Deep Learning-Based Satellite Image Segmentation for Land Cover Classification: Model Design and Comparative Analysis

Gaurav Manish (B22CS079), Arjun Bhattad (B22AI051), Raunak Singh (B22CS085)

1 Introduction

Satellite imagery offers a critical vantage point for understanding land cover patterns, which is essential for urban planning, environmental monitoring, and agricultural management. With the increasing availability of high-resolution satellite data, automated segmentation using deep learning has become a powerful tool to extract valuable insights from these images. This project focuses on designing and evaluating several state-of-the-art segmentation models for the task of land cover classification.

2 Dataset Exploration

The dataset used in this project is the **DeepGlobe Land Cover Challenge** dataset, available on Kaggle. The dataset consists of three folders; however, only the `train` folder contains both satellite images and their corresponding segmentation masks.

2.1 Dataset Structure

The dataset is structured as follows:

- The `train` folder contains satellite images and their respective masks.
- Masks are provided in PNG format and correspond to specific images in the dataset.
- Each mask encodes land cover categories using a predefined color scheme.

2.2 Preprocessing Steps

To prepare the dataset for training, several preprocessing steps were applied:

- **Data Collection:** Image and mask paths were extracted and verified to ensure each image had a corresponding mask.
- **Data Splitting:** The dataset was randomly split into training (80%) and validation (20%) sets.
- **Label Encoding:** Mask images, originally in RGB format, were mapped to class indices using a predefined color-to-class correspondence.
- **Data Augmentation:** Training images were resized and transformed using random flips and rotations to enhance generalization.
- **Normalization:** Images were normalized using standard mean and standard deviation values for improved model convergence.

3 Approaches and Models

The following segmentation models were implemented and tested in the study. For each model, a brief architecture description is provided along with an illustrative diagram.

3.1 FPN (Feature Pyramid Network)

FPN utilizes a top-down architecture with lateral connections to extract multi-scale features. This design enhances the model's ability to segment objects of various sizes.

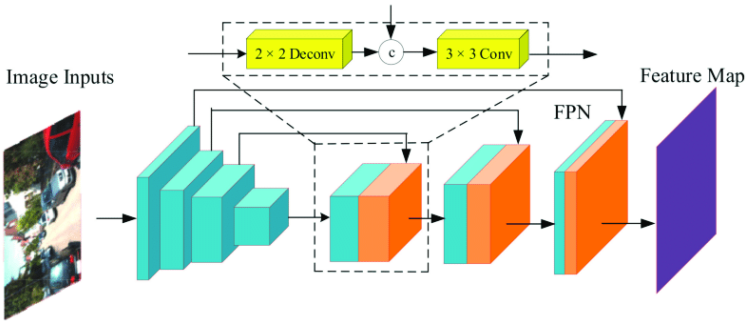

Figure 1: Architecture diagram of the FPN model.

3.2 DeeplabV3+

DeeplabV3+ extends the standard DeeplabV3 model by integrating an encoder-decoder structure, which aids in capturing detailed boundary information and improves segmentation quality.

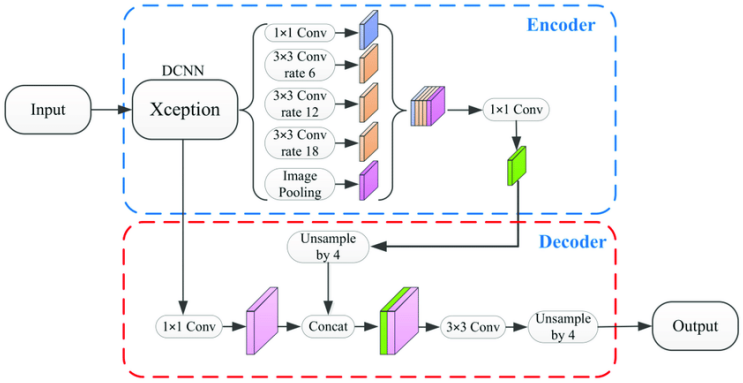

Figure 2: Architecture diagram of the DeeplabV3+ model.

3.3 Unet

Unet is a popular encoder-decoder network that employs skip connections to preserve spatial context. It is particularly effective for segmenting biomedical and remote sensing images.

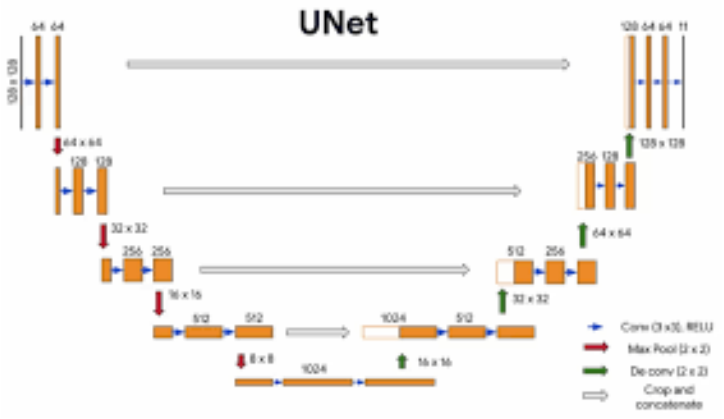

Figure 3: Architecture diagram of the Unet model.

3.4 Unet++

Unet++ is an improved version of Unet that features nested skip connections. This design bridges the semantic gap between the encoder and decoder feature maps, resulting in more precise segmentation.

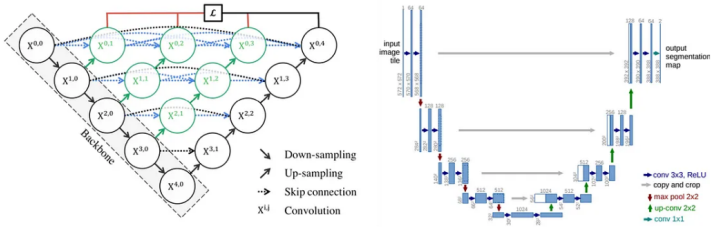

Figure 4: Architecture diagram of the Unet++ model.

3.5 Vision Transformer (ViT) for Segmentation

The Vision Transformer (ViT) model applies transformer-based attention mechanisms to capture global contextual information. This approach is a recent development in segmentation tasks, demonstrating promising results in handling complex spatial relationships.

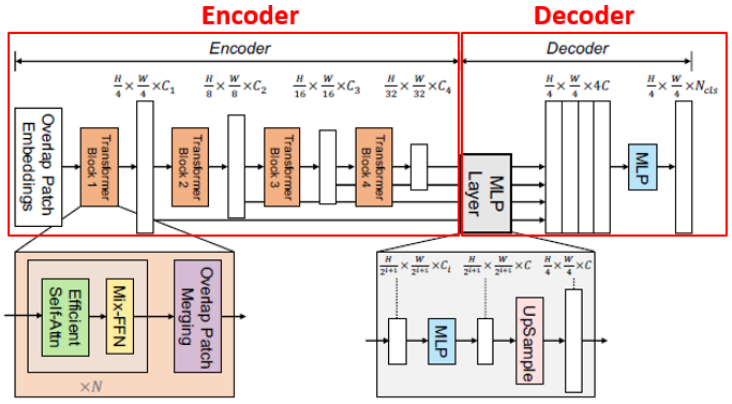

Figure 5: Architecture diagram of the Vision Transformer (ViT) model for segmentation.

3.6 Scratch Implementation of DeeplabV3+

In this approach, a DeepLabV3+ model is implemented from scratch using a ResNet101 backbone. The network uses an Atrous Spatial Pyramid Pooling (ASPP) module on the deepest encoder features and fuses these with low-level features from an earlier layer via a decoder. The following table details the encoder–decoder architecture and dimension changes assuming an input image of size $3 \times 512 \times 512$:

Layer	Input Size	Output Size	Operation
Layer0	512×512×3	128×128×64	Conv + BN + ReLU
Layer1	128×128×64	128×128×256	ResNet Block
Layer2	128×128×256	64×64×512	ResNet Block
Layer3	64×64×512	32×32×1024	ResNet Block
Layer4	32×32×1024	16×16×2048	ResNet Block
ASPP	16×16×2048	16×16×256	Atrous Conv
Interpolation	16×16×256	128×128×256	Bilinear Upsampling
Low-Level	128×128×256	128×128×48	1×1 Conv
Decoder	128×128×(48+256)	128×128×7	Conv Layers
Output	128×128×7	512×512×7	Bilinear Upsampling

Table 1: Encoder-Decoder Architecture of DeepLabV3+

Each model was fine-tuned on the satellite image dataset, and performance was evaluated using standard metrics such as Dice Loss, mean Intersection over Union (mIoU), and Pixel Accuracy.

4 Experimental Results

The performance of each model was recorded and compared using the following metrics:

Model	**Dice Loss**	**mIoU**	**Pixel Accuracy**
FPN	0.34	0.64	87.31%
DeeplabV3+	0.36	0.65	87.94%
Unet	0.39	0.58	86.90%
Unet++	0.37	0.59	87.05%
Scratch DeeplabV3+	0.36	0.66	87.56%

Table 2: Comparison of Segmentation Models

5 Inference and Reasoning

5.1 Performance Analysis

Scratch Implementation of DeeplabV3+:
The scratch version of DeeplabV3+ achieved a Dice Loss of 0.36, an mIoU of 0.66, and a Pixel Accuracy of 87.56%. These results indicate a slight improvement over the standard implementation, highlighting the benefits of custom adjustments and optimizations tailored to satellite imagery.

Unet and Unet++ Models:
Unet and Unet++ performed reliably, with Dice Loss values of 0.39 and 0.37, respectively. The mIoU scores were 0.58 for Unet and 0.59 for Unet++, while Pixel Accuracy stood at 86.90% and 87.05%. The nested skip connections in Unet++ provided enhanced semantic feature integration, beneficial for segmenting complex land cover structures.

FPN and Standard DeeplabV3+:
The FPN model demonstrated competitive performance with a Dice Loss of 0.34, an mIoU of 0.64, and a Pixel Accuracy of 87.31%. This suggests that its multi-scale feature extraction capability is effective for land cover segmentation. The standard DeeplabV3+ model recorded a Dice Loss of 0.36, an mIoU of 0.65, and a Pixel Accuracy of 87.94%, reinforcing its role as a strong baseline for segmentation tasks.

These evaluations underscore the impact of model architecture on segmentation performance within the DeepGlobe Land Cover Classification dataset.

6 Conclusion

The comparative study reveals that traditional CNN-based models such as U-Net and DeepLabV3+ provide reliable performance in handling satellite imagery segmentation tasks. Additionally, the scratch implementation of DeepLabV3+ indicates that custom architectural modifications can lead to incremental improvements. Future work may include further hyperparameter tuning, ensemble methods, or integrating multi-modal data to enhance segmentation accuracy.

7 Contributions

The following outlines the contributions of each team member:

- **Gaurav Manish**: Responsible for implementing the U-Net and U-Net++ models, as well as contributing to the report writing.
- **Arjun Bhattad**: Focused on implementing the Feature Pyramid Network (FPN) and developing a DeepLabV3+ model from scratch, in addition to assisting with the report writing.
- **Raunak Singh**: Worked on implementing DeepLabV3+ and developing a DeepLabV3+ model from scratch, along with contributing to the report writing.

All work was carried out in a collaborative manner.