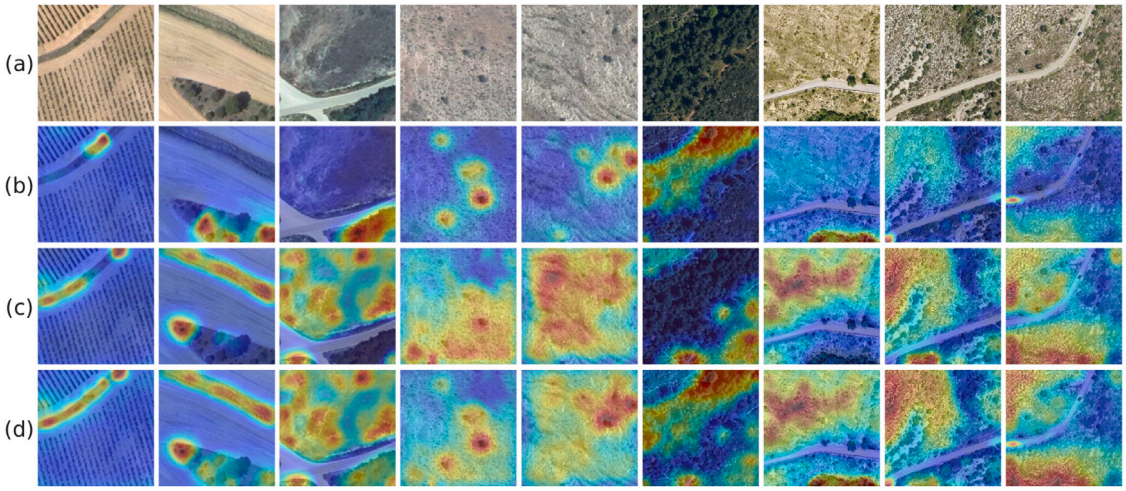


Fig. 12. Visualization of the generated CAMs: (a) Input image; (b) LB; (c) ARB; and (d) Dual-branch integrated CAM.

target regions that existing WSSS methods have to overcome. As a result, the performance of our network in segmenting woody vegetation RS images has been significantly improved. Additionally, our network attains the best performance on UOPONA and MiniFrance which are acquired by UAVs.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data

Acknowledgements

The research was supported by Innovation and Entrepreneurship Project of Jiangsu Province.

References

Ahn, J., Cho, S., Kwak, S., 2019. Weakly supervised learning of instance segmentation with inter-pixel relations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2209–2218. http://dx.doi.org/10.1109/CVPR.2019.00231.

Ahn, J., Kwak, S., 2018. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4981–4990. http://dx.doi.org/10.1109/CVPR.2018.00523.

Allen-Zhu, Z., Li, Y., 2020. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. http://dx.doi.org/10.48550/arXiv.2012.09816, arXiv preprint arXiv:2012.09816.

Alzu'bi, A., Alsmadi, L., 2022. Monitoring deforestation in Jordan using deep semantic segmentation with satellite imagery. Ecol. Inform. 70, 101745. http://dx.doi.org/10.1016/j.ecoinf.2022.101745.

Bearman, A., Russakovsky, O., Ferrari, V., Fei-Fei, L., 2016. What's the point: Semantic segmentation with point supervision. In: European Conference on Computer Vision (ECCV). Springer, pp. 549–565. http://dx.doi.org/10.1007/978-3-319-46478-7_34.

Castillo Navarro, J., Le Saux, B., Boulch, A., Audebert, N., Lefèvre, S., 2020. MiniFrance. http://dx.doi.org/10.21227/b9pt-8x03.

Chapin, F.S., Matson, P.A., Vitousek, P.M., 2011. Principles of Terrestrial Ecosystem Ecology. Spinger.

Chen, B., Wang, L., Fan, X., Bo, W., Yang, X., Tjahjadi, T., 2023. Semi-FCMNet: Semi-supervised learning for forest cover mapping from satellite imagery via ensemble self-training and perturbation. Remote Sens. 15 (16), http://dx.doi.org/10.3390/rs15164012, URL: https://www.mdpi.com/2072-4292/15/16/4012.

Chen, Q., Yang, L., Lai, J.-H., Xie, X., 2022. Self-supervised image-specific prototype exploration for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4288–4298. http://dx.doi.org/10.1109/CVPR52688.2022.00425.

Dai, J., He, K., Sun, J., 2015. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 1635–1643. http://dx.doi.org/10.1109/ICCV.2015.191.

De Petris, S., Boccardo, P., Borgogno-Mondino, E., 2019. Detection and characterization of oil palm plantations through modis EVI time series. Int. J. Remote Sens. 40 (19), 7297–7311. http://dx.doi.org/10.1080/01431161.2019.1584689.

Fisher, A., Danaher, T., Gill, T., 2017. Mapping trees in high resolution imagery across large areas using locally variable thresholds guided by medium resolution tree maps. Int. J. Appl. Earth Obs. Geoinf. 58, 86–96. http://dx.doi.org/10.1016/j.jag.2017.02.004.

Flood, N., Watson, F., Collett, L., 2019. Using a U-net convolutional neural network to map woody vegetation extent from high resolution satellite imagery across

queensland, Australia. Int. J. Appl. Earth Obs. Geoinf. 82, 101897. http://dx.doi.org/10.1016/j.jag.2019.101897.

Guo, M.-H., Lu, C.-Z., Hou, Q., Liu, Z., Cheng, M.-M., Hu, S.-m., 2022. SegNeXt: Rethinking convolutional attention design for semantic segmentation. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (Eds.), Advances in Neural Information Processing Systems (NeurIPS), Vol. 35. Curran Associates, Inc., pp. 1140–1156, URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/08050f40fff41616ccfc3080e60a301a-Paper-Conference.pdf.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778. http://dx.doi.org/10.48550/arXiv.1512.03385.

Hou, Q., Zhou, D., Feng, J., 2021. Coordinate attention for efficient mobile network design. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13713–13722. http://dx.doi.org/10.1109/CVPR46437.2021.01350.

Jiang, P.-T., Hou, Q., Cao, Y., Cheng, M.-M., Wei, Y., Xiong, H.-K., 2019. Integral object mining via online attention accumulation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2070–2079. http://dx.doi.org/10.1109/ICCV.2019.00216.

Jiang, P.-T., Yang, Y., Hou, Q., Wei, Y., 2022. L2g: A simple local-to-global knowledge transfer framework for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16886–16896. http://dx.doi.org/10.1109/CVPR52688.2022.01638.

Kweon, H., Yoon, S.-H., Kim, H., Park, D., Yoon, K.-J., 2021. Unlocking the potential of ordinary classifier: Class-specific adversarial erasing framework for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6994–7003. http://dx.doi.org/10.1109/ICCV48922.2021.00691.

La Rosa, L.E.C., Sothe, C., Feitosa, R.Q., de Almeida, C.M., Schimalski, M.B., Oliveira, D.A.B., 2021. Multi-task fully convolutional network for tree species mapping in dense forests using small training hyperspectral data. ISPRS J. Photogramm. Remote Sens. 179, 35–49. http://dx.doi.org/10.1016/j.isprsjprs.2021.07.001.

Lee, J., Kim, E., Yoon, S., 2021a. Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4071–4080. http://dx.doi.org/10.1109/CVPR46437.2021.00406.

Lee, S., Lee, M., Lee, J., Shim, H., 2021b. Railroad is not a train: Saliency as pseudo-pixel supervision for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5495–5505. http://dx.doi.org/10.1109/CVPR46437.2021.00545.

Li, L., Mu, X., Chianucci, F., Qi, J., Jiang, J., Zhou, J., Chen, L., Huang, H., Yan, G., Liu, S., 2022. Ultrahigh-resolution boreal forest canopy mapping: Combining uav imagery and photogrammetric point clouds in a deep-learning-based approach. Int. J. Appl. Earth Obs. Geoinf. 107, 102686. http://dx.doi.org/10.1016/j.jag.2022.102686.

Lin, D., Dai, J., Jia, J., He, K., Sun, J., 2016. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3159–3167. http://dx.doi.org/10.1109/CVPR.2016.344.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3431–3440. http://dx.doi.org/10.1109/CVPR.2015.7298965.

Nasiri, V., Beloiu, M., Darvishsefat, A.A., Griess, V.C., Maftei, C., Waser, L.T., 2023. Mapping tree species composition in a caspian temperate mixed forest based on spectral-temporal metrics and machine learning. Int. J. Appl. Earth Obs. Geoinf. 116, 103154. http://dx.doi.org/10.1016/j.jag.2022.103154.

Pedrayes, O.D., Lema, D.G., García, D.F., Usamentiaga, R., Alonso, Á., 2021. Evaluation of semantic segmentation methods for land use with spectral imaging using sentinel-2 and PNOA imagery. Remote Sens. 13 (12), 2292. http://dx.doi.org/10.3390/rs13122292.

Purwanto, A.D., Wikantika, K., Deliar, A., Darmawan, S., 2022. Decision tree and random forest classification algorithms for mangrove forest mapping in sembilang national park, Indonesia. Remote Sens. 15 (1), 16. http://dx.doi.org/10.3390/rs15010016.

Puthumanaillam, G., Verma, U., 2023. Texture based prototypical network for few-shot semantic segmentation of forest cover: Generalizing for different geographical regions. Neurocomputing 538, 126201. http://dx.doi.org/10.1016/j.neucom.2023.03.062.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer, pp. 234–241. http://dx.doi.org/10.1007/978-3-319-24574-4_28.

Ru, L., Du, B., Wu, C., 2021. Learning visual words for weakly-supervised semantic segmentation. In: IJCAI, Vol. 5. p. 6. http://dx.doi.org/10.1007/s11263-022-01586-9.

Rüetschi, M., Weber, D., Koch, T.L., Waser, L.T., Small, D., Ginzler, C., 2021. Countrywide mapping of shrub forest using multi-sensor data and bias correction techniques. Int. J. Appl. Earth Obs. Geoinf. 105, 102613. http://dx.doi.org/10.1016/j.jag.2021.102613.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al., 2015. Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. 115, 211–252. http://dx.doi.org/10.1007/s11263-015-0816-y.

Schindler, J., Dymond, J.R., Wiser, S.K., Shepherd, J.D., 2021. Method for national mapping spatial extent of southern beech forest using temporal spectral signatures. Int. J. Appl. Earth Obs. Geoinf. 102, 102408. http://dx.doi.org/10.1016/j.jag.2021.102408.

Schmitt, M., Prexl, J., Ebel, P., Liebel, L., Zhu, X.X., 2020. Weakly supervised semantic segmentation of satellite images for land cover mapping challenges and opportunities. ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci. V-3-2020, 795–802. http://dx.doi.org/10.5194/isprs-annals-V-3-2020-795-2020.

Shafeian, E., Fassnacht, F.E., Latifi, H., 2021. Mapping fractional woody cover in an extensive semi-arid woodland area at different spatial grains with sentinel-2 and very high-resolution data. Int. J. Appl. Earth Obs. Geoinf. 105, 102621. http://dx.doi.org/10.1016/j.jag.2021.102621.

Shen, R., Chen, X., Chen, L., He, B., Yuan, W., 2021. Regional evaluation of satellite-based methods for identifying leaf unfolding date. ISPRS J. Photogramm. Remote Sens. 175, 88–98. http://dx.doi.org/10.1016/j.isprsjprs.2021.02.021.

Trenčanová, B., Proença, V., Bernardino, A., 2022. Development of semantic maps of vegetation cover from UAV images to support planning and management in fine-grained fire-prone landscapes. Remote Sens. 14 (5), 1262. http://dx.doi.org/10.3390/rs14051262.

Waldeland, A.U., Trier, Ø.D., Salberg, A.-B., 2022. Forest mapping and monitoring in africa using sentinel-2 data and deep learning. Int. J. Appl. Earth Obs. Geoinf. 111, 102840. http://dx.doi.org/10.1016/j.jag.2022.102840.

Wang, Y., Zhang, J., Kan, M., Shan, S., Chen, X., 2020. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12275–12284. http://dx.doi.org/10.1109/CVPR42600.2020.01229.

Wang, L., Zhou, Y., Hu, Q., Tang, Z., Ge, Y., Smith, A., Awada, T., Shi, Y., 2021. Early detection of encroaching woody juniperus virginiana and its classification in multi-species forest using UAS imagery and semantic segmentation algorithms. Remote Sens. 13 (10), 1975. http://dx.doi.org/10.3390/rs13101975.

Wei, Y., Feng, J., Liang, X., Cheng, M.-M., Zhao, Y., Yan, S., 2017. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1568–1576. http://dx.doi.org/10.1109/CVPR.2017.687.

Wu, W., Fan, X., Qu, H., Yang, X., Tjahjadi, T., 2022. Tcdnet: Tree crown detection from UAV optical images using uncertainty-aware one-stage network. IEEE Geosci. Remote Sens. Lett. 19, 1–5. http://dx.doi.org/10.1109/LGRS.2022.3214281.

Xiao, H., Su, F., Fu, D., Lyne, V., Liu, G., Pan, T., Teng, J., 2021. Optimal and robust vegetation mapping in complex environments using multiple satellite imagery: Application to mangroves in southeast Asia. Int. J. Appl. Earth Obs. Geoinf. 99, 102320. http://dx.doi.org/10.1016/j.jag.2021.102320.

Xie, J., Hou, X., Ye, K., Shen, L., 2022. CLIMS: cross language image matching for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4483–4492. http://dx.doi.org/10.1109/CVPR52688.2022.00444.

Xu, L., Ouyang, W., Bennamoun, M., Boussaid, F., Xu, D., 2022. Multi-class token transformer for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4310–4319. http://dx.doi.org/10.1109/CVPR52688.2022.00427.

Yang, X., Xiao, X., Qin, Y., Wang, J., Neal, K., 2021. Mapping forest in the southern great plains with ALOS-2 PALSAR-2 and landsat 7/8 data. Int. J. Appl. Earth Obs. Geoinf. 104, 102578. http://dx.doi.org/10.1016/j.jag.2021.102578.

Zhou, H., Fu, L., Sharma, R.P., Lei, Y., Guo, J., 2021. A hybrid approach of combining random forest with texture analysis and VDVI for desert vegetation mapping based on UAV RGB data. Remote Sens. 13 (10), 1891. http://dx.doi.org/10.3390/rs13101891.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., 2016. Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2921–2929. http://dx.doi.org/10.48550/arXiv.1512.04150.