

Vision Transformers for Burned Area Delineation

Original

Vision Transformers for Burned Area Delineation / Rege Cambrin, Daniele; Colomba, Luca; Garza, Paolo. - 3343:(2023).
(Intervento presentato al convegno MACLEAN: MACHine Learning for EArth ObservatioN (workshop @ECML/PKDD2022) tenutosi a Grenoble (FR) nel 19/09/2022).

Availability:

This version is available at: 11583/2971022 since: 2023-02-15T10:51:18Z

Publisher:

CEUR-WS

Published

DOI:

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

Vision Transformers for Burned Area Delineation

Daniele Rege Cambrin^{1,*}, Luca Colomba¹ and Paolo Garza¹

¹Politecnico di Torino, Corso Duca degli Abruzzi, 24, Torino, 10129, Italy

Abstract

The automatic identification of burned areas is an important task that was mainly managed manually or semi-automatically in the past. In the last years, thanks to the availability of novel deep neural network architectures, automatic segmentation solutions have been proposed also in the emergency management domain. The most recent works in burned area delineation leverage on Convolutional Neural Networks (CNNs) to automatically identify regions that were previously affected by forest wildfires. A largely adopted segmentation model, U-Net, demonstrated good performances for the task under analysis, but in some cases a high overestimation of burned areas is given, leading to low precision scores. Given the recent advances in the field of NLP and the first successes also in the vision domain, in this paper we investigate the adoption of vision transformers for semantic segmentation to address the burned area identification task. In particular, we explore the SegFormer architecture with two of its variants: the smallest MiT-B0 and the intermediate one MiT-B3. The experimental results show that SegFormer provides better predictions, with higher precision and F1 score, but also better performance in terms of the number of parameters with respect to CNNs.

Keywords

Earth Observation, Deep Learning, Semantic Segmentation,

1. Introduction

The preservation and continuous monitoring of natural resources is a fundamental topic that, over the years, pushed by climate change and rapid succession of natural hazards, assumed higher and higher relevance among the research community and society in general. The availability of sensors with high resolution, in conjunction with the usage of aircraft and satellites, enables the acquisition of national- and global-scale information in a short amount of time. Moreover, thanks to the recent advances in computer vision and the high availability of data in the remote sensing domain, such a topic represents an active field of research with a strong community being involved.

The Earth Observation domain involves several different tasks, ranging from land monitoring and land cover change characterization [1], change detection [2], damage estimation [3] and many others. Deep learning-based methodologies demonstrated state-of-the-art performances over a multitude of these tasks (e.g., [4, 5]).


MACLEAN: *MACHine Learning for EArth ObservatioN Workshop 2022, in conjunction with ECML/PKDD (European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases), September 19, 2022, Grenoble, France*

*Corresponding author.

✉ daniele.regecambrin@polito.it (D. R. Cambrin); luca.colomba@polito.it (L. Colomba); paolo.garza@polito.it (P. Garza)

ORCID 0000-0002-5067-2118 (D. R. Cambrin); 0000-0003-2911-4522 (L. Colomba); 0000-0002-1263-7522 (P. Garza)

© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

Among the Earth Observation domain, the field of emergency management plays an important role for public authorities as well as governments in handling natural hazards, trying to limit societal and environmental damages as much as possible with timely intervention and proper restoration. Handling natural hazards also involves precise identification of affected areas, damage estimation and restoration process planning. Such mentioned operations are often performed in-situ by human operators, requiring a great amount of time and effort to quantify the negative impact of the concluded catastrophic event. The availability of remote sensing data and satellite imagery enables the development of automatic recognition systems to delimit affected areas and provide initial damage assessments for operators and authorities.

In this context, we concentrate our analyses on forest fires. More specifically, we propose our work in the field of semantic segmentation and automatic burned area identification from Copernicus Sentinel-2 L2A acquisitions, a European multi-spectral imaging mission with a resolution up to 10m (depending on the spectral band): given a post-fire multispectral acquisition from Sentinel-2, the goal is to precisely identify the region affected by the already-extinguished forest fire. This paper explores the application of one of the most recent advances in deep learning and computer vision: transformer-based architectures for semantic segmentation. In particular, we assess the performances of the SegFormer [6] architecture on an open dataset in comparison with a standard thresholding baseline and a CNN-based state-of-the-art architecture, namely U-Net [7]. The considered model proved superior performance compared to both methods. Our source code is available at <https://github.com/DarthReca/vit-burned-detection>.

The paper is structured as follows: Section 2 introduces the related works in the field of deep learning for automatic burned area detection, Section 3 introduces the vision transformers for semantic segmentation and the SegFormer model, whereas Section 4 is the experimental section, in which the quantitative results of the considered methodologies are compared. Finally, Section 5 concludes the paper.

2. Related work

The burned area identification problem, also named as burned area delineation problem, is a well-known and tackled challenge in remote sensing literature. The aforementioned issue consists in identifying, given a multispectral input acquisition, the areas previously affected by forest wildfire and currently damaged. Such information is useful to (i) quantify damages, both environmental and economical, for public authorities and (ii) plan the restoration process. In scientific literature, before the advent of modern computer vision methodologies, researchers tackled the problem with the analysis of burned area indexes. Specifically, by gathering and combining information from several spectral bands which are sensitive to humidity and vegetation, it is possible to highlight regions affected by the hazardous event. Some examples of such indexes are Normalized Burn Ratio (NBR) [8], Normalized Burn Ratio 2 (NBR2) [9], Burned Area Index (BAI), Burned Area Index for Sentinel-2 (BAIS2) [10] and delta Normalized Burn Ratio 2 (dNBR2) [11]. Some of them, such as the latter, perform the comparison of the burned area index before and after the wildfire to improve performances and detect drastic changes in vegetation but are heavily sensitive to the presence of agricultural areas and crops. Index-based methodologies for burned area delineation are often coupled with automatic or

semi-automatic [12, 13] thresholding algorithms, such as the Otsu method [14]. One of the main complications of threshold-based techniques is the choice of the most adequate threshold, varying the vegetation type, environmental and lighting condition, making it difficult to determine a unique, universal value [15] for every region worldwide. Given the recent developments in deep learning, image segmentation tasks are tackled with convolutional neural networks. Models such as U-Net [7] and DeepLab [16] proved their effectiveness in numerous fields, ranging from the biomedical field [17] to autonomous driving [18] and remote sensing [19], including burned area delineation [20]. More recently, several researchers, given the excellent performances achieved in the field of NLP by the self-attention mechanism and the transformer architecture [21], started exploring the adoption of transformer-based architectures in the vision domain, too. Starting from the original Vision Transformer (ViT) [22], several architectures were developed in the context of image classification and segmentation: Swin Transformer [23], DeiT [24], and SegFormer [6]. Many applications in similar tasks, such as fire detection [25, 26], proved their effectiveness. Hence, in this paper, we explore the adoption of SegFormer architecture for burned area delineation, comparing the achieved performances with U-Net and threshold-based techniques.

3. Transformer-based burned area identification

3.1. Problem statement and model

Given a set of labelled satellite images of size $W \times H$, each one associated with a binary mask representing the information about the burned/unburned pixels, the goal consists in training a classification model that can then be used to predict the class label (burned/unburned) for all pixels of new images, i.e., we are interested in training a model that solves the semantic segmentation task.

Previous works (e.g., [20]) addressed this problem using convolutional neural networks, while, in this paper, we exploit a vision transformer model called SegFormer [6]. We decided to use this model because it can have fewer parameters, be computationally lighter and more noise resistant than U-Net [7] and other vision transformer architectures. Looking at the SegFormer architecture, we can notice it is different from other vision transformers because of (1) the hierarchical encoder that outputs multiscale features and (2) the absence of positional encoding. It is also important to note the output size is not equal to the input size, so it is necessary to upsample the output images. We choose to use bilinear interpolation according to the original implementation. SegFormer was designed specifically for semantic segmentation, optimizing the computationally expensive parts. In Table 1 we report the number of parameters of different instances of SegFormer and U-Net. Section 4 quantifies their impact on the quality of the predictions.

Table 1

Models comparison by number of parameters (expressed in millions).

	SegFormer-B0	SegFormer-B1	SegFormer-B2	U-Net	SegFormer-B3	SegFormer-B4	SegFormer-B5
# parameters	3.8	15.9	27.5	31.0	47.3	64.1	81.4

The first approach that we used to address the burned area identification problem consists in finetuning a pre-trained SegFormer on our task providing as input $W \times H$ labelled images of burned/unburned areas. Then, we apply the trained model to new images to perform predictions. Furthermore, we explored a second approach which we called *Crop&Recompose*, in which the training phase was done on images of size $N \times N$, being N smaller than the reference size $W \times H$, i.e., $N \leq W$ and $N \leq H$ (to be more comfortable with calculations we choose N submultiple of W and H). The second solution was proposed to verify the positive or negative impact of smaller patches during the training phase of SegFormer model in terms of precision of the predictions. In the second case we have smaller crops, and hence less context, but more images (in terms of images analyzed by the network at training time). However, the final goal consists in segmenting the original images of size $W \times H$, thus requiring recomposing the output to match the original input. The model is trained on smaller images of size $N \times N$ using the same architecture discussed before. Then, we apply the following approach to segment the new images, which are of size $W \times H$:

1. The original image of size $W \times H$ is cropped into M patches of size $N \times N$;
2. The M new images are passed through the model to perform the predictions;
3. The output composed of the predictions for the M images is recomposed into a single prediction/image of size $W \times H$.

Losses. Different loss functions were evaluated. To address the unbalanced problem of burned area delineation, we initially considered the dice loss and then we explored the possibility to use compound losses to reach a better stability point. In particular, we evaluated (i) the Dice loss, (ii) the Focal Loss, and (iii) the DiceFocal loss, a compound loss consisting of the former two combined.

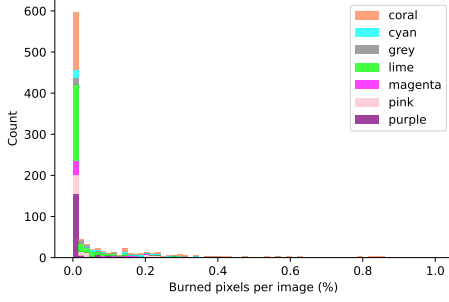
4. Experiments

4.1. Experimental settings

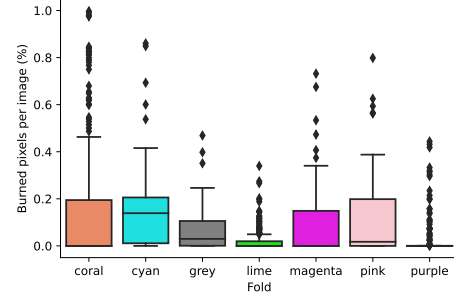
4.1.1. Dataset

We adopted an open dataset [27] of satellite imagery consisting of several areas of interest (AoIs) spread mainly across Europe. Data is of variable resolution, up to 5000x5000 pixels. The dataset delimits burned areas with a discrete severity level, ranging from 0 (undamaged) to 4 (completely destroyed). It is composed of images acquired from Sentinel-2 in combination with data provided by Copernicus Emergency Management Service, which contains manually and semi-automatically annotated damage severity maps of burned regions hit by past wildfires. Each Sentinel-2 acquisition has 12 channels. The dataset contains post-fire images of 73 different AoIs, which were aggregated in 7 different folds according to their geographical position. We choose to assign as name a color arbitrarily. For training, validation, and test we adopted the folds reported in [27]. In this paper, we explore the burned area delineation problem and consequently we binarize the target labels into unburned/burned classes, accordingly to our problem statement. As such, all values in range $[1, 4]$ were encoded into the burned class. We

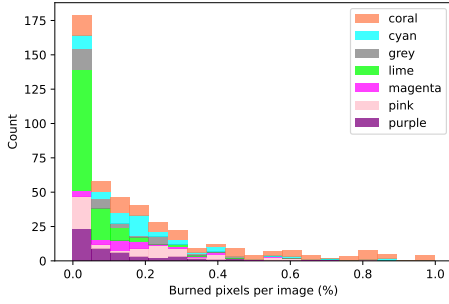
Figure 1: Left: distribution of the percentage of burned pixels per image. Right: distribution of the percentage of burned pixels per image for each fold.



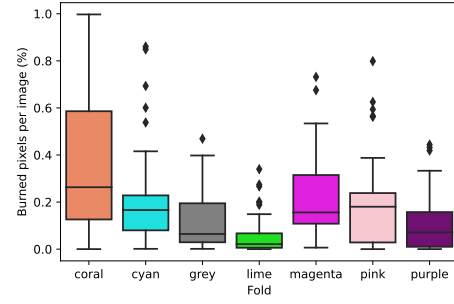
(a) Complete dataset: distribution of burned pixels per image (%)



(b) Complete dataset: percentage of burned pixels per image for each fold



(c) Images with at least one burned pixel: distribution of burned pixels per image (%)



(d) Images with at least one burned pixel: percentage of burned pixels per image for each fold

set the reference image resolution of 512×512 ($W \times H$) pixels, cropping bigger acquisitions into several images due to hardware limitations.

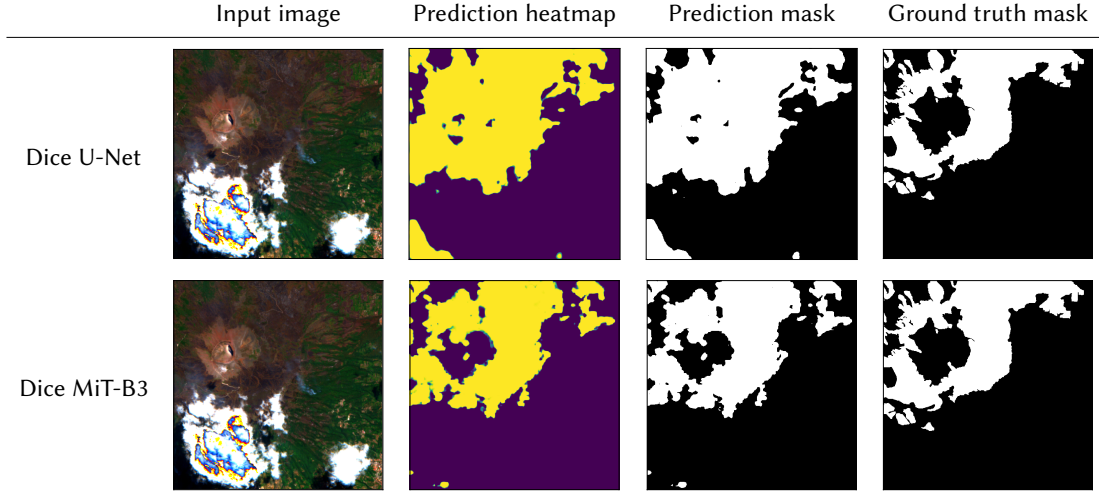
Figure 1 shows the high imbalanced distribution of target labels. Many images contain few pixels assigned to the burned class. Looking at the box plot, we can also see folds suffering from higher imbalance, having the majority of samples below 0.5 in the complete dataset. Thus, we exclude the cropped images without any burned pixel from the dataset, mitigating the class imbalance. In the ablated dataset the *coral* fold is the most complete one with percentages from 0 to 1, while the others have the majority of the samples below 0.6 and *lime* even below 0.2. The assumption is reasonable because we expect our system will be applied to areas we know there have been wildfires (several public services usually provide this information).

The sole exception is the Crop&Recompose method, for which we train with the complete dataset, whereas testing is performed with only crops containing at least one burned pixel.

4.1.2. Parameter and Experimental setting

The experiments were run on a single Tesla V100. We used the SegFormer implementation of HuggingFace [28] with a pre-trained encoder on Imagenet-1K, but because the original model has only 3 channels (RGB), we replicated the weights for all the 12 channels of the satellite images 4 times cyclically. This allowed us to leverage the pre-trained model even if the number

Figure 2: U-Net and SegFormer-B3 with dice loss predictions for the same image of *lime* fold



of input channels is different. The applied mapping (satellite image band, RGB channel) is as follows: (B01,R), (B02,R), (B03,G), (B04,B), (B05,G), (B06,B), (B07,R), (B08,G), (B09,B), (B10,R), (B11,G), and (B12,B).

The following augmentations were adopted for generalization with probability 0.5: random rotation with an angle in $[-50^\circ, 50^\circ]$, random vertical and horizontal flipping, and random shear with an angle in $[-20^\circ, 20^\circ]$. Image resolution is set to 512×512 except for the Crop&Recompose method, in which a size of 64×64 was used. We used the AdamW optimizer as in [6] and the starting learning rate was set to 0.001. A decreasing scheduler was chosen to reduce the LR by a factor of 10 every 15 epochs in conjunction with an early stopping mechanism on validation loss, with a tolerance of 10^{-4} and patience of 50 epochs. The maximum number of epochs is 200 and the batch size is 8.

4.1.3. Model tuning

While for the dice loss the weights are self-computed according to definition [29], focal loss needs to be tuned for correct usage. Knowing the number of positive pixels is about 4 times smaller than negatives, we chose $\alpha = 0.2$, while for γ we tested (1, 2, 5) values as suggested in [30] and we selected 5. As for DiceFocal loss, we chose $\alpha = 0.5$. As for *Crop&Recompose*, we trained using the DiceFocal loss with the previously selected parameters, considering both the entire dataset and the filtered dataset, keeping only crops with at least one burned pixel. According to our experiments, we achieved better results by including all the available crops, even though the class imbalance is furtherly worsened.

4.2. Comparison

The results in Table 2 show how for each model the worst results are obtained in the *lime* and *grey* folds, while the better ones are in the *pink* and *purple* folds. The *lime* fold performances are due to the presence of volcanic areas that frequently create conditions similar to those caused

Table 2
Summary of test metrics for each tested model and loss

		coral	cyan	grey	lime	magenta	pink	purple	mean	std
F1 score	Crop&Recompose MiT-B3	0.909	0.791	0.778	0.694	0.897	0.917	0.895	0.840	0.086
	Dice MiT-B3	0.899	0.790	0.762	0.712	0.877	0.909	0.899	0.835	0.080
	Dice U-Net	0.895	0.797	0.817	0.506	0.883	0.907	0.894	0.814	0.142
	DiceFocal MiT-B0	0.898	0.787	0.755	0.671	0.884	0.927	0.908	0.833	0.096
	DiceFocal MiT-B3	0.891	0.805	0.788	0.721	0.883	0.923	0.907	0.845	0.075
	Focal MiT-B3	0.694	0.732	0.661	0.650	0.830	0.716	0.859	0.734	0.081
	Otsu	0.678	0.644	0.410	0.151	0.655	0.535	0.374	0.492	0.193
Precision	Crop&Recompose MiT-B3	0.881	0.806	0.852	0.657	0.854	0.903	0.948	0.843	0.094
	Dice MiT-B3	0.898	0.828	0.859	0.655	0.866	0.898	0.897	0.843	0.087
	Dice U-Net	0.829	0.790	0.704	0.356	0.801	0.848	0.861	0.741	0.177
	DiceFocal MiT-B0	0.876	0.810	0.864	0.613	0.879	0.926	0.935	0.843	0.110
	DiceFocal MiT-B3	0.901	0.823	0.876	0.693	0.871	0.893	0.922	0.854	0.078
	Focal MiT-B3	0.931	0.880	0.877	0.753	0.910	0.990	0.948	0.898	0.075
	Otsu	0.626	0.558	0.275	0.084	0.518	0.453	0.235	0.393	0.198
Recall	Crop&Recompose MiT-B3	0.938	0.777	0.717	0.736	0.944	0.932	0.847	0.842	0.099
	Dice MiT-B3	0.901	0.755	0.685	0.779	0.888	0.920	0.902	0.833	0.092
	Dice U-Net	0.972	0.804	0.973	0.869	0.985	0.974	0.930	0.930	0.068
	DiceFocal MiT-B0	0.920	0.766	0.670	0.742	0.888	0.929	0.882	0.828	0.101
	DiceFocal MiT-B3	0.880	0.787	0.716	0.751	0.894	0.955	0.892	0.839	0.088
	Focal MiT-B3	0.553	0.626	0.531	0.572	0.762	0.561	0.785	0.627	0.104
	Otsu	0.739	0.761	0.803	0.801	0.889	0.653	0.905	0.793	0.087

by wildfires. The table also presents the performances achieved by a threshold-based technique (Otsu) on the NBR2 index [20], demonstrating the superior performances achieved by deep learning models.

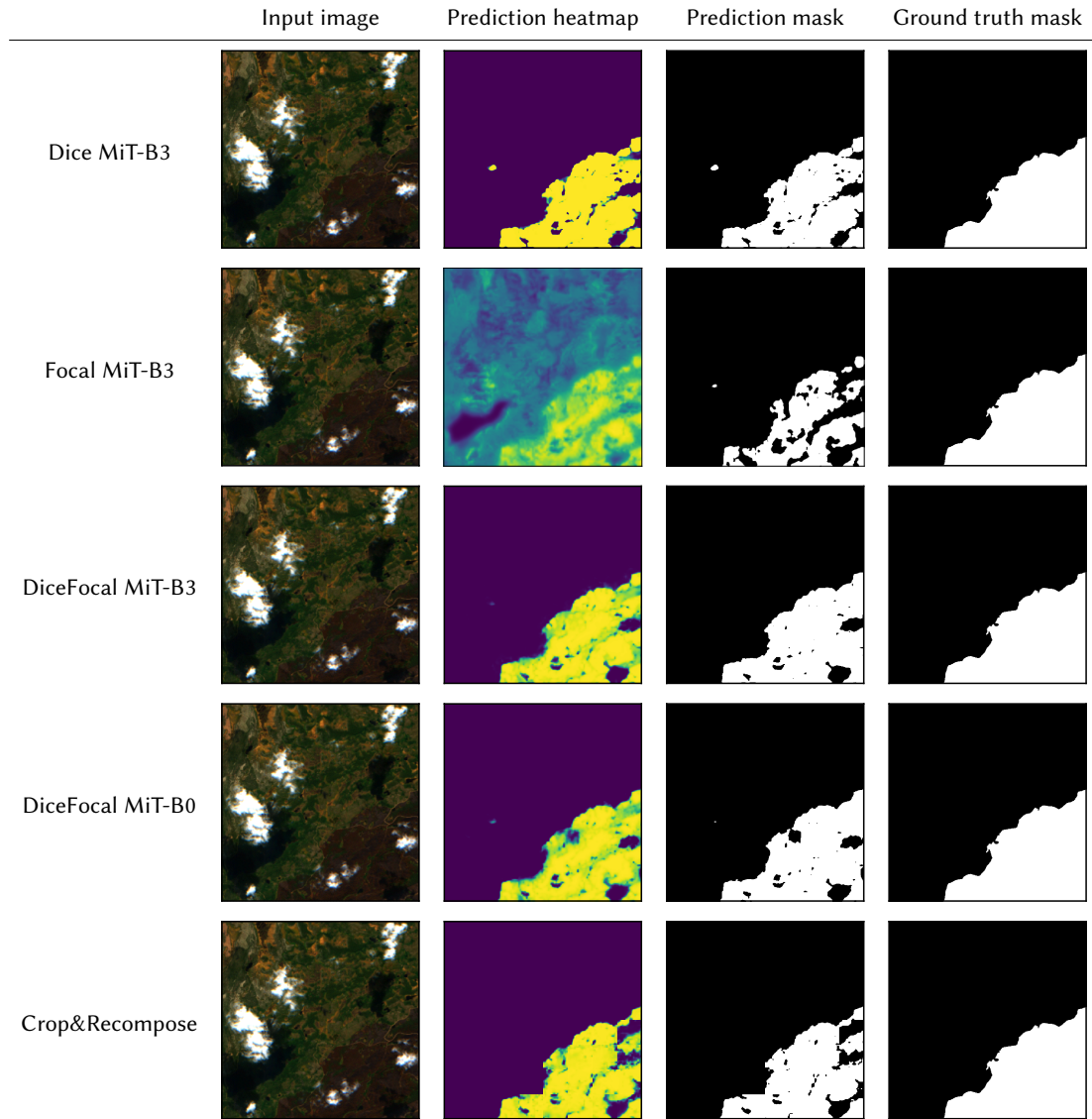
The F1 mean results confirm the superiority of SegFormer in all its variants over U-Net, nevertheless using a loss composed of dice and focal achieves better results compared to the sole usage of dice loss and focal loss. MiT-B3 with DiceFocal loss shows an improved F1 score (+3% than U-Net, +1% than MiT-B3 with dice loss), greatly increasing the precision (+11% than U-Net, +1% than MiT-B3 with dice loss). U-Net achieves higher recall but lower precision because it tends to overestimate the burned area.

The smallest model (MiT-B0) shows degraded performance compared to MiT-B3, but considering the low number of parameters (12 times less than MiT-B3), it can still be considered a good competitor.

Analyzing the standard deviation, SegFormer achieves more stable results for precision and F1 score. This trend is boosted by the use of the DiceFocal loss. The comparison in Figure 2 shows how it is more precise than U-Net, while the comparison in Figure 3 shows the effects of the different losses and models on the same image, highlighting how the cloud presence can generally negatively affect the prediction. The heatmap of the model trained with focal loss underlines the uncertainty of its prediction, but when combined with dice loss the model achieves better results, solving most of the underestimation problems. The benefits can still be seen when switching to a lighter model, despite the lower performance.

The Crop&Recompose method performs worse than the SegFormer trained on 512x512 images because the contextual information provided by 64x64 images is not sufficient to generalize well enough in complex cases (see Table 2).

Figure 3: Outputs of different models on the same image of *grey* fold.



5. Conclusion

In this paper we investigated how a novel vision transformer architecture, SegFormer, can be a good substitute for known CNN-based architectures in the context of remote sensing and burned area delineation, providing not only better results, but also better performance in terms of computational cost and number of parameters. Furthermore, we analyzed the effectiveness of several loss functions and different versions of the SegFormer architecture, achieving superior results in terms of precision and F1 score with respect to state-of-the-art models.

As future works, we plan to apply self-supervised learning and multi-modal transformers on the combinations of different satellite acquisitions, such as Sentinel-1 and Sentinel-2.

References

- [1] M. C. Hansen, T. R. Loveland, A review of large area monitoring of land cover change using Landsat data, *Remote Sensing of Environment* 122 (2012) 66–74. Landsat Legacy Special Issue.
- [2] A. Asokan, J. Anitha, Change detection techniques for remote sensing applications: a survey, *Earth Science Informatics* 12 (2019) 143–160.
- [3] H. Ma, Y. Liu, Y. Ren, J. Yu, Detection of Collapsed Buildings in Post-Earthquake Remote Sensing Images Based on the Improved YOLOv3, *Remote Sensing* 12 (2020).
- [4] A. Tavera, E. Arnaudo, C. Masone, B. Caputo, Augmentation Invariance and Adaptive Sampling in Semantic Segmentation of Agricultural Aerial Images, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2022, pp. 1656–1665.
- [5] W. Zhang, P. Tang, L. Zhao, Fast and accurate land-cover classification on medium-resolution remote-sensing images using segmentation models, *International Journal of Remote Sensing* 42 (2021) 3277–3301.
- [6] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, P. Luo, SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers, in: *Advances in Neural Information Processing Systems*, volume 34, Curran Associates, Inc., 2021, pp. 12077–12090.
- [7] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional Networks for Biomedical Image Segmentation, 2015.
- [8] D. Roy, L. Boschetti, S. Trigg, Remote sensing of fire severity: assessing the performance of the normalized burn ratio, *IEEE Geoscience and Remote Sensing Letters* 3 (2006) 112–116.
- [9] E. Roteta, A. Bastarrika, M. Padilla, T. Storm, E. Chuvieco, Development of a Sentinel-2 burned area algorithm: Generation of a small fire database for sub-Saharan Africa, *Remote Sensing of Environment* 222 (2019) 1–17.
- [10] F. Filippini, BAIS2: Burned Area Index for Sentinel-2, *Proceedings* 2 (2018).
- [11] J. D. Miller, A. E. Thode, Quantifying burn severity in a heterogeneous landscape with a relative version of the delta Normalized Burn Ratio (dNBR), *Remote Sensing of Environment* 109 (2007) 66–80.
- [12] W. Bin, L. Ming, J. Dan, L. Suju, C. Qiang, W. Chao, Z. Yang, Y. Huan, Z. Jun, A Method of Automatically Extracting Forest Fire Burned Areas Using Gf-1 Remote Sensing Images, in: *IGARSS 2019*, 2019, pp. 9953–9955.
- [13] G. P. Emang, Y. Touge, S. Kazama, Evaluating Trees Crowns Damage for the 2017 Largest Wildfire in Japan Using Sentinel-2A NDMI, in: *IGARSS 2020*, 2020, pp. 6794–6797.
- [14] N. Otsu, A threshold selection method from gray-level histograms, *IEEE transactions on systems, man, and cybernetics* 9 (1979) 62–66.
- [15] L. Saulino, A. Rita, A. Migliozi, C. Maffei, E. Allevato, A. P. Garonna, A. Saracino, Detecting Burn Severity across Mediterranean Forest Types by Coupling Medium-Spatial Resolution Satellite Imagery and Field Data, *Remote Sensing* 12 (2020).
- [16] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille, DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40 (2018) 834–848.
- [17] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, J. Wu, UNet

- 3+: A Full-Scale Connected UNet for Medical Image Segmentation, in: ICASSP 2020, 2020, pp. 1055–1059.
- [18] M. Siam, M. Gamal, M. Abdel-Razek, S. Yogamani, M. Jagersand, H. Zhang, A Comparative Study of Real-Time Semantic Segmentation for Autonomous Driving, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2018, pp. 700–70010.
 - [19] T. Kattenborn, J. Leitloff, F. Schiefer, S. Hinz, Review on convolutional neural networks (CNN) in vegetation remote sensing, *ISPRS Journal of Photogrammetry and Remote Sensing* 173 (2021) 24–49.
 - [20] A. Farasin, L. Colomba, G. Palomba, G. Nini, C. Rossi, Supervised Burned Areas delineation by means of Sentinel-2 imagery and Convolutional Neural Networks, in: Proceedings of the 17th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2020), 2020, pp. 24–27.
 - [21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is All You Need, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17, Curran Associates Inc., Red Hook, NY, USA, 2017, p. 6000–6010.
 - [22] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, 2020.
 - [23] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin Transformer: Hierarchical Vision Transformer using Shifted Windows, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9992–10002.
 - [24] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jegou, Training data-efficient image transformers and distillation through attention, in: Proceedings of the 38th International Conference on Machine Learning, volume 139, PMLR, 2021, pp. 10347–10357.
 - [25] R. Ghali, M. A. Akhloufi, M. Jmal, W. Souidene Mseddi, R. Attia, Wildfire segmentation using deep vision transformers, *Remote Sensing* 13 (2021) 3527.
 - [26] M. Shahid, K.-I. Hua, Fire detection using transformer network, in: Proceedings of the 2021 International Conference on Multimedia Retrieval, Association for Computing Machinery, New York, NY, USA, 2021, p. 627–630.
 - [27] L. Colomba, A. Farasin, S. Monaco, S. Greco, P. Garza, D. Apiletti, E. Baralis, T. Cerquitelli, Satellite Burned Area Dataset, 2022. URL: <https://zenodo.org/record/6597139>.
 - [28] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, HuggingFace’s Transformers: State-of-the-art Natural Language Processing, 2019.
 - [29] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, M. Jorge Cardoso, Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations, in: Deep learning in medical image analysis and multimodal learning for clinical decision support, Springer, 2017, pp. 240–248.
 - [30] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal Loss for Dense Object Detection, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2999–3007.