Soft Computing Fusion with Applications

www.scfa.reapress.com

Soft. Comput. Fusion. Appl. Vol. 1, No. 4 (2024) 229–241.

Paper Type: Original Article

Deep Learning-Based Segmentation for Land Management Using Satellite Imagery

Mohammad Sheihan Javaid*

Faculty of Engineering and Technology, Jamia Millia Islamia, New Delhi, India; sheihanjd20@gmail.com.

Citation:

Received: 15 March 2024 *Revised: 20 June 2024* *Accepted: 04 October 2024*	Javaid, M. J. (2024). Deep learning-based segmentation for land management using satellite imagery. *Soft computing fusion with applications, 1*(4), 229-241.

Abstract

The Land Use and Land Cover (LULC) segmentation represents a critical challenge in environmental monitoring and sustainable development. Traditional Convolutional Neural Networks (CNNs) excel at local pattern recognition but struggle with comprehensive spatial understanding, while transformer architectures capture global contexts at significant computational expense. This research introduces an innovative hybrid model that strategically combines the strengths of CNNs and vision transformers. We develop a segmentation approach that transcends existing methodological limitations by efficiently extracting local features through CNNs and leveraging transformers' ability to comprehend long-range dependencies. The proposed framework achieves a high accuracy of nearly 95 % and a mean Intersection over Union (IoU) of nearly 91% with reduced computational complexity, making advanced geospatial analysis more accessible. This approach advances technical capabilities and empowers researchers and policymakers with precise, timely insights into landscape dynamics, enabling more informed environmental decision-making across diverse geographical contexts.

Keywords: Vision transformer, Multipath feature fusion network, Conditional random fields, Land use and land cover.

1 | Introduction

The swift evolution of remote sensing technologies has made high-resolution satellite imagery widely available, creating unique opportunities for the analysis and management of land resources. The Land Use and Land Cover (LULC) segmentation, a prominent application of such imagery, is critical in land management domains, including urban planning, agricultural monitoring, environmental conservation, and disaster response. Precisely segmenting satellite images into distinct land cover types—such as urban areas, forests, water bodies, and agricultural lands—is vital for well-informed decision-making in these fields [1]. This need has spurred increasing efforts toward developing reliable and efficient LULC segmentation models that can address the inherent complexities of the task. As the demand for accurate, high-resolution LULC maps grows, particularly with the increasing availability of satellite data, new methodologies and innovations in image analysis are becoming indispensable to meeting these challenges [2].

Corresponding Author: sheihanjd20@gmail.com

https://doi.org/10.22105/scfa.v1i4.61

Transformers, which transformed natural language processing, have also been adapted for computer vision tasks, leading to the creation of vision transformers. These models capture long-range dependencies in images, unlike traditional Convolutional Neural Networks (CNNs), which primarily focus on local regions [3]. Vision transformers have demonstrated superior performance in capturing global context and object relationships in images, making them highly valuable for tasks like LULC segmentation, where the spatial distribution and interrelationships of land types across large areas are crucial. Similarly, Segmentation Transformers (SETRs) have shown remarkable success in image segmentation tasks. However, SETRs often face limitations in handling large images due to the high computational demands of the self-attention mechanism, which compares every pixel to all other pixels, creating scalability issues. As a result, optimizing these models for large-scale image datasets, such as those produced by satellite imaging, remains a key area of active research.

CNNs, while powerful, often struggle to handle boundaries in remote sensing images accurately. Traditional methods treat all boundaries the same, ignoring their diverse nature. However, boundaries in remote sensing are crucial, as they represent transitions between different land cover types. These transitions can vary widely, from sharp to gradual, and can be influenced by factors like shadows in high-resolution images. Current methods often use boundary information as an intermediate feature, limiting its impact on the final classification. It is essential to associate boundaries with specific categories to achieve optimal results. Refining boundary information and categorizing transitions between land cover types can significantly improve segmentation accuracy and preserve fine-grained details in satellite images.

To improve LULC classification, researchers have explored various deep-learning techniques. Patch-based approaches, which divide images into small segments and classify each as a unit, can introduce misclassification errors if patches contain pixels from multiple classes. This approach overlooks spatial coherence within patches, leading to inaccuracies in boundary areas, where multiple land cover types may converge. In contrast, pixel-based approaches classify each pixel individually, reducing misclassification risks and providing a more detailed, fine-grained classification of land cover types. While unsupervised learning methods, such as those utilized by Kussul et al. [4], typically require substantial training data and are constrained by their reliance on inherent image features without label guidance, supervised learning methods, as demonstrated by Torres et al. [5], have shown effectiveness with smaller datasets, offering greater flexibility and adaptability when labeled data is limited.

Most methods treat all image boundaries similarly, ignoring their diverse nature. Treating Image boundaries similarly can lead to information loss. Boundaries in remote sensing images are significant, as they represent transitions between different land cover types. These transitions vary widely, from sharp to gradual, and can be influenced by factors like shadows. Current methods often use boundary information as an intermediate feature, limiting its impact on the final classification. To improve accuracy, it is essential to associate boundaries with specific categories. Traditional methods like patch-based and pixel-based approaches have limitations. Patch-based methods can misclassify patches with multiple classes, while pixel-based methods may overlook contextual information. Unsupervised methods require large datasets, while supervised methods can be more effective with smaller datasets.

The main contribution of this paper can be summarized as the following.

Proposal of an architecture that uses a modified lightweight Vit encoder with patch embeddings and Multi-Head Self-Attention (MHSA) mechanism to capture and process multiple inputs at a time, which are connected to the Feature Pyramid Network (FPN) decoder for performing semantic segmentation using features extracted by the transformer. The FPN decoder helps in the drastic reduction of the original parameters while still maintaining the required efficiency.

The architecture leverages the power of a transformer with an attention mechanism that gives several advantages over the traditional CNN method for capturing features, as it has global context awareness, dynamic feature selection property, parallelization efficiency, and multiple scale feature integration.

The use of semi-supervised Conditional Random Fields (CRFs) has proven to be a significant step as it helps to refine the predicted segmentation masks by improving the alignment of object boundaries and reducing noise. It leverages labeled data (Supervised) and unlabelled data (Unsupervised) to enhance segmentation accuracy. This paper contributes to the semantic segmentation of satellite imagery for proper LULC mapping by proposing a lightweight model architecture.

The rest of the paper is structured as follows:

The literature survey is discussed in Section 2. Section 3 describes the methodology, including dataset, data preprocessing, and model architecture. Section 4 discusses the results and outcomes of the method used. Section 5 includes evaluation metrics used. Finally, Section 7 concludes this paper.

2|Literature Survey

Towards the advancements of segmentation of satellite images, Unet has been a remarkable model for segmentation tasks. However, the proposed model, mUnet, is based on a modified U-Net architecture designed for pixel-level semantic segmentation. This model follows an encoder-decoder, or ladder-like, structure that uses convolutional layers to encode features, followed by decoding layers that reconstruct a segmented output. Compared to traditional methods, mUnet is advantageous due to its lower number of trainable parameters, making it more efficient, and demonstrates superior segmentation performance on high-resolution, 3-band FCC satellite imagery [6].

Multipath Feature Fusion Network (MPFFNet) is a novel method for precise LULC classification, especially for high-resolution satellite images. It effectively merges deep learning techniques with traditional image processing, particularly Gabor filters. By combining both strengths, MPFFNet surpasses previous methods, especially in fine-grained classification tasks, delivering superior results [7]. The study focuses on pixel-level land cover classification using satellite imagery for monitoring and change detection applications. It employs U-Net with different ResNet backbones (ResNet18, ResNet34, ResNet101) for segmentation tasks. Among the configurations, the highest mean Intersection over Union (IoU) of 85.1 was achieved with U-Net and ResNet101, highlighting its effectiveness for high-resolution segmentation and detailed land cover mapping.

The paper introduces a novel approach to classify land cover accurately in medium-resolution satellite images. It leverages the power of the Swin Transformer, a state-of-the-art deep-learning architecture renowned for its ability to capture long-range dependencies within data. The Swin Transformer makes it particularly well-suited for analyzing medium-resolution images, where traditional methods often struggle due to limited spatial information. The paper incorporates an improved Swin Unet model designed explicitly for image segmentation tasks to enhance performance. Additionally, the authors integrate preprocessing, image enhancement, and spectral selection techniques to optimize the model's input data. By combining these advanced techniques, the proposed method significantly improves classification accuracy, making it a valuable tool for various Earth observation applications.

This work employs a Fully Convolutional Network (FCN-8) with VGG-16 weights for semantic segmentation of high-resolution satellite images into four LULC categories: Forest, built-up, farmland, and water. A non-overlapping grid-based approach is proposed to enhance segmentation accuracy. The FCN-8 model researches low-resolution features onto high-resolution pixel space for dense classification. Tested on the Gaofen-2 dataset, the model achieved an average accuracy of 91.0% and an IoU of 84.2% [8]. It combines a Denoising Diffusion Probabilistic Model (DDPM) for refined semantic features and a vision transformer for global context. DDPM for LULC segmentation and feature-level fusion between DDPM and Transformer for improved segmentation accuracy [9]. The model uses a diffusion-based U-Net to effectively capture multiscale features, including detailed context and edge information, making it ideal for high-resolution segmentation. It also includes a lightweight classification module with a spatial-channel attention mechanism that helps the model focus on the most important spatial and channel features. Incorporating unsupervised

pre-trained components addresses class imbalance and speeds up training. This approach not only improves accuracy but also reduces complexity and training time [10].

3 | Methodology

This research aimed to develop a powerful yet efficient semantic segmentation model designed explicitly for LULC mapping. The challenge was balancing model complexity and computational efficiency while still achieving high-quality, accurate results. To accomplish this, we proposed a hybrid model architecture that combines the strengths of a lightweight transformer with a CNN decoder. The transformer component uses MHSA and a Multi-Layer Perceptron (MLP) to capture global and local spatial details within each image effectively. These layers enable the model to understand broader contextual relationships and finer details simultaneously. The CNN decoder then hierarchically processes and stacks these features, structuring them to enhance the model's ability to distinguish between different classes. Finally, the processed features are passed to the segmentation head, which produces the output mask in the desired shape, allowing for accurate LULC mapping while keeping computational demands manageable. *Fig. 1* shows the workflow component required for practical model training and predictions.

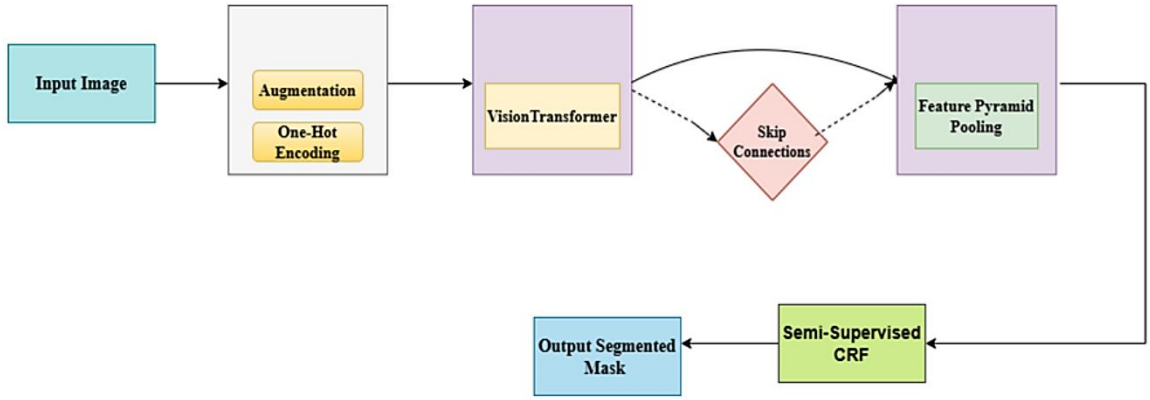

Fig. 1. The workflow component required for practical model training and predictions.

3.1 | Deep Globe Land Cover Dataset

This dataset offers high-resolution satellite images, each with dimensions of 2448x2448 pixels, which allows for fine-grained segmentation of various land cover classes. The dataset is organized into train and test folders, each containing images paired with their corresponding segmentation masks [11]. These masks serve as ground truth for evaluating segmentation performance. The dataset defines multiple land cover categories, each represented by unique RGB values in the segmentation masks. *Table 1* below provides details on the class labels, corresponding RGB values, and a brief description of each class:

Table 1. Class labels and their corresponding RGB values of the dataset.

Class Labels	RGB Values	Description
Urban land	(0,255, 255)	Artificial, built-up areas with human artifacts (Ignoring roads)
Agriculture land	(255, 255, 0)	Farms, plantations, cropland, orchards, vineyards, nurseries, and confined feeding operations
Rangeland	(255, 0, 255)	Non-forest, non-farm, green land, grass
Forest land	(0, 255, 0)	Land with significant tree cover, including clear-cuts
Water	(0, 0, 255)	Rivers, oceans, lakes, wetlands, ponds
Barren land	(255, 255, 255)	Mountain, land, rock, desert, beach, no vegetation
Non-observed area	(0, 0, 0)	Clouds and other unclassified areas

3.2 | Land Cover Dataset

The dataset delineates various land cover categories, each encoded with distinct RGB values in the segmentation masks [10]. These color-coded masks enable pixel-level classification, where each color represents a specific land cover class. Below is a comprehensive *Table 2* that outlines the class labels, RGB values, and corresponding descriptions:

Table 2. Class labels and their corresponding RGB values of the dataset.

Class Labels	RGB Values	Description
Background	(255, 255, 255)	Unlabelled or background pixels
Buildings	(255, 0, 0)	Artificial structures like houses, buildings, and factories
Woodland	(0, 255, 0)	Areas with significant tree cover, including forests and parks
Water	(0, 0, 255)	Bodies of water like rivers, lakes, oceans, and ponds
Non-observed area	(0, 0, 0)	Areas obscured by clouds or not classified in the dataset

3.3 | Technical Significance of Class Encoding

Pixel-level accuracy

Each pixel in the segmentation mask corresponds directly to a specific land cover class, ensuring high granularity in classification tasks. The pixel allows models to learn fine distinctions between visually similar but semantically distinct regions.

Color-coding for efficient segmentation

The distinct RGB values assigned to each class help distinguish between different land cover types. The RGB values facilitate the training of deep learning models, where accurate pixel classification is critical.

Balanced representation

The diversity of classes, ranging from natural environments like forests and water bodies to human-made areas such as urban and agricultural land, ensures a comprehensive representation of land cover types. The diversity of classes is particularly important for generalizing segmentation models to diverse geospatial datasets.

Non-observed

Including a "Non-observed Area" class (Represented by black) ensures that regions obscured by clouds or other anomalies do not interfere with the training process, improving model robustness.

3.4 | Data Preprocessing

Data preprocessing was a crucial step in the training pipeline, significantly enhancing the dataset's quality and diversity. This step ensured that the model could generalize effectively across various real-world scenarios by simulating different conditions through data augmentation techniques.

Spatial transformations

To introduce spatial variability, we applied transformations such as:

I. Cropping: Randomly extracting sections to help the model learn from different parts of the image.

II. Flipping and rotating: Altering image orientation ensures the model recognizes patterns from multiple perspectives.

III. Resizing: Adjusting the dimensions to meet model input requirements while preserving image structure.

These augmentations helped the model handle changes in the position and orientation of objects within the images.

3.5 | Intensity and Contrast Adjustments

We incorporated techniques to simulate varying lighting conditions:

I. Brightness and gamma adjustments: Modulating brightness and luminance to mimic different exposure levels.

II. Contrast Limited Adaptive Histogram Equalization (CLAHE): Enhancing local contrast for better clarity in darker regions.

These adjustments improved the model's robustness to varying lighting conditions, ensuring effective performance under different illumination scenarios.

3.6 | Colour and Noise Variations

To simulate environmental changes and sensor noise, we used:

I. Hue and saturation adjustments: Modifying color tones to account for vegetation and water appearance variations.

II. Gaussian blur and noise: Mimicking atmospheric distortion and sensor noise to make the model resilient to degraded images.

3.7 | Coarse Dropout for Robust Feature Learning

By randomly erasing parts of the images using coarse dropout, we forced the model to focus on critical features, preventing over-reliance on specific regions. These augmentation techniques enhanced the model's ability to handle diverse input variations, ultimately improving generalization and segmentation performance.

4 | Model Architecture

In this research, we implemented a hybrid architecture for semantic segmentation, combining a modified lightweight transformer for feature extraction with an FPN decoder and CRF for post-processing. This combination enhanced segmentation quality by refining edges and improving boundary definition, resulting in smoother, more accurate masks. Query: Describe the architecture represented in *Fig. 2*.

Fig. 2 illustrates a deep learning architecture for image segmentation, likely applied to satellite imagery for land management. It employs a Vision Transformer Encoder to extract hierarchical features from the input image. These features are fed into an FPN Decoder, aggregating multiscale information to generate a detailed segmentation mask. Skip connections between the encoder and decoder help preserve fine-grained details, and a CRF is used post-processing to refine the segmentation output.

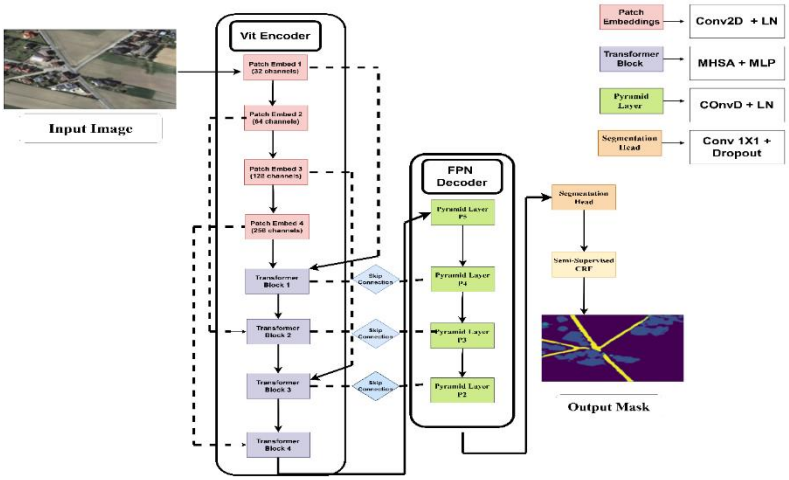

Fig. 2. Model architecture.

4.1 | Vision Transformer Encoder

At the core of the architecture is a modified transformer, which replaces traditional CNN-based encoders. Transformers excel at capturing global context through their attention mechanisms. However, to ensure efficiency, several optimizations were made:

I. Reduced number of blocks: A smaller number of transformer blocks was used, reducing computational overhead while retaining the core advantages of the attention mechanism.

II. Drop-path regularization: Drop-path randomly turns off specific pathways during training, improving model robustness and preventing overfitting.

III. Lightweight design: Careful adjustments were made to the architecture to reduce the number of parameters, ensuring a balance between computational efficiency and feature extraction capability.

4.2 | Feature Pyramid Network Decoder

Following the encoder, we utilize an FPN decoder to create a hierarchical structure of feature maps at different resolutions. The FPN decoder refines and integrates these feature maps using skip connections, which help preserve fine-grained details and broader contextual information from the encoder outputs. This hierarchical approach ensures the model can recognize objects at multiple scales, enhancing segmentation accuracy and detail.

4.3 | Segmentation Head

The processed feature maps are then passed through multiple segmentation heads, each tailored to predict segmentation masks at specific levels of detail. These masks capture objects at varying scales, ensuring comprehensive object identification across the image. Finally, the individual predictions from each segmentation head are combined and upscaled to match the original input resolution. This step results in a high-quality segmentation mask that precisely delineates objects and boundaries in the image.

4.4 | Semi-Supervised Conditional Random Field

We apply a CRF as a post-processing step to further enhance the segmentation quality. The CRF acts as a spatial filter, refining the segmentation mask by aligning boundaries more accurately with object edges and reducing noise. This semi-supervised approach allows the CRF to maintain consistency across the segmentation, yielding more polished and accurate segmentation results, and is given by the equation. This hybrid approach integrates transformers, pyramid networks, and CRF post-processing to create a powerful and flexible model for high-quality semantic segmentation across diverse scales and contexts.

$$E(x) = \underbrace{\sum_{i \in L} U(x_i) + \sum_{i \in U} U(x_i)}_{\text{Labeled and Unlabeled Data Unary Potentials}} + \underbrace{\sum_{i,j} \psi(x_i, x_j)}_{\text{Pairwise Potentials for Smoothness}}. \quad (1)$$

Evaluation metrics

I. Jaccard index: In multi-class semantic segmentation, the Jaccard index, or IoU, measures the overlap between the predicted and ground truth segmentation masks for each class and tends to give equal weight to False Positives and False Negatives. Here is how it would be computed for N classes step by step:

$$\text{Jaccard index} = \frac{1}{N} \sum \frac{TPc}{TPc + FNc + FPc}. \quad (2)$$

II. Dice Coefficient: The dice coefficient metric is closely related to the Jaccard index, which measures the similarity between the ground truth masks and the predicted mask.

$$\text{Dice Coefficient} = \frac{2 \cdot \sum(TP_c)}{\sum(2 \cdot TP_c + FP_c + FN_c)}. \quad (3)$$

Loss function

The loss function I have used here is the total weighted loss of both weighted cross entropy and dice loss to reach an optimal solution, as it considers the weighted sum of both components. The weighted cross-entropy penalizes class imbalance by assigning different weights to each class. At the same time, the Dice Loss measures the overlap between the predicted and ground truth masks, ensuring that the model improves in terms of segmentation quality.

$$\text{Weighted Cross-Entropy Loss} = \frac{1}{C}\sum_{c=1}^{C}\left(-w_c \cdot \left(\sum_{i=1}^{N} g_{i,c}\log(p_{i,c}) + (1-g_{i,c})\log(1-p_{i,c})\right)\right).$$

$$\text{Dice Loss} = \frac{1}{C}\sum_{c=1}^{C}\left(1 - \frac{2\cdot\sum_{i=1}^{N} p_{i,c}\cdot g_{i,c}}{\sum_{i=1}^{N} p_{i,c} + \sum_{i=1}^{N} g_{i,c} + \epsilon}\right).$$

$$\text{Total Loss} = \lambda_1 \cdot \frac{1}{C}\sum_{c=1}^{C}\left(-w_c \cdot \left(\sum_{i=1}^{N} g_{i,c}\log(p_{i,c}) + (1-g_{i,c})\log(1-p_{i,c})\right)\right) + \lambda_2 \cdot \frac{1}{C}\sum_{c=1}^{C}\left(1 - \frac{2\cdot\sum_{i=1}^{N} p_{i,c}\cdot g_{i,c}}{\sum_{i=1}^{N} p_{i,c} + \sum_{i=1}^{N} g_{i,c} + \epsilon}\right).$$

5 | Results and Discussion

The result of this proposed study is represented as follows:

5.1 | Deep Globe Dataset

When the model was trained on the deep globe dataset, the results metrics gave the following results, as shown in *Fig. 3*, followed by the bar graph, as shown in *Fig. 4*.

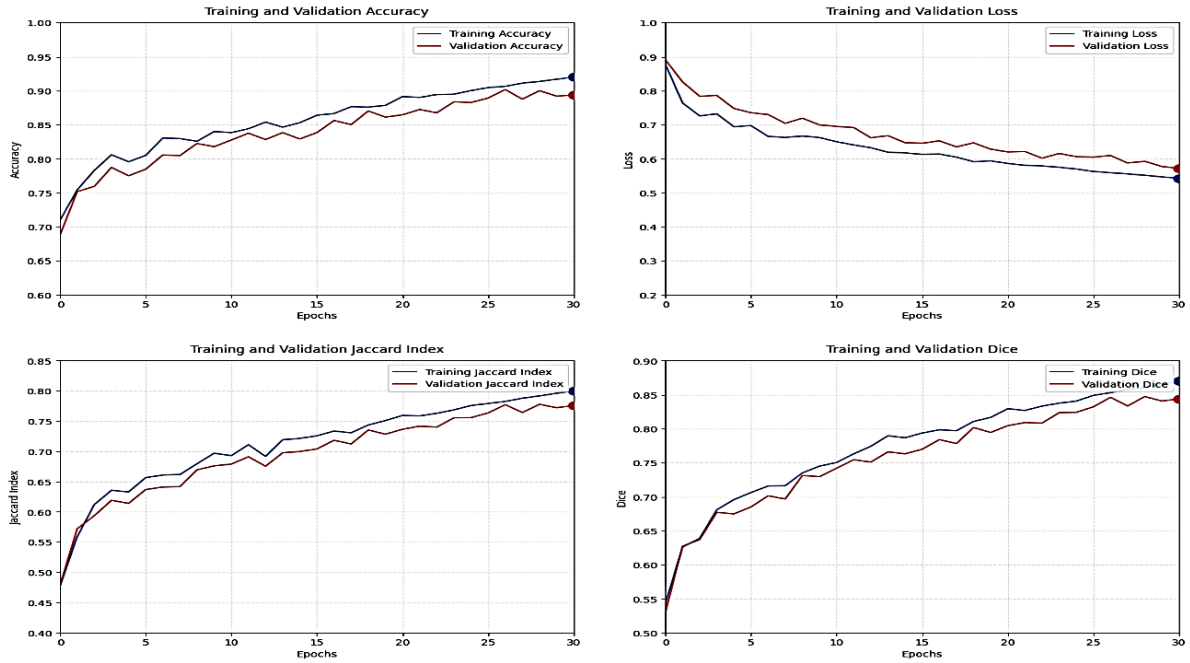

Fig. 3. Plotting training vs. Testing metrics graph for the deep globe dataset.

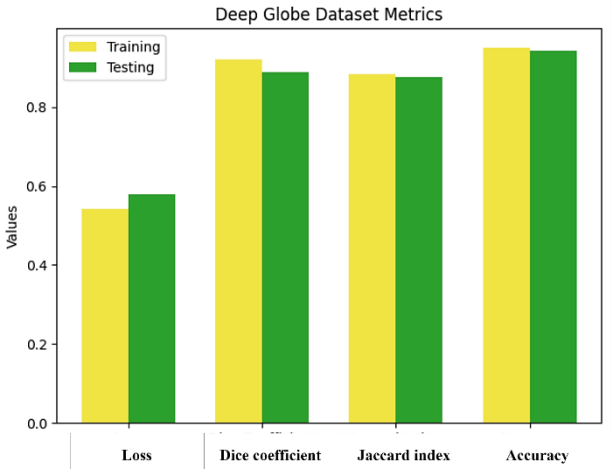

Fig. 4. Metric bar graph depicting results.

The result of both training and testing samples from the dataset is shown in *Table 3*.

Table 3. Result metric table for deep globe dataset.

Metric	Training Samples	Testing Samples
Loss	0.5421	0.5789
Dice coefficient	0.9201	0.8893
Jaccard index	0.8823	0.8756
Accuracy	0.951	0.943

Fig. 5 shows the visualization of results obtained with the original image, ground truth masks, and the predicted masks.

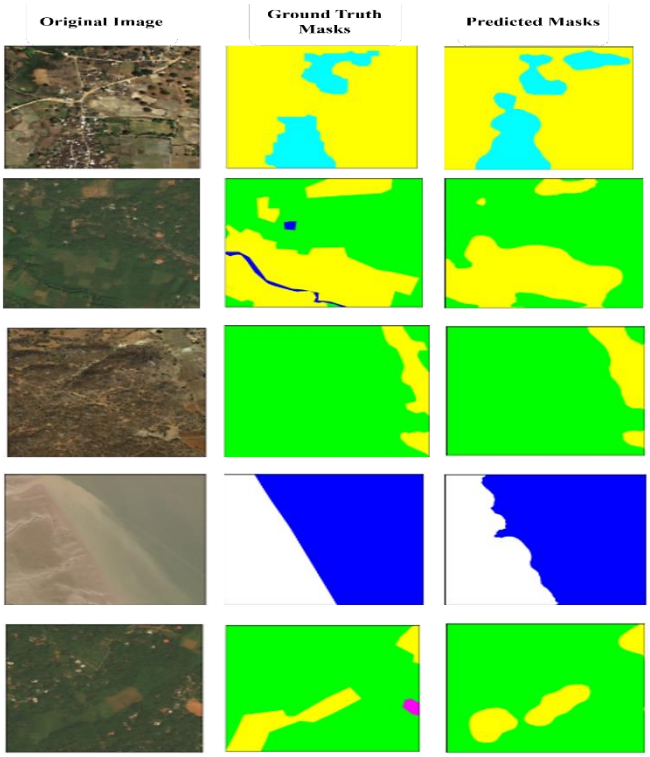

Fig. 5. Visualization of results on the deep globe dataset.

5.2 | Land Cover Dataset

When the model was trained on the deep globe dataset, the results metrics gave the following results, as shown in *Fig. 6*, followed by the bar graph, as shown in *Fig. 7*.

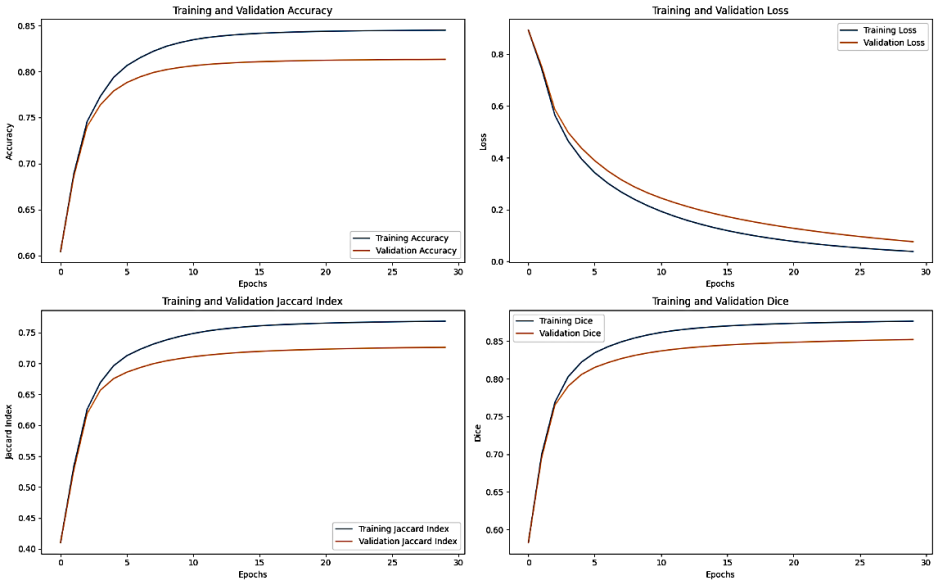

Fig. 6. Plotting training vs. Testing metrics graph for the land cover dataset.

Fig. 7. Metric bar graph depicting results.

The results of both training and testing samples from the dataset are shown in *Table 4*. The result metric table for the land cover dataset is presented in *Table 4*.

Table 4. Result metric table for land cover dataset.

Metric	Training Samples	Testing Samples
Loss	0.419	0.443
Dice coefficient	0.9208	0.8997
Jaccard index	0.9001	0.8829
Accuracy	0.954	0.948

The proper visualization of the model trained on the land cover dataset with its original image, ground truth masks, and the predicted masks is shown in *Fig. 8* for the dataset.

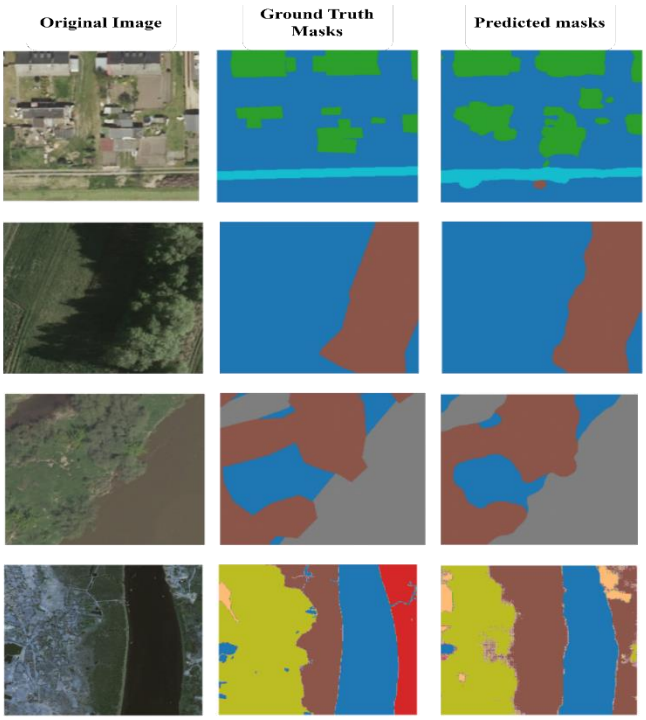

Fig. 8. Visualization of results on land cover dataset.

The discussion of the proposed problem is shown in *Figs. 4-8*. The deep learning model, integrating a Vision Transformer encoder, FPN decoder, and CRF model, and trained with the Adam optimizer and a combined loss function of weighted cross-entropy and Dice loss with L2 regularization, demonstrated robust performance on both the Deep Globe and Land Cover datasets. Both datasets exhibited similar trends, with training metrics consistently outperforming testing metrics, suggesting effective learning but potential overfitting. The model achieved impressive results across all metrics, particularly in accuracy, Dice coefficient, and Jaccard index, indicating strong pixel-level classification capabilities. While the Deep Globe dataset presented a slightly higher loss (0.5421 vs. 0.419), both datasets showcased the model's ability to extract meaningful features and generate accurate predictions. To further enhance performance, future research could explore advanced regularization techniques, more sophisticated data augmentation strategies, refined hyperparameter tuning, and innovative model architectures, such as incorporating self-attention mechanisms or exploring hybrid approaches combining CNNs with transformers.

6 | Comparative Analysis

Table 5 highlights the performance of various segmentation models based on multiple metrics, including the Accuracy Index, Dice Coefficient, F1-score, Precision, Recall, and Accuracy. The mUnet model demonstrates a Jaccard Index 70.6, with solid Precision and Recall values (87.13 and 85.66, respectively), but lacks Dice Coefficient data. MPFFNET shows a higher Jaccard Index of 81.02 and a notable F1-score of 92.1, indicating strong performance. The Unet++ with ResNet101 improves further with a Jaccard Index of 87.1, a Dice Coefficient of 88.23, and the highest accuracy (92.06%). DDPM-SegFormer exhibits impressive results with a Jaccard Index of 83.57 and a high Precision (91.72) alongside strong Recall (90.23). The Swin-Unet Transformer model also demonstrates competitive performance, with a Jaccard Index of 85.9 and a high F1-score of 93.89. Finally, the Proposed Model outperforms all others with a Jaccard Index of 90.01, a Dice Coefficient of 92.08, and the highest Accuracy of 95.1%, showcasing its highly reliable segmentation performance across all metrics.

Table 5. Comparative analysis of various models with the proposed model.

Model Used	Jaccard Index	Dice Coefficient	F1-score	Precision	Recall	Accuracy
mUnet	70.6	-	69.84	87.13	85.66	88.05
MPFFNET	81.02	-	92.1	86.94	87.32	89.0
Unet++ with resnet101	87.1	88.23	87.95	85.23	88.19	92.06
DDPM-SegFormer	83.57	85.14	90.97	91.72	90.23	93.89
Swin- Unet transformer	85.9	-	93.89	94.1	89.2	91.2
Proposed Model	90.01	92.08	93.1	91.1	94.00	95.1

7 | Conclusion

This study developed a hybrid deep learning model by integrating a Vision Transformer encoder**,** a**n** FPN decoder**,** and a Semi-Supervised CRF for boundary refinement. This architecture balanced segmentation accuracy and computational efficiency, with the ViT capturing long-range dependencies through global attention and the FPN aggregating multiscale features for precise boundary delineation. The CRF further enhanced segmentation quality by refining edge boundaries, ensuring smoother and more accurate predictions. The model was trained using the Adam optimizer with a combined loss function of weighted cross-entropy and Dice loss**,** complemented by L2 regularization to prevent overfitting. Training results showed consistently higher metrics than testing, suggesting effective learning but highlighting potential overfitting. Despite this, the model demonstrated strong pixel-level classification capabilities, achieving high scores in accuracy**,** Dice coefficient, and Jaccard index, indicating robust feature extraction and prediction. Future iterations will focus on reducing overfitting through advanced data augmentation, optimized regularization techniques, and exploring transfer learning to enhance adaptability across diverse geographic landscapes. Expanding the dataset to cover more varied terrain types and challenging environmental conditions will further improve the model's generalizability, solidifying it as a comprehensive solution for semantic segmentation across dynamic real-world scenarios.

7.1 | Future Direction

I. Scaling to a multi-spectral dataset: The results shown above were trained and tested on the RGB dataset, but this is not the scenario in every case. The satellite images should be in a multi-spectral range for more advanced and detailed segmentation and better LULC mapping.

II. Advanced loss functions: Exploring more sophisticated loss functions, such as focal loss for class imbalance and Dice coefficient loss for better overlap precision, could enhance model performance further. The loss functions will refine the pixel-level classification capabilities, particularly in challenging datasets.

III. Transfer learning: Implementing transfer learning using pre-trained models on larger, domain-specific datasets will speed up convergence and improve performance, especially when adapting the model to new regions or environmental conditions. Transfer learning will ensure that the model can handle various semantic segmentation tasks.

IV. Geographic adaptability: Future iterations will enhance the model's adaptability to diverse geographic landscapes, such as urban environments, agricultural fields, and natural terrains like forests and water bodies. This will help the model generalize better across varying terrain types and environmental challenges.

V. Expansion of dataset diversity: Expanding the dataset to include a broader range of geographic regions and more complex environmental conditions, including seasonal variations, cloud cover, and occlusions, will improve the model’s robustness and generalizability.

VI. Boundary refinement improvements: Further optimizing the CRF component for better edge detection in complex regions will be explored to enhance the accuracy of boundary delineation, especially in areas with fine-grained details or ambiguous transitions between classes.

VII. Real-time application testing: Finally, real-time applications in dynamic environments will be tested to evaluate how well the model performs in operational settings, ensuring it is ready for deployment in real-world applications such as land cover mapping, environmental monitoring, and autonomous systems.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. All associated costs were covered by the author.

References

[1] da Silva, M., PLP Correa, S., AR Schaefer, M., CS Reis, J., M Nunes, I., dos Santos, J. A., & N Oliveira, H. (2024). From tradition to transformation: Deep and self-supervised learning approaches for remote sensing in agriculture and environmental change. *Nunes, Ian and Dos Santos, Jefersson Alex and N. Oliveira, Hugo, from tradition to transformation: Deep and self-supervised learning approaches for remote sensing in agriculture and environmental change*. https://dx.doi.org/10.2139/ssrn.5063928

[2] Fan, J., Shi, Z., Ren, Z., Zhou, Y., & Ji, M. (2024). DDPM-SegFormer: Highly refined feature land use and land cover segmentation with a fused denoising diffusion probabilistic model and transformer. *International journal of applied earth observation and geoinformation*, *133*, 104093. https://doi.org/10.1016/j.jag.2024.104093

[3] Yao, J., Zhang, B., Li, C., Hong, D., & Chanussot, J. (2023). Extended vision transformer (ExViT) for land use and land cover classification: A multimodal deep learning framework. *IEEE transactions on geoscience and remote sensing*, *61*, 1–15. https://doi.org/10.1109/TGRS.2023.3284671

[4] Kussul, N., Drozd, S., Yailymova, H., Shelestov, A., Lemoine, G., & Deininger, K. (2023). Assessing damage to agricultural fields from military actions in Ukraine: An integrated approach using statistical indicators and machine learning. *International journal of applied earth observation and geoinformation*, *125*, 103562. https://doi.org/10.1016/j.jag.2023.103562

[5] Torres, A. F., Walker, W. R., & McKee, M. (2011). Forecasting daily potential evapotranspiration using machine learning and limited climatic data. *Agricultural water management*, *98*(4), 553–562. https://doi.org/10.1016/j.agwat.2010.10.012

[6] Yuan, H., Zhang, Z., Rong, X., Feng, D., Zhang, S., & Yang, S. (2023). MPFFNet: LULC classification model for high-resolution remote sensing images with multi-path feature fusion. *International journal of remote sensing*, *44*(19), 6089–6116. https://doi.org/10.1080/01431161.2023.2261153

[7] Huang, L., Jiang, B., Lv, S., Liu, Y., & Fu, Y. (2024). Deep-learning-based semantic segmentation of remote sensing images: A survey. *IEEE journal of selected topics in applied earth observations and remote sensing*, *17*, 8370–8396. https://doi.org/10.1109/JSTARS.2023.3335891

[8] Jiang, T., Xu, T., & Li, X. (2022). VA-TransUNet: A u-shaped medical image segmentation network with visual attention. *Proceedings of the 2022 11th International conference on computing and pattern recognition* (pp. 128-135). Association for computing machinery. https://doi.org/10.1145/3581807.3581826

[9] Pokhariyal, S., Patel, N. R., & Govind, A. (2023). Machine learning-driven remote sensing applications for agriculture in India—A systematic review. *Agronomy*, *13*(9), 2302. https://doi.org/10.3390/agronomy13092302

[10] Wang, D., Yang, R., Zhang, Z., Liu, H., Tan, J., Li, S., … ., & Su, P. (2023). P-swin: Parallel swin transformer multi-scale semantic segmentation network for land cover classification. *Computers & geosciences*, *175*, 105340. https://doi.org/10.1016/j.cageo.2023.105340

[11] Demir, D. B., & Musaoglu, N. (2023). Automatic classification of selected corine classes using deep learning based semantic segmentation. *The international archives of the photogrammetry, remote sensing and spatial information sciences*, *48*, 71–75. https://doi.org/10.5194/isprs-archives-XLVIII-M-3-2023-71-2023