Weakly Supervised Burned Area Mapping in the Brazilian Pantanal Using Multispectral Satellite Imagery

Maximilian Jaderson de Melo[1,2], Thiago Edgar Bauce Venancio[2], Lucas Yuri Dutra de Oliveira[3], Mário de Araújo Carvalho[2], José Marcato Junior[3], Wesley Nunes Gonçalves[2]

[1]Federal Institute of Education, Science, and Technology of Mato Grosso do Sul, Naviraí, MS, Brazil - maximilian.melo@ifms.edu.br
[2]Faculty of Computing, Federal University of Mato Grosso do Sul, Campo Grande, MS, Brazil - {maximilian.melo,thiago.bauce,mario.carvalho,wesley.goncalves}@ufms.br
[3]Faculty of Engineering, Architecture, and Urbanism and Geography, Federal University of Mato Grosso do Sul, Campo Grande, MS, Brazil - {lucas.oliveira,jose.marcato}@ufms.br

Keywords: Deep Learning, Weakly Supervised Semantic Segmentation, Satellite Imagery, Wildfire Area Mapping

1. Introduction

The Pantanal biome is a region of critical importance for biodiversity conservation, scientific research, and sustainable development. Nevertheless, it is frequently threatened by recurring wildfires (Leal Filho et al., 2021). Furthermore, the use of satellite imagery enables the observation of large areas with high temporal frequency and adequate spatial resolution, making it a valuable resource for critical tasks such as land cover classification through semantic segmentation techniques (Seydi et al., 2022, Hu et al., 2021, Cho et al., 2022).

Traditional approaches to semantic segmentation in remote sensing typically rely on fully supervised learning (Seydi et al., 2022, Hu et al., 2021, Cho et al., 2022), which demands large volumes of high-quality, pixel-level annotations, requiring specialists to manually annotate each pixel of an image, identifying the location and type of elements of interest. This meticulous and time-consuming process must be repeated for every image in the dataset, which then serves as the training data for a deep learning model. Although high-quality annotations are crucial for model performance, the extensive volume of pixel-level data required makes this process impractical, especially in remote or data-scarce environments.

To address this limitation, we explore a Weakly Supervised Semantic Segmentation (WSSS) approach that relies only on image-level labels, significantly reducing annotation costs. Specifically, we benchmark Self-supervised Equivariant Attention Mechanism (SEAM) (Wang et al., 2020) and Puzzle-CAM (Jo and Yu, 2021), two recent methods that enhance Class Activation Maps (CAMs) by incorporating spatial consistency and complementary localization cues. These techniques enable the generation of reliable pseudo-labels for effective burned area mapping while substantially reducing the need for costly and labor-intensive pixel-level annotations.

The main contributions of this study are as follows.

- We adapt state-of-the-art WSSS frameworks, SEAM and Puzzle-CAM, for the task of wildfire mapping in the Pantanal biome.
- We demonstrate that these methods enable a significant reduction in manual labeling requirements by leveraging image-level annotations while maintaining competitive segmentation performance.
- We show that this approach facilitates the upscaling of training datasets, making it a practical and cost-effective solution for large-scale environmental monitoring and fire impact assessment.

2. Materials and Methods

In this study, we benchmark two state-of-the-art WSSS methods for the task of burned area mapping in satellite imagery. Both approaches are originally designed for natural image datasets. By relying only on image-level labels, these models aim to generate high-quality CAMs that can serve as pseudo-labels for training segmentation networks. This section describes the core components of each method, the dataset preparation process, and the experimental setup used to assess their effectiveness in identifying burned regions in the Pantanal biome.

2.1 Dataset

Our study focuses on fire-affected areas in the southern Pantanal (Mato Grosso do Sul) using PlanetScope multispectral imagery (16 bits imagery, four bands: B, G, R, NIR) at approximately $3m$ of ground sample distance. The images are provided in surface reflectance and orthorectified format. Specialists manually annotated burned and unburned regions using QGIS software.

The burned areas corresponded to 225,483.97 ha in total, configuring into 676,452 pixels marked as regions of interest. A total of 1,502 fire or burning areas/events were recorded within this region.

We divided the images into non-overlapping patches of 256×256 pixels, resulting in $5,690$ training samples, 595 validation samples, and 595 test samples. All spectral bands were normalized to 8-bit values to ensure consistency. The dataset includes diverse fire-related conditions such as recent burn scars, regenerating vegetation, and areas with partial smoke coverage, offering a challenging setting for wildfire mapping.

For image-level labeling, we adopted a binary annotation policy: any patch containing at least one burned pixel was labeled as “burned,” while all others were labeled as “background.” This approach facilitated the use of WSSS techniques by simplifying the annotation process while retaining relevant supervision for model training. Figure 1 presents representative image patches from the dataset. The first row presents the multispectral patches downscaled to 8-bit values and histogram-equalized to enhance visual interpretation. The second row displays the corresponding ground truth masks, where burned areas are highlighted in red. According to the adopted annotation policy, the first three samples are labeled as burned, whereas the last sample is labeled as background.

2.2 Weakly Supervised Segmentation Methods

SEAM (Wang et al., 2020) is a WSSS method that enhances CAMs by addressing their common weakness: producing sparse and incomplete activation maps. It achieves this by introducing a self-supervised equivariant attention mechanism that encourages the model to generate spatially consistent CAMs even when the input image undergoes geometric transformations. The core idea is that the class activation for a specific object or region should remain consistent in both the original image and its transformed version.

Puzzle-CAM (Jo and Yu, 2021) introduces a Siamese network that adopts a "puzzle-solving" strategy to improve weakly supervised semantic segmentation. The method divides the input image into several non-overlapping patches, treating each patch as a piece of a puzzle. The network then learns to classify both the full image and its individual patches, while enforcing consistency between the Class Activation Maps (CAMs) of the whole image and the reassembled patches. By constraining the CAMs of the parts to align with those of the original image, Puzzle-CAM prevents the network from relying solely on the most discriminative regions. Instead, it encourages the model to capture contextual cues across the entire object, leading to more complete, spatially coherent, and accurate activation maps, even without pixel-level supervision.

2.3 Approach To Work With Four Bands

Both WSSS methods were originally designed for standard RGB images with three input channels. To leverage the full spectral information of multispectral data, particularly the NIR (Near-Infrared) band which is sensitive to vegetation health and fire impact, we modified the backbone to accept four-channel inputs. This required adapting the first convolutional layer of the backbones.

Specifically, the original convolutional layer, which has weights for three input channels, was replaced with a new convolution layer configured for four input channels while keeping the same number of output channels. The weights for the first three channels were initialized with the pre-trained RGB weights, and the weights of the new NIR channel were either copied from the existing channels or initialized appropriately to preserve learned representations. This ensures that the network can immediately exploit the pre-trained features while incorporating the additional spectral information.

After this modification, the first stage of the backbone processes four-channel inputs, while the subsequent stages remain unchanged. This simple yet effective adaptation allows the network to fully exploit multispectral inputs, enhancing its ability to discriminate between burned and unburned areas in complex landscapes like the Pantanal.

2.4 Experimental Setup

We evaluated the WSSS methods on the burned area dataset described in Section 2.1, with both quantitative (Tables 1 and 2) and qualitative (Figure 2) comparisons.

Our pipeline consists of two main stages: (i) pseudo-label generation using Weakly Supervised Semantic Segmentation (WSSS) methods, where CAMs are produced from image-level annotations and refined into pixel-level masks; and (ii) training a Fully Supervised Semantic Segmentation (FSSS) model with these pseudo-labels, enabling the network to learn detailed spatial patterns and improve segmentation accuracy.

For the first stage (pseudo-label generation), we employed widely used backbones in the literature, such as ResNet (He et al., 2016), and a more advanced architecture, ResNeSt (Zhang et al., 2022), to enhance efficiency in the task. As summarized in Table 1, ResNet-34 was used for Classifier and Puzzle-CAM, ResNet-38d for SEAM, and ResNeSt-101 for Puzzle-CAM.

In the second stage, we adopted SegFormer (Xie et al., 2021) with the MiT-B5 encoder as the segmentation backbone, chosen for its strong performance in a variety of semantic segmentation tasks and its efficiency in processing high-resolution inputs. The model was trained using the pseudo-labels generated in the first stage, enabling it to learn fine-grained spatial patterns from weak supervision. To establish an upper-bound reference, we also trained the same SegFormer configuration using the available pixel-level ground truth annotations (FSSS), allowing for a direct performance comparison between weakly supervised and fully supervised training.

First-stage models were trained with default hyperparameters, adjusting batch size and image resolution to fit GPU memory. Second-stage models were trained for 40,000 iterations with a batch size of 16, using the AdamW optimizer and a learning rate of 6×10^{-5}. All experiments were conducted on an NVIDIA RTX 3090 GPU using PyTorch and the MMSegmentation toolbox (Contributors, 2020). Performance was measured using the Intersection over Union (IoU) metric:

$$IoU = \frac{TP}{TP + FP + FN}, \tag{1}$$

where TP, FP and FN are the count of true positive, false positive, and false negative pixels, respectively.

3. Results

We first evaluated the quality of the pseudo-labels generated by the WSSS approaches. Specifically, we trained and compared three WSSS models: a single baseline classifier, SEAM, and Puzzle-CAM. The first three rows in Table 1 present a quantitative comparison of these models under equivalent backbone configurations. The fourth row highlights the impact of using a stronger backbone within the Puzzle-CAM framework.

Method	Backbone	Pseudo-label
Classifier	ResNet-34	65.620
SEAM	ResNet-38d	66.607
Puzzle-CAM	ResNet-34	69.280
Puzzle-CAM	ResNest-101	**86.455**

Table 1. Pseudo-label quality comparison (IoU %) for WSSS methods.

As shown in Table 1, Puzzle-CAM (ResNet-34) outperformed both baseline classifier and SEAM, achieving an IoU of 69.28%This improvement demonstrates the effectiveness of

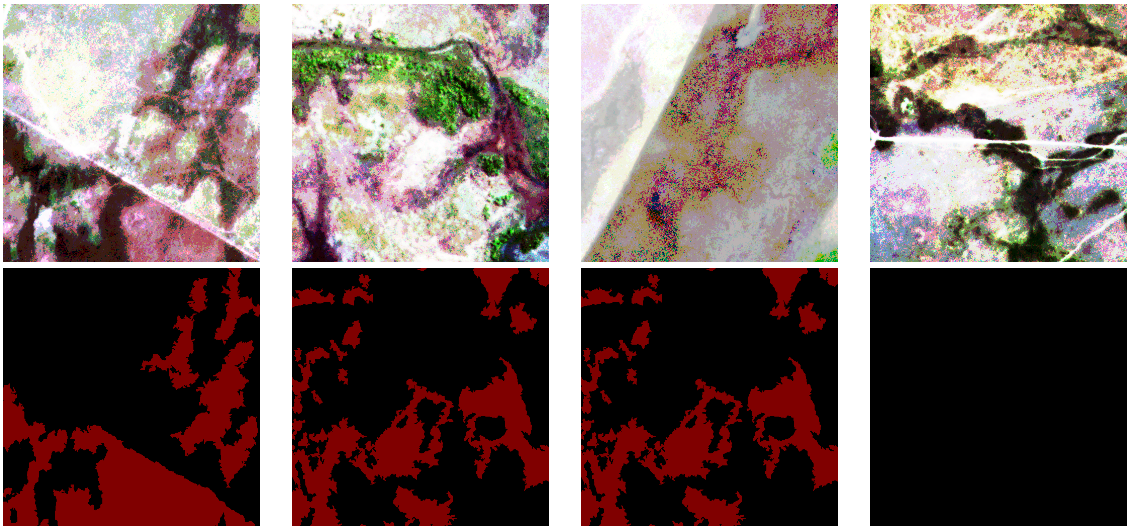

Figure 1. Example RGB image patches from the burned area dataset.

Puzzle-CAM's region-based feature separation strategy in producing more complete and spatially accurate CAMs for burned area detection. SEAM also improved over the baseline, reflecting the benefit of spatial consistency constraints.

A particularly notable result is observed when Puzzle-CAM is combined with a stronger backbone (ResNeSt-101), where the pseudo-label quality increases substantially to 86.46% IoU. This represents an improvement of over 12 percentage points compared to the same method with ResNet-34 for the burned area class, underscoring the significant influence of backbone representational capacity on WSSS performance. The gain suggests that richer feature representations, coupled with effective CAM refinement strategies, can lead to pseudo-labels of a quality that approaches those typically associated with more costly annotation methods.

Method	Supervision	Background	Burn. Area
SegFormer	F	96.31	91.95
SEAM	I	72.14	61.37
Puzzle-CAM-RN34	I	79.98	67.76
Puzzle-CAM-R101	I	**89.05**	**79.99**

Table 2. Final Segmentation performance comparison (Iou %) between FSSS and WSSS-trained models.

Table 2 presents the final segmentation performance of the FSSS baseline and the models trained using pseudo-labels generated by different WSSS approaches (Table 1). As expected, the FSSS SegFormer model achieved the highest IoU (96.31% for the background and 91.95% for the burned area). However, the Puzzle-CAM with a ResNest-101 backbone (Puzzle-CAM-R101) WSSS model achieved a strong 79.99% IoU for burned areas, significantly reducing the gap to fully supervised performance. Overall, these results demonstrate the potential of WSSS methods and confirm that stronger backbones play a crucial role in WSSS frameworks for burned area mapping.

Finally, Figure 2 illustrates a qualitative comparison between the ground truth (red contours) and SegFormer trained with Puzzle-CAM-R101 pseudo-labels (yellow contours). The results show strong spatial agreement in many burned regions, demonstrating the model's ability to accurately delineate fire-affected areas. However, some discrepancies remain, with instances of both over-segmentation—where burned areas are overestimated—and under-segmentation—where small or fragmented burned patches are missed. These patterns highlight the challenges of capturing complex burn boundaries in heterogeneous landscapes such as the Pantanal.

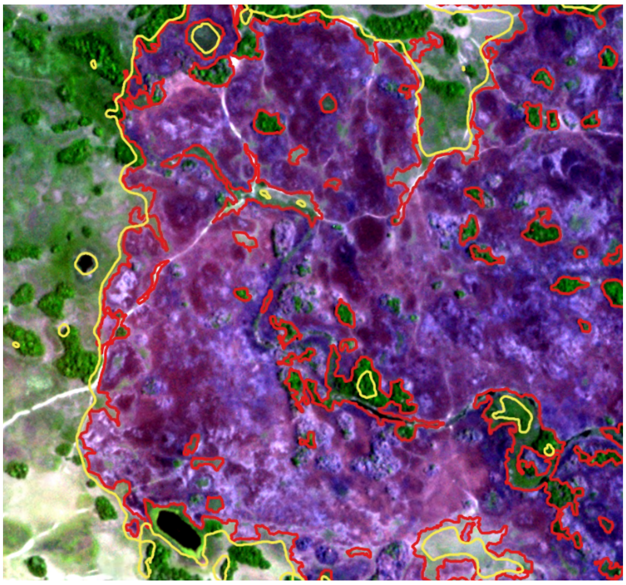

Figure 2. Qualitative results of WSSS. Ground truth (red) is contrasted against the WSSS model (yellow).

4. Conclusion

In this study, we demonstrated the effectiveness of adapting WSSS methods, such as Puzzle-CAM for weakly supervised burned area mapping in the southern Pantanal using multispectral (RGB-NIR) satellite imagery. By leveraging a stronger backbone architecture (ResNeSt-101), we achieved substantial improvements in both pseudo-label quality, which was reflected in the final segmentation performance, significantly narrowing the gap to fully supervised methods.

These findings indicate that combining advanced deep learning architectures with weak supervision offers a promising strategy to reduce the dependence on costly pixel-level annotations. Our results highlight the potential of combining advanced network architectures with weak supervision strategies as a viable and efficient solution for large-scale environmental monitoring in remote and data-scarce regions. As future work, we aim to further improve model performance and robustness, to develop a solution capable of mapping all wildfire events across the entire Pantanal biome. Additionally, we plan to extend the proposed approach to other types of ecological disturbances, such as flooding and deforestation.

References

Cho, A. Y., Park, S.-e., Kim, D.-j., Kim, J., Li, C., Song, J., 2022. Burned area mapping using Unitemporal Planetscope imagery with a deep learning based approach. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16, 242–253.

Contributors, M., 2020. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. `https://github.com/open-mmlab/mmsegmentation`.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Hu, X., Ban, Y., Nascetti, A., 2021. Uni-temporal multispectral imagery for burned area mapping with deep learning. *Remote sensing*, 13(8), 1509.

Jo, S., Yu, I.-J., 2021. Puzzle-cam: Improved localization via matching partial and full features. *2021 IEEE International Conference on Image Processing (ICIP)*, IEEE, 639–643.

Leal Filho, W., Azeiteiro, U. M., Salvia, A. L., Fritzen, B., Libonati, R., 2021. Fire in Paradise: Why the Pantanal is burning. *Environmental Science & Policy*, 123, 31–34.

Seydi, S. T., Hasanlou, M., Chanussot, J., 2022. Burnt-Net: Wildfire burned area mapping with single post-fire Sentinel-2 data and deep learning morphological neural network. *Ecological Indicators*, 140, 108999.

Wang, Y., Zhang, J., Kan, M., Shan, S., Chen, X., 2020. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., Luo, P., 2021. Segformer: Simple and efficient design for semantic segmentation with transformers. *Neural Information Processing Systems (NeurIPS)*.

Zhang, H., Wu, C., Zhang, Z., Zhu, Y., Lin, H., Zhang, Z., Sun, Y., He, T., Mueller, J., Manmatha, R. et al., 2022. Resnest: Split-attention networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2736–2746.