AUTOMATED LANDCOVER SEGMENTATION

Sakshi Sinha
Department of Computer Science
University of Adelaide
sakshi.sinha@student.adelaide.edu.au

ABSTRACT

This report presents mid-term progress on developing deep learning models for semantic segmentation of Aurizn's multispectral satellite imagery, tailored to Australian landscapes. Phase 1 established baselines using RGB data, evaluating U-Net and DeepLabV3 (ResNet-101) on the DeepGlobe dataset, revealing class-specific performance and segmentation challenges. Phase 2 is underway, focusing on multispectral adaptation, exploring various techniques, and addressing encountered challenges. Future work includes Vision-Language Model exploration and model refinement.

1 INTRODUCTION

1.1 MOTIVATION

The increasing availability of detailed satellite imagery has significantly advanced remote sensing applications. Precise segmentation of these imagery is essential for various domains, including urban planning, environmental monitoring, disaster response, and resource management [1]. For example, accurate identification of land cover types (vegetation, buildings, roads, water), as illustrated in Figure 1, facilitates improved urban development strategies [2], monitoring deforestation and ecosystem health [3], assessing post-disaster damage [4], and optimising agricultural practices [5]. Inaccurate segmentation can lead to flawed analyses and ineffective decision-making in these crucial areas.

Moreover, the rich information provided by accurate segmentation enables more refined insights and a deeper understanding of the observed environment. Hence, accurate satellite image segmentation is crucial, yet open-source research tailored to Australia's unique landscapes is lacking. While private studies may exist, Aurizn requires this project to develop accessible, context-specific models.

While deep learning (DL) advancements in image segmentation have been substantial, unique challenges arise with specialised datasets. Aurizn's high-resolution multispectral dataset, acquired over Australia, differs from much of the existing research, which primarily focuses on RGB imagery [6]. Multispectral data offers richer information beyond the visible spectrum, allowing for better differentiation of features often indistinguishable in RGB images [7]. However, to effectively utilize this data, specialised techniques are necessary to capture the complex relationships between spectral bands [8].

1.2 RESEARCH FOCUS

This research focuses on how to optimise deep learning (DL) models for accurate semantic segmentation of Aurizn's high-resolution panchromatic satellite imagery, with a focus on key land cover classes (roads, buildings, vegetation, water).

1.3 PROJECT AIMS

This project will focus on addressing Aurizn's need for precise and efficient land cover segmentation in its high-resolution multispectral imagery. By developing and implementing tailored deep learning solutions, the project aims to empower Aurizn and its stakeholders to gain actionable insights, facilitating smarter, data-driven decisions throughout its operations. This will promote more sustainable and efficient practices, fully leveraging the rich information contained in its multispectral data.

Figure 1: Example of land cover segmentation in satellite imagery. Image sourced from [9].

2 LITERATURE REVIEW

2.1 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, which Aurizn faces, must be addressed to enhance image analysis and decision-making [6]. This project aims to tackle these challenges by exploring and adapting advanced DL models to Aurizn's specific multispectral dataset.

2.2 Deep Learning for Semantic Segmentation

2.2.1 Convolutional Neural Networks

CNN architectures like U-Net [11], DeepLabv3+ [12], and HRNet [13, 14] have shown promise 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, necessitate further investigation [15]. A core objective of this project is to determine the effectiveness of CNNs, particularly pre-trained CNNs, on Aurizn's datset and compare their performance against transformer-based models. This comparison will directly inform the selection of the most suitable architecture for Aurizn's operational needs.

2.2.2 Transformer-based Architectures

Transformers, including ViTs and Swin Transformers, offer improved long-range dependency modeling through self-attention [16]. The emergence of models like SAM, with its zero-shot and few-shot capabilities, is particularly relevant for Aurizn's dataset, where labeled data may be limited [17]. This project will explore and adapt SAM and other transformer-based models to Aurizn's multispectral imagery, assessing their ability to generalise and perform effectively in this specialised domain.

2.3 MULTISPECTRAL IMAGE SEGMENTATION IN REMOTE SENSING

Multispectral satellite imagery captures reflected energy beyond the visible spectrum, as illustrated in Figure 2, provides crucial information for detailed land cover analysis [7]. While offering richer data for improved classification compared to RGB, it also presents unique challenges. The high dimensionality of multispectral data increases computational cost and can hinder model performance. Varying spectral resolutions and bandwidths across sensors require careful preprocessing, including atmospheric correction and radiometric calibration [8, 6].

This project will focus on effectively utilising spectral information, addressing the complexities of multispectral data through techniques like spectral attention, multi-branch networks, and pan-sharpening [18].

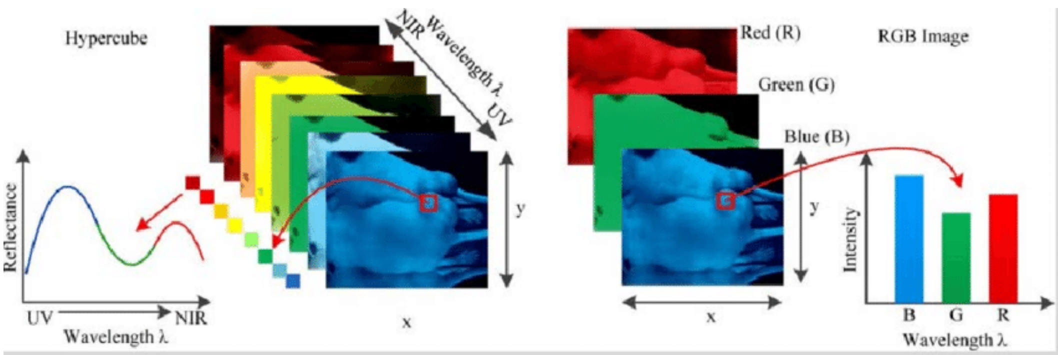

Figure 2: Difference between RGB and multispectral imaging. Image sourced from [19].

2.4 TRANSFER LEARNING AND FINE-TUNING

2.4.1 BENEFITS OF TRANSFER LEARNING

Transfer learning has been instrumental in remote sensing applications, allowing models pre-trained on large datasets to be adapted to specific tasks with limited labeled data. This is particularly beneficial for segmentation, where labeled data acquisition is expensive and time-consuming [20].

2.4.2 FINE-TUNING PRETRAINED RGB MODELS ON MULTISPECTRAL IMAGES

Adapting RGB-trained models to multispectral data requires careful adjustments, including band-wise feature fusion and spectral attention [21, 22, 23]. A key objective is to investigate the effectiveness of fine-tuning pre-trained RGB models on Aurizn's multispectral dataset, comparing their performance with baseline HRNet model to find the most efficient approach.

2.5 VISION-LANGUAGE MODELS (VLMS) FOR REMOTE SENSING

VLMs like GeoChat offer new possibilities for remote sensing through natural language interaction [24, 25]. As an exploratory objective, this project will explore the application of VLMs to Aurizn's data for tasks like temporal understanding, object counting, and detailed captioning. This will demonstrate the potential of VLMs to Aurizn and its clients, showcasing their ability to extract complex information from satellite imagery.

3 RESEARCH PLAN AND PROGRESS

The project design includes 3 phases, with the planned timeline illustrated in Figure 3. As of the mid-term review, Phase 1 has been completed, and Phase 2 is currently in progress. The following section will detail the methodology and experiments conducted in each phase, starting with the baseline establishment in Phase 1, followed by the adaptation to multispectral data in Phase 2, and concluding with the planned exploration of Vision-Language Models (VLMs) in Phase 3.

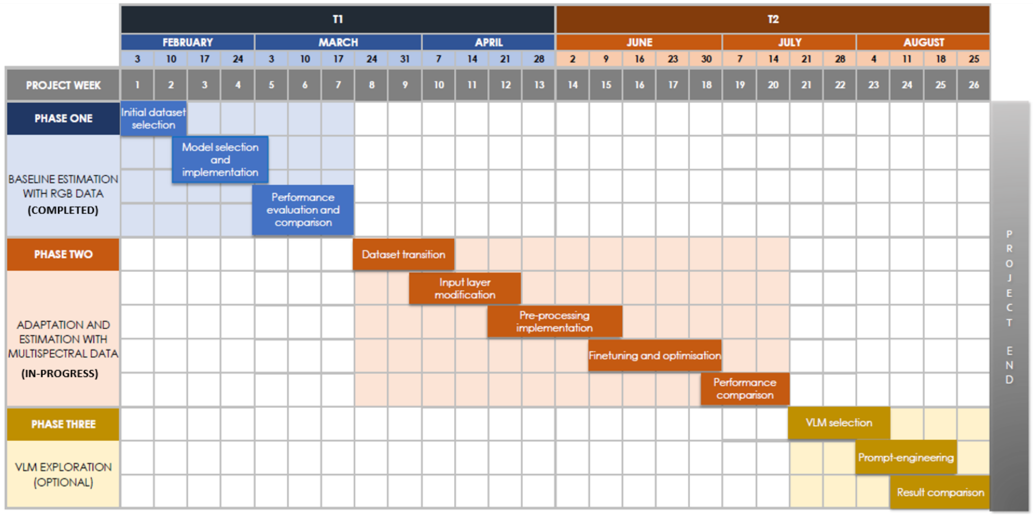

Figure 3: Project Timeline.

4 Methodology and Experiments

4.1 Phase 1: Baseline Establishment with RGB Data [Completed]

The initial phase aimed to establish a performance baseline using RGB datasets to evaluate model performance and identify challenges. The following steps were undertaken:

- **Dataset Exploration and Selection:** Several RGB datasets were explored for initial testing:
  - Dataset 1 (Kaggle - DataMes): A small dataset with 72 images grouped into 6 larger tiles. The classes included buildings, land, roads, vegetation, water, and unlabeled areas. However, the dataset was too small for training large models and fine-tuning pre-trained models, limiting generalisation capabilities [26].
  - Dataset 2 (Kaggle - DeepGlobe Land Cover Classification): This dataset offered 803 high-resolution images with 7 classes. Despite challenges with masks lacking clear boundaries, which initially resulted in suboptimal model performance, it was selected for model training due to its larger size and high-resolution imagery, providing a reasonable compromise for initial experimentation [27].
  - Dataset 3 (LoveDA): This dataset provided a larger and higher-resolution dataset with 5987 images and 7 classes. However, issues such as inconsistent pixel values across images, memory constraints, and the lack of Australian-specific imagery were encountered, making it less suitable for the initial phase [28].
- **Model Selection and Implementation:** Two models, DeepLabV3 and U-Net, were selected for initial experimentation. Both models were tested with various ResNet backbones.
  - **Evaluation Metrics:** The models were evaluated using standard segmentation metrics: Dice coefficient, Intersection over Union (IoU), overall accuracy, precision, and recall.
  - **Class-wise Performance:** Table 1 presents the class-wise metrics for U-Net and DeepLabV3 (ResNet-101 backbone) on the DeepGlobe test set. U-Net outperformed DeepLabV3 in most classes, particularly for Water and Agriculture, but DeepLabV3 showed better performance for Urban and Forest.
  - **Challenges with Vegetation and Water:** Both models struggled with vegetation (Forest) and Water compared to other classes, with DeepLabV3 facing greater challenges for Water. Additionally, both models performed poorly on Rangeland, indicating significant segmentation challenges for this class.

- **Visual Inspection:** visual inspection revealed that even classes with high metric values failed to segment properly, as shown in Figure 4 for the best U-Net model performance, emphasising the limitations of relying solely on quantitative metrics.

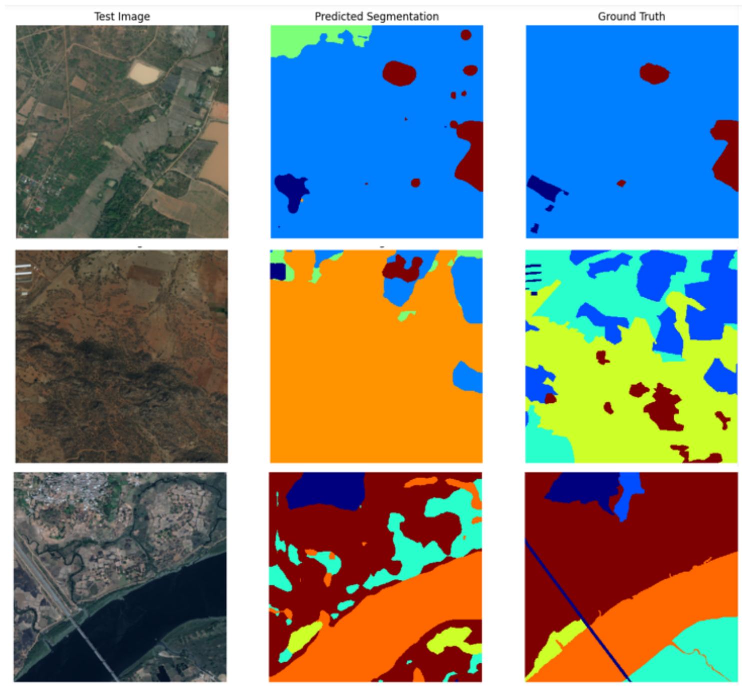

Figure 4: Segmentation results of the best U-Net model (ResNet-101 backbone) on the DeepGlobe test set. Colors in the masks represent classes: blue (Urban), cyan (Agriculture), green (Rangeland), yellow (Forest), orange (Water), red (Barren).

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.

- **Model Selection for Multispectral Adaptation:** The limitations of RGB dataset necessitated a transition to multispectral data. DeepLabV3 and U-Net were selected for further adaptation in Phase 2 due to their established performance in Phase 1, with plans to also explore transformer-based models to potentially improve segmentation accuracy for challenging classes.

4.2 Phase 2: Adaptation and Optimisation for Multispectral Data [In Progress]

Phase 2 focuses on adapting the models for multispectral data. The initial steps involved transitioning to panchromatic data before incorporating multispectral data.

Model	Val Dice	Dice (Road)	Dice (Building)	Dice (Vegetation)	Dice (Water)
Segformer	0.6039	0.5611	0.5716	0.5746	0.7488
UNet	0.5319	0.4553	0.4517	0.5805	0.6123
Swin Transformer	0.3187	0.2578	0.2161	0.5090	0.0443
Segformer-UNet	0.4719	0.5530	0.5451	0.4750	0.7074

Table 2: Performance metrics of different models on panchromatic data. The metrics include overall Dice coefficient and class-wise Dice scores for road, building, vegetation, and water classes.

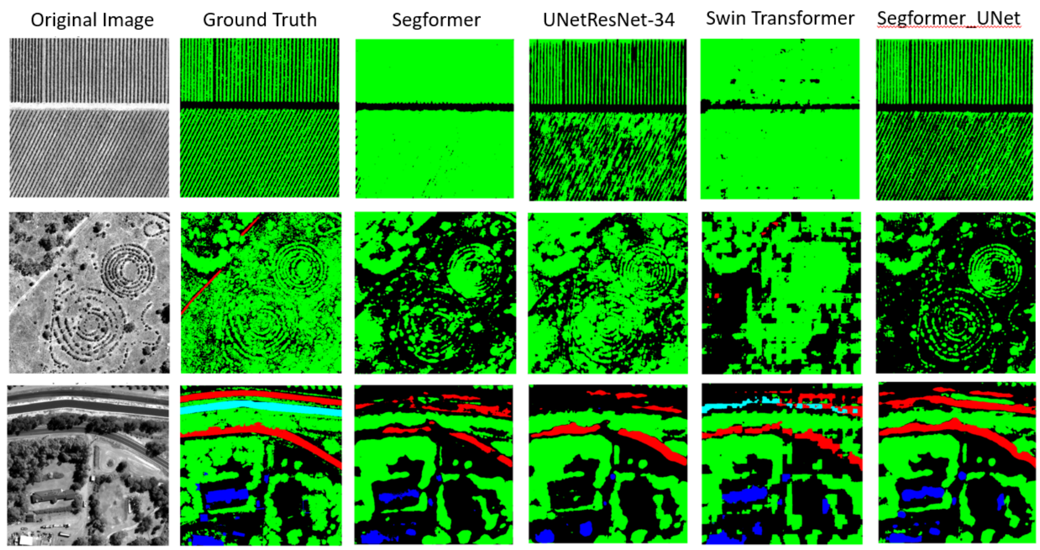

Figure 5: Semantic segmentation comparison on satellite imagery: (1) Original Image, (2) Ground Truth, (3) Segformer, (4) UNetResNet-34, (5) Swin Transformer, (6) SegformerUNet. Colors represent classes: black (nodata/ground/clutter), red (roads), blue (buildings), green (vegetation), cyan (water).

- **Dataset Transition:** The project transitioned to a dataset comprising panchromatic and multispectral images. The dataset included 512x512 panchromatic images, 128x128x8 multispectral images, representing the same location, and corresponding ground truth masks. The original 7 classes were relabeled into 5 classes: nodata/ground/clutter, road, building, vegetation, and water.
- **Model Performance on Panchromatic Data:** U-Net (ResNet-34 backbone), SegFormer [29], and Swin Transformer [30] models were tested on panchromatic data, with their input layers modified to handle single-channel input. Refer to Figure 5 for visual comparisons and Table 2 for quantitative metrics.
  - **U-Net:**
    - Excelled in capturing fine details, particularly vegetation (highest Dice score: 0.5805).
    - Struggled with class identification for roads and buildings.
  - **SegFormer:**
    - Demonstrated superior class identification, especially for roads (0.5611) and buildings (0.5716).
    - Achieved the highest overall Dice score (0.6039) and best water segmentation (0.7488).
  - **Swin Transformer:**
    - Produced patchy segmentation layers due to its patch-based nature.
    - Resulted in the lowest overall Dice score (0.3187).

    * Patchiness persisted despite attempts to mitigate it with skip connections and data augmentation.
  - **SegFormer-U-Net:**
    * Combined the strengths of U-Net and SegFormer.
    * Performed well on roads (0.5530), buildings (0.5451), and water (0.7074).
    * Performed best in visualisation despite not having the highest overall Dice score (0.4719).
    * Balanced detailed segmentation and class accuracy more effectively than other models.
  - **Overall Summary:**
    * SegFormer outperformed other models in terms of Dice scores.
    * SegFormer-U-Net showed promise in balancing detail and class accuracy.
    * Panchromatic data lacked sufficient spectral information for accurate segmentation of water and vegetation.

- **Input Layer Modification:** The SegFormer-U-Net hybrid model, selected for its overall good performance in panchromatic segmentation, has been used as the base for multispectral adaptation. Three techniques are being explored to incorporate multispectral data:
  - **Upsampling multispectral images using convolutional layers:** This is a straightforward approach to utilise multispectral data. It has been implemented in two ways: either as part of the model architecture or as a separate preprocessing step. Figure 6 shows example images generated with separate resampling. Experimentation is ongoing to optimise as well as integrate this layer directly into the model.

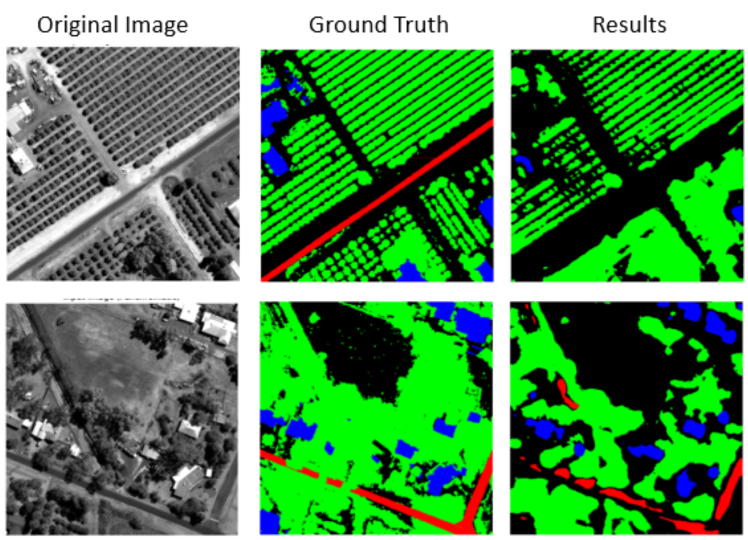

Figure 6: Result generated by model trained using convolution layer upsampled ms image

  - **Pan-sharpening multispectral images to match panchromatic resolution:** Bayesian pan-sharpening is currently used [31]. This process is computationally expensive. The initial results, shown in Figure 7, are promising, but require further optimisation.
  - **Developing a fusion model that processes panchromatic and multispectral data separately before concatenation:** This approach aims to preserve more data. However, it results in feature maps of different sizes. Two resizing strategies were investigated: resizing during data loading and resizing feature maps within the model. Resizing during data loading yielded better initial results compared to bilinear interpolation at the fusion level. Further experimentation with different interpolation methods and convolutional fusion layers is underway.

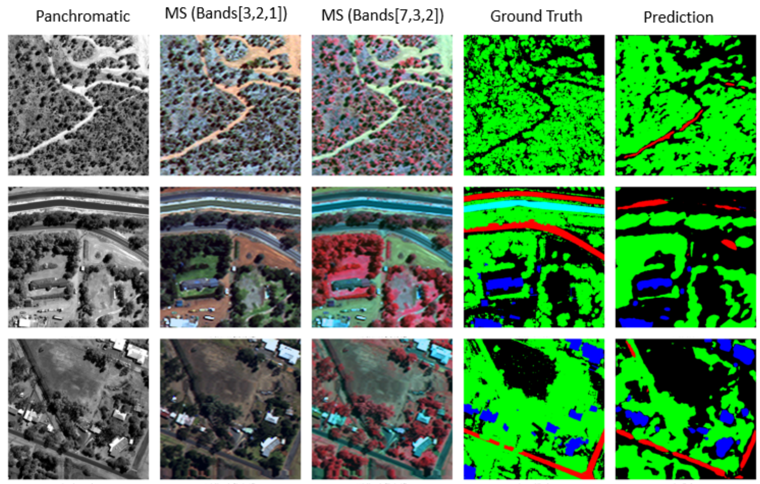

Figure 7: Result generated by model trained using Baysian pan-sharpened ms image

The input layer of the hybrid SegFormer-U-Net model was modified to accept 9-channel input (8 multispectral bands + 1 panchromatic band). Pretrained weights were used for the RGB channels, while the remaining channels were trained from scratch.

- **Preprocessing Implementation:** Various preprocessing techniques, including data augmentation, learning rate optimisation, and optimizer selection, are being investigated. Initial results indicate that band selection experiments, in which all bands were used, outperformed dropping bands. Further experimentation with preprocessing techniques is ongoing.

4.3 Next Steps and Future Directions

Future work for phases two and three will focus on the following areas:

- **Multispectral Data Optimisation:**
  - Experiment with advanced pan-sharpening techniques and optimise the pan-sharpening process to reduce computational costs.
  - Explore new fusion methods for multispectral and panchromatic data, including alternative interpolation techniques and convolutional fusion layers.
  - Investigate alternative approaches to resolution matching, such as downsampling the panchromatic image to the multispectral resolution, rather than upsampling the multispectral data.
  - Test additional transformer-based models and architectures specifically designed for multispectral data.
  - Conduct further band selection experiments, specifically testing the impact of dropping RE and NIR2 bands.
- **Vision-Language Model (VLM) Exploration:**
  - Investigate the potential of VLMs for advanced tasks, including temporal understanding, referring segmentation, scene understanding, counting, and detailed image captioning.
  - Evaluate the feasibility of using VLMs to enhance the semantic understanding of the segmented satellite imagery.
  - Explore fine-tuning techniques for VLMs in the context of our specific dataset.
- **Refinement and Evaluation:**
  - Continue to refine the SegFormer-U-Net hybrid model, focusing on balancing detailed segmentation and class identification.

– Conduct thorough evaluations of all techniques using appropriate metrics, addressing the challenges encountered in panchromatic experiments.
– Address the limitations observed in panchromatic data, particularly regarding vegetation and water segmentation, through improved multispectral data utilisation.

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] 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.

[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] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] X. Li, C. Wen, Y. Hu, Z. Yuan, and X. X. Zhu, "Vision-language models in remote sensing: Current progress and future trends," *IEEE Geoscience and Remote Sensing Magazine*, 2024.

[25] C. Liu, J. Zhang, K. Chen, M. Wang, Z. Zou, and Z. Shi, "Remote sensing temporal vision-language models: A comprehensive survey," *arXiv preprint arXiv:2412.02573*, 2024.

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

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

[28] 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.

[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," *CoRR*, vol. abs/2105.15203, 2021.

[30] 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.

[31] 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.