THE UNIVERSITY of ADELAIDE

Automated Landcover Segmentation

Sakshi Sinha (a1898508)

The University of Adelaide

4536_COMP_SCI_7306B Artificial Intelligence and Machine Learning
Industry Project Part B

Supervisors:
Dr. Mark McDonnell (Aurizn)
Dr. Qi Wu (The University of Adelaide)

---

1 Abstract

Accurate land cover segmentation from high-resolution satellite imagery is critical for applications such as urban planning, environmental monitoring, and disaster response. This industry project, conducted for Aurizn, aimed to optimize deep learning models for semantic segmentation of Aurizn's MS and PAN dataset, focusing on five land cover classes: no data, road, building, vegetation, and water. A hybrid SegFormer-U-Net model with Canny edge detection was developed, leveraging MS-PAN data to achieve robust segmentation. Feature engineering with spectral indices enhanced segmentation accuracy, while scale-dependent performance was observed: larger-scale images (1024×1024) improved segmentation of contiguous classes like vegetation, and smaller-scale images (256×256) excelled for fine features like roads. The Canny Hybrid model achieved a class-wise Dice score of 0.5843, outperforming Sobel-based and ensemble approaches. Strategic data augmentation and validation across Australian regions, combined with targeted preprocessing to address class imbalance and MS-PAN dimension alignment, ensured reliable outcomes. These advancements enable Aurizn to monitor Australia's ecosystems with greater precision, supporting sustainable practices. Future work includes advanced fusion architectures and dataset diversification.

2 INTRODUCTION

The increasing availability of high-resolution satellite imagery has significantly advanced remote sensing, making precise land cover segmentation vital for applications such as urban planning, environmental monitoring, disaster response, and resource management [1, 2]. Accurate identification of land cover type (such as vegetation, buildings, roads, etc.) enables improved urban development strategies [2], effective monitoring of deforestation and ecosystem health [3], enhanced post-disaster damage assessment [4], and optimized agricultural practices [5]. Figure 1 shows an example of land cover segmentation in satellite imagery.

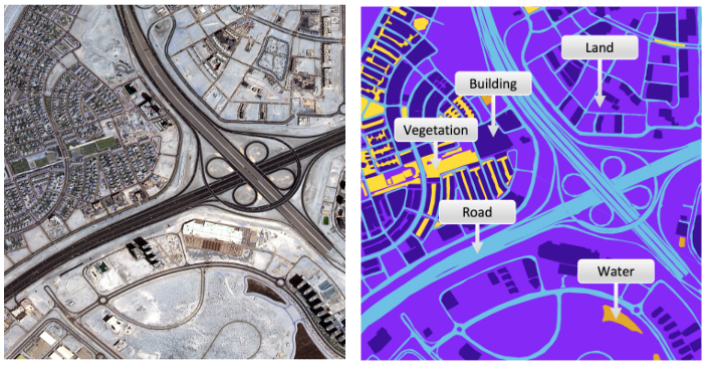


Figure 1: Example of vegetation cover change in Ibbaluru forest, Bengaluru, India. Image sourced from [6].

For Aurizn, the practical value of this capability is substantial. Australia's unique landscapes lack extensive open-source research tailored to local conditions, and accurate mapping is critical for sustainable land and resource management. Aurizn's high-resolution multispectral (MS) dataset offers richer information than standard RGB imagery [7], but also requires specialized processing techniques to fully exploit its potential [8]. By integrating MS data with PAN imagery, there is an opportunity to leverage both fine spatial detail and diverse spectral

information for improved segmentation accuracy. The type of work Aurizn wishes to perform includes vegetation tracking, similar to projects that have successfully monitored changes in vegetation cover (densification) in the Ibbaluru forest in Bengaluru, India (Figure 2).

This project focused on the development and optimization of deep learning models for semantic segmentation of Aurizn's MS-PAN dataset, targeting key land cover classes: no data, road, building, vegetation, and water. This research successfully delivered open-source, high-performing, and context-specific models capable of supporting actionable decision-making for urban planning, environmental monitoring, and sustainable resource management.

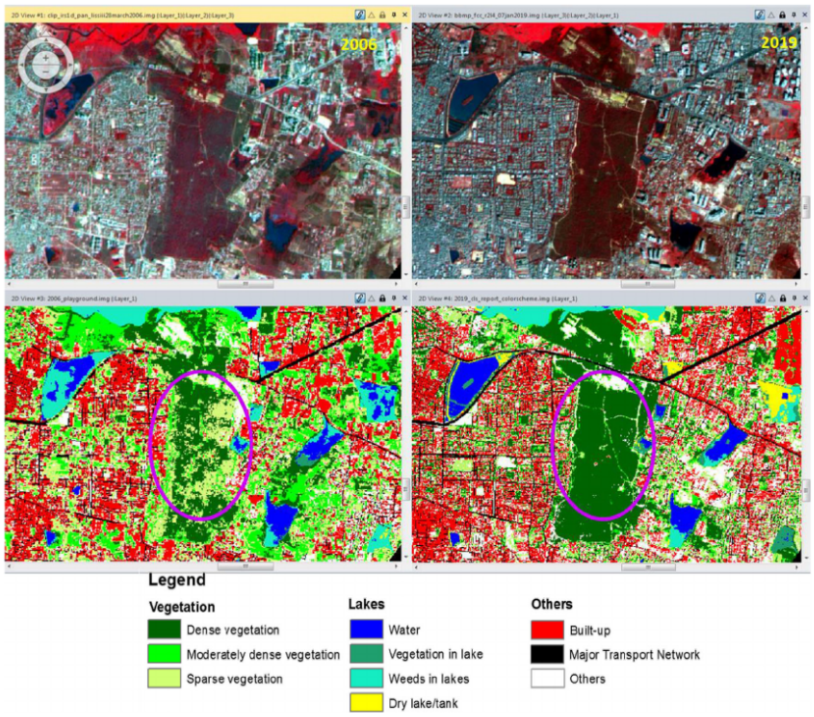


Figure 2: Example of land cover segmentation in satellite imagery. Image sourced from [9].

This report is structured to provide a comprehensive overview of the project. In Section 3, the report reviews key literature to establish the background and context of the research. Section 4 outlines the original project plan and adaptive modifications made during execution. Section 5 details the datasets, data processing, model architectures, and evaluation metrics used. Section 6 presents the experimental setup, along with the outcomes and model performance across different scales and post-processing methods. Section 7 brings together the key findings and their implications. Section 8 discusses the main challenges encountered and lessons learned. Section 10 proposes avenues for further research.

3 LITERATURE REVIEW

3.1 Introduction to Semantic Segmentation In Remote Sensing

Semantic segmentation is a fundamental task in computer vision, particularly relevant in remote sensing applications such as urban planning, land cover mapping, disaster response, and resource management [1]. The goal of semantic segmentation is to classify each pixel in an image into predefined categories, such as vegetation, buildings, roads, and water bodies. Recent

advancements in DL have significantly improved segmentation accuracy compared to traditional machine learning and rule-based methods [10].

While DL-based approaches offer substantial improvements, their effectiveness depends on overcoming key challenges like data scarcity and class imbalance. These challenges, which Aurizn faced, were addressed to enhance image analysis and decision-making [11]. This project aimed to tackle these challenges by exploring and adapting advanced DL models to Aurizn's specific MS dataset.

3.2 Deep Learning for Semantic Segmentation

3.2.1 Convolutional Neural Networks

CNN architectures like U-Net [12], DeepLabv3+ [13], and HRNet [14, 15] showed promising results in remote sensing, offering efficient feature extraction and fine-grained spatial detail capture. However, their limitations in capturing long-range dependencies, especially in high-resolution imagery, necessitated further investigation [16]. A core objective of this project was to determine the effectiveness of CNNs, particularly pre-trained CNNs, on Aurizn's dataset and compare their performance against transformer-based models. This comparison directly informed the selection of the most suitable architecture for Aurizn's operational needs.

3.2.2 Transformer-based Architectures

Transformers, including ViTs and Swin Transformers, offered improved long-range dependency modeling through self-attention [17]. The emergence of models like SAM, with its zero-shot and few-shot capabilities, was particularly relevant for Aurizn's dataset, where labeled data was limited [18]. However, SAM's requirement for prompting made it less suitable for automated large-scale processing, leading us to explore alternatives.

SegFormer's hierarchical design and efficiency in handling high-resolution images positioned it as a promising transformer variant for remote sensing applications [**?**]. Swin Transformers' success in capturing both local and global features through shifted windowing demonstrated the potential for transformer-CNN complementarity [19].

This project explored and adapted transformer-based models to Aurizn's MS imagery, assessing their ability to generalise and perform effectively in this specialised domain.

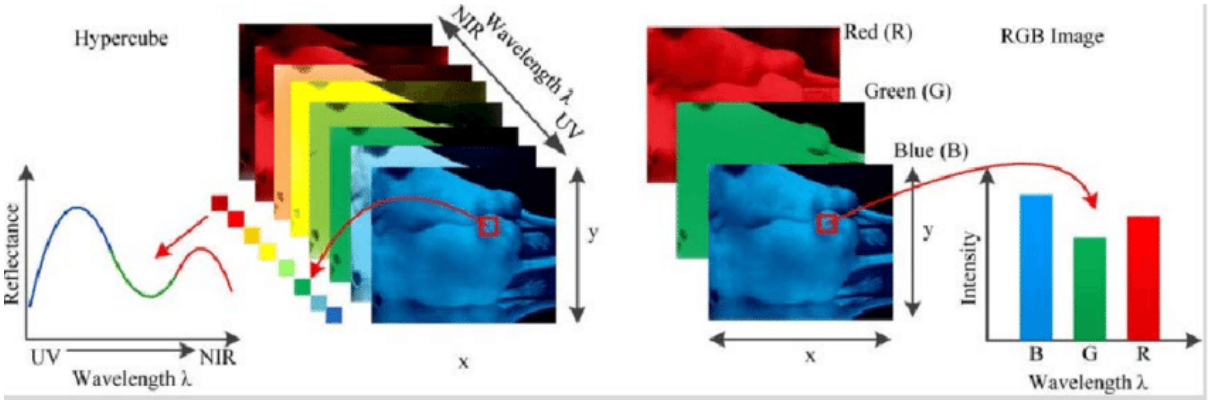


Figure 3: Difference between RGB and MS imaging. Image sourced from [20].

3.3 MS Image Segmentation in Remote sensing

MS satellite imagery captures reflected energy beyond the visible spectrum, as illustrated in Figure 3, provided crucial information for detailed land cover analysis [7]. While offering richer data for improved segmentation compared to RGB, it also presented unique challenges that directly informed our preprocessing strategy. The high dimensionality of MS data increased computational cost and could hinder model performance, leading us to implement strategic feature engineering with spectral indices [21]. The high dimensionality of MS data increased computational cost and could hinder model performance. Varying spectral resolutions and bandwidths across sensors required careful preprocessing, including atmospheric correction and radiometric calibration [8, 11].

This project focused on effectively utilising spectral information, addressing the complexities of MS data through techniques like spectral attention, multi-branch networks and pan-sharpening [22]. It explored both traditional pan-sharpening as a preprocessing step and deep learning-based methods that learn the optimal fusion within the network itself [23].

3.4 Transfer Learning and Fine-Tuning

3.4.1 Benefits of Transfer Learning

Transfer learning was instrumental in remote sensing applications, allowing models pre-trained on large datasets to be adapted to specific tasks with limited labeled data. This was particularly beneficial for segmentation, where labeled data acquisition is expensive and time-consuming [24, 25]. This approach was utilised to leverage the rich features learned from large pre-trained datasets, thereby mitigating the challenges of limited labeled data in this project.

3.4.2 Fine-Tuning Pretrained RGB Models on MS Images

Adapting models trained in RGB to MS data required careful adjustments, including band-wise feature fusion and spectral attention [26, 27, 28]. This is an important task due to the domain shift between natural images and satellite imagery. Common approaches to adapting pre-trained RGB models include modifying the first convolutional layer to handle the extra bands or replicating the MS bands to fit the RGB model's input size. A key objective of this project was to investigate the effectiveness of fine-tuning pre-trained RGB models on Aurizn's MS dataset, comparing their performance to find the most efficient approach.

3.5 Edge Detection in Remote Sensing Applications

Edge detection techniques have proven valuable in remote sensing for enhancing structural features and improving segmentation accuracy. Research has shown that incorporating edge information as additional input channels can significantly improve segmentation performance for infrastructure mapping [29], with multi-stage edge detectors generally outperforming simpler gradient-based methods for linear feature extraction [30, 31].

3.6 Hybrid and Ensemble Architectures

Recent research has shown promising results from combining different architectural approaches to leverage their complementary strengths. Studies on CNN-Transformer hybrids have demonstrated that combining local feature extraction capabilities of CNNs with global context mod-

eling of transformers can achieve superior performance compared to individual architectures [17]. However, ensemble methods in remote sensing have shown mixed results, with some studies reporting improved robustness while others noting increased computational overhead without proportional performance gains [11, 32]. The effectiveness of ensemble approaches depends heavily on model diversity and fusion strategies, with suboptimal combinations sometimes underperforming simpler architectures.

3.7 Scale-Dependent Performance in Remote Sensing

Limited research exists on the systematic evaluation of scale-dependent performance in land cover segmentation, representing a gap that this project addresses. While some studies have noted varying performance across different spatial resolutions, few have systematically analyzed how image scale affects different land cover classes [33].

3.8 Research Gaps and Project Contributions

The literature reveals several gaps that this project addresses:

- **Limited Hybrid Architectures:** While CNN-Transformer combinations exist, few studies have systematically evaluated these hybrid models for MS remote sensing applications.
- **Scale-Dependent Analysis:** Insufficient research on how image scale systematically affects different land cover classes in semantic segmentation.
- **Australian-Specific Datasets:** A lack of comprehensive studies on land cover segmentation tailored to Australian landscapes and conditions.
- **MS-PAN Fusion Optimization:** Limited work on learnable, end-to-end fusion methods for combining MS and PAN data within segmentation networks.

This project contributes to filling these gaps by developing and validating a novel hybrid architecture, conducting systematic multi-scale analysis, and providing insights specific to Australian remote sensing applications.

4 Research Plan vs Progress

4.1 Original Project Plan

The project was originally structured into three sequential phases as follows:

- Phase 1: Baseline Adaptation with RGB Data
- Phase 2: Adaptation and Estimation with MS-PAN Data
- Phase 3: VLM Exploration [Optional]

4.2 Project Progress and Strategic Adjustments

As illustrated in Figure 4, a significant strategic adjustment was made following a review of Phase 2 results and direct feedback from Aurizn. While Phases 1 and 2 were executed according to the initial plan, Aurizn expressed a clear preference to prioritise a more thorough validation of the MS-PAN models over the originally optional Phase 3. The final six weeks of the project were therefore re-allocated to extensive testing of the existing models across different image scales and the development of post-processing techniques, ensuring a more robust and comprehensive set of results that directly addressed the client's needs.

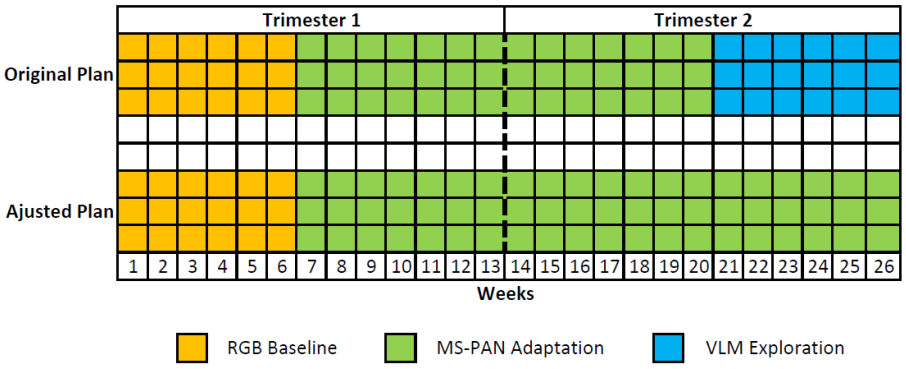


Figure 4: Project Timeline: Original Plan vs. Adjusted Progress

5 Methodology

This section outlines the systematic approach taken to establish a performance baseline, and adapt models for MS data. It details the datasets used, the model architectures selected, and the various strategies explored for data fusion and segmentation. The overall workflows for the RGB baseline and the MS-PAN adaptation are illustrated in Figure 5 and Figure 6, respectively.

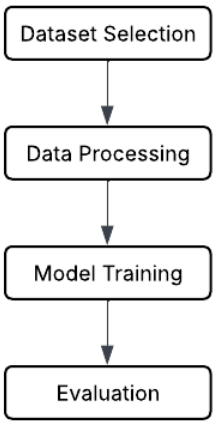


Figure 5: RGB baseline methodology workflow

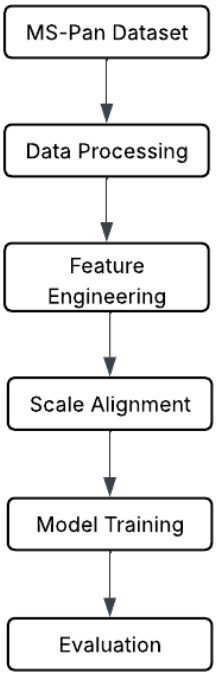


Figure 6: MS-PAN methodology workflow

5.1 Dataset

5.1.1 RGB Dataset

For the RGB phase of this project, a thorough exploration of three potential datasets was conducted to identify the most suitable resource for our land cover segmentation model. This process began with Dataset 1 (Kaggle - DataMes), a small collection of 72 images. While it provided foundational classes like buildings, land, and roads, its limited size was a significant constraint that prevented effective training and generalisation [34].

Next, Dataset 2 (LoveDA) was considered. This dataset offered a much larger pool of 5,987 high-resolution images across 7 classes. However, several critical issues made it less ideal for our purpose. These included inconsistent pixel values and memory constraints [35].

Ultimately, Dataset 3 (Kaggle - DeepGlobe Land Cover segmentation) was selected as the primary resource for initial experimentation. Although it presented challenges with less-defined mask boundaries, its considerable size of 803 high-resolution images across 7 classes offered a superior balance. The larger volume of data provided a robust foundation for training and fine-tuning, overcoming the limitations faced with the other two datasets and making it the most viable option [36].

5.1.2 MS-PAN Dataset

This dataset was provided by Aurizn for the next phase of the project. It contained both PAN and MS images. This dataset included 512×512 PAN images, $128 \times 128 \times 8$ MS images (representing the same location), and corresponding ground truth masks. The original 7 classes were re-labeled into 5: nodata/ground/clutter, road, building, vegetation, and water. The images were sourced from four Australian regions: Cowra, Young, Goulburn, and Griffith.

5.2 Data Processing

5.2.1 Data Split, Augmentation, and Oversampling

To ensure the model's robustness and to address class imbalance, a series of data processing steps were applied to both the RGB and MS-PAN datasets.

For the RGB dataset, the data was split into training set (80%) and a validation set (20%). To prevent overfitting and to improve the model's generalisation capabilities, various data augmentation techniques were applied, including random rotations, flips, and color jittering.

A similar approach was taken with the MS-PAN dataset, which was split into training and validation sets. Specifically, the data from 'Cowra', 'Young', and 'Goulburn' was used for training, while the 'Griffith' region was reserved exclusively for validation. To address class imbalance, a combination of oversampling on the minority classes and undersampling on the majority classes was applied. For data augmentation, only horizontal and vertical flips were used to expand the dataset, as techniques like color jittering and brightening resulted in degraded performance in this case and were therefore removed. This was crucial for achieving high performance across all five re-labeled classes. Please see Figure 7 for the before and after class distribution graph.

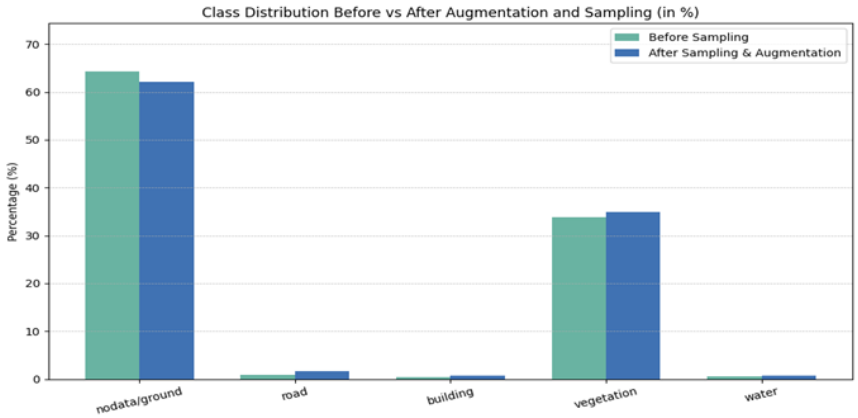


Figure 7: Class distribution for MS-PAN dataset before and after balancing techniques.

5.2.2 Feature Engineering

To enhance the information available to the model, additional features were engineered from the raw data. For the MS images, the Normalized Difference Red Edge (NDRE) and Normalized Difference Built-up Index (NDBI) were calculated and appended as additional channels [37]. This brought the MS data from 8 channels to 10, providing the model with more information about vegetation and built-up areas. These indices are calculated using specific bands from the MS data: NIR (Near-Infrared), Red Edge, and SWIR (Short-Wave Infrared), which are crucial for assessing vegetation health and identifying built-up areas. The formulas for these indices are as follows:

$$NDRE = \frac{NIR - RedEdge}{NIR + RedEdge + 10^{-8}} \tag{1}$$

$$NDBI = \frac{SWIR - NIR}{SWIR + NIR + 10^{-8}} \tag{2}$$

For the PAN images, Sobel and Canny edge detection algorithms were applied [31, 30]. The Sobel operator uses a pair of convolution kernels to compute the horizontal (G_x) and vertical (G_y) gradients of the image intensity function, which are then combined to find the absolute magnitude of the gradient at each point. The Canny edge detector, a multi-stage process, also uses a gradient calculation similar to Sobel, but follows it with non-maximum suppression and hysteresis thresholding to produce a refined, single-pixel-wide edge map, with a σ value of 1.0. These resulting edge features, converted to a 'float32' data type, were added as new channels to the PAN images, assisting the model in identifying boundaries and structures.

5.3 Model Architecture

5.3.1 Model Selection and Implementation for RGB

To establish a performance baseline for RGB segmentation, two popular models, U-Net and DeepLabV3, were selected and implemented with a ResNet-101 backbone. The architecture takes a 3-channel RGB input and produces a 7-channel output corresponding to the 7 landcover classes. For Unet, The encoder-decoder structure with skip connections facilitates detailed segmentation, leveraging the pre-trained ResNet101 backbone for robust feature extraction. Whereas For Deeplabv3 use of atrous convolutions and an Atrous Spatial Pyramid Pooling module to capture multi-scale context, which is essential for accurately segmenting objects of various sizes.

5.3.2 Model Selection for MS Adaptation

The limitations of the RGB dataset necessitated a transition to MS data. U-Net was selected for further adaptation in Phase 2 due to its established performance in Phase 1, with plans to also explore transformer-based models to potentially improve segmentation accuracy for challenging classes.

Scale Alignment The first step in adapting the models for MS-PAN data was to address the resolution difference between the MS (128×128) and PAN (512×512) images. Three primary techniques were explored for pansharpening and alignment:

- **Bayesian Pan-sharpening:** This technique [22] was explored and showed promising initial results but was computationally intensive and was not extensively utilised due to time constraints.
- **Interpolation:** Interpolating to directly resize either the MS or PAN image before feeding it to the model was also considered. However, this approach resulted in a loss of information and consistently led to poor performance.
- **Learnable Convolutional Layers:** The final and most successful approach was to use learnable convolutional layers for upsampling and downsampling. This was achieved using 'Conv2d' for downsampling the PAN image to match the MS resolution and 'ConvTranspose2d' for upsampling the fused feature maps to match the original PAN and label dimensions. This method was chosen as it performed best and allowed the model

to intelligently preserve details, but its effectiveness is dependent on a constant MS:PAN resolution ratio.

Input Layer Modification Each pre-trained model, originally trained on RGB data, underwent an input layer modification to handle the multi-channel MS-PAN input. The pre-trained RGB weights were retained as is for the RGB channels, while the additional channels were initialised randomly. Following this initialisation, all weights were updated during the training process, allowing the model to adapt and optimise its performance across the full MS-PAN spectrum. This approach ensured compatibility with the diverse spectral information while leveraging the pre-trained knowledge for the RGB portion of the data.

Multi-Modal Architecture Experiments Three separate architectural approaches were tested to leverage the combined MS and PAN data.

Hybrid SegFormer-U-Net The proposed architecture integrates a SegFormer [38] and a U-Net, connected through learned upsampling and downsampling blocks to align the spatial scales of different input modalities. The panchromatic (PAN) image is first downsampled to match the resolution of the multispectral (MS) image, after which the PAN and MS features are concatenated to form a fused feature representation.

This fused representation is processed in parallel by two segmentation heads: a SegFormer and a U-Net. The SegFormer contributes strong global context understanding and superior class differentiation, particularly for complex classes such as roads and buildings, while the U-Net excels at capturing fine spatial details. The outputs from these two heads are then combined using an attention mechanism, which adaptively emphasizes the most informative features from each branch to produce the final segmentation mask (see Figure 8).

The Hybrid SegFormer U-Net was selected after extensive evaluation against a variety of transformer-CNN combinations and standalone architectures, including standalone SegFormer, standalone U-Net, U-Net-SegFormer variants, and Swin Transformer [19] models. This hybrid design consistently outperformed the alternatives by effectively balancing local detail preservation with global semantic consistency. Its adaptability enables it to handle the diverse spatial and spectral characteristics of MS-PAN data, while the attention fusion further enhances segmentation consistency across all land cover classes.

Ensemble of Hybrid Models An alternative approach involved training two separate hybrid models. One model was trained to predict all five classes, while the second was a specialised model trained only to identify Water and Roads. The outputs of these two models were then stacked and fed into a final set of convolutional layers, which acted as a refinement block, intelligently combining the predictions for a more robust final output. A visual representation of this architecture is shown in Figure 9.

Ensemble of Binary Classifiers This architecture involved training a separate, specialised binary segmentation model for each of the five classes. Each binary model was designed to use a specific combination of input channels (e.g., MS bands, NDRE, NDBI, Sobel, Canny) that were most relevant for that particular class. The final predictions from each of these binary models were then stacked together and passed through a refinement block, similar to the ensemble hybrid model, to generate the final multi-class output. A visual representation of this architecture is shown in Figure 10.

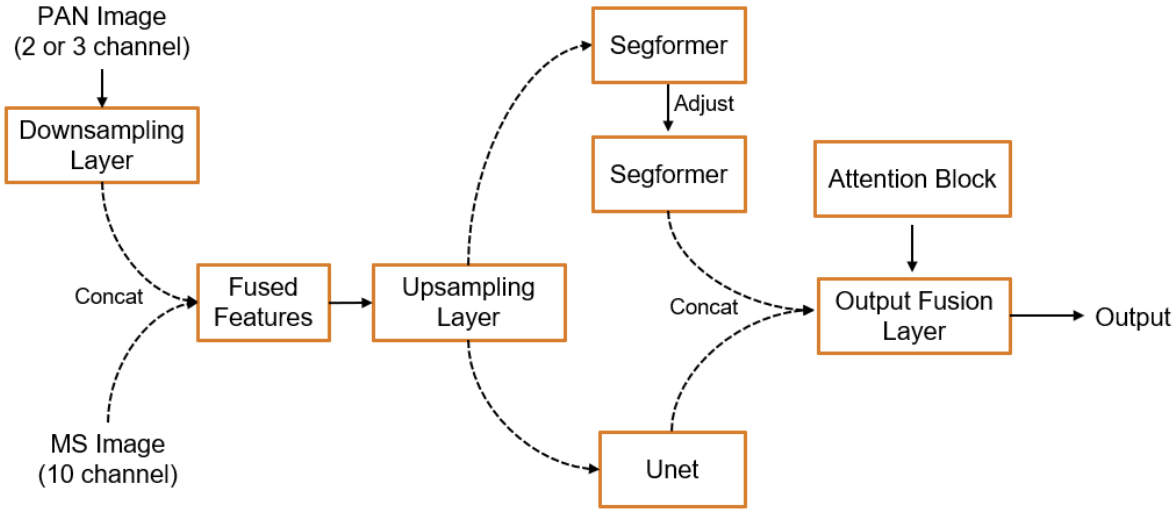


Figure 8: Unet-Segformer Hybrid Model Architecture.

5.4 Loss Functions

The Combined Segmentation Loss used throughout this project combines Dice Loss [39] and Cross-Entropy Loss [40] to balance pixel-wise accuracy with overall segmentation quality:

$$\text{Combined Loss} = \alpha \times \text{Dice Loss} + (1 - \alpha) \times \text{Cross-Entropy Loss} \tag{3}$$

Where α represents the weighting ratio between the two loss components. The Dice Loss focuses on overlap between predicted and ground truth regions, addressing class imbalance by emphasising the intersection over union, while Cross-Entropy Loss ensures pixel-wise classification accuracy. This combination leverages the complementary strengths of both loss functions for robust segmentation performance.

5.5 Evaluation

5.5.1 RGB Evaluation

The models were evaluated using standard segmentation metrics: *Dice coefficient*, *Intersection over Union (IoU)*, *overall accuracy*, *precision*, and *recall* [41]. The formulas for these metrics are as follows, where TP is True Positives, TN is True Negatives, FP is False Positives, and FN is False Negatives:

$$\text{Dice} = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{4}$$

$$\text{IoU} = \frac{TP}{TP + FP + FN} \tag{5}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{7}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{8}$$

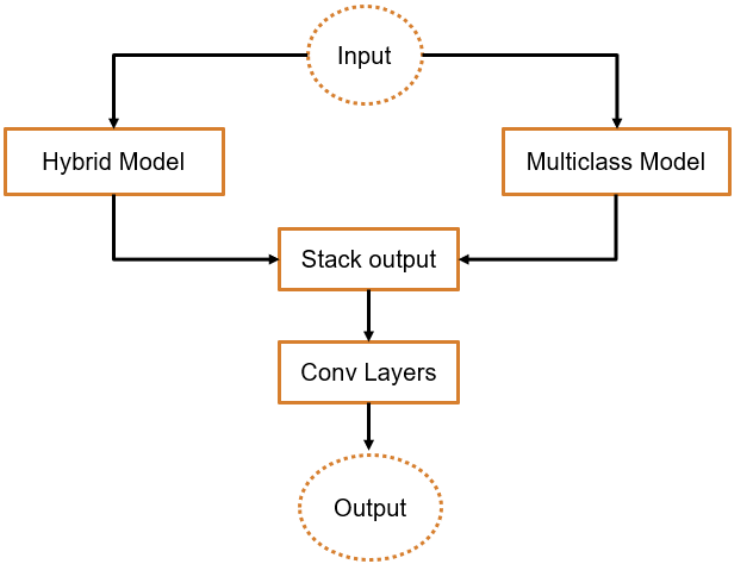


Figure 9: Ensemble Model Architecture.

Class-wise analysis revealed that U-Net generally outperformed DeepLabV3, with both models struggling with vegetation and water classes, and performing particularly poorly on Rangeland. Visual inspection also showed that even for classes with high metric values, the segmentation was often flawed, highlighting the limitations of relying solely on quantitative metrics.

5.5.2 MS-PAN Evaluation

For the MS-PAN models, a custom implementation of the *class-wise Dice coefficient* was used. This metric was designed to address scenarios where certain classes might not be present in a given image's ground truth. Instead of computing a global average, the class-wise Dice score was calculated for each batch by summing the intersection and union of pixels across all images for a particular class. This ensured that only classes with a presence in the ground truth were included in the final score, providing a more robust and accurate representation of the model's performance on the available data. The formula for the class-wise Dice score, where B is the batch size, and c is the class, is as follows:

$$\text{Dice}_c = \frac{2\sum_{b=1}^{B} |\text{Prediction}_{b,c} \cap \text{GroundTruth}_{b,c}|}{\sum_{b=1}^{B} (|\text{Prediction}_{b,c}| + |\text{GroundTruth}_{b,c}|)} \tag{9}$$

5.6 Environment

All model training and testing were conducted on an AWS SageMaker instance ml.g4dn.2xlarge, provided by Aurizn. This instance offers an accelerated computing environment with a single Tesla T4 GPU, providing 15360 MiB of GPU memory, 8 vCPUs, and 32 GiB of system memory. The environment was configured with NVIDIA driver version 550.163.01 and CUDA version 12.4 to support the computational demands of the deep learning models.

6 Experimental Results

This section presents the results from the various experiments conducted throughout the project, with a focus on the key findings and the insights gained from each experiment.

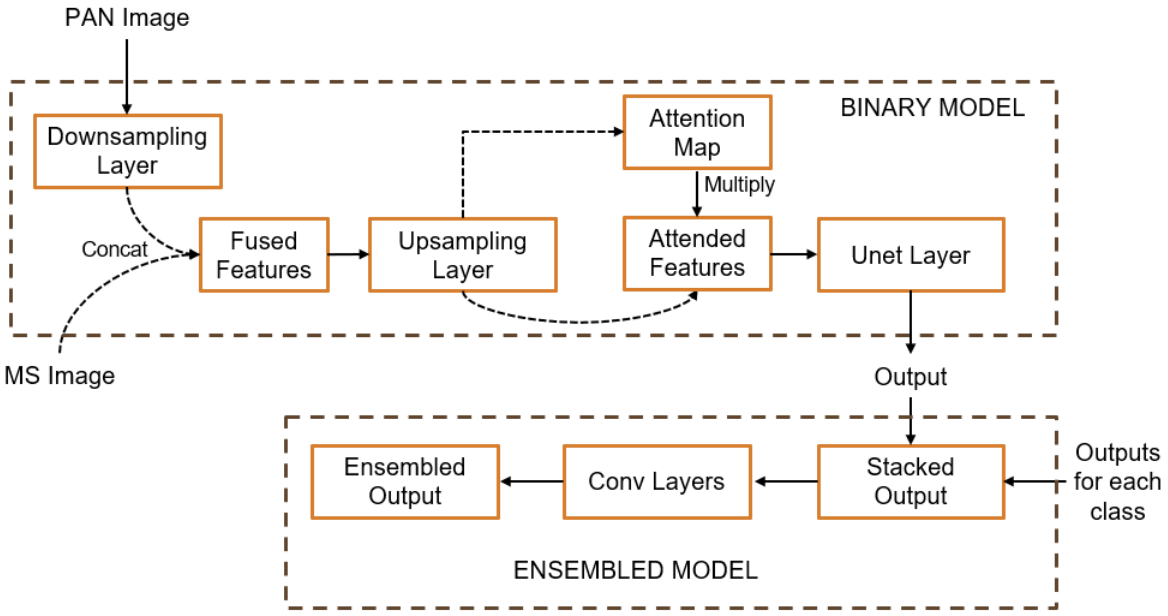


Figure 10: Ensemble Binary Model Architecture.

6.1 RGB Experiments

The initial phase of the project focused on establishing a performance baseline. Two popular segmentation models, U-Net and DeepLabV3, were trained with a ResNet-101 backbone to evaluate their effectiveness on RGB imagery.

6.1.1 Experimental Setup

For both the U-Net and DeepLabV3 models, an AdamW optimizer was used with a learning rate of $1e^{-4}$ and a weight decay of $1e^{-4}$. The models were trained for 20 epochs using the Combined Segmentation Loss with a 0.5 Dice to 0.5 Cross-Entropy ratio.

6.1.2 Result Discussion

As shown in the table below, U-Net generally outperformed DeepLabV3 across most metrics and classes, demonstrating its superior capability for this specific task. DeepLabV3 did show a stronger performance for specific classes such as Urban and Forest, but overall struggled with classes like Water and Rangeland, leading to a poorer overall result compared to U-Net. Both models faced challenges with vegetation (Forest) and water, and performed particularly poorly on Rangeland, indicating significant segmentation challenges for this class. Visual inspection, as illustrated in the figure 11, revealed that even for classes with high metric values, the segmentation was often flawed, emphasising the limitations of relying solely on quantitative metrics.

6.2 MS-PAN Experiments

This phase explored several multi-modal architectures to leverage the combined MS and PAN data. The experiments focused on the impact of different edge detection techniques and ensemble methods.

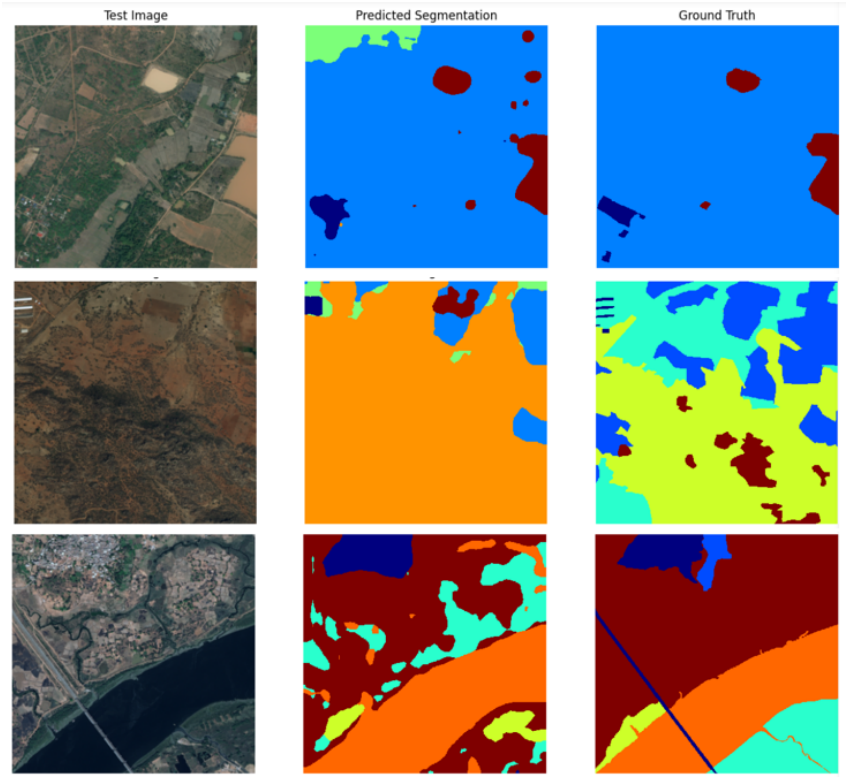


Figure 11: Segmentation results of the best U-Net model (ResNet-101 backbone) on the DeepGlobe test set. Colors in the masks represent classes: blue (Urban), cyan (Agriculture), green (Rangeland), yellow (Forest), orange (Water), red (Barren).

Model (ResNet-101)	Metric	Urban	Agriculture	Rangeland	Forest	Water	Barren
U-Net	Dice	0.7028	0.9367	0.1896	0.7233	0.8759	0.7257
	IoU	0.5418	0.8809	0.1047	0.5665	0.7792	0.5694
	Precision	0.6039	0.9403	0.2794	0.5739	0.8373	0.8310
	Recall	0.8406	0.9331	0.1435	0.9780	0.9181	0.6482
DeepLabV3	Dice	0.7748	0.8788	0.0645	0.7600	0.5762	0.5212
	IoU	0.6324	0.7838	0.0334	0.6130	0.4047	0.3525
	Precision	0.8056	0.8309	0.4274	0.8725	0.4550	0.4762
	Recall	0.7463	0.9325	0.0349	0.6733	0.7853	0.5756

Table 1: Class-wise metrics comparison of U-Net and DeepLabV3 (ResNet-101 backbone) on the DeepGlobe dataset.

6.2.1 Training Setup

All multi-class models were trained with a learning rate of $1e^{-4}$, a weight decay of $1e^{-4}$, and an AdamW optimizer. The scheduler for these models was CosineAnnealingWarmRestarts. The Combined Segmentation Loss with a 0.7 Dice to 0.3 Cross-Entropy ratio was used for all multi-class experiments.

Hybrid SegFormer-U-Net Two variants of the Hybrid SegFormer-U-Net model were trained to compare the effectiveness of using Canny edge detection versus Sobel edge detection as a feature channel. Both models were trained with an aggressive sampling strategy: an oversam-

pling factor of 20, an undersampling factor of 80, and an augmentation factor of 2. The batch size was 2, and the models were trained for 10 epochs. The scheduler parameters were $T_0 = 5$, $T_{mult} = 2$, and $\eta_{min} = 1e^{-6}$. This aggressive sampling strategy resulted in a training time of approximately 45 minutes per epoch.

Ensemble of Hybrid Models All three models within this ensemble approach were trained using a less aggressive sampling strategy, with an oversampling factor of 5, an undersampling factor of 5, and an augmentation factor of 2.2. For the two individual hybrid models, a batch size of 2 was used, and each was trained for 10 epochs.

The final ensemble model's refinement block was trained to intelligently combine the outputs of these two models. For this refinement block, the learning rate was $1e^{-3}$, and the weight decay was $1e^{-4}$. The scheduler had parameters of $T_0 = 3$, $T_{mult} = 2$, and $\eta_{min} = 1e^{-5}$. All three models were quick to train, with each epoch taking approximately 10 minutes.

Ensemble of Binary Classifiers This architecture involved training a separate binary model for each class. Each individual binary model was trained for 10 epochs with a batch size of 2 and class weights of $[0.5, 2.0]$. The optimizer was AdamW with a learning rate of $3e^{-4}$ and weight decay of $1e^{-4}$. A CosineAnnealingWarmRestarts scheduler with $T_0 = 5$, $T_{mult} = 2$, and $\eta_{min} = 1e^{-6}$ was used. For the final stacked ensemble refinement, the training was for 10 epochs with a batch size of 2. An AdamW optimizer with a learning rate of $1e^{-3}$ and a weight decay of $1e^{-4}$ was used for this step, along with a CosineAnnealingWarmRestarts scheduler with $T_0 = 3$, $T_{mult} = 2$, and $\eta_{min} = 1e^{-5}$.

6.2.2 Result Discussion

The performance of the different multi-modal architectures on the MS-PAN dataset is detailed in Table 2. My Canny Hybrid model achieved the highest overall Dice score of 0.5843, with strong performance on Road (0.6048), Vegetation (0.6109), and Water (0.7046) classes. Visual analysis using Figure 14 confirms that this model excels at predicting linear features like roads, accurately tracing edges and maintaining continuity. Vegetation is well-segmented with clear boundaries, and water areas are consistently identified. However, the model struggles with Building class predictions (0.4158 Dice), occasionally misclassifying edges as roads or over-segmenting small structures. The No Data class (0.5851 Dice) is well-predicted in open areas but can be confused with Vegetation in mixed land cover regions.

The Sobel Hybrid model, with an overall Dice score of 0.5394, performed adequately but fell short of the Canny model. Figure 14 shows it predicts Vegetation (0.5798 Dice) in broad patches and Water (0.6674 Dice) with decent accuracy. However, it fails significantly with Road (0.4075 Dice) and Building (0.4282 Dice), where the broader gradient edges lead to noisy predictions, over-segmenting green areas and missing fine building outlines. The No Data class (0.6139 Dice) is reasonably handled but shows inconsistencies near vegetation boundaries.

The Ensemble Hybrid model, with an overall Dice score of 0.4764, showed strengths in Water prediction (0.6005 Dice), as seen in Figure 13, where it accurately identifies large water bodies. It also predicted Vegetation (0.5185 Dice) reasonably well. However, it failed with Road (0.2734 Dice) and Building (0.3222 Dice). Figure 13 reveals excessive blue over-prediction in water areas, obscuring roads, and fragmented building segments that lack cohesion, suggesting inconsistencies in the fusion of the hybrid models' outputs.

The Ensemble Binary model, with the lowest overall Dice score of 0.3566, performed poorly across most classes. Figure 12 highlights its failure to predict Building (0.1996 Dice)

and Water (0.2786 Dice), with fragmented patches that poorly align with the ground truth. It predicted No Data (0.5193 Dice) and Vegetation (0.4889 Dice) with some success in sparse areas but failed with Road (0.2966 Dice), where linear features are broken or absent. This poor performance suggests that training a separate binary model for each class without knowledge of the other classes hinders the model's ability to learn and differentiate between features effectively.

Overall, the Canny Hybrid model was the most robust, leveraging Canny edge detection to enhance segmentation accuracy for roads, vegetation, and water. The Sobel Hybrid excelled in broad areas but struggled with fine details. The Ensemble Hybrid performed well for water but faltered with roads and buildings due to fusion inconsistencies, while the Ensemble Binary model's fragmented predictions indicate challenges across all classes.

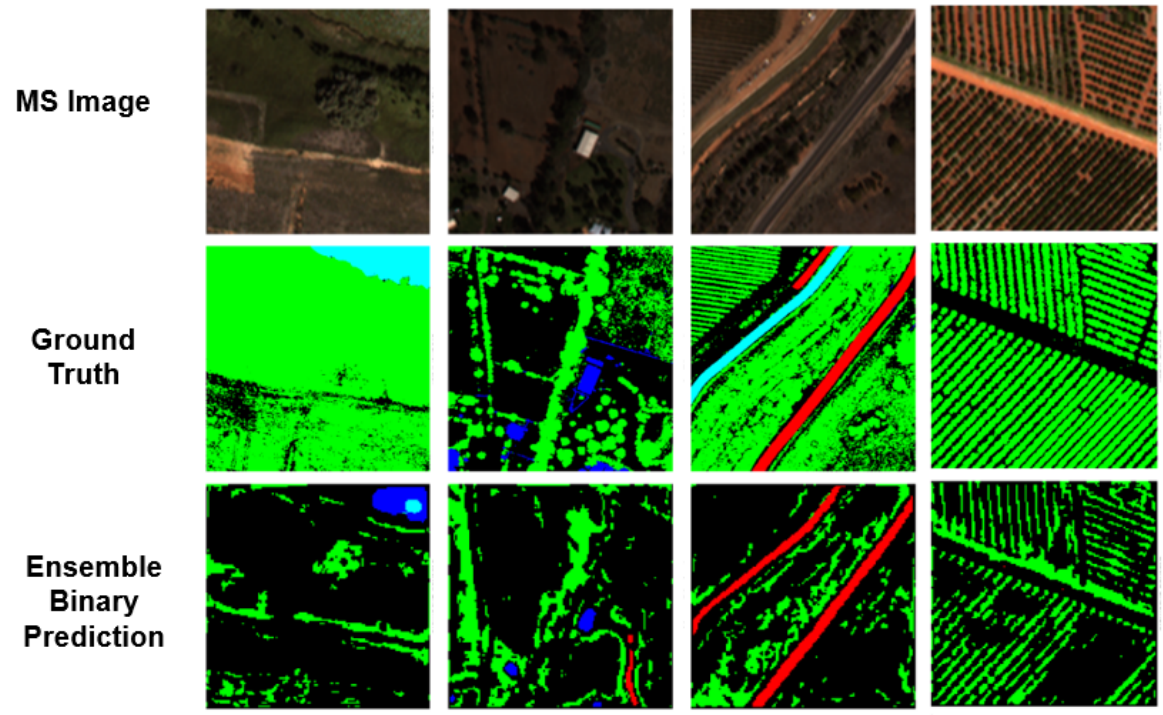


Figure 12: Visual comparison of MS Image, Ground Truth, and Ensemble Binary model predictions on the MS-PAN dataset.

Model	**Overall Dice**	**Nodata Dice**	**Road Dice**	**Building Dice**	**Vegetation Dice**	**Water Dice**
Sobel Hybrid	0.5394	0.6139	0.4075	0.4282	0.5798	0.6674
Canny Hybrid	**0.5843**	**0.5851**	**0.6048**	**0.4158**	**0.6109**	**0.7046**
Ensemble	0.4764	0.5272	0.2734	0.3222	0.5185	0.6005
Stacked Binary	0.3566	0.5193	0.2966	0.1996	0.4889	0.2786

Table 2: Class-wise Dice coefficient comparison of the multi-modal architectures on the MS-PAN dataset.

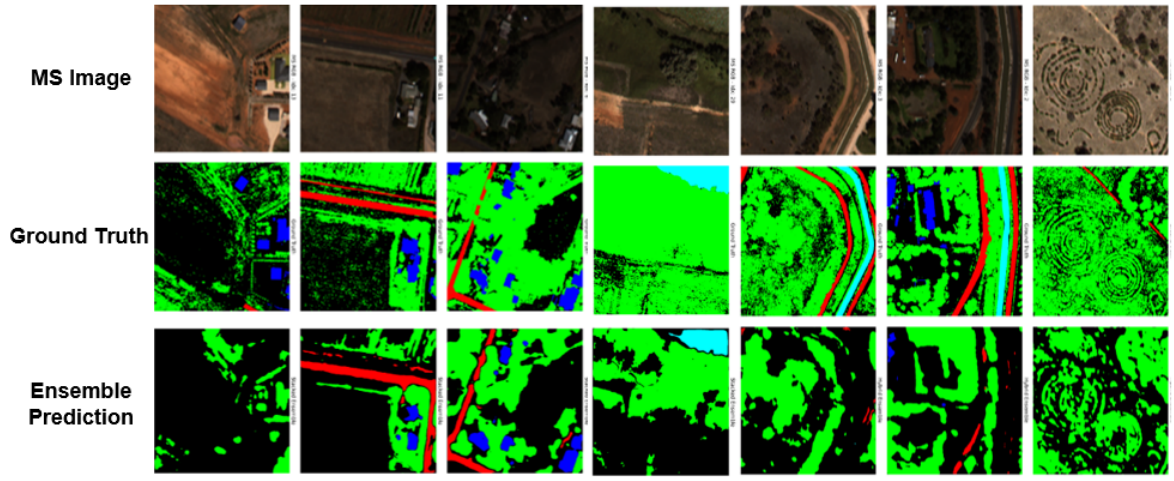


Figure 13: Visual comparison of MS Image, Ground Truth, and Ensemble Hybrid model predictions on the MS-PAN dataset.

6.2.3 Model Performance Across Varying Image Sizes

To evaluate the models' robustness to different image scales, they were tested on datasets with varying resolutions. The models were originally trained on images with a PAN size of 512×512 pixels. For this experiment, two new test datasets were created from the Griffith region tiles: a *larger scale* dataset with 1024×1024 PAN and 256×256 MS images, and a *smaller scale* dataset with 256×256 PAN and 64×64 MS images. The pan-to-MS resolution ratio of 4:1 was maintained for all scales.

The process for creating these datasets involved using the PAN image as a reference and then aligning the MS and ground truth label images to its coordinate reference system via reprojection. This technique ensures that the spatial relationship between the different image types is preserved. The images were then tiled into the desired resolutions, with a filtering process applied to discard any tiles that were edge cases or contained insufficient valid data.

The results showed that model performance varied significantly depending on the class and the scale of the input image, a consistent finding across all models tested. For example, graphical comparison of the Canny Hybrid model is shown in Figure 15. These figures illustrate that the models perform better with *larger-scale images* (1024×1024) for larger, more contiguous classes like *Vegetation* and *No Data*. This is likely because the larger context allows the model to better identify and segment these extensive regions. Conversely, the models perform better with *smaller-scale image*s (256×256) for the minority classes, specifically *Road*, *Water*, and *Building*. This suggests that when the input image is smaller, the model can focus more on the fine-grained details and edges that are characteristic of these smaller structures, leading to more accurate predictions. The visual comparison in Figure 16 further illustrates this, showing clearer roads and buildings in the smaller scale predictions.

6.2.4 Post-Processing Analysis

To further enhance the performance of the best-performing model, the Canny Hybrid, a series of morphological post-processing operations were tested. These techniques, including Dilation, Closing, Opening, and a sequence of all three, were applied to the raw segmentation masks to

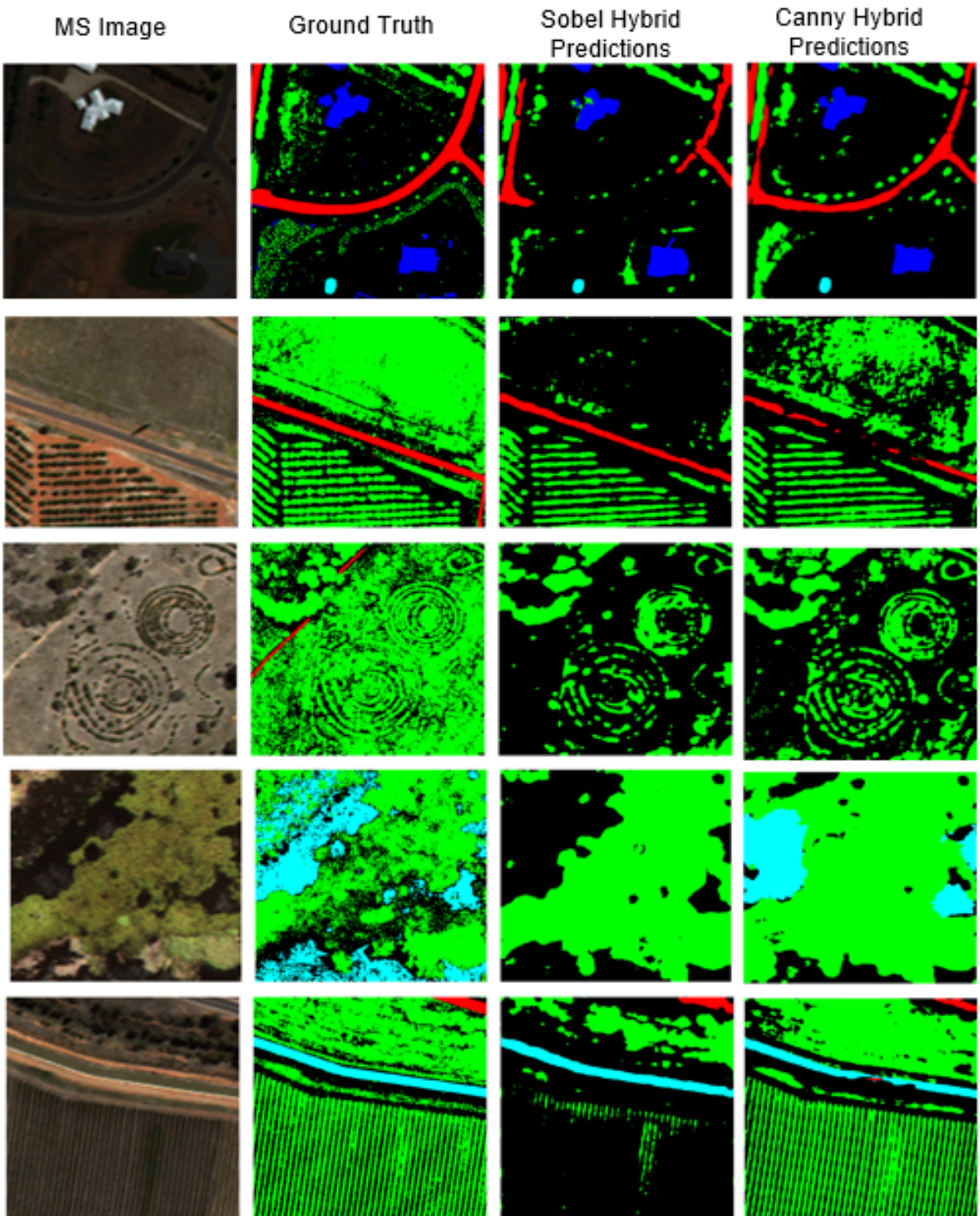


Figure 14: Visual comparison of MS Image, Ground Truth, Sobel Hybrid, and Canny Hybrid model predictions on the MS-PAN dataset.

refine the final predictions [42].

Morphological Operations

- **Dilation:** This operation adds pixels to the boundaries of objects in an image. Its primary purpose is to expand foreground regions, which can help connect fragmented segments

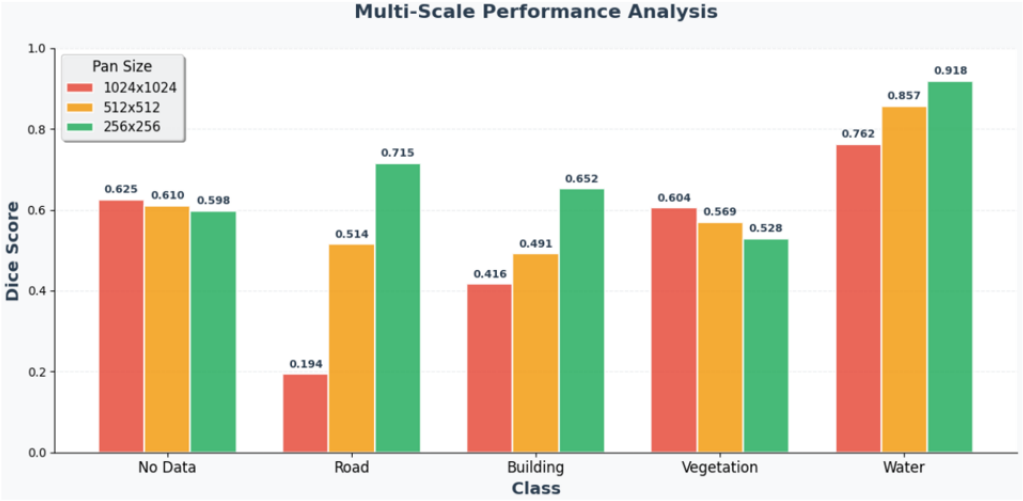
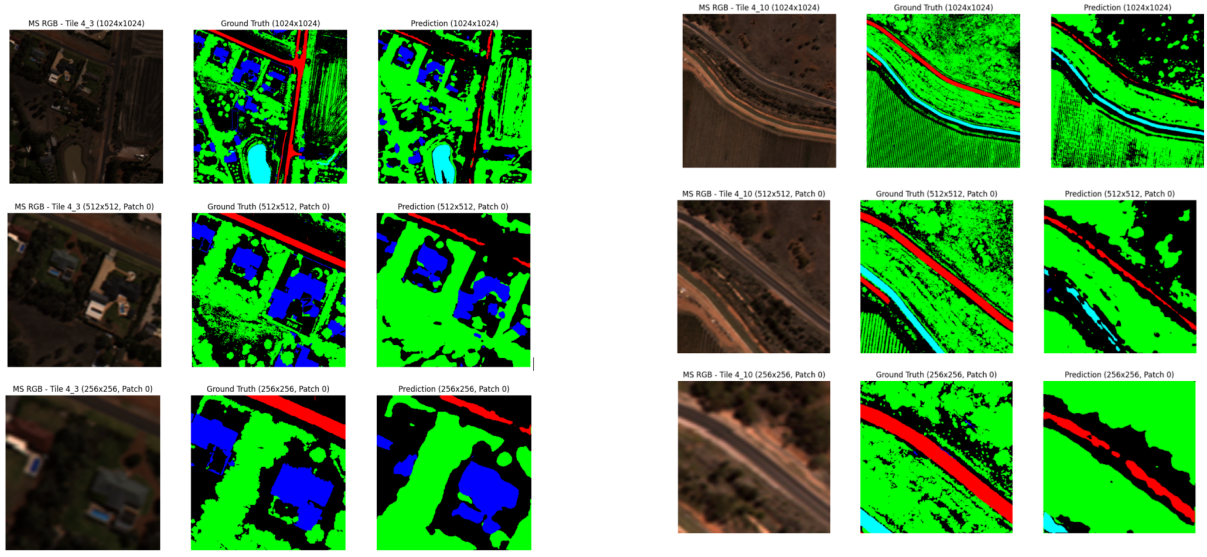


Figure 15: Multi-scale analysis showing how the Canny Hybrid model's Dice Score changes across three different image sizes for each class.

and fill in small holes. The effect is to make the objects appear larger and more contiguous.

- **Closing:** A combination of dilation followed by erosion. It is effective at closing small gaps and holes within a foreground object while preserving its overall shape.
- **Opening:** A combination of erosion followed by dilation. It is used to remove small, isolated regions or spurious noise from the foreground while generally preserving the size and shape of larger objects.
- **Sequence:** This is a sequential application of Dilation, followed by Closing, and then Opening. This combines the benefits of each operation: Dilation expands the segments, Closing fills any small internal gaps, and Opening removes any small, isolated noise that may have been created or exaggerated by the previous steps.

Result Discussion The performance of the Canny Hybrid model with and without morphological post-processing is illustrated in the visualisation (see Figure 17) and in the performance graph (see Figure 18). The raw model excels at predicting linear features like roads and large-area features like water bodies, with moderate success in segmenting vegetation and buildings, though it faces challenges in distinguishing no data or ground areas from surrounding regions. Applying Dilation improves the model's overall performance, particularly enhancing the connectivity of road segments and the continuity of vegetation patches, as seen in the visualisation where fragmented areas are better linked. Water boundaries also benefit from slight expansion, but this comes at the cost of reduced accuracy in no data or ground areas, where over-expansion into adjacent classes is noticeable. Building segmentation sees a minor improvement, though small structures can appear over-segmented. The Closing operation provides a modest overall improvement, effectively filling small gaps in road networks and water bodies, which helps maintain their shape, as observed in the visualisation. However, it shows limited impact on vegetation and building predictions, where retained noise slightly affects accuracy, and the distinction of no data or ground areas remains difficult due to preserved imperfections. Opening



(a) (b)

Figure 16: Comparison of Canny Hybrid model predictions, where (a) and (b) show different images at different scales.

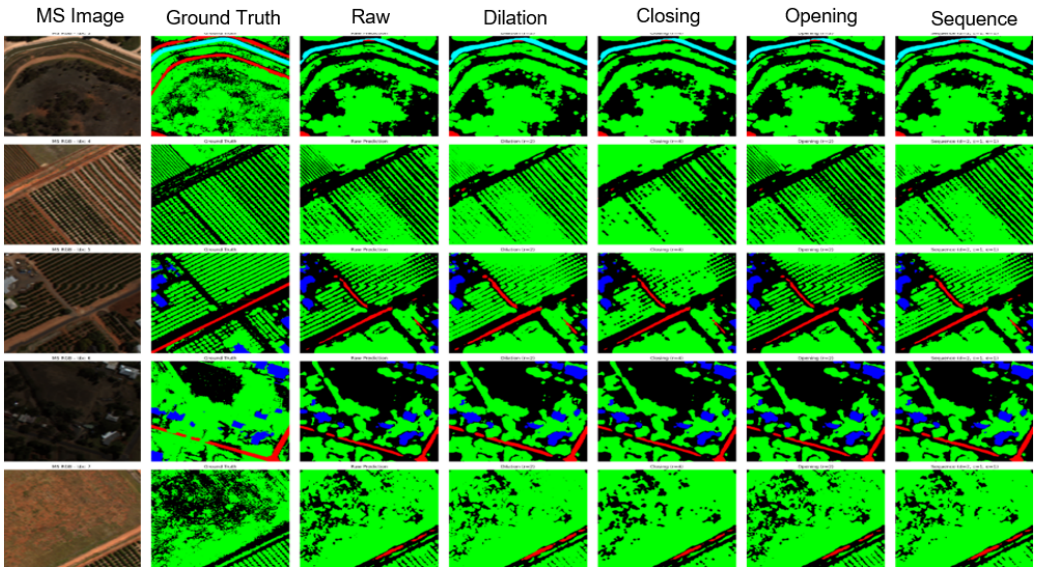


Figure 17: Performance comparison of the Canny Hybrid model various morphological post-processing methods.

performs similarly to the raw model, offering slight gains in identifying no data or ground areas and refining water edges by removing noise, as depicted in the visualisation. However, it weakens the prediction of roads, vegetation, and buildings, as the erosion step erodes fine details, leading to loss of road continuity and building outlines. The Sequence approach, which applies Dilation, Closing, and Opening in order, improves vegetation and road segmentation by connecting segments and filling gaps, while water prediction remains stable, as shown in the visualisation. Building segmentation sees a slight uplift, but the no data or ground class suffers from over-processing, where noise reappears after the initial expansion and refinement steps. Overall, Dilation stands out as the most effective post-processing method, enhancing road and

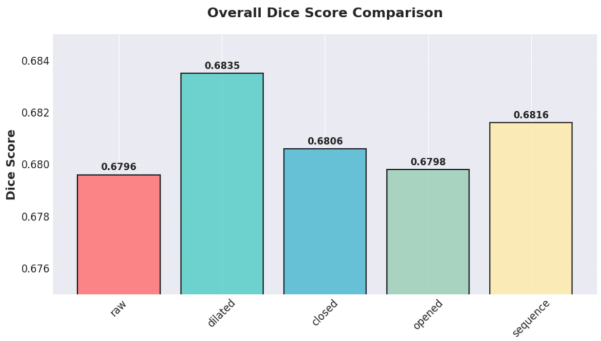


Figure 18: Morphological post-processing method comparison

vegetation segmentation by connecting fragmented regions, though it introduces some over-expansion in no data or ground areas. The visualisation confirms that while these techniques refine certain features, they also present trade-offs, with no single method universally improving all class predictions.

7 Conclusion

This project successfully developed and validated a multi-modal approach for land cover segmentation, marking a significant improvement over standard RGB-based methods. As detailed in Section 5.3.2, the Hybrid SegFormer-U-Net architecture with Canny edge detection was the robust solution developed. The superior performance of the Canny edge detection variant, as evidenced by the results in Section 6.2.2, confirms its effectiveness. This success stems from its ability to effectively fuse fine-grained spatial and rich spectral information, and the inclusion of engineered features like NDRE and NDBI, which proved crucial for boosting segmentation accuracy (see Section 5.2.2).

A key finding of this research, which can be observed in Figure 15, was the class-dependent nature of optimal image scale. It was discovered that larger images are more effective for segmenting broad, continuous classes such as vegetation, while smaller-scale images are better suited for fine features like roads and buildings. The investigation into data alignment also showed that learnable convolutional layers offer a superior method for preserving detail compared to simple interpolation, especially for fixed scale factors (see Section5.3.2).

In essence, this project highlights the importance of a holistic strategy for land cover segmentation. The Canny model's superior performance, combined with the evidence for a multi-scale inference approach and the clear benefits of tailored spectral indices, provides a strong framework for future research. This comprehensive methodology, which balances architectural innovation with data-driven insights, is essential for achieving reliable and accurate segmentation in real-world applications.

8 Technical Challenges

The project faced several key challenges that significantly impacted model development and performance. These issues highlight the complexities of working with real-world geospatial data.

8.1 Outdated and Inconsistent Datasets

Both the RGB and MS datasets used had significant limitations due to their age. The MS dataset was from 2016, and the RGB dataset was similarly old, which meant the land cover information was outdated and not reflective of recent changes. The original labeling in the MS dataset was based on LiDAR data, which, while useful for elevation, led to inconsistencies in land cover segmentation. This was compounded by manual updates that still lacked clear definitions. For example, both grass and vegetation were merged into a single ”vegetation” class, creating ambiguity.

8.2 Ambiguous Class Definitions

A major challenge was the lack of clear and consistent class definitions, which resulted in labeling errors. For the ”road” class, a key issue was the inconsistent segmentation of dirt roads. In some instances, they were correctly labeled as roads, but in others, they were categorised as ”no road.” This ambiguity in the ground truth masks made it difficult for the models to learn a robust and generalisable definition of a road.

8.3 Data Imbalance

A major challenge was the unequal class representation, particularly within the MS-PAN dataset. This led to biased training, where the model performed well on majority classes but poorly on minority ones. To address this, a combination of oversampling and undersampling was employed to create a more balanced training set.

8.4 MS Warping Issues

An extensive analysis was conducted to investigate warping issues in the MS-PAN dataset. These issues, which affected the alignment between the MS images and their corresponding ground truth masks, were found to be present only in a few images within the validation dataset. This had a minor impact on the evaluation metrics but highlighted the need for careful data curation.

8.5 Memory Management

To effectively utilise the limited memory of the AWS SageMaker instance, several memory management strategies were implemented. These included reducing the batch size to as low as 2, regularly clearing the GPU cache, and utilising mixed precision training to significantly reduce the memory footprint of the models without a notable loss in performance.

9 Ethical Considerations

The development and deployment of land cover segmentation models, especially those intended for real-world applications come with significant ethical considerations.

9.1 Geographic Bias

As discussed in Section 8.1, the MS-PAN dataset used in this study was collected from only four Australian regions (Cowra, Young, Goulburn, and Griffith) in 2016. This limited geographic and temporal coverage means the model is tuned to specific Australian land cover types, climate conditions, and urban layouts. Consequently, it may generalise poorly to regions with different environmental or architectural characteristics, for example, tropical rainforests in Southeast Asia, arid landscapes in Africa, or medieval European cities.

Such geographic bias can lead to systematic misclassifications when applied to different geographies, which could impact decision making in environmental management or disaster recovery. As proposed in Section 10.1, diversifying datasets to include varied climatic zones, vegetation types, and urban morphologies is critical for improving fairness and generalisability.

To mitigate the risks, transparent documentation of dataset composition, collection dates, and known limitations should be provided, following best practices such as "Datasheets for Datasets" [43].

9.2 Dual-Use Technology

Land cover segmentation technology can be a dual-use technology, meaning it can be used for both beneficial and harmful purposes. While it can aid in environmental conservation, sustainable development, and post-disaster recovery, it can also be used for surveillance, resource exploitation, or military applications. As developers, it is crucial to be aware of these dual-use implications and to promote the technology's use for ethical and socially beneficial purposes [44].

9.3 Transparency and Accountability

The lack of interpretability and explainability in many deep learning models makes it difficult to understand the rationale behind a given segmentation. In high-stakes applications, such as identifying safe evacuation routes or assessing environmental damage, this lack of transparency could lead to a loss of trust. This problem is made more complex by the difficulty of determining where responsibility lies for a model's errors whether with the AI developers who created the algorithm, the conservators or analysts who deployed and interpreted its output, or the algorithm itself. This challenge in assigning accountability poses significant legal and ethical risks. As proposed in Section 10.2,integrating Explainable AI (XAI) techniques can improve trust and aid in debugging model behaviour [45].

9.4 Data Privacy and Indigenous Land Rights

The use of high-resolution satellite imagery raises significant concerns regarding data privacy and consent. While this data does not typically contain personally identifiable information, it can be used to infer sensitive details about private property and daily routines [46]. This blurs the line between public observation and private surveillance, making it essential to govern the collection and use of such data with clear privacy and data protection policies.

A critical ethical consideration, particularly in Australia, is the potential impact on Indigenous land rights and cultural heritage. AI models used for land mapping can influence decisions about resource management and land use, which have profound implications for Indigenous

communities. These models may fail to recognize traditional land management practices, potentially leading to misclassifications that undermine Indigenous sovereignty and cultural practices [47]. It is therefore crucial to engage with Indigenous communities and ensure that the technology supports, rather than hinders, their rights to self-determination and cultural preservation.

10 Future Work

Building upon the insights and challenges identified in this project, several avenues for future work could significantly improve performance and address current limitations.

10.1 Data and Labelling Enhancement

Future research should prioritise expanding the dataset with diverse geographic and environmental data to improve generalisation beyond Australian contexts. This includes collecting samples from tropical, arid, and urbanized regions worldwide. Additionally, adopting automated or semi-automated labelling tools with human oversight could reduce errors and ensure consistent mask boundaries, enhancing training accuracy and reliability.

10.2 Advanced Models and Interpretability

Exploring advanced architectures and improving model transparency can unlock new capabilities. Graph Neural Networks (GNNs), for instance, could model the spatial relationships between land cover classes, which would help improve contextual accuracy by distinguishing, for example, roads in forests from those in cities. The use of VLMs could enable segmentation based on textual descriptions, supporting more complex tasks like identifying specific vegetation types. Further research could also test specialised advanced architectures like the SAM to assess its effectiveness on multi-modal land cover data and its ability to generate high-quality segmentation masks with minimal prompting. Finally, integrating XAI techniques, such as feature importance heatmaps, could reveal how a model prioritises its inputs. This would not only enhance the model's trustworthiness but also provide valuable insights for debugging and improving performance, especially in high-stakes applications.

10.3 Multi-Scale Refinement

The scale-dependent performance observed suggests developing adaptive models that dynamically adjust resolution based on class characteristics. A multi-stage approach, where models specialise in different scales and class types, could optimise segmentation across varied landscapes.

10.4 Advanced Fusion Architectures

Further investigation into more sophisticated fusion techniques is warranted. This could involve using cross-attention mechanisms between the MS and PAN streams, or developing a more generalised fusion block that can handle variable resolution ratios without being explicitly trained on them.

References

[1] J. Cheng, C. Deng, Y. Su, Z. An, and Q. Wang, "Methods and datasets on semantic segmentation for unmanned aerial vehicle remote sensing images: A review," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 211, pp. 1–34, 2024.

[2] T. Arulananth, P. Kuppusamy, R. K. Ayyasamy, S. M. Alhashmi, M. Mahalakshmi, K. Vasanth, and P. Chinnasamy, "Semantic segmentation of urban environments: Leveraging u-net deep learning model for cityscape image analysis," *Plos one*, vol. 19, no. 4, p. e0300767, 2024.

[3] A. Alzu'bi and L. Alsmadi, "Monitoring deforestation in jordan using deep semantic segmentation with satellite imagery," *Ecological Informatics*, vol. 70, p. 101745, 2022.

[4] M. Rahnemoonfar, T. Chowdhury, and R. Murphy, "Rescuenet: A high resolution uav semantic segmentation dataset for natural disaster damage assessment," *Scientific data*, vol. 10, no. 1, p. 913, 2023.

[5] Z. Cai, Q. Hu, X. Zhang, J. Yang, H. Wei, Z. He, Q. Song, C. Wang, G. Yin, and B. Xu, "An adaptive image segmentation method with automatic selection of optimal scale for extracting cropland parcels in smallholder farming systems," *Remote Sensing*, vol. 14, no. 13, p. 3067, 2022.

[6] G. R. Kasaragod, S. Trivedi, K. Ramesh, S. R. Subramoniam, H. Ravishankar, and V. A., "Assessment of vegetation cover of bengaluru city, india, using geospatial techniques," *Journal of the Indian Society of Remote Sensing*, vol. 49, 11 2020.

[7] L. Ramos and A. D. Sappa, "Multispectral semantic segmentation for land cover classification: An overview," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.

[8] K. Zhang, F. Zhang, W. Wan, H. Yu, J. Sun, J. Del Ser, E. Elyan, and A. Hussain, "Panchromatic and multispectral image fusion for remote sensing and earth observation: Concepts, taxonomy, literature review, evaluation methodologies and challenges ahead," *Information Fusion*, vol. 93, pp. 227–242, 2023.

[9] A. J. Davies, "Semantic segmentation of aerial imagery using u-net in python," Jan 2025.

[10] Y. Mo, Y. Wu, X. Yang, F. Liu, and Y. Liao, "Review the state-of-the-art technologies of semantic segmentation based on deep learning," *Neurocomputing*, vol. 493, pp. 626–646, 2022.

[11] J. Lv, Q. Shen, M. Lv, Y. Li, L. Shi, and P. Zhang, "Deep learning-based semantic segmentation of remote sensing images: a review," *Frontiers in Ecology and Evolution*, vol. 11, p. 1201125, 2023.

[12] X. Wang, Z. Hu, S. Shi, M. Hou, L. Xu, and X. Zhang, "A deep learning method for optimizing semantic segmentation accuracy of remote sensing images based on improved unet," *Scientific reports*, vol. 13, no. 1, p. 7600, 2023.

[13] H. Peng, C. Xue, Y. Shao, K. Chen, J. Xiong, Z. Xie, and L. Zhang, "Semantic segmentation of litchi branches using deeplabv3+ model," *Ieee Access*, vol. 8, pp. 164546–164555, 2020.

[14] S. Seong and J. Choi, "Semantic segmentation of urban buildings using a high-resolution network (hrnet) with channel and spatial attention gates," *Remote Sensing*, vol. 13, no. 16, p. 3087, 2021.

[15] J. Bai, C. Jia, S. Yu, L. Sun, L. Zhang, Z. Chang, and A. Hou, "Building extraction from high-resolution remote sensing images using improved hrnet method," in *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*, pp. 7982–7985, IEEE, 2024.

[16] F. Fogel, Y. Perron, N. Besic, L. Saint-André, A. Pellissier-Tanon, M. Schwartz, T. Boudras, I. Fayad, A. d'Aspremont, L. Landrieu, *et al.*, "Open-canopy: A country-scale benchmark for canopy height estimation at very high resolution," *arXiv preprint arXiv:2407.09392*, 2024.

[17] Y. Zhang, M. Huang, Y. Chen, X. Xiao, and H. Li, "Land cover classification in high-resolution remote sensing: using swin transformer deep learning with texture features," *Journal of Spatial Science*, pp. 1–25, 2024.

[18] X. Ma, Q. Wu, X. Zhao, X. Zhang, M.-O. Pun, and B. Huang, "Sam-assisted remote sensing imagery semantic segmentation with object and boundary constraints," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[19] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *CoRR*, vol. abs/2103.14030, 2021.

[20] S. Koundinya, H. Sharma, M. Sharma, A. Upadhyay, R. Manekar, R. Mukhopadhyay, A. Karmakar, and S. Chaudhury, "2d-3d cnn based architectures for spectral reconstruction from rgb images," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 844–851, 2018.

[21] N. Hikmah and P. Manurung, "Application of spectral indices and deep learning (convolutional neural network model) on land cover change analysis," *Applied Environmental Science*, vol. 3, no. 1, 2025.

[22] T. Wang, F. Fang, F. Li, and G. Zhang, "High-quality bayesian pansharpening," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 227–239, 2018.

[23] J. Kaur, "Revolutionizing pan sharpening in remote sensing with cutting-edge deep learning optimization," in *2024 Second International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI)*, pp. 1357–1362, IEEE, 2024.

[24] Y. Ma, S. Chen, S. Ermon, and D. B. Lobell, "Transfer learning in environmental remote sensing," *Remote Sensing of Environment*, vol. 301, p. 113924, 2024.

[25] D. Yan, H. Zhang, G. Li, X. Li, H. Lei, K. Lu, L. Zhang, and F. Zhu, "Improved method to detect the tailings ponds from multispectral remote sensing images based on faster r-cnn and transfer learning," *Remote Sensing*, vol. 14, no. 1, p. 103, 2021.

[26] J. Ouyang, P. Jin, and Q. Wang, "Multimodal feature-guided pre-training for rgb-t perception," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.

[27] A. Pendota and S. S. Channappayya, "Are deep learning models pre-trained on rgb data good enough for rgb-thermal image retrieval?," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4287–4296, 2024.

[28] M. Noman, M. Naseer, H. Cholakkal, R. M. Anwer, S. Khan, and F. S. Khan, "Rethinking transformers pre-training for multi-spectral satellite imagery," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27811–27819, 2024.

[29] H. Ghandorh, W. Boulila, S. Masood, A. Koubaa, F. Ahmed, and J. Ahmad, "Semantic segmentation and edge detection—approach to road detection in very high resolution satellite images," *Remote Sensing*, vol. 14, no. 3, p. 613, 2022.

[30] I. Sobel, G. Feldman, *et al.*, "A 3x3 isotropic gradient operator for image processing," *a talk at the Stanford Artificial Project in*, vol. 1968, pp. 271–272, 1968.

[31] J. Canny, "A computational approach to edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-8, no. 6, pp. 679–698, 1986.

[32] S. Xu, W. Xiao, L. Ruan, W. Chen, and J. Du, "Assessment of ensemble learning for object-based land cover mapping using multi-temporal sentinel-1/2 images," *Geocarto International*, vol. 38, no. 1, p. 2195832, 2023.

[33] R. Li, X. Gao, F. Shi, and H. Zhang, "Scale effect of land cover classification from multi-resolution satellite remote sensing data," *Sensors*, vol. 23, no. 13, p. 6136, 2023.

[34] H. Shinde, "Semantic segmentation of satellite imagery," Feb 2024.

[35] J. Wang, Z. Zheng, A. Ma, X. Lu, and Y. Zhong, "LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation," Oct. 2021.

[36] B. Ashwath, "Deepglobe land cover classification dataset," Nov 2020.

[37] V. Henrich, G. Krauss, C. Götze, and C. Sandow, "IDB - www.indexdatabase.de, entwicklung einer datenbank für fernerkundungsindizes," 2012. Talk given at AK Fernerkundung, Bochum, 4-5 October 2012.

[38] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *CoRR*, vol. abs/2105.15203, 2021.

[39] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*, pp. 565–571, Ieee, 2016.

[40] 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*, pp. 234–241, Springer, 2015.

[41] L.-C. Chen, G. Papandreou, F. Schroff, and K. Murphy, "Rethinking atrous convolution for semantic image segmentation," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 506–514, IEEE, 2017.

[42] R. Mondal, M. S. Dey, and B. Chanda, "Image restoration by learning morphological opening-closing network," *Mathematical Morphology-Theory and Applications*, vol. 4, no. 1, pp. 87–107, 2020.

[43] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii, and K. Crawford, "Datasheets for datasets," *Communications of the ACM*, vol. 64, no. 12, pp. 86–92, 2021.

[44] M. Brundage, S. Avin, J. Clark, H. Toner, P. Eckersley, B. Garfinkel, A. Dafoe, P. Scharre, T. Zeitzoff, B. Filar, *et al.*, "The malicious use of artificial intelligence: Forecasting, prevention, and mitigation," *arXiv preprint arXiv:1802.07228*, 2018.

[45] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[46] K. Crawford, *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press, 2021.

[47] L. Tuhiwai Smith, *Decolonizing methodologies: Research and indigenous peoples*. Zed books, 2012.