
EVALUATING THE EFFICACY OF CONVOLUTIONAL NEURAL
NETWORKS AND TRANSFORMER-BASED MODELS FOR
BUSHFIRE BURNED AREA DETECTION

Author: Chaojun Tang

Supervisor: Prof. Richard Sinnott

Faculty of Engineering and Information Technology

The University of Melbourne

October 2023

ABSTRACT

Bushfires are particularly prevalent in Australia. They pose significant economic and safety challenges. A key component in the mitigation of these challenges is the precise monitoring of burnt areas. This study aims to evaluate the potential of modern computer vision models in detecting burned areas, especially in the context of noisy data, in contrast to the limitation of traditional methods. The contribution of this research includes designing a bushfire data collection pipeline and collecting a dataset covering 2019 Australian Black Summer bushfire events. Several prominent computer vision models, including Convolutional Neural Network(CNN)-based models like U-Net and Mask R-CNN, and transformer-based models like SAM and SegFormer were tested against this dataset. SegFormer-b0 emerged as the top performer in the presence of noise such as clouds and smoke, followed closely by a modified U-Net model that utilized all available spectral data from satellite imagery. Despite the promising findings, the study identified certain limitations, including model backbone inconsistencies and potential improvements in data labeling quality. This research provides the basis for future studies focusing on Australia's bushfire observations and monitoring.

Keywords Bushfires · Burned Area Segmentation · Computer Vision · Satellite Imagery

For detailed source codes and datasets related to this research, please refer to the dedicated GitHub repository at: https://github.com/TangChao729/Burned_Area_Segmentation.

Declaration

I certify that:

- this thesis does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any university; and that to the best of my knowledge and belief it does not contain any material previously published or written by another person where due reference is not made in the text.
- The thesis is 7301 words in length (excluding text in images, table, bibliographies and appendices).

Signed: Chaojun Tang

Date: 27 October 2023

Contents

1	Introduction	1
1.1	Aim of the research	3
2	Literature Review	4
2.1	Traditional method: Index-product	4
2.2	CNN-based computer vision models	5
2.3	Transformers, the new competitor	6
2.4	Limitations of Previous Research and the Path Forward	7
3	Burned Area Dataset Preparation	9
3.1	Region of interests	9
3.2	Data Collection	9
3.3	Data Labelling	11
3.4	Dataset Splitting	12
3.5	Data Augmentation	13
3.6	Dataset Summary	13
4	Models	14
4.1	U-net	14
4.2	Mask R-CNN	15
4.3	YOLOv8	15
4.4	Segment Anything Model (SAM)	16
4.5	SegFormer	17
4.6	Accuracy evaluation methods	18
4.7	Training Configuration and Setups	19
5	Result and Discussion	20
5.1	Training results	20
5.2	Detailed Examination of Predicted Masks	23
5.3	Overall Model Efficacy	26
5.4	Comparison with Previous Works	27
5.5	Limitations	27
6	Concluding Remarks	29
7	Future Directions	30

A	Sample Images	31
B	True Masks	32
C	Index-product predications	33
D	U-Net predications	34
E	U-Net all spectral predications	35
F	Mask R-CNN predications	36
G	YOLOv8 predications	37
H	SAM predications	38
I	SegFormer b0 predications	39
J	SegFormer b3 predications	40

List of Figures

1	Impacted areas and severity during Black Summer	1
2	RGB true color vs SWIR false color vs NBR grey-scale	2
3	NBR differentiates for burned area generalization	4
4	Burned Area Grids	9
5	Sentinel-2 wavelength and bands	11
6	Burned area mask generation.	12
7	U-Net architecture	14
8	Mask R-CNN architecture	15
9	SAM architecture	16
10	SegFormer architecture	17
11	Predictions visual comparison	23

List of Tables

1	Comparison between Index-product and Computer Vision Techniques . .	6
2	Black Summer burned area satellite imagery samples, SWIR false color. .	10
3	Architecture and Training Parameters	19
4	Models Evaluation Results	20
5	Model Parameters, Accuracy, and AIC Metrics	26

1 Introduction

Bushfires pose a severe threat to Australia, with catastrophic events like the Black Friday and Black Summer bushfires serving as grim reminders. These fires wreaked havoc, causing not only extensive damage but also taking numerous lives. The Black Summer season, in particular, left an indelible mark: it burned approximately 18.6 million hectares as illustrated in Figure 1, destroyed over 3,000 homes, and claimed 34 lives [1, 2]. The financial ramifications were profound, with damages amounting to AUD 103 billion [3].

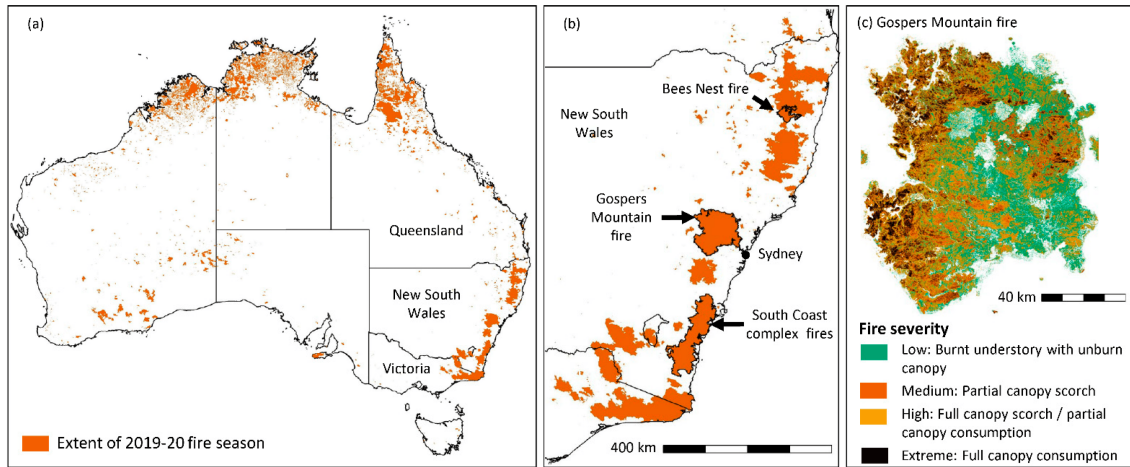


Figure 1: Impacted areas and severity during Black Summer [4]

In the aftermath of such devastating bushfires, there was a surge in efforts from multiple parties. Key players, like the Minderoo Foundation, pledged to achieve rapid fire detection by 2025 [5]. Other research organizations have engaged in extensive studies on bushfire analytics [6] and many universities launching bushfire-centric studies [7–11]. The Australian government underscored the role of space-based Earth observation in bushfire management, spanning pre-fire preparations, real-time monitoring, and post-fire assessments.

Burnt area detection, as part of the post-fire assessments, is pivotal for several reasons. Firstly, it aids in understanding the ecological impact, informing both biodiversity preservation and habitat restoration efforts. Additionally, pinpointing burnt areas helps in understanding environmental challenges like soil erosion and water contamination. From an infrastructural standpoint, knowing the precise affected regions assists governments in prioritizing rehabilitation and safety checks, ensuring both rapid recovery and the safety of residents and recovery personnel. Furthermore, it also guides research for future fire prevention strategies. In summary, understanding the full extent of bushfire damage is paramount not just for immediate recovery, but also for the longer-term resilience and preparedness of affected regions.

Satellite imagery stands out as a particularly reliable data source due to its extensive coverage and frequent revisit intervals. It provides additional support for retrieving bushfire data including pin-point location and scale across remote regions. When set against field data capture techniques, satellite methods are both time-saving and cost-effective. The capability of modern satellites has made a profound impact, particularly with multi-band detection systems that capture diverse wavelengths. As shown in Figure 2, a combination of Near-Infrared (NIR) and Shortwave-Infrared (SWIR) false color images highlight burnt areas with a dark red and brown color, with the processed normalized burned ratio (NBR) shown in grey-scale index highlighting the burned area with dark colors, while traditional RGB band show fewer features for the burnt area.

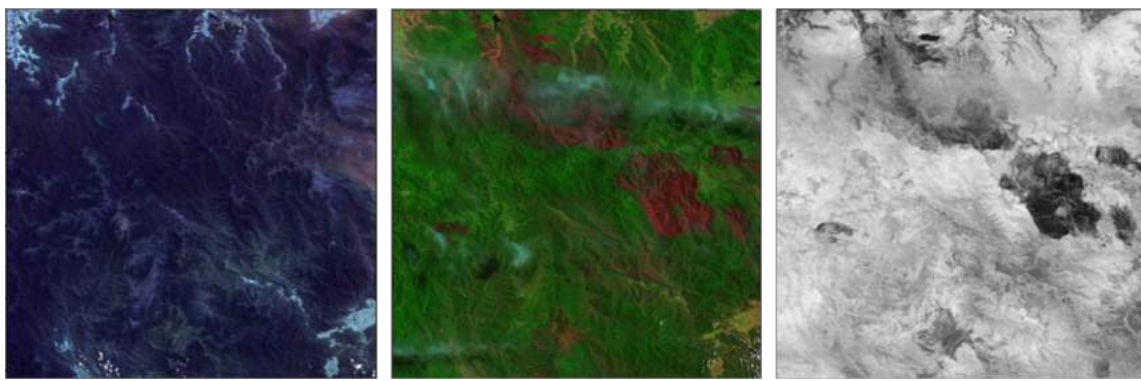


Figure 2: RGB true color vs SWIR false color vs NBR grey-scale

The progression in satellite imaging has benefited through advancements in image processing. While rule-based methods, such as normalized burned ratio (NBR) and modified burned area index (BAI), have their merits, they possess inherent limitations, e.g., they require both pre-event data and post-event data to get the delta differences. They also require data cleaning including removing background noise pixels [12, 13]. Machine learning approaches, especially deep learning algorithms offer flexibility, precision, and reduced data dependency. The recent onset of transformer-based models in computer vision further accentuates the potential of such algorithms for satellite image analysis.

1.1 Aim of the research

While bushfires are not unique to Australia, the specific conditions and characteristics of Australian vegetation, terrains, and fires necessitate dedicated research. Addressing this, this research carefully assembled a novel dataset exclusively sourced from the Australian terrain, catering to the nuances of local bushfires. Furthermore, we introduced a comprehensive pipeline for data extraction from the Sentinel-2 satellite, detailing the processes from region selection to bounding box geo-location and subsequent data processing together with sophisticated analysis and labeling techniques.

Second, this research embraces the challenge of interpreting images comprising cloud and smoke cover to deliver faster actionable insights, in contrast to other research restricted to clean, cloud-free data, which introduces a limitation that often leads to data-sourcing delays.

Third, this research conducts a comprehensive evaluation of diverse machine learning models, ranging from CNN-based architectures such as U-net and Mask R-CNN, to emerging transformer-based models like SegFormer and Segment Anything. The objective of this research is to identify the most effective model by leveraging key metrics including F1-score, IoU, Matthews Correlation Coefficient, model parameters, and FLOPs, among others. Importantly, all models are benchmarked against a consistent dataset.

Last but not least, we also probe the feasibility of leveraging all spectral bands from Sentinel-2, diverging from the conventional three-channel inputs. This offers an innovative perspective to satellite image analysis.

Currently (Oct 2023), pressing bushfire situations are unfolding across New South Wales and Queensland, Australia with 86 fires actively burning, impacting an area of more than 5000 hectares [14]. As climate change presents new, dynamic scenarios, fire seasons have been observed to become longer and to affect larger areas than ever before [15]. The recent declaration of an El Niño year for 2024 further elevates the bushfire risk [16]. This research hopes to pioneer advanced methodologies for rapid and precise burnt area detection, aiming not just to advance scientific understanding but also to empower stakeholders with timely, actionable insights to better combat and mitigate the devastating impacts of future bushfire events.

2 Literature Review

Research on the use of satellite imagery for bushfire studies predominantly follows two categories: index-based methods and machine learning methods - most notably deep learning-based methods.

2.1 Traditional method: Index-product

The use of multi-spectral sensors for bushfire detection was pioneered by Garcia et al in 1991 [17]. They proposed a multi-band product approach to extract information and create a burned area index. By differentiating between two index maps with a set threshold, they could identify burned areas, as shown in Figure 3. This resulted in the development of indices like the normalized difference vegetation index and the burned area index.

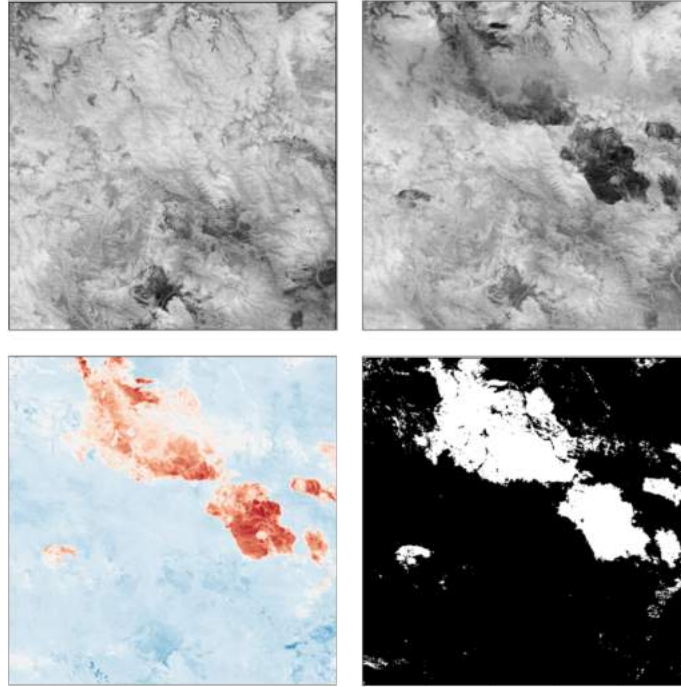


Figure 3: NBR differentiates for burned area generalization

However, the index method has several inherent challenges. For accuracy, the data must be largely free of clouds and reflections, which often mandates the patching and fusing of images over several days through rigorous data cleansing. This procedure introduces inevitable delays. Additionally, the requirement of both pre-fire and post-fire datasets increases the complexity of data retrieval. The calibration process for determining burnt areas is influenced by variation in vegetation density. This introduces another layer of complexity.

Modern methodologies, exemplified by Lizundia et al [18], alleviate some of these challenges by integrating sophisticated processing pipelines with diverse satellite data and use of machine learning for fine-tuning threshold adjustments.

To sum up, while index-based methodologies offer consistent results, computational efficiency, and interpretability, they have constraints that limit their agility and adaptability for real-time bushfire analysis.

2.2 CNN-based computer vision models

The deep learning approach, on the other hand, poses different merits. Different from machine learning, deep learning approaches utilize relatively large neural networks and enhanced feature extraction techniques.

The cornerstone of deep learning in computer vision is Convolutional Neural Networks (CNNs). CNNs operate by using kernels to extract features and employ deep, fully-connected neural networks for pixel-level classification. As burnt areas often have distinguishing textures, patterns, and colors compared to their surroundings, CNNs can efficiently learn these unique features from labeled datasets and predict them in unseen data.

Numerous studies have championed CNNs like U-Net for their effectiveness in satellite imagery studies. Knopp et al. [19] utilized U-Net model and Sentinel-2 data for burned area detection. They achieved a high accuracy of 94%. Hu et al. [20] compared various CNN-based models along with the Corinthia bushfire and Fågelsjö-Lillåsen bushfire data, sourced from Sentinel-2 and Landsat-8 data. The results showed that U-Net stood out with a decent performance, compared to HRNet and Fast-SCNN. Pan et al. [21] studied the effectiveness of U-Net in classifying and segmenting unplanned urban settlements from satellite images. They observed an average accuracy of 88%. Other studies such as cloud/cloud shadow segmentation [22], water identification [23], and burned area segmentation [24] all demonstrated the effectiveness of the U-Net architecture for satellite imagery analysis.

Mask R-CNN is another noteworthy candidate. It is recognized for its robustness and dual capabilities for object detection and segmentation. Zhao et al [25] used mask R-CNN for building boundaries segmentation using satellite images. Quoc et al. [26] studied the effectiveness of computer vision models based on agriculture satellite images, and showed that Mask R-CNN out-performed U-Net with an accuracy of 95.3%.

In contrast to the mathematical index-product methods, deep learning techniques exhibit resilience against data inconsistencies. They can function without pre-fire and post-fire frames, deducing burnt areas based solely on the nuances learned from the training dataset.

Moreover, with their capacity to process vast amounts of data, they can discern burnt areas against varying vegetative backdrops without intensive calibration.

Table 1 presents a comparative assessment of the traditional index-product methods against deep learning-based approaches:

Table 1: Comparison between Index-product and Computer Vision Techniques

	Index-product	Computer Vision (ML/DL)
Advantages		
	Clear results	Identifies complex features
	Economical use of resources	Improves with data
	Established methodology	Adaptable to diverse tasks
Disadvantages		
	Requires clean data	Needs extensive datasets
	Depends on temporal series	Requires significant resources
	Needs periodic adjustments	"Black box" challenge

However, CNNs aren't without their challenges. The necessity for vast amounts of labeled data, the "black box" nature of predictions, and the re-calibration barrier for some models remain pertinent issues. Details such as model architecture and training pipeline will be explained further in section 4.

2.3 Transformers, the new competitor

The inception of the attention mechanism in 2016, alongside the introduction of the transformer architecture, marked a breakthrough moment in deep learning [27]. Originally heralded in the Natural Language Processing domain, its relevance has since permeated other AI spheres. Vision transformers (ViT) embody this transition into computer vision, offering a novel methodology for visual tasks [28].

At the heart of ViT is the self-attention mechanism. This enables models to discern relationships over vast spatial expanses. Compared to the sliding convolution kernel mechanism of CNN-based computer vision models, the self-attention mechanism helps the model to capture global features. This is particularly useful for detecting burned areas in satellite imagery, where contextual relationships (e.g., the presence of specific vegetation around a burned patch) can be crucial for accurate detection. Moreover, the self-attention

mechanism enables the model to capture burnt areas even though the burned area may have partial cloud coverage. This capability is paramount for burned area detection.

This new model inspired many recent computer vision studies. NVIDIA's SegFormer [29] is one model that is highly regarded. Utilizing a ViT backbone, attention mechanism, and hierarchical fusion, SegFormer is able to capture fine-grained details from both low-level features and high-level features, by considering the entire context of an input sequence when producing an output.

Meta's Segment Anything Model (SAM) [30] is another similar mechanism. SAM is implemented with a "prompt" mechanism, where it can receive extra information input used to help with the segmentation task. The prompt system provides a unique way to guide the model's segmentation, offering fine-grained control over the output. In combination with transfer learning features and a pre-trained backbone, SAM is able to be fine-tuned with a relatively small size dataset while maintaining decent results.

The advent of transformer-based models in computer vision is relatively new. Studies, such as that by Keselimi et al. [31], have demonstrated that transformer-based vision models surpass their CNN counterparts, especially in studies involving satellite imagery applied to deforestation. Similarly, Horvath et al. [32] employed transformer-based models for satellite image manipulation detection, highlighting the prowess of ViT models in handling sparse data features. Yet, there is a noticeable gap in the literature concerning the application of transformer-based models to bushfire studies. This presents a promising opportunity for exploration.

2.4 Limitations of Previous Research and the Path Forward

Research into bushfire detection and analysis has made significant progress over the years. These past efforts have given us valuable insights into how burned areas can be derived and segmented. But as technology keeps advancing and the challenges faced become more complex, it is clear that methods for studying bushfires also need to evolve.

- **Original Australian Bushfire Dataset:** Prior studies have leaned heavily on generic datasets, which might not fully capture the intricacies of bushfires in specific regions. Given the unique nature of Australian bushfires and their distinct vegetation, climate, and topography – it is imperative to use data sourced directly from Australia. Relying on an Australian-specific dataset ensures that the research is tailored to the challenges and patterns native to the region, thereby enhancing the accuracy and applicability of findings in local bushfire management and intervention.

-
- **Dealing with Clouds and Smoke:** Most current research (both index-products and computer vision) utilise clear, cloud-free images. The dependency on pristine, cloud-free data is a major limitation, especially when rapid action is required during bushfires.
 - **Model Diversity:** The variety of choices of deep learning models presents both a challenge and an opportunity. On the one hand, the sheer diversity of architectures, from CNNs to transformers, promises a wide array of tools. On the other, the lack of a comprehensive comparative study leaves practitioners in a quandary, uncertain about the best tool for the job. Moreover, given the recent introduction of vision transformer models, limited research exists on their application to satellite burned area segmentation.
 - **Spectral Band Utilization:** Satellite imagery, like Sentinel-2, offers a wealth of information in the form of multiple spectral bands. Most models take three-channel images as inputs (RGB) as these are most readable to human eyes. Utilizing all bands could potentially enhance model performance and interpretability.

This research addresses these issues. Exploring such aspects not only bolsters the relevance of our work but also provides innovation in the field. By addressing these challenges head-on, we aim to offer a more holistic, timely, and region-specific approach to bushfire detection.

3 Burned Area Dataset Preparation

3.1 Region of interests

This study centers around areas affected by the 2019 Black Summer bushfires, mainly covering southeastern coastal zones encompassing Victoria (VIC), the Australian Capital Territory (ACT), New South Wales (NSW), and southern Queensland (QLD), as these were the areas with highest bushfire density. The areas under study are visually represented in Figure 4, with pink polygons highlighting the most affected regions.

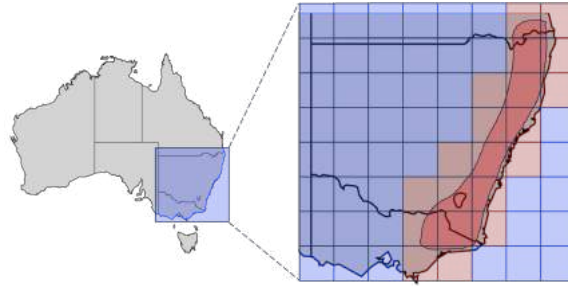


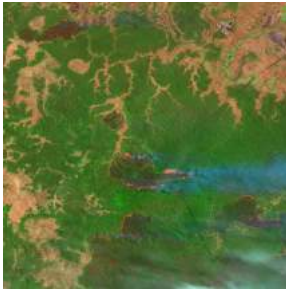
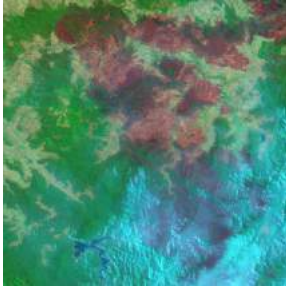

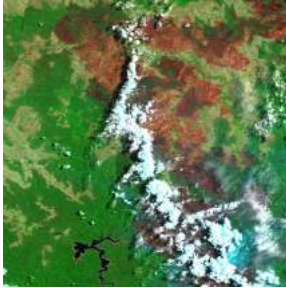

Figure 4: Burned Area Grids

3.2 Data Collection

Bushfire data was sourced from the 2020 National Operational Bushfire Boundaries dataset available on data.gov.au [33]. This provides each bushfire’s ignition date, geolocation, and approximate impact area given as simple polygons. During the Black Summer bushfires, 1487 fires were recorded. To standardize the representation of the diverse fire sizes, we employed a grid system, where each cell symbolized an approximate area of 8,200km². 20 such cells captured the bushfire episodes during the period, as depicted in Figure 4. For every grid, we collected a consecutive series of data throughout the study duration from 20 days before a fire event until 25 days afterward. Repeated or similar images are removed.

The data was procured from the Sentinel-2 L1C satellite imagery database. This offers several advantages, such as sharp resolution and regular updates. While other sources like Landsat were considered, Sentinel was ultimately the preferred choice. The queried data from Sentinel-2 comprises 14 bands, covering visible (B1 to B4), visible to near-infrared (B5 to B8a), short-wave infrared (B9 to B12) wavelengths, and a cloud classification layer. All data are saved in tiff format to accommodate the multi-spectral format. A metadata file that contains each tiff file’s bounding box location, capture date, and cloud percentage is saved. Sample images (in SWIR to show the burned area) are shown in Table 2.

Table 2: Black Summer burned area satellite imagery samples, SWIR false color.

Type	Sample	Description
Clear burned area		Image with no or very little noise, burned area clearly showing.
Light smoke		Image with light smoke noise, burned area still showing.
Heavy smoke		Image with heavy smoke noise, burned area boundaries hard to differentiate.
Light cloud		Ground partially blocked by scatted cloud, still showing the burned area.
Heavy cloud		Ground almost fully blocked by heavy cloud, burned area barely seen under cloud.

3.3 Data Labelling

The multiple spectral features of satellite imagery are highly beneficial for burned area detection, as some bands have higher correlations with burned area, as shown in the Sentinel-2 Wavelength and Bands chart in Figure 5. The labeling process utilizes several band combination outputs, specifically:

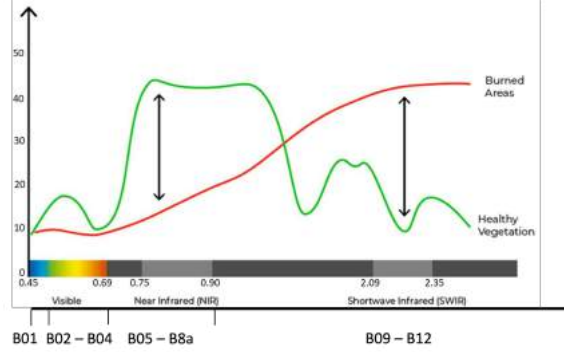


Figure 5: Sentinel-2 wavelength and bands [34]

Short Wavelength Infrared False Color (SWIR):

By assessing moisture content in soil and vegetation, SWIR is invaluable for drought studies, hydrology, and fire detection. It is especially responsive to moisture variations. It is calculated through:

$$\text{SWIR Composite} = \text{Combine}(B12, B09, B08) \quad (1)$$

Normalized Burn Ratio (NBR):

NBR is a trusted index for spotting burned areas. A higher NBR value typically denotes healthy vegetation, whereas a lower value indicates a recently burned zone [34]. It is calculated through:

$$\text{NBR} = \frac{B12 - B8A}{B12 + B8A} \quad (2)$$

Normalized Difference Vegetation Index (NDVI):

NDVI is a well-established index used to assess the health and density of vegetation over various landscapes [34]. A high NDVI value indicates healthy, dense vegetation, whereas a lower value often points to barren lands. It is calculated through:

$$\text{NDVI} = \frac{B08 - B04}{B08 + B04} \quad (3)$$

Band combinations such as RGB, SWIR, NDVI, and NBR yield unique visual patterns helpful in identifying burned regions within the captured data. A comparison of these combinations from the same tiff file and the truth mask generation is presented in Figure 6:

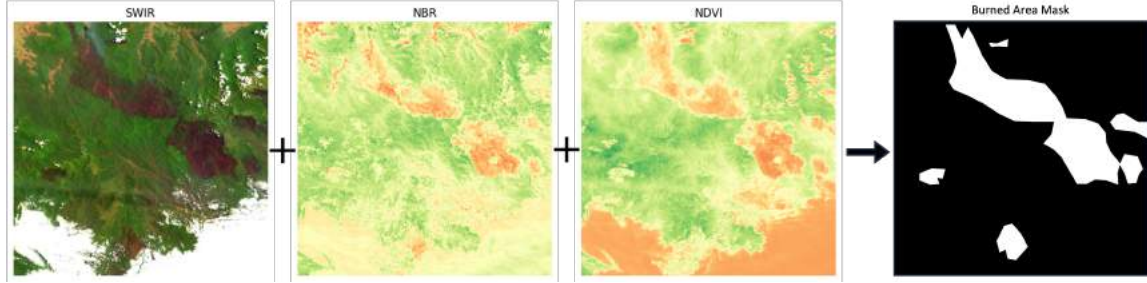


Figure 6: Burned area mask generation.

Ground truth image segmentation was determined through the analysis of these band combination outputs with sketches produced based on the VGG image annotator. The original polygon data of the burned area masks are saved in JSON format and later transformed to both image and text formats.

3.4 Dataset Splitting

Upon securing the mask data, the data and corresponding masks were further processed into several formats to cater to varying model input requirements.

Image data:

- 3-channel JPEG files, encompassing bands 4, 8a, and 12. These bands symbolize the visible light, near-infrared, and short-wave infrared wavelengths respectively.
- TIFF files with the original 13 bands, excluding the cloud classification layer due to error data.

Mask data:

- 1-channel JPEG files.
- Text files delineating the coordinates of each burned area polygon.

Subsequent to the above, the dataset was systematically divided into training, validation, and test sets with a ratio of 70/15/15%.

To guarantee that models remained uninformed about the test images and their intrinsic terrain details, we manually selected (15%) certain regions and their entire time series data for testing and visualization.

The residual data underwent a manual inspection, based on its noise level percentage (previously retrieved image metadata) between 0% (indicating pristine data with no clouds) to 100% (representing nearly obscured ground terrain). Following this classification, the data was split into training and validation sets at a ratio of 70% and 15% with similar noise levels. This approach ensured an unbiased validation dataset, while the test dataset remained entirely unknown to the model.

3.5 Data Augmentation

Data augmentation plays a critical role in enhancing the diversity of training samples and has been extensively applied to boost the performance of machine learning models. For the dataset used in this study, data augmentation was uniformly applied across all instances. However, given the unique 13-band nature of the original data, certain augmentations like color manipulations were not considered, as they might introduce unscientific distortions and risk interference between bands.

Moreover, since the dataset was derived exclusively from the Sentinel-2 satellite, it was essential to maintain the authenticity and integrity of the satellite imagery. Therefore, blurring techniques, which could alter the fixed resolution of these images were avoided.

Instead, to amplify the dataset without compromising its integrity, rotational augmentation was chosen. The dataset was augmented with rotations at 90 degrees, 180 degrees, and 270 degrees. Rotational augmentation was observed to enhance the model's in-variance to orientation changes and potentially increase its generalization capabilities, since when one image was rotated, the terrain and burned area still remained in the same correlation. An early experiment was conducted in this research showing improvement of the performance of the model when adding data augmentation.

3.6 Dataset Summary

The final dataset comprised 1980 images: 1380 in the training set, 300 in the validation set, and 300 for testing. 6 images were handpicked for visualization. Dataset splitting and data augmentation were conducted beforehand to form a universal dataset, to ensure that all training models access the same datasets, minimizing the impact of randomness.

4 Models

4.1 U-net

In recent research, U-Net [35] has established itself as a pioneering deep learning model for satellite image segmentation tasks [19–24]. Burned areas in satellite imagery often present varied spatial scales and possess intricate boundaries, necessitating a model that can simultaneously capture both fine and coarse details. U-Net’s encoder-decoder design, as shown in Figure 7, excels in this regard. The encoder efficiently extracts multi-scale features through convolutional blocks using 3×3 convolutions, ReLu activations, batch normalization, and max-pooling. This process down-scales feature maps and doubles the feature channels. The decoder then up-samples these maps to the input size using transpose convolution, integrates feature maps from corresponding encoder stages (leveraging U-Net’s effective skip connections), and applies further convolutions with ReLu and batch normalization. The final output is a 1×1 convolution layer that predicts pixel-level probabilities for a given area.

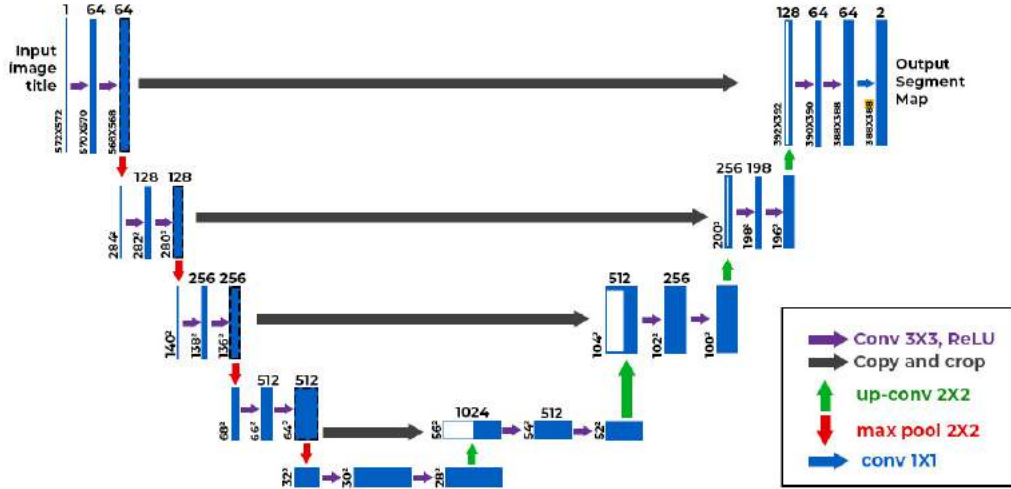


Figure 7: U-Net architecture [35]

13-channel U-Net variant:

To leverage the rich spectral information inherent in satellite imagery, we introduced a variation to the traditional 3-channel inputs U-Net model. This novel adaptation accepts 13 channels, embracing the entirety of the spectral bands present in the dataset. Such a comprehensive intake could improve the model’s capability to differentiate subtle variations within the burned regions, especially those elements that might be imperceptible within a limited channel spectrum.

4.2 Mask R-CNN

Mask R-CNN [36] realizes a refinement of Faster R-CNN, which is a premier model for object detection and instance segmentation. Well-suited for diverse applications, it excels in tasks like dynamic entity detection [37], building boundary regularization [38], and ship tracking [25]. Notably, its application in burned area detection has shown promising results [39]. Beyond accuracy, Mask R-CNN's strength lies in its ability to handle multi-class imagery, offering both bounding boxes and segmentation masks. This is especially crucial in satellite images featuring scattered burned scars of varying sizes to support comprehensive fire impact detection.

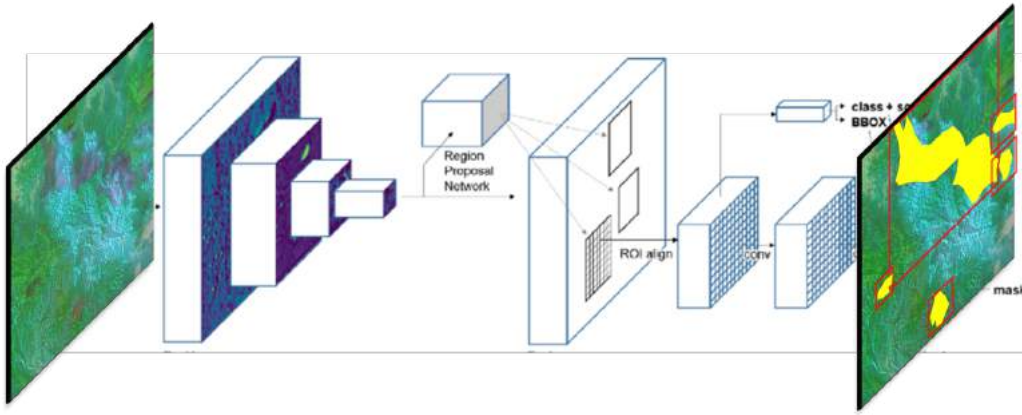


Figure 8: Mask R-CNN architecture [36]

Structurally, Mask R-CNN is a modification of the Faster R-CNN paradigm, as shown in Figure 8. Central to its design is the Region Proposal Network (RPN) that proposes prospective object bounding boxes. Each such proposition is scrutinized by a classifier for object identification and a regressor used for bounding box fine-tuning. The distinctiveness of Mask R-CNN emerges from its dedicated mask prediction branch. Functioning in tandem with the main architecture, this branch processes each Region of Interest (RoI), and uses a convolutional layer followed by a sigmoid activation function to predict the mask for each class at a pixel level. Then, within each classified bounding box, it generates the segmentation.

4.3 YOLOv8

Originating from the same Faster R-CNN architecture, YOLOv8 is the latest in the "You Only Look Once" series. While YOLO's inception was primarily aligned with object

detection, its segmentation prowess has been progressively improved. Hence, this study includes YOLOv8 for comparison.

4.4 Segment Anything Model (SAM)

Segment Anything Model (SAM) is a new image segmentation model introduced by Kirillov, Alexander, et al. (Meta AI Research) in 2023 [30]. As part of the vision transformers family, SAM utilizes a vision transformer encoder in its core to cope with an extremely large dataset (11 million images and 1 billion masks). It aims to become a model able to segment most objects even from complex features. Though a newcomer, SAM has demonstrated its capability in image segmentation. However, its capability to segment satellite images is yet to be investigated.

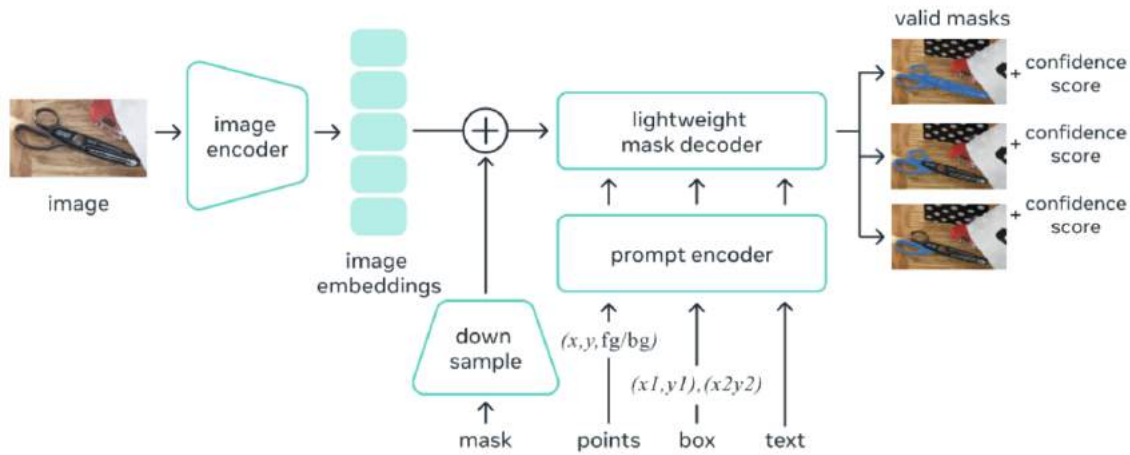


Figure 9: SAM architecture [30]

SAM's distinctive architecture (Figure 9) uses dual inputs: embedded image features and embedded prompts. These prompts can assume various forms, be it a point, a bounding box, an area, or potentially even a sentence. Drawing from both image and prompt embeddings, SAM's mask decoder proceeds to delineate the segmentation.

Despite its extensive training on vast datasets, SAM's application to satellite imagery remains limited. Fine-tuning the whole model is challenging given SAM's size. Even its compact version, SAM-b, uses 94.7 million parameters. A pragmatic solution to fine-tuning SAM is to focus exclusively on training its mask decoder. This approach simplifies the process and also optimizes the performance given limited computational resources.

4.5 SegFormer

SegFormer was introduced by Xie et al [29] as a product of NVIDIA Research Lab. It is a part of the vision transformer family. After its initial debut, SegFormer quickly drew the attention of computer vision scientists due to its robust performance and relatively light architecture. Notably, it has been successfully employed in intricate tasks like segmenting seismic facies in geology [40] and detecting foliar diseases in agriculture [41]. Both research showed that SegFormer was able to capture features from sparse data, hence making SegFormer a potential candidate for this research.

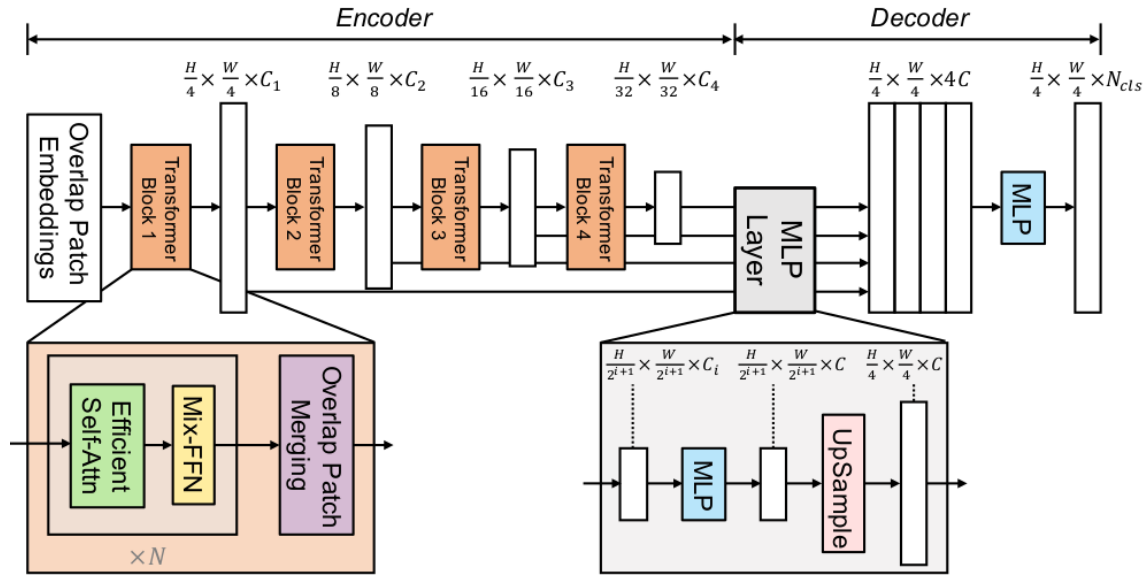


Figure 10: SegFormer architecture [29]

SegFormer maintains the encoder-decoder architecture of transformers (Figure 10), while also using the concept of down-sampling from traditional CNN models. The embedding of different down-sampled images is gathered at a single multi-layer perception decoder and concatenated into a long embedding for the decoder. Contrasting with conventional CNNs that employ fixed-sized kernels—and are constrained by their localized focus—SegFormer taps into the attention mechanism. This ensures the model discerns feature associations on a broader scale.

Backbone choice

This research selected MiT-b0 and MiT-b3 as the backbone of the SegFormer model. MiT-b0 has 3.7 million parameters and is the most lightweight model among all candidates.

4.6 Accuracy evaluation methods

When performing semantic segmentation for burned area detection using computer vision models, choosing the right evaluation metrics is crucial. Each metric illuminates distinct facets of a model's efficacy.

Baseline - In this research, the baseline constitutes the standard Normalized Burn Ratio (NBR) difference calculation, as outlined in Garcia (1991) [17]. The best threshold is established on the training dataset, which is subsequently applied to the testing dataset for burned area delineation, then compared against manually labeled masks to determine baseline accuracy.

F1-Score - is the harmonic mean of precision and recall, providing a combined metric that considers both false positives and false negatives. It is particularly useful for datasets with imbalances, like cases where the burned pixels are vastly outnumbered by non-burned ones. This makes the F1-Score crucial in contexts where both types of errors are significant. In binary segmentation tasks, where there is a distinct foreground and background classification, the F1-Score is equivalent to the Dice coefficient. Mathematically, the F1-Score is given by:

$$\text{F1-score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (4)$$

$$\text{where precision} = \frac{TP}{TP + FP} \quad \text{and recall} = \frac{TP}{TP + FN} \quad (5)$$

(TP = True Positives, FP = False Positives, FN = False Negatives).

Intersection over Union (IoU) - quantifies the overlap between predicted and actual segmentation by dividing the intersection area by the union area of both. In this research, it's pivotal for gauging the difference between the model's predictions of burned areas and the ground truth. Mathematically, it is represented as:

$$\text{IoU} = \frac{TP}{TP + FP + FN} \quad (6)$$

Matthews Correlation Coefficient (MCC) - is a correlation coefficient between observed and predicted binary classifications. It returns a value between -1 and 1, where 1 indicates perfect prediction, -1 indicates total disagreement, and 0 indicates no better than random prediction. It's a robust metric when classes are imbalanced. Given that burned areas might cover only a fraction of the total area in many satellite images, MCC provides an

informative metric for this study. Mathematically, it is represented as:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (7)$$

Akaike Information Criterion - offers a means to balance model complexity with its performance. The formula is given by:

$$AIC = -2 \cdot \ln(L) + 2 \cdot k \quad (8)$$

For burned area segmentation tasks, the likelihood L can be replaced using the model’s average test accuracy as a proxy, as it provides a practical estimation of how well the model fits unseen data.

4.7 Training Configuration and Setups

A brief summary of each model’s training configuration is shown in Table 3 below:

Table 3: Architecture and Training Parameters

Architecture		Complexity Metrics			Training Parameters		
Main	Backbone	Param	Flops	Model Size	Epochs	LR	Optimizer
U-Net	Custom	31.2M	219.7B	119.5MB	500*	1e-4*	AdamW
U-Net all-band	Custom	31.3M	220.1B	119.5MB	500	1e-4	AdamW
Mask RCNN	ResNet-50	45.6M	280.3B	175.2MB	500	1e-4	AdamW
YOLO	yolov8m	27.3M	110.2B	52.4MB	500	1e-4	AdamW
SegFormer	MiT-b0	3.7M	6.76B	14.4MB	500	1e-5	AdamW
SegFormer	MiT-b3	45.2M	71.4B	179.1MB	500	1e-5	AdamW
SAM	ViT-base	93.7M	743.9B	375.2MB	200*	1e-5	AdamW

Notes*: Models are trained with 500 epochs and has a 100 epoch patience early stop mechanism. All learning rates are reduced to 10% of the starting rate in linear. SAM is trained for 200 epochs with a 50 epoch patience. Model parameters and flops are observed by using THOP PyTorch-OpCounter with batch size of 1.

All research experiments are taken place at Google Colab, with the following specifications:

CPU: Intel Xeon CPU @2.20 GHz

RAM: 80 GB

GPU: Nvidia A100 40GB

5 Result and Discussion

5.1 Training results

Table 4: Models Evaluation Results

Model	F1-score	IoU	MCC	Average
Traditional (Baseline)				
Index-product	0.5300	0.4020	0.4827	0.4716
CNN-based models				
U-net	0.8431	0.7343	0.8353	0.8042
U-net all band	0.8800	0.7897	0.8728	0.8475
Mask RCNN	0.6126	0.7850	0.6110	0.6695
YOLO	0.6466	0.7765	0.6394	0.6875
Transformer-based models				
SegFormer b0	0.8593	0.7638	0.8533	0.8255
SegFormer b3	0.8977	0.8165	0.8904	0.8682
SAM	0.6899	0.5631	0.6907	0.6479

Training results are shown in Table 4 above. In terms of the F1-score, SegFormer with MiT-b3 backbone achieves the highest score (0.8977), suggesting a superior balance of precision and recall. On the other end of the spectrum, Mask RCNN with ResNet-50 scored the lowest score (0.6126), indicating potential shortcomings in distinguishing true positives from false positives and negatives.

In terms of the IoU metric. SegFormer with MiT-b3 backbone again excels (0.8165), pointing to its efficiency in spatial localization. Conversely, SAM with SAM-ViT-base struggles in this metric (0.5631), suggesting possible challenges in accurate boundary delineation.

In terms of the Matthews Correlation Coefficient (MCC), which provides a balanced measure of binary classification performance, SegFormer with MiT-b3 outperforms U-Net all bands by a small margin (0.8904 vs 0.8728), while Mask RCNN performed the worst (0.6110). This indicates the robustness of SegFormer in handling both positive and negative classes, whereas Mask RCNN faces challenges in predicting the negative class in unbalanced datasets.

5.1.1 Index-product versus Computer Vision

The index-product acts as our baseline, achieving an F1-score of 0.5300. Significantly, every computer vision model surpassed this baseline, with the IoU (0.4020) and MCC (0.4827) scores further reinforcing this result.

Such an outcome is anticipated given the limitations of the index-product calculation, particularly when confronted with unclear data. Specifically, in the presence of clouds, the index-product method faces challenges. In contrast, computer vision models, once trained on datasets with such uncertainties, become adept at distinguishing noise from the actual burned areas. Consequently, these models consistently outpace the baseline, a result that aligns with logical expectations.

5.1.2 U-Net versus U-Net All Spectral

Both U-Net models show strong performance across F1, IoU and, MCC metrics. Notably, the U-Net variant utilizing all spectral bands demonstrates better results overall. This implies that tapping into the full spectrum provides the model with a more comprehensive understanding of the burned area, enhancing its segmentation abilities. Even though humans might find it challenging to discern information from all 13 bands, computer vision models manage to effectively harness these extra details. This mirrors the concept behind index-product compositions, where multiple bands are combined to make burned areas more distinguishable. Similarly, computer vision models can integrate insights from multiple bands to extract more nuanced details from the imagery.

5.1.3 Mask R-CNN vs YOLOv8

Mask R-CNN and YOLOv8 show relatively lower scores in both F1 (0.6126 and 0.6466, respectively) and MCC (0.6110 and 0.6394, respectively), indicating potential difficulties in distinguishing true positives from false positives. However, they obtained decent IoU scores (0.7850 and 0.7765), which suggests that while they managed to localize the burned areas effectively, they failed at delineating precise boundaries. The parallel performance trends between the two can be expected given that they both employ a ResNet-50 backbone and have architectural roots in the Fast R-CNN model. The stronger performance in IoU aligns with their design intent, as both Fast R-CNN and YOLO were primarily developed for object detection, where a higher IoU is a sought-after metric.

5.1.4 Segment Anything

While SAM exhibited sub-optimal performance across all metrics, with an F1-score of 0.6899, IoU of 0.5631, and MCC of 0.6907, it is important to contextualize these results. These metrics suggest that SAM faced difficulties in localizing the burned areas and delineating their boundaries accurately. However, this under-performance might not necessarily reflect the inherent capabilities of the SAM model. Given SAM’s complexity and larger number of parameters, it is plausible that the dataset size was insufficient for effective training. Larger models often require more extensive data to leverage their full potential, and without this, they risk overfitting or failing to capture broader patterns. It is also possible that certain hyper-parameters or training configurations might not have been optimal for this specific task.

5.1.5 SegFormer

SegFormer emerged as the top performer in this research. Specifically, the variant with the MiT-b3 backbone led the metrics with the highest F1-score (0.8977), IoU (0.8165), and MCC (0.8904). Interestingly, the SegFormer with the MiT-b0 backbone, despite having a significantly smaller parameter count of only 3.7M, produced results that were close to those of its MiT-b3 counterpart. This highlights the efficiency and potency of transformer-based segmentation models. The standout performance of SegFormer can likely be attributed to the self-attention mechanism that is integral to transformers. This mechanism facilitates a holistic understanding of image features, enabling the model to discern and process intricate image scenarios effectively, such as landscapes with scattered clouds overlaying burned areas. Through training, the self-attention mechanism captures and prioritizes global image features, which likely bolsters the model’s performance.

5.1.6 CNN-based vs. Transformer-based

When comparing the two paradigms, it is clear that SegFormer, representing the transformer-based models, outshines its CNN-based counterparts. Notably, it surpasses even the U-Net model that utilizes the full spectral input. This underscores the effectiveness of the self-attention mechanism, particularly in segmentation tasks involving images with intricate features.

5.2 Detailed Examination of Predicted Masks

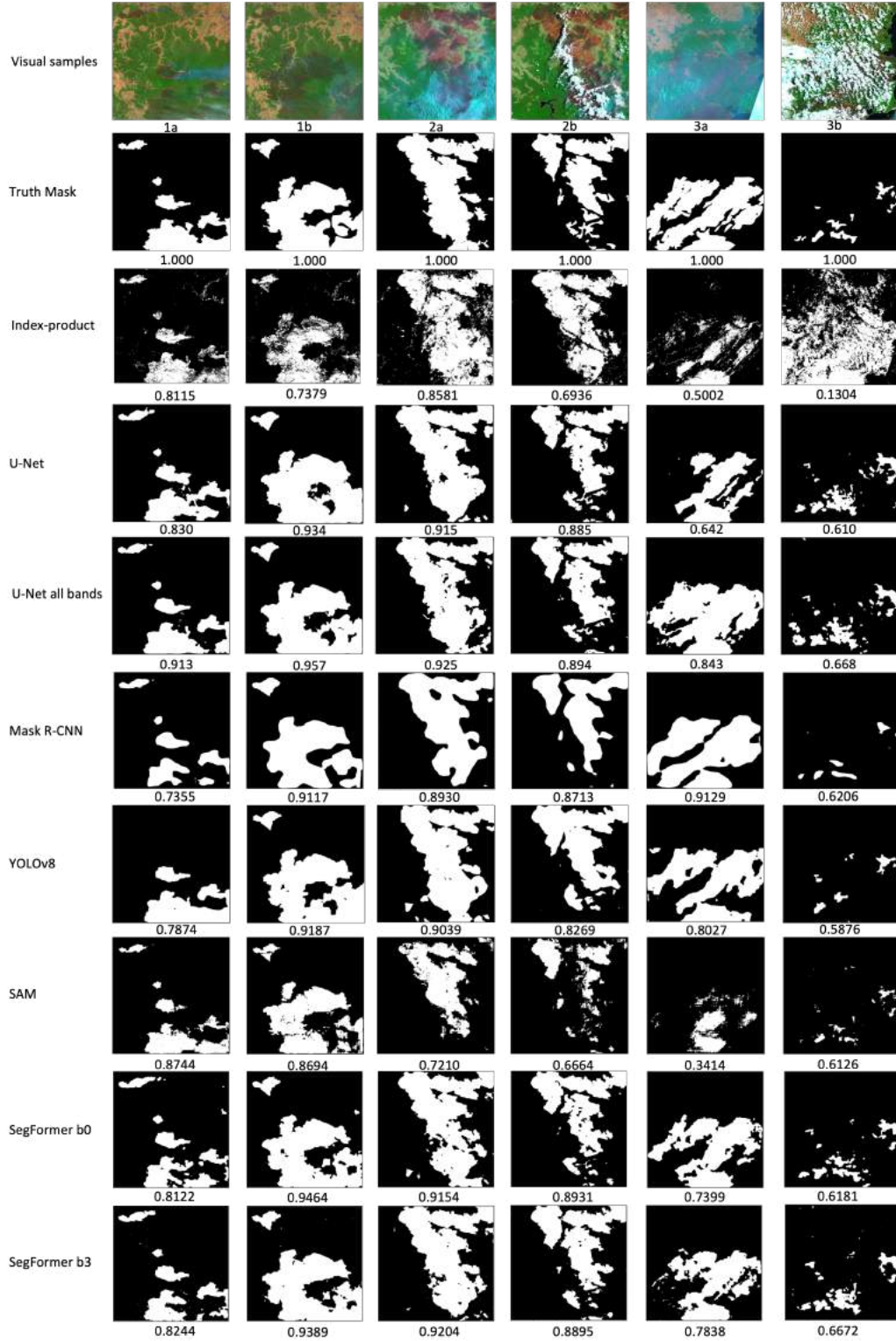


Figure 11: Predictions visual comparison (Refer to appendix A-J for larger images)

Figure 11 presents the predicted masks generated by various models. Selected visual samples were chosen to closely examine performance in diverse scenarios:

- **Images 1a and 1b:** These represent ideal conditions with clear skies, devoid of cloud cover, and minimal noise interference.
- **Image 2a:** This displays a light smoke layer obscuring the burned area. However, the near-infrared band's ability to penetrate such obstructions allows the underlying burned regions to remain visible.
- **Image 2b:** This has a moderate cloud cover positioned directly over the burned area, introducing a modest challenge to the models in accurately delineating the affected regions.
- **Image 3a:** While free from cloud interference, it is characterized by pronounced smoke disturbances. This makes discerning the burned terrain from its surroundings quite challenging.
- **Image 3b:** Showcases fragmented and dense cloud cover, imposing a significant challenge for models attempting to accurately predict the masked regions.

5.2.1 Limitation of Index-product

Upon examining the prediction results of the index-product, it is evident that the index-product excels (with an F1-score of approximately 0.8200) in scenarios where the data is free of clouds. Nevertheless, its efficacy is notably compromised when faced with obstructions such as clouds and smoke, as seen in images 2b and 3a. The performance drops to an F1-score of around 0.6000. In more challenging conditions like image 3b, which is dominated by dense cloud cover, the score plummets to a mere 0.1304. These observations underscore the vulnerabilities of the index-product. While it can yield commendable results on clear, noise-free data, it exhibits several limitations under less-than-ideal conditions.

5.2.2 Resilience of CV Models Against Noise

Conversely, computer vision models consistently outshine the index-product in challenging scenarios. Even the least successful of these models manages an F1-score of 0.5876 when contending with extensive cloud coverage. The convolutional layers intrinsic to these models excel at extracting hierarchical features, allowing them to discern patterns even amidst noise. This makes them particularly adept at tasks where the subject of interest (in this case, burned areas) can be obfuscated by various interferences like clouds or smoke.

5.2.3 Performance of U-Net with all spectral

A closer look at the prediction outcomes reveals that the U-Net model, equipped with full spectral input, emerges as the top performer across the sample images. It not only achieves precise segmentation of the burned areas in images 1a and 1b, with excellent F1-scores of 0.913 and 0.957 respectively, but also demonstrates resilience against cloud interference, as showcased in images 2b and 3b. Compared with the original U-Net model, the modified U-Net consistently surpasses its predecessor. A case in point is image 3a, characterized by dense smoke that blocked the burned area. The enhanced model successfully segments the burned area, securing an F1-score of 0.843. In contrast, the original model manages a less impressive score of 0.642. Such results prove the benefits of utilizing a comprehensive spectral input, underscoring its potential to enhance prediction accuracy.

5.2.4 Performance of Mask R-CNN and YOLOv8

While Mask R-CNN and YOLOv8 garnered only mediocre results in the earlier sections — with average accuracies of 0.6695 and 0.6875, respectively - their performance on the sample images presents a contrasting narrative. In several instances, these models demonstrated high precision, at times even matching or surpassing the leading models. One potential reason could be the unrepresentative nature of the sample images, implying that while the models excelled with these samples, they struggled with others, possibly those data with noise. This divergence hints at issues with model generalization. Further investigation is required to assess the validity of this supposition. Regardless, the results underscore that while both models possess potential in burned area segmentation tasks, they remain overshadowed by other CNN-based counterparts.

5.2.5 Performance of SAM

In spite of SAM being the most parameter-rich model, pre-trained on an extensive database, its performance was underwhelming, especially in predicting burned masks in sample images. It lagged in most of the assessments, specifically images 2a, 2b, 3a, and 3b, suggesting its vulnerability to noisy data. This lackluster performance, particularly when compared to simpler models, raises pertinent questions. One possibility is that SAM wasn't adequately trained due to data limitations. This leads to overfitting during the training iterations. This compromises its ability to generalize to new and diverse data. The intricate nature of large models necessitates rigorous hyperparameter tuning. It is plausible that sub-optimal configurations and setups prevented SAM from unlocking its full potential. Moreover, while SAM's architecture is intricate and advanced, it may not be the ideal

choice for tasks such as burned area detection. Simpler, task-specific models might be better suited for such challenges.

5.2.6 Performance of SegFormer

The SegFormer with the MiT-b0 backbone, despite being the most compact model among the contenders, displayed an impressive performance that rivaled leading models. Its efficiency with noise-free data, as seen in image 1b, was almost on par with the U-Net using all spectra, scoring 0.9464. Even in the presence of noise, such as in image 2b, it maintained commendable performance, securing the second-best score of 0.8931, just trailing the U-Net with all spectra. This exemplary performance accentuates the promise of transformer-based models. Furthermore, the slightly superior performance of its sibling model, SegFormer with the MiT-b3 backbone, reinforces the argument that SegFormer is well-suited for this task.

5.3 Overall Model Efficacy

Table 5: Model Parameters, Accuracy, and AIC Metrics

Model	Param (M)	Avg Accuracy (%)	AIC score
U-Net	31.2	80.42	53.62
U-Net all band	31.3	84.76	53.72
Mask RCNN	45.6	66.95	82.79
YOLO	27.3	68.75	46.13
SegFormer-b0	3.7	82.55	-1.42
SegFormer-b3	45.2	86.82	81.47
SAM	93.7	64.79	179.05

Utilizing the Akaike Information Criterion to gauge the balance between model complexity and performance (Table 5, the lower the better), the SegFormer with the MiT-b0 backbone emerges as the standout choice. With superior performance and a mere 3.7 million parameters, it demonstrates the highest efficiency and stands out as the most commendable model of all.

5.4 Comparison with Previous Works

Compared to prior research, the U-Net model consistently exhibits top-tier performance, reaffirming its effectiveness for satellite imagery tasks, as previously suggested by various research [19–24]. Our customized U-Net, which incorporates all spectral bands, showcases a marginal enhancement, indicating the value of this exploration.

While Mask R-CNN and YOLOv8 did not match the performance of U-Net, they still demonstrated strengths in certain sample tasks. In previous research findings [25, 37, 38], Mask R-CNN was shown to be able to carry segmentation tasks on vehicles and building boundaries. However, both Mask R-CNN and YOLOv8 might struggle with complex segmentation tasks, such as delineating burned area boundaries.

SegFormer, although previously untested on burned area satellite imagery, has demonstrated its prowess in detailed segmentation tasks. This mirrors suggestions from earlier studies [40, 41], crediting the model’s self-attention mechanism, which efficiently captures sparse features.

SAM, despite being the largest model with extensive pre-training, delivered subpar results. This indicates that its fine-tuning process might warrant optimization to fully harness SAM’s capabilities.

5.5 Limitations

Several limitations emerged during the course of this research:

Dataset Size: The dataset amassed for this research comprises 495 original images, augmented to 1980. While certain models, like SAM (which has the most tunable parameters), may greatly benefit from an expanded dataset, the limited size raises concerns about the generalizability of the models. The constrained size challenges the assertion that all models generalize effectively.

Labeling Challenges: The labeling task forms a cornerstone of this study, particularly when marking the noisy background data. We labeled this data manually, drawing on insights from multiple index products. Nonetheless, labeling proved ambiguous in certain instances. For instance, areas under cloud cover or those depicting near-recovered burn scars could lead to mislabeling. Furthermore, labeling became difficult for images depicting burn zones when blocked by clouds, leading to the delineation of numerous tiny burned area polygons.

Future studies would benefit from investing more time in the labeling phase and perhaps using help from the expertise of professionals.

Training Hyperparameters: The chosen hyperparameters, while tailored to our objectives, might not be ultimately optimal. The potential exists for other parameters to better suit each model. However, due to resource and time constraints, a comprehensive exploration remained necessary.

Backbone Selection: This research focused on comparing the model combinations without a uniform backbone for comparison. The lack of a common backbone might introduce some biases, given that backbones provide foundational structures and pre-trained weights to neural network architectures. Different backbones can significantly influence a model’s performance. However, selecting a standard backbone poses challenges, especially when comparing distinct model structures ranging from CNN-based to transformer-based architectures. Each model inherently carries unique characteristics, rendering it challenging, if not impossible, to identify a universally applicable backbone.

6 Concluding Remarks

Bushfires present significant threats in Australia, leading to economic losses and endangering lives. Monitoring burned areas plays a pivotal role in bushfire observation, assisting in real-time strategies and post-fire recovery plans. Satellite imagery, with its multi-spectral features and vast coverage area, stands as a vital tool for detecting burned areas. While traditional index-product-based methods have limitations, modern computer vision models offer an alternative, demonstrating robustness even with noisy data, making them valuable for real-time strategic decisions.

The research presented offers several contributions:

- We collected an original dataset centered on the 2019 Australia Black Summer bushfire incidents, sourced from the Sentinel-2 database. These were manually labeled and formed the foundation for this research.
- The standardized collection process we employed can serve as a blueprint for future Australian fire data collection researchers.
- We evaluated multiple computer vision models, encompassing CNN-based models like U-Net and Mask R-CNN, and transformer-based models like SAM and SegFormer. Our findings underscore that computer vision models exhibit superior performance compared to index-product methods, especially in the presence of noisy backgrounds such as clouds and smoke. Notably, SegFormer-b0 emerged as the top performer in our benchmarks, when considering its accuracy and complexity, followed closely by a modified U-Net model.
- The modified U-Net model's use of all satellite image spectral hints at its potential applicability to further improve other models.
- Despite its insights, our research has some limitations. Notably, the models we examined do not share a consistent backbone, complicating direct comparisons. Further challenges arise from the limited dataset size and potential improvements in data labeling quality.

In summary, while the index-product method remains viable for clean data, U-Net distinguishes itself among CNN models. Meanwhile, SegFormer, as part of the up-rise of transformer-based models, exhibited commendable performance. Through these findings, this study aspires to lay the foundation for subsequent research endeavors centered on Australian bushfire observation.

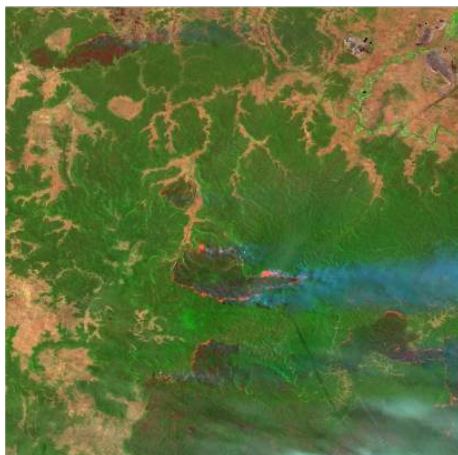
7 Future Directions

1. **Leveraging Additional Information Channels:** The U-Net model, when incorporated with all spectral inputs, demonstrated marked improvements in segmentation accuracy, particularly in scenarios with low to medium noise. This encourages the exploration of refined computer vision models to accommodate more information channels. A potential candidate for such a modification is SegFormer, the standout performer from our research. Given its self-attention mechanism, there is a presumption that it can more effectively harness the additional information from these channels.

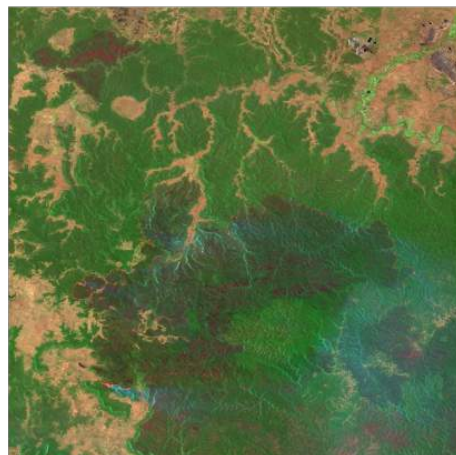
Additionally, instead of employing a convolution kernel that potentially averages out and thereby loses data across channels, a more effective strategy might be to use image tiling as highlighted in the original vision transformer paper [28]. The resultant tiled images can be concatenated into a long tensor, which can then be guided by the self-attention mechanism. Although this might substantially increase the model's complexity, given the relatively small size of SegFormer-b0, it might be a viable pathway.

2. **Utilizing Time-Series Data:** An alternate strategy to counteract noisy backgrounds might lie in fully leveraging time-series data. Given that the collected data derives from a consistent coordinate bounding box but spans different days, it might be possible to employ methods like Recurrent Neural Networks. Such an approach could utilize the prediction from a prior frame to aid the prediction of the subsequent frame, potentially offering clarity even in heavy noisy scenarios, providing the ability to "guess" the burned area under the noise.
3. **Diversifying Data Sources:** Building upon the previous point, future endeavors could diversify their data sources. An example would be Sentinel-3, which offers the advantage of a faster revisit interval (every 24 hours), albeit at the expense of reduced resolution. This approach could enhance the richness of the dataset, especially when there are limited fire events.

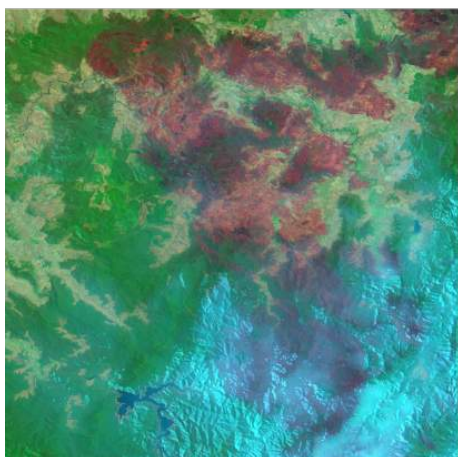
A Sample Images



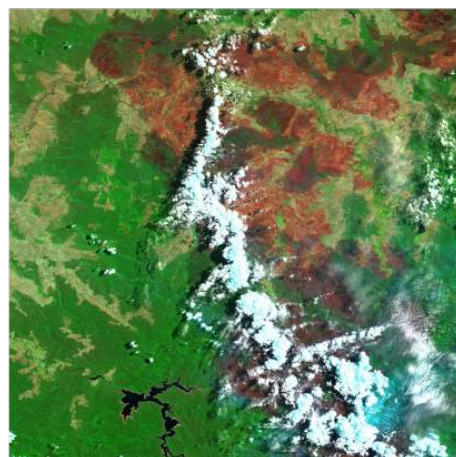
(a) Sample image 1a



(b) Sample image 1b



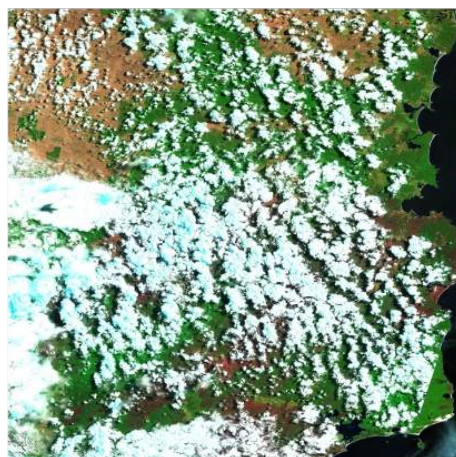
(c) Sample image 2a



(d) Sample image 2b



(e) Sample image 3a



(f) Sample image 3b

B True Masks



(a) true mask 1a



(b) true mask 1b



(c) true mask 2a



(d) true mask 2b

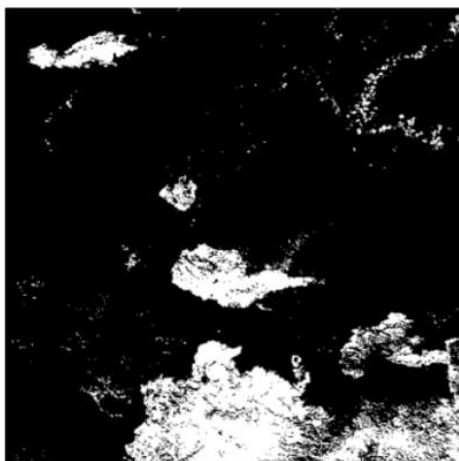


(e) true mask 3a

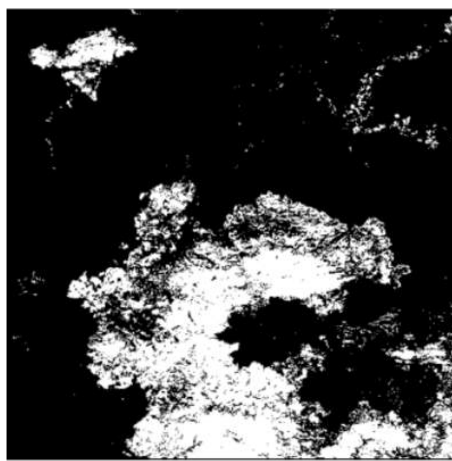


(f) true mask 3b

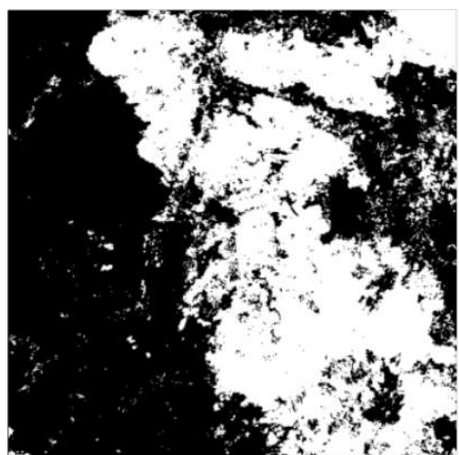
C Index-product predication



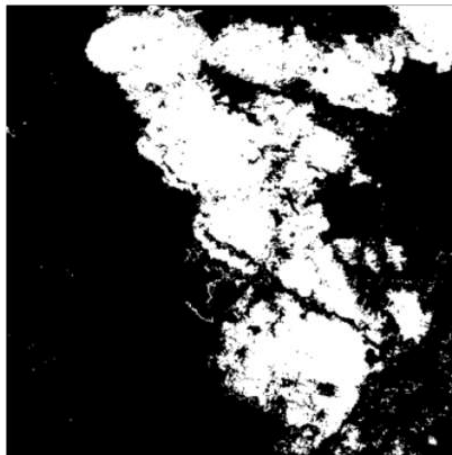
(a) index-product predication 1a



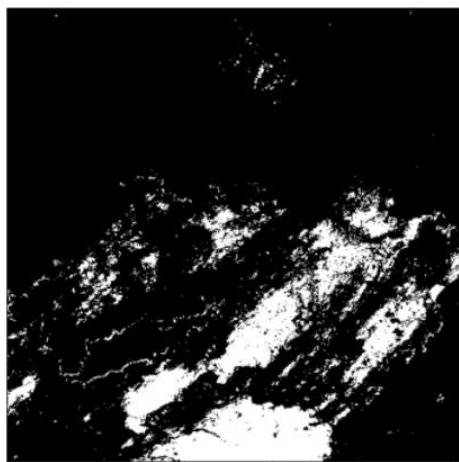
(b) index-product predication 1b



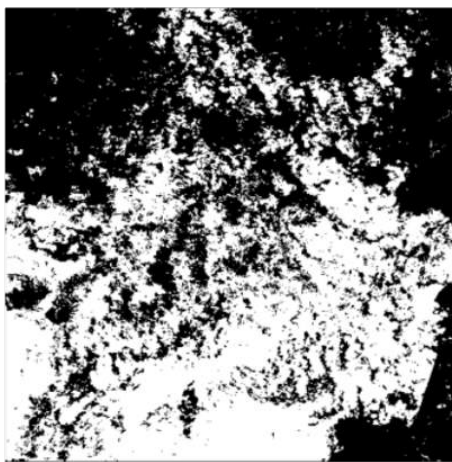
(c) index-product predication 2a



(d) index-product predication 2b



(e) index-product predication 3a



(f) index-product predication 3b

D U-Net predictions



(a) U-Net prediction 1a



(b) U-Net prediction 1b



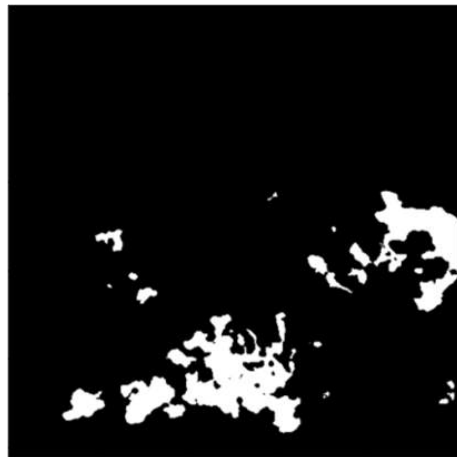
(c) U-Net prediction 2a



(d) U-Net prediction 2b



(e) U-Net prediction 3a



(f) U-Net prediction 3b

E U-Net all spectral predications



(a) U-Net all spectral prediction 1a



(b) U-Net all spectral prediction 1b



(c) U-Net all spectral prediction 2a



(d) U-Net all spectral prediction 2b



(e) U-Net all spectral prediction 3a



(f) U-Net all spectral prediction 3b

F Mask R-CNN predictions



(a) Mask R-CNN prediction 1a



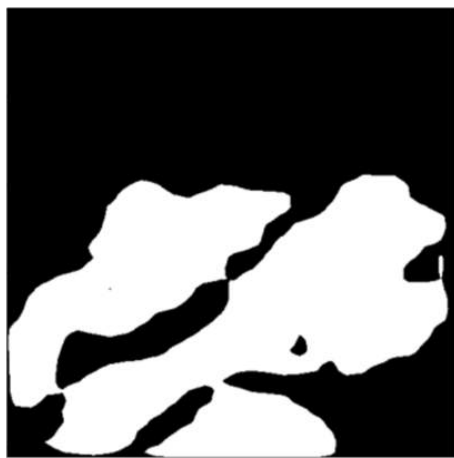
(b) Mask R-CNN prediction 1b



(c) Mask R-CNN prediction 2a



(d) Mask R-CNN prediction 2b



(e) Mask R-CNN prediction 3a



(f) Mask R-CNN prediction 3b

G YOLOv8 predication



(a) YOLOv8 predication 1a



(b) YOLOv8 predication 1b



(c) YOLOv8 predication 2a



(d) YOLOv8 predication 2b



(e) YOLOv8 predication 3a



(f) YOLOv8 predication 3b

H SAM predications



(a) SAM predication 1a



(b) SAM predication 1b



(c) SAM predication 2a



(d) SAM predication 2b



(e) SAM predication 3a



(f) SAM predication 3b

I SegFormer b0 predictions



(a) SegFormer b0 prediction 1a



(b) SegFormer b0 prediction 1b



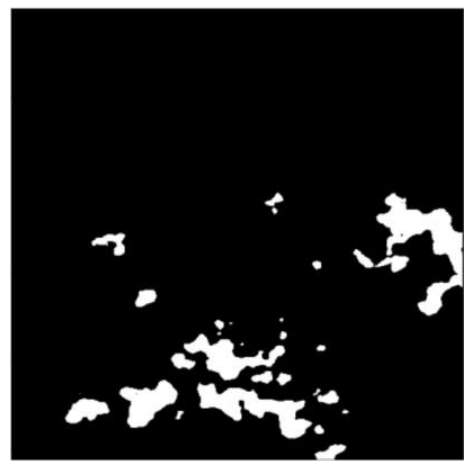
(c) SegFormer b0 prediction 2a



(d) SegFormer b0 prediction 2b



(e) SegFormer b0 prediction 3a



(f) SegFormer b0 prediction 3b

J SegFormer b3 predication



(a) SegFormer b3 predication 1a



(b) SegFormer b3 predication 1b



(c) SegFormer b3 predication 2a



(d) SegFormer b3 predication 2b



(e) SegFormer b3 predication 3a



(f) SegFormer b3 predication 3b

References

- [1] Bureau of Meteorology, “Special climate statement 72 – dangerous bushfire weather in spring 2019 leads into summer,” 2020. [Online]. Available: <http://www.bom.gov.au/climate/current/statements/scs72.pdf>
- [2] A. B. Corporation. (2020) Bushfire season 2019–2020 in numbers. [Online]. Available: <https://www.abc.net.au/news/2020-02-19/australia-bushfires-how-heat-and-drought-created-a-tinderbox/11976134>
- [3] I. A. Group, “Climate change action report 2020,” 2020. [Online]. Available: <https://www.iag.com.au/sites/default/files/Documents/Reports/2020-Climate-Change-Action-Report.pdf>
- [4] B. E. O. Taskforce, “Report on the role of space-based earth observations to support planning, response and recovery for bushfires,” Australian Space Agency, space.gov.au, Tech. Rep., May 2020. [Online]. Available: <https://www.industry.gov.au/sites/default/files/2020-12/bushfire-earth-observation-taskforce-report.pdf>
- [5] J. Purtill. (2023, 8) Rapid bushfire detection was promised after the black summer fires. it may have hit a roadblock. ABC Sciences. [Online]. Available: <https://www.abc.net.au/news/science/2023-08-30/bushfires-can-be-detected-from-space-within-minutes-of-ignition/102765624>
- [6] (2023, 10) Actionable insights for bushfire response. ESRI Australia. [Online]. Available: <https://esriaustralia.com.au/gis-in-bushfire-management>
- [7] (2020, 11) Bushfire research capability. The University of Queensland. [Online]. Available: https://www.uq.edu.au/research/files/69925/Bushfire_Research_Capability.pdf
- [8] (2021, 8) Ozfuel pre-phase a study, australian forest fuel monitoring from space. The Australian National University. [Online]. Available: https://inspace.anu.edu.au/files/ANU%20OzFuel%20Pre-Phase%20A_Aug%202021.pdf
- [9] K. Perera, R. Tateishi, K. Akihiko, and S. Herath. (2021) A combined approach of remote sensing, gis, and social media to create and disseminate bushfire warning contents to rural australia.
- [10] (2023, 10) Remote sensing: Observing the earth from above. The University of Tasmania. [Online]. Available: <https://www.utas.edu.au/courses/cse/units/kgg103-remote-sensing-observing-the-earth-from-above>
- [11] (2023, 10) Forecasting bushfire risk: Integrating a new ground-based sensor network, remote sensing, and weather data to forecast forest fuel dryness. The University

-
- of Melbourne. [Online]. Available: <https://www.unimelb.edu.au/mdap/research/2021-collaborations/forecasting-bushfire-risk>
- [12] C. Key and N. Benson, “Measuring and remote sensing of burn severity: The cbi and nbr,” in *Proceedings of the Joint Fire Science Conference and Workshop*, vol. 2000, Boise, ID, USA, 15–17 June 1999, pp. 284–285.
- [13] M. Martín, I. Gómez, and E. Chuvieco, “Performance of a burned-area index (baim) for mapping mediterranean burned scars from modis data,” in *Proceedings of the 5th International Workshop on Remote Sensing and GIS Applications to forest fire management: Fire effects assessment*, Zaragoza, Spain, 16–18 June 2005, pp. 193–198.
- [14] (2023, 10) Nsw fire zone ‘still a very dangerous area’ despite downgrades. 9NEWS. [Online]. Available: <https://www.9news.com.au/national/nsw-bushfires-emergency-warnings-cessnock-bega-valley/92d2b2a5-bc9f-48f6-80c6-d320a07a04a0>
- [15] (2019) Global assessment report on disaster risk reduction 2019. Accessed: 7 January 2020. [Online]. Available: https://gar.undrr.org/sites/default/files/reports/2019-05/full_gar_report.pdf
- [16] (2023, 9) El niño under way in the tropical pacific. Bureau of Meteorology. [Online]. Available: <http://www.bom.gov.au/climate/enso/outlook/#:~:text=An%20El%20Ni%C3%B1o%20has%20been,%C2%B0C%20warmer%20than%20average>
- [17] M. L. García and V. Caselles, “Mapping burns and natural reforestation using thematic mapper data,” *Geocarto International*, vol. 6, no. 1, pp. 31–37, 1991.
- [18] J. Lizundia-Loiola, M. Franquesa, A. Khairoun, and E. Chuvieco, “Global burned area mapping from sentinel-3 synergy and viirs active fires,” *Remote Sensing of Environment*, 2022, available online 7 October 2022. [Online]. Available: <https://doi.org/10.1016/j.rse.2022.113298>
- [19] L. Knopp, M. Wieland, M. Rättich, and S. Martinis, “A deep learning approach for burned area segmentation with sentinel-2 data,” *Remote Sensing*, vol. 12, no. 15, p. 2422, July 2020, received: 18 June 2020 / Revised: 24 July 2020 / Accepted: 25 July 2020 / Published: 28 July 2020. [Online]. Available: <https://doi.org/10.3390/rs12152422>
- [20] X. Hu, Y. Ban, and A. Nascetti, “Uni-temporal multispectral imagery for burned area mapping with deep learning,” *Remote Sensing*, vol. 13, no. 8, p. 1509, April 2021, received: 7 February 2021 / Revised: 12 March 2021 / Accepted: 5 April 2021 / Published: 14 April 2021. [Online]. Available: <https://doi.org/10.3390/rs13081509>

-
- [21] Z. Pan, J. Xu, Y. Guo, Y. Hu, and G. Wang, "Deep learning segmentation and classification for urban village using a worldview satellite image based on u-net," *Remote Sensing*, vol. 12, no. 10, p. 1574, 2020.
- [22] M. Wieland, Y. Li, and S. Martinis, "Multi-sensor cloud and cloud shadow segmentation with a convolutional neural network," *Remote Sensing of Environment*, vol. 230, p. 111203, 2019.
- [23] M. Wieland and S. Martinis, "A modular processing chain for automated flood monitoring from multi-spectral satellite data," *Remote Sensing*, vol. 11, no. 19, p. 2330, 2019.
- [24] P. P. de Bem, O. A. de Carvalho Júnior, O. L. F. de Carvalho, R. A. T. Gomes, and R. Fontes Guimarães, "Performance analysis of deep convolutional autoencoders with different patch sizes for change detection from burnt areas," *Remote Sensing*, vol. 12, no. 16, p. 2576, 2020.
- [25] K. Zhao, J. Kang, J. Jung, and G. Sohn, "Building extraction from satellite images using mask r-cnn with building boundary regularization," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 247–251.
- [26] T. T. P. Quoc, T. T. Linh, and T. N. T. Minh, "Comparing u-net convolutional network with mask r-cnn in agricultural area segmentation on satellite images," in *2020 7th NAFOSTED Conference on Information and Computer Science (NICS)*. IEEE, 2020, pp. 124–129.
- [27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008. [Online]. Available: <https://arxiv.org/abs/1706.03762>
- [28] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
- [29] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *arXiv preprint arXiv:2105.15203*, 2021, accepted by NeurIPS 2021. [Online]. Available: <https://doi.org/10.48550/arXiv.2105.15203>
- [30] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick,

-
- “Segment anything,” *arXiv preprint arXiv:2304.02643*, 2023. [Online]. Available: <https://doi.org/10.48550/arXiv.2304.02643>
- [31] M. Kaselimi, A. Voulodimos, I. Daskalopoulos, N. Doulamis, and A. Doulamis, “A vision transformer model for convolution-free multilabel classification of satellite imagery in deforestation monitoring,” *IEEE Transactions on Neural Networks and Learning Systems*, 2022.
- [32] J. Horváth, S. Baireddy, H. Hao, D. M. Montserrat, and E. J. Delp, “Manipulation detection in satellite images using vision transformer,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1032–1041.
- [33] “National bushfire recovery agency datasets,” https://data.gov.au/data/organization/nbra?q=bushfire&sort=score+desc%2C+metadata_modified+desc, Australian Government, data.gov.au, 2023, accessed: 2023-10-19.
- [34] E. Alcaras, D. Costantino, F. Guastaferro, C. Parente, and M. Pepe, “Normalized burn ratio plus (nbr+): A new index for sentinel-2 imagery,” *Remote Sensing*, vol. 14, no. 7, 2022. [Online]. Available: <https://www.mdpi.com/2072-4292/14/7/1727>
- [35] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.
- [36] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask r-cnn,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.
- [37] J. W. Johnson, “Adapting mask-rcnn for automatic nucleus segmentation,” *arXiv preprint arXiv:1805.00500*, 2018.
- [38] X. Nie, M. Duan, H. Ding, B. Hu, and E. K. Wong, “Attention mask r-cnn for ship detection and segmentation from remote sensing images,” *Ieee Access*, vol. 8, pp. 9325–9334, 2020.
- [39] Z. Guan, X. Miao, Y. Mu, Q. Sun, Q. Ye, and D. Gao, “Forest fire segmentation from aerial imagery data using an improved instance segmentation model,” *Remote Sensing*, vol. 14, no. 13, 2022. [Online]. Available: <https://www.mdpi.com/2072-4292/14/13/3159>
- [40] Z. Wang, Q. Wang, Y. Yang, N. Liu, Y. Chen, and J. Gao, “Seismic facies segmentation via a segformer-based specific encoder–decoder–hypercolumns scheme,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–11, 2023.

-
- [41] J. Deng, X. Lv, L. Yang, B. Zhao, C. Zhou, Z. Yang, J. Jiang, N. Ning, J. Zhang, J. Shi *et al.*, “Assessing macro disease index of wheat stripe rust based on segformer with complex background in the field,” *Sensors*, vol. 22, no. 15, p. 5676, 2022.