

Evaluating Deep Segmentation Models for Wildfire Mapping Using Sentinel-2 Satellite Imagery

Avaliação de Modelos de Segmentação Profunda para Mapeamento de Incêndios Florestais Usando Imagens de Satélite Sentinel-2

Victor Hugo Lara Souza¹, Rodrigo Smarzaró¹, Leandro Henrique Furtado Pinto Silva¹, João Fernando Mari¹

Abstract: Wildfires pose serious threats to ecosystems, human health, and urban infrastructure, making accurate and efficient mapping of burned areas essential for mitigation and recovery. In this context, the use of computer vision techniques has shown highly effective results for rapid and accurate delineation of burned areas. However, these techniques require considerable computational cost, and their performance is sensitive to model selection and hyperparameter adjustments. Therefore, this work evaluates deep learning models for the semantic segmentation of regions affected by wildfires using multispectral images from the CEMS-Wildfire dataset, obtained from the Sentinel-2 satellite. We evaluated the combination of three traditional architectures: U-Net, Feature Pyramid Network (FPN), and SegFormer, combined with different encoders (ResNet-50, EfficientNet B2, and MiT-B0), under different data augmentation strategies (none, mild, strong) and loss functions (Binary cross-entropy, Dice loss, and hybrid BCE + Dice loss). U-Net achieved the best performance with the MiT B0 encoder, trained using Dice loss and moderate augmentation, resulting in an IoU of 0.7790 and an F1-score of 0.8758.

Keywords: wildfires — semantic segmentation — deep learning — data augmentation

Resumo: Incêndios florestais representam graves ameaças aos ecossistemas, à saúde humana e à infraestrutura urbana, tornando o mapeamento preciso e eficiente de áreas queimadas essencial para mitigação e recuperação. Nesse contexto, o uso de técnicas de visão computacional tem mostrado resultados altamente eficazes para uma delimitação rápida e precisa de áreas queimadas. Entretanto, essas técnicas demandam custo computacional considerável além de terem o desempenho sensível à escolha do modelo e ajustes de hiperparâmetros. Assim, este trabalho avalia modelos de aprendizado profundo para segmentação semântica de regiões afetadas por incêndios florestais usando imagens multiespectrais do conjunto de dados CEMS-Wildfire e obtidas a partir do satélite Sentinel-2. Realizamos a avaliação da combinação de três arquiteturas tradicionais, U-Net, Feature Pyramid Network (FPN) e SegFormer, combinadas com diferentes encoders (ResNet-50, EfficientNet B2, e MiT-B0), em diferentes estratégias de aumento de dados (none, mild, strong) e funções de perda (Binary cross-entropy, Dice loss, e hybrid BCE + Dice loss). O melhor desempenho foi alcançado pelo U-Net com encoder MiT B0, treinado com perda Dice e aumento moderado, atingindo um IoU de 0,7790 e um F1-score de 0,8758.

Palavras-Chave: incêndios florestais — segmentação semântica — aprendizado profundo — aumento de dados

¹ Instituto de Ciências Exatas e Tecnológicas, Universidade Federal de Viçosa (UFV), Brazil

*Corresponding author: joaofmari@gmail.com

DOI: <http://dx.doi.org/10.22456/2175-2745.150922> • Received: 14/10/2025 • Accepted: 03/11/2025

CC BY-NC-ND 4.0 - This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

1. Introduction

Wildfires have proven to be a significant threat to the environment, human health, and urban planning. In recent years, the number of fires has increased due to the effects of climate change and direct human activities. In Brazil, recent data

show that in August 2024 alone, 11,396,079 hectares were burned [1, 2]. These figures highlight the importance of using efficient methods for delineating burned areas, aiming at damage reduction, prevention, and the recovery of affected regions.

Manual analysis of satellite images is inefficient and can be very time-consuming, highlighting the need for the use of deep learning techniques for the segmentation of burned areas [3, 4]. Satellite images, such as those from Sentinel-2, offer high resolution, making them an ideal tool for monitoring wildfires through computer vision techniques [5].

Advances in convolutional neural network (CNN) architectures, such as U-Net and its modern variants, have demonstrated strong effectiveness for semantic segmentation in remote sensing. Architectures such as U-Net have stood out for delineating features in satellite images and are widely adopted [6, 7]. Despite this, many challenges remain, including variations in terrain and lighting, as well as limited datasets. In particular, the segmentation of forest fires proves to be challenging due to spectral variability across different regions, the presence of clouds, and atmospheric artifacts.

In this context, the selection of an appropriate dataset is essential. The CEMS-Wildfire dataset [8], although little explored in the literature, provides high-quality annotations, geographical diversity, and variability in atmospheric conditions, offering images and burned area masks collected across Europe. Our goal is to address the gap in systematic studies by applying and comparing various architectures for binary segmentation on this dataset.

In this study, we evaluated several semantic segmentation architectures (U-Net, Feature Pyramid Network (FPN), and SegFormer), each one combined with different pre-trained encoders: ResNet-50, EfficientNet-B2, and MiT-B0. All architecture-decoder pairings were implemented through the Segmentation Models PyTorch (SMP) framework [9], which handles the adaptation between encoders and decoders transparently. We assessed the performance of each fine-tuned model on the CEMS-Wildfire dataset using standard evaluation metrics, including Intersection over Union (IoU), F1-score, and accuracy. Our study aims to provide a solid foundation for future research on forest fire segmentation using deep learning.

This paper is organized as follows. Section 2 reviews related work and outlines the state of the art. Section 3 describes the materials and methods employed in this study. Section 4 presents and discusses the experimental results, while Section 5 concludes the paper.

2. Related works

Image segmentation is a fundamental task in computer vision and is especially useful for wildfire monitoring, where it helps in detecting and mapping burned areas. Deep learning methods, mainly convolutional neural networks (CNNs), have been widely applied to this problem with good results.

Khryashchev et al. [6] proposed a U-Net with a ResNet-34 encoder for wildfire segmentation in high-resolution aerial images. Using two RGB datasets and extensive data augmentation, the model achieved strong performance across multiple metrics, highlighting the potential of CNNs for early wildfire detection.

Colomba et al. [10] released a publicly available Sentinel-2 dataset for burned area segmentation, including 73 annotated images with five severity levels, providing a valuable resource for wildfire damage assessment.

Arnaudo et al. [8] introduced the CEMS-Wildfire dataset, built from Sentinel-2 imagery and Copernicus EMS annotations, which also provides land cover maps and cloud masks for multitask learning. They further proposed a multitask learning framework with UPerNet and SegFormer, where land cover classification acts as an auxiliary task to improve the robustness of burned area segmentation.

Al-Dabbagh et al. [11] proposed a wildfire detection workflow using uni-temporal Sentinel-2 imagery, introducing a manually labeled dataset with 21,690 images from Turkey. They evaluated several segmentation models, such as U-Net, LinkNet, Attention ResU-Net, and DeepLabV3+, and found U-Net with a ResNet-50 encoder to be the most effective.

Sismanis et al. [4] proposed an unsupervised method for burned area mapping from Sentinel-2 imagery, combining adaptive thresholding of spectral indices (NBR, NDVI, NDWI) with mean-shift segmentation. Their results were comparable to official CEMS products, showing the potential of low-cost approaches when labeled data are scarce.

Our work differs from the previous studies by providing a comprehensive and systematic comparison of several combinations of deep learning architectures and encoders, along with an evaluation of data augmentation strategies and loss. The models were evaluated on the CEMS-Wildfire dataset, which remains relatively unexplored, using standardized splits and metrics (IoU, F1-score, and accuracy). This systematic and reproducible analysis offers a broader performance perspective and can help guide future research and operational efforts in wildfire mapping.

3. Material and Methods

3.1 Dataset

We used the CEMS-Wildfire Dataset¹ [8], which contains imagery acquired by the Sentinel-2 satellite and annotated by the Copernicus Emergency Management Service (CEMS). The dataset was originally designed for multitask learning, providing annotations for burned area segmentation, land-cover classification, and cloud detection. In this study, we used only the visible RGB bands of Sentinel-2 (B2: Blue, B3: Green, and B4: Red), each with a spatial resolution of 10 m, which were properly stacked to compose an RGB image.

In this work, we focused exclusively on the binary segmentation task, employing the masks where each pixel is labeled as either burned area or non-burned area. The imagery used has a spatial resolution of 10 m for the selected bands, with precise co-registration between images and masks. The dataset spans diverse landscapes, vegetation types, and atmospheric conditions, thereby capturing high variability and enhancing the robustness of model evaluation.

¹<https://github.com/MatteoM95/CEMS-Wildfire-Dataset>

The complete dataset comprises 433 pairs of images and masks, which we partitioned into 281 for training, 53 for validation, and 99 for testing. Each image contains regions impacted by wildfires, as detected and mapped by CEMS, providing a reliable reference for the burned area delineation task.

3.2 Preprocessing and Data Augmentation

To improve the generalization ability of the models and enhance robustness to variations in input data, we adopted different data augmentation strategies during training. All transformations were implemented using the Albumentations library, with a fixed random seed to ensure reproducibility. Images were normalized using the ImageNet dataset statistics and resized to a fixed resolution of (512×512) as defined in each experiment.

We evaluated three augmentation settings with increasing intensity: (i) No Augmentation (No DA), (ii) Mild Augmentation (DA-1), (iii) and Strong Augmentation (DA-2).

No Augmentation (No DA): This baseline setup included only image resizing and normalization. No transformations were applied to alter the visual appearance of the images.

Mild Augmentation (DA-1): This configuration introduced light geometric and photometric changes, including horizontal and vertical flips, 90-degree rotations, affine transformations (scaling, translation, rotation), and brightness/contrast adjustments. These transformations simulate minor variations in viewpoint and lighting conditions while preserving semantic content.

Strong Augmentation (DA-2): In addition to all transformations from the mild setting, this setup incorporated hue/saturation/value (HSV) shifts, Gaussian noise, elastic deformations, grid and optical distortions, and local contrast enhancement through CLAHE. These augmentations aim to simulate more challenging real-world conditions and structural distortions in the data.

For validation and testing, we applied only resizing and normalization, ensuring a fair and consistent evaluation without any augmentation-induced biases.

3.3 Architectures and Encoders

Architectures for semantic segmentation define how features are decoded and combined to produce the final mask. In this work, we use three architectures (U-Net, FPN, and SegFormer), each combined with different pretrained encoders.

U-Net [12] is a convolutional neural network designed for image segmentation. It is divided into two symmetrical parts, the encoder and the decoder. The encoder reduces the image resolution and extracts features, while the decoder restores the spatial resolution and generates the segmentation mask.

Feature Pyramid Network [13] improves feature maps for segmentation by merging high-level (semantically rich, lower-resolution) features with low-level (detail-rich, higher-resolution) features. It uses a top-down pathway and lateral connections to combine these multi-scale features.

Segformer [14] is a transformer-based model that combines lightweight multilayer perceptrons for decoding with hierarchical vision transformers. SegFormer captures long-range dependencies without the need for convolutions or positional encodings.

The encoder (backbone) is responsible for extracting multi-scale features from the input image. We evaluated three pre-trained encoders (ResNet-50, EfficientNet-B2, and MiT-B0) combined with each architecture.

ResNet [15] introduces residual connections to address the vanishing gradient problem in deep networks. It uses a stack of residual blocks that allow the network to learn identity mappings and deeper representations without degradation. ResNet50 provides a good balance between depth and performance.

EfficientNet-B2 [16] presents a compound scaling method that uniformly scales network depth, width, and input resolution. This approach allows the model to achieve high accuracy while maintaining computational efficiency. EfficientNet-B2 is a mid-scale variant in the family, offering an effective balance between accuracy and resource usage, making it suitable for tasks involving high-resolution imagery, such as wildfire segmentation.

MiT-B0 [14] is the smallest variant of the Mix Transformer architecture, designed as an efficient encoder for segmentation tasks in SegFormer. Unlike convolution-based encoders, MiT-B0 uses hierarchical transformers to capture long-range dependencies and global context without relying on positional encodings. It employs overlapping patch embeddings and a progressive reduction in spatial resolution, enabling multi-scale feature extraction with fewer parameters. This lightweight design provides an excellent trade-off between computational efficiency and segmentation accuracy, which is especially beneficial for large-scale remote sensing applications.

3.4 Training Procedure

We conducted all experiments on a binary semantic segmentation task, where each pixel is classified as foreground or background. Models were implemented in PyTorch using encoder-decoder architectures provided by the SMP library².

We evaluated three segmentation models (U-Net, FPN, and SegFormer), each combined with three encoder backbones: ResNet-50, EfficientNet-B2, and MiT-B0. For each model configuration, we trained variants using three data augmentation settings (none, mild, strong) and three loss functions, resulting in a total of 54 distinct experimental conditions.

Training was performed using the Adam optimizer with a fixed learning rate of 1×10^{-5} , a batch size of 8, and early stopping with a patience of 20 epochs. The learning rate was scheduled using a reduction-on-plateau strategy, which decreases the rate when the validation loss plateaus. Each model was trained for up to 400 epochs, and the checkpoint

²https://github.com/qubvel-org/segmentation_models.pytorch

with the lowest validation loss was retained for evaluation [17].

We compared three loss functions: (1) binary cross-entropy (BCE), implemented with label smoothing (0.01); (2) Dice loss, computed from logits and optimized for region-level consistency; and (3) a hybrid BCE + Dice loss that combines pixel-level and structural supervision.

Predictions were obtained by applying a sigmoid activation to the model output, followed by thresholding at 0.5. During training and validation, we computed standard metrics including pixel accuracy, IoU, and F1-score. To ensure consistent and reliable evaluation, we used only metrics provided by SMP, with macro reduction applied in all cases.

This training setup enables a controlled comparison of architectural choices, data augmentation strategies, and loss formulations under a unified framework, forming the basis for the results presented in Section 4.

3.5 Model Evaluation

We evaluated the performance of all trained models on the test set using standard metrics widely adopted in binary semantic segmentation: Intersection over Union (IoU), F1-score, and pixel-wise accuracy. All metrics were computed using functions from the SMP library, with macro reduction applied. In macro reduction, the confusion matrix components (true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN)) are first accumulated over all test images, and the final metrics are computed from the aggregated values.

Intersection over Union (IoU) quantifies the overlap between the predicted segmentation and the ground truth masks, and is defined as:

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}. \quad (1)$$

Higher IoU values indicate stronger agreement between prediction and ground truth.

The F1-score (Dice Coefficient) evaluates the harmonic mean of precision and recall, emphasizing a balanced trade-off between false positives and false negatives:

$$\text{F1-score} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}}. \quad (2)$$

This metric is particularly robust in imbalanced settings, where the foreground class occupies a small portion of the image.

Pixel accuracy reflects the overall proportion of correctly classified pixels, computed as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \quad (3)$$

Although accuracy is less sensitive to class imbalance, it remains a useful general-purpose indicator of model performance across the full image domain.

3.6 Computational Resources and Reproducibility

All experiments were conducted on a workstation equipped with an Intel Core i5 processor (3.0 GHz), 32 GB of RAM, and an NVIDIA Titan Xp GPU with 12 GB of memory. The development environment was based on Python 3.12 and PyTorch 2.7.1 with CUDA 12.6 support, using the Segmentation Models PyTorch (SMP) library version 0.5 and Albumentations version 2.0.8 for data augmentation. Additional libraries employed included scikit-learn 1.7, scikit-image 0.25.2, Pillow 11.0, and Matplotlib 3.10.3 for data processing, visualization, and evaluation.

4. Results and Discussion

Table 1 presents the results of all trained models across three architectures (U-Net, FPN, and SegFormer), each combined with three different encoder backbones (ResNet-50, EfficientNet-B2, and MiT-B0). For each configuration, results are reported under three data augmentation strategies (No DA, DA-1, and DA-2), as described in Section 3.2). The table includes evaluation for three different loss functions (Cross Entropy, Dice, and a hybrid CE + Dice) measured using the metrics IoU, F1-score, and pixel accuracy. All metrics were computed on the test set using macro reduction, as described in Section 3.5. The best results for each model–encoder combination, considering all data augmentation strategies and loss functions, are underlined for each metric, whereas the overall best results for each model are emphasized in bold.

In addition to Table 1, Figure 2 illustrates the same results using line charts for both IoU (top row) and F1-score (bottom row). Each chart corresponds to one of the three architectures (U-Net, FPN, and SegFormer), evaluated with different encoders (ResNet-50, EfficientNet-B2, and MiT-B0), represented by distinct colors. The loss functions (Cross Entropy, Dice, and CE + Dice) are indicated by different markers, while the x-axis shows the three data augmentation strategies (No DA, DA-1, and DA-2). Thus, each curve represents a specific encoder–loss configuration, with colors distinguishing the encoders and markers identifying the loss functions.

The best performance was obtained by U-Net with the MiT-B0 encoder, trained with Dice loss and mild data augmentation (DA-1), reaching an IoU of 0.7790 and an F1-score of 0.8758. FPN also reached its best result with the MiT-B0 encoder and mild augmentation, but using Cross Entropy loss, with an IoU of 0.7657 and an F1-score of 0.8673. SegFormer showed the lowest top result among the three models; unlike U-Net and FPN, its best configuration used the EfficientNet-B2 encoder with the combined CE + Dice loss, achieving an IoU of 0.7519 and an F1-score of 0.8584.

Among the tested encoders, MiT-B0 consistently outperformed ResNet-50 and EfficientNet-B2 in terms of both IoU and F1-score across U-Net and FPN. While EfficientNet-B2 also showed good results, its improvements over ResNet-50 were smaller and less consistent. SegFormer performed better with EfficientNet-B2, but did not surpass the top results of U-Net or FPN. All models showed high pixel accuracy, above

Table 1. Performance metrics for different models and encoders using three loss functions (Cross Entropy, Dice, CE+Dice), with and without data augmentation (DA).

Model + Encoder		DA	Cross Entropy			Dice			CE + Dice		
			IoU	F1-score	Acc.	IoU	F1-score	Acc.	IoU	F1-score	Acc.
U-Net	ResNet-50	No DA	0.6906	0.8170	0.9660	0.6726	0.8043	0.9619	0.6679	0.8009	0.9621
		DA-1	0.7234	0.8395	0.9697	<u>0.7505</u>	<u>0.8575</u>	<u>0.9736</u>	0.7300	0.8440	0.9706
		DA-2	0.7124	0.8320	0.9687	0.7164	0.8348	0.9687	0.6896	0.8163	0.9642
	EfficientNet-B2	No DA	0.7125	0.8321	0.9681	0.7091	0.8298	0.9673	0.7104	0.8307	0.9675
		DA-1	0.7377	0.8491	0.9715	<u>0.7409</u>	<u>0.8512</u>	<u>0.9719</u>	0.7370	0.8486	0.9715
		DA-2	0.7099	0.8303	0.9678	0.7061	0.8277	0.9667	0.7149	0.8337	0.9684
	Mit-B0	No DA	0.7617	0.8647	0.9762	0.7576	0.8621	0.9746	0.7554	0.8607	0.9749
		DA-1	0.7428	0.8524	0.9729	0.7790	0.8758	0.9774	0.7417	0.8517	0.9724
		DA-2	0.7464	0.8548	0.9734	0.7508	0.8577	0.9732	0.7522	0.8586	0.9738
FPN	ResNet-50	No DA	0.6474	0.7860	0.9599	0.6558	0.7921	0.9595	0.6583	0.7940	0.9612
		DA-1	0.7159	0.8344	<u>0.9697</u>	<u>0.7176</u>	<u>0.8356</u>	0.9691	0.7029	0.8255	0.9672
		DA-2	0.6787	0.8086	0.9644	0.6297	0.7728	0.9544	0.6889	0.8158	0.9652
	EfficientNet-B2	No DA	0.7316	0.8450	0.9716	0.7084	0.8293	0.9678	0.7276	0.8423	0.9707
		DA-1	0.7496	0.8569	0.9738	0.7493	0.8567	0.9736	<u>0.7534</u>	<u>0.8594</u>	<u>0.9741</u>
		DA-2	0.7062	0.8278	0.9676	0.6925	0.8184	0.9645	0.7016	0.8246	0.9663
	Mit-B0	No DA	0.7635	0.8659	0.9765	0.7430	0.8525	0.9725	0.7621	0.8650	0.9756
		DA-1	0.7657	0.8673	0.9765	0.7401	0.8506	0.9720	0.7510	0.8578	0.9743
		DA-2	0.7167	0.8349	0.9698	0.7187	0.8363	0.9695	0.7107	0.8309	0.9680
SegFormer	ResNet-50	No DA	0.6646	0.7985	0.9628	0.6794	0.8091	0.9654	0.6766	0.8071	0.9651
		DA-1	0.7393	0.8501	<u>0.9729</u>	<u>0.7396</u>	<u>0.8503</u>	0.9722	0.7269	0.8419	0.9707
		DA-2	0.7001	0.8236	0.9674	0.6837	0.8121	0.9644	0.6899	0.8165	0.9647
	EfficientNet-B2	No DA	0.7326	0.8457	0.9717	0.7341	0.8466	0.9712	0.7416	0.8517	0.9729
		DA-1	0.7437	0.8530	0.9733	0.7420	0.8519	0.9722	0.7519	0.8584	0.9742
		DA-2	0.7129	0.8324	0.9689	0.6937	0.8191	0.9646	0.7076	0.8288	0.9675
	Mit-B0	No DA	0.7049	0.8269	0.9674	0.7280	0.8426	0.9703	0.7375	0.8489	0.9722
		DA-1	0.7198	0.8371	0.9697	0.7358	0.8478	0.9714	0.7364	0.8482	0.9719
		DA-2	<u>0.7452</u>	<u>0.8540</u>	<u>0.9735</u>	0.7169	0.8351	0.9687	0.7299	0.8438	0.9709

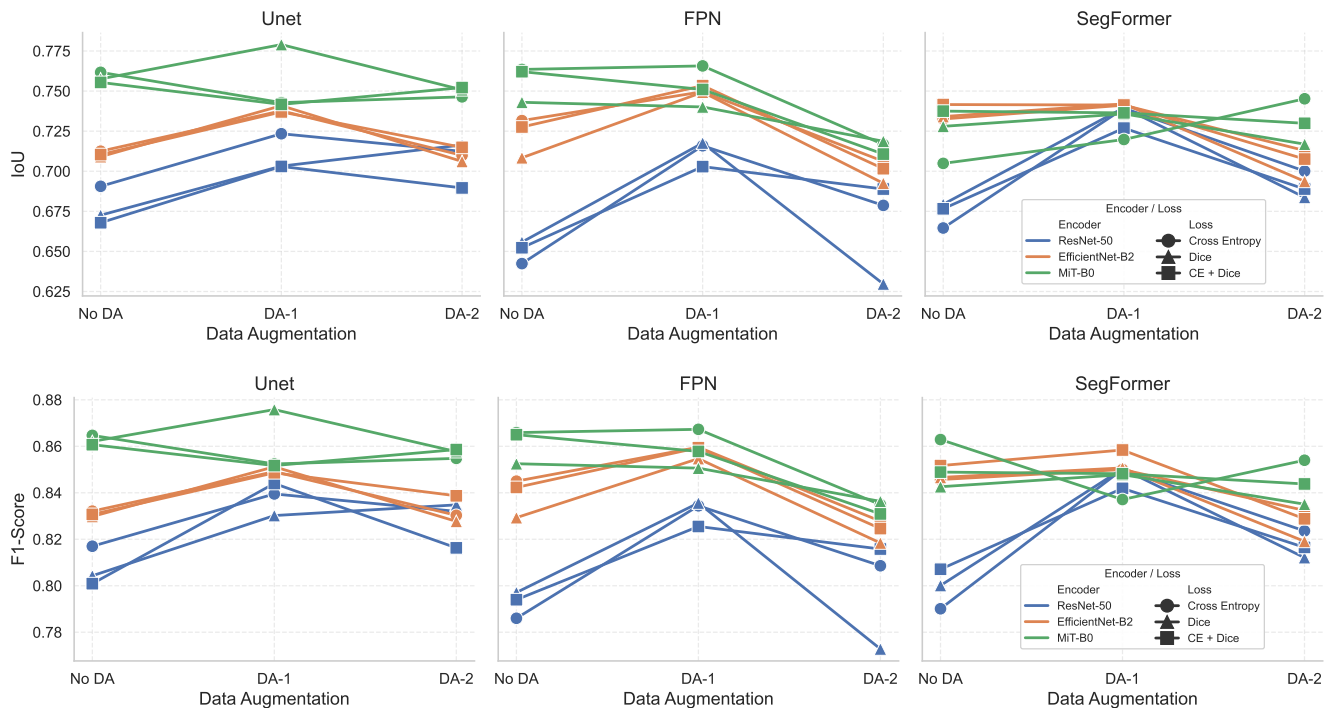


Figure 1. IoU (top row) and F1-score (bottom row) for the three architectures (U-Net, FPN, and SegFormer). Within each plot, results are shown for different encoders (ResNet-50, EfficientNet-B2, and MiT-B0, represented by colors) and loss functions (Cross Entropy, Dice, and CE + Dice, represented by markers). The x-axis corresponds to the data augmentation strategies (No DA, DA-1, and DA-2).

0.96, but this metric is less informative in unbalanced cases like this one, where the background class occupies most of the image. These results confirm that the choice of the model-encoder combination plays an important role in segmentation performance and show that transformer-based encoders like MiT-B0 can be very effective.

4.1 Impact of the Loss Function

For U-Net, the Dice loss yielded the best results across all encoders, with MiT-B0 reaching an IoU of 0.7790 and an F1-score of 0.8758 (DA-1), the best overall results. Considering the encoders, ResNet-50 achieved its highest scores with Dice in all architectures. EfficientNet-B2 performed best with CE + Dice in FPN with IoU of 0.7534 and F1-score of 0.8594 (DA-1) and in SegFormer with IoU of 0.7519 and F1-score of 0.8584 (DA-1), although Dice was superior for U-Net with IoU of 0.7409 and F1-score of 0.8512 (DA-1). MiT-B0 obtained its top results with Cross Entropy in FPN with IoU of 0.7657 and F1-score of 0.8673 (DA-1) and in SegFormer with IoU of 0.7452 and F1-score of 0.8540 (DA-2), while for U-Net the best configuration was with Dice. These findings indicate that the optimal loss function depends on the specific combination of architecture and encoder.

4.2 Impact of Data Augmentation

Data augmentation proved to be an important factor in improving segmentation quality, though its effect varied with the intensity of the transformations. A clear gain was observed

when moving from no augmentation (No DA) to the mild augmentation (DA-1) across nearly all models. For example, U-Net with MiT-B0 under DA-1 reached the best IoU (0.7790) and F1-score (0.8758), while FPN with MiT-B0 also peaked under DA-1. However, the more aggressive DA-2 strategy often led to inconsistent or even degraded performance, with the exception of SegFormer with MiT-B0 under DA-2. In summary, mild augmentation (DA-1) offered the most consistent trade-off between robustness and stability, consistently producing the best results across architectures, while strong augmentation (DA-2) should be applied cautiously, as it rarely provided additional benefits and could harm performance.

4.3 Qualitative Analysis

Figure 2 presents four randomly selected samples from the test set, illustrating the segmentation results of the best-performing model, namely UNet with MiT-B0 encoder, trained using Dice loss and strong data augmentation. Each column corresponds to one sample, while each row shows: (a) the RGB input image, (b) the ground truth mask, (c) the predicted segmentation mask, and (d) the comparison map, where true positives (TP) are shown in green, false positives (FP) in red, false negatives (FN) in orange, and true negatives (TN) in black.

5. Conclusion

In this work, we conducted a systematic evaluation of deep learning architectures for wildfire semantic segmentation us-

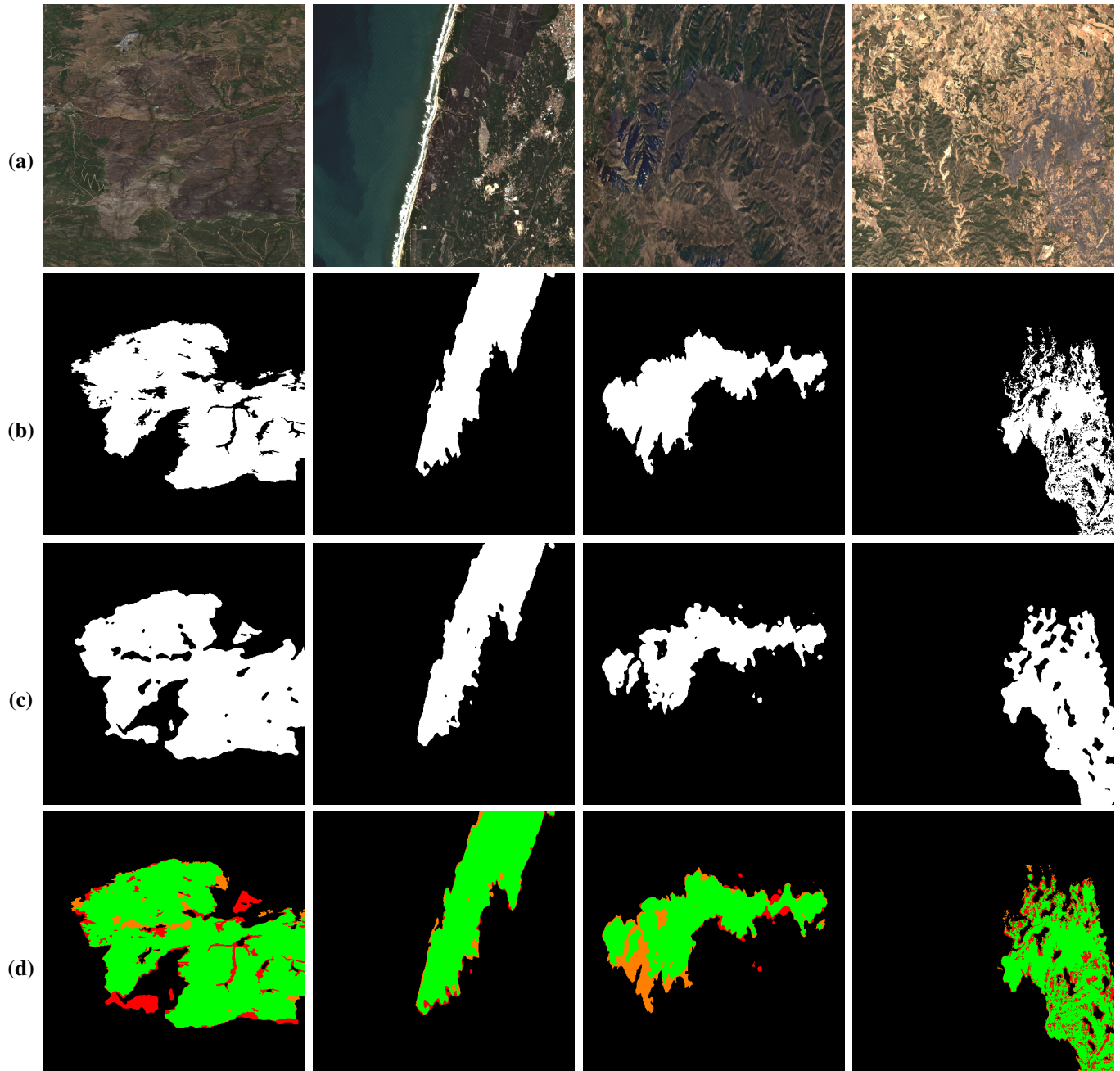


Figure 2. Five sample images randomly selected from the test set to present the results. The images were segmented by the best overall model + decoder (UNet + Mit-B0 trained with Dice loss function and strong data augmentation). (a) RGB input image, (b) Ground truth, (c) segmented image, (d) segmentation map, highlighting TP (green), TN (black), FP (red), and FN (orange).

ing Sentinel-2 imagery and the CEMS-Wildfire dataset. We analyzed the performance of U-Net, FPN, and SegFormer with different encoders, loss functions, and data augmentation strategies, providing a comprehensive comparison of their strengths and limitations. The results showed that U-Net combined with the MiT-B0 encoder and trained with Dice loss under mild augmentation (DA-1) achieved the best overall performance, with IoU and F1-score values of 0.7790 and

0.8758, respectively.

Our findings highlight three main aspects: (i) transformer-based encoders such as MiT-B0 consistently outperformed traditional convolutional backbones like ResNet-50 and EfficientNet-B2 in U-Net and FPN, confirming their potential for remote sensing segmentation tasks; (ii) the choice of loss function proved to be dependent on the model–encoder combination, with Dice loss being particularly effective for U-Net, while hy-

brid and cross-entropy losses benefited other setups; (iii) data augmentation played a crucial role in improving generalization, with mild augmentation (DA-1) consistently enhanced results, whereas stronger augmentations (DA-2) often introduced instability or degraded performance.

Overall, this study contributes a rigorous benchmark for wildfire segmentation and reinforces the importance of carefully selecting model architecture, encoder, and training strategies. Future work may explore multiclass segmentation to identify different levels of wildfire damage, domain adaptation across different geographical regions and land cover types, and the integration of temporal information. These directions have the potential to further improve the accuracy and robustness of automated wildfire mapping, supporting timely disaster response and environmental monitoring.

Acknowledgements

We would like to thank CAPES, CNPq, and FAPEMIG for the financial support. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. V. H. L. Souza received a scholarship from PIBIC/CNPq.

Author Contributions

Conceptualization: V.H.L.S., R.S., L.H.F.P.S., and J.F.M. Methodology: V.H.L.S., R.S., L.H.F.P.S., and J.F.M. Code Development: V.H.L.S., and J.F.M. Experiments: V.H.L.S., R.S., L.H.F.P.S., and J.F.M. Discussion of Results: V.H.L.S., R.S., L.H.F.P.S., and J.F.M. Writing - Original Draft: V.H.L.S., R.S., L.H.F.P.S., and J.F.M. Writing - Review & Editing: V.H.L.S., R.S., L.H.F.P.S., and J.F.M. Supervision: R.S., L.H.F.P.S., and J.F.M.

References

- [1] REQUIA, W. J. Fires in brazil: health crises and the failure of government action. *The Lancet Regional Health–Americas*, Elsevier, v. 39, 2024.
- [2] MAPBIOMAS. *MapBiomass Brasil*. 2025. <<https://plataforma.brasil.mapbiomas.org/monitor-do-fogo>>. Accessed: 2025-09-01.
- [3] PHIRI, D. et al. Sentinel-2 data for land cover/use mapping: A review. *Remote sensing*, Mdpi, v. 12, n. 14, p. 2291, 2020.
- [4] SISMANIS, M. et al. An unsupervised burned area mapping approach using sentinel-2 images. *Land*, MDPI, v. 12, n. 2, p. 379, 2023.
- [5] SILVA, L. H. F. P. et al. Cloud removal with compact diffusion models: A residual block-based approach. *Remote Sensing Applications: Society and Environment*, Elsevier, p. 101680, 2025.
- [6] KHRYASHCHEV, V.; LARIONOV, R. Wildfire segmentation on satellite images using deep learning. In: IEEE. *2020 Moscow Workshop on Electronic and Networking Technologies (MWENT)*. [S.l.], 2020. p. 1–5.
- [7] BRAND, A.; MANANDHAR, A. Semantic segmentation of burned areas in satellite images using a u-net-based convolutional neural network. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Copernicus GmbH, v. 43, p. 47–53, 2021.
- [8] ARNAUDO, E. et al. Robust burned area delineation through multitask learning. In: SPRINGER. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. [S.l.], 2023. p. 436–447.
- [9] IAKUBOVSKII, P. *Segmentation Models Pytorch*. [S.l.]: GitHub, 2019. <https://github.com/qubvel/segmentation_models.pytorch>.
- [10] COLOMBA, L. et al. A dataset for burned area delineation and severity estimation from satellite imagery. In: *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. [S.l.: s.n.], 2022. p. 3893–3897.
- [11] AL-DABBAGH, A. M.; ILYAS, M. Uni-temporal sentinel-2 imagery for wildfire detection using deep learning semantic segmentation models. *Geomatics, Natural Hazards and Risk*, Taylor & Francis, v. 14, n. 1, p. 2196370, 2023.
- [12] RONNEBERGER, O.; FISCHER, P.; BROX, T. U-net: Convolutional networks for biomedical image segmentation. In: SPRINGER. *International Conference on Medical image computing and computer-assisted intervention*. [S.l.], 2015. p. 234–241.
- [13] LIN, T.-Y. et al. Feature pyramid networks for object detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2017. p. 2117–2125.
- [14] XIE, E. et al. Segformer: Simple and efficient design for semantic segmentation with transformers. In: *Advances in Neural Information Processing Systems (NeurIPS)*. [S.l.: s.n.], 2021. v. 34, p. 12077–12090.
- [15] HE, K. et al. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2016. p. 770–778.
- [16] TAN, M.; LE, Q. V. Efficientnet: Rethinking model scaling for convolutional neural networks. In: *Proceedings of the 36th International Conference on Machine Learning (ICML)*. [S.l.]: PMLR, 2019. v. 97, p. 6105–6114. ICML 2019.
- [17] FERREIRA, J. et al. Evaluating multiple combinations of models and encoders to segment clouds in satellite images. In: INSTICC. *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 3: VISAPP*. [S.l.]: SciTePress, 2024. p. 233–241. ISBN 978-989-758-679-8.