RAJALAKSHMI ENGINEERING COLLEGE
An AUTONOMOUS Institution
Affiliated to ANNA UNIVERSITY, Chennai

25 years OF ACADEMIC EXCELLENCE

Department of Computer Science and Engineering

AERIAL IMAGERY SEGMENTATION USING UNET AND VISION TRANSFORMER

Mrs. M. Divya M.E.
SUPERVISOR,
Assistant Professor
Department of Computer Science and Engineering

SHAGHABETH HUSSAIN M
(220701256)

Problem Statement and Motivation

Problem Statement:

Accurately segmenting aerial images into meaningful land cover classes (like buildings, roads, vegetation, etc.) is a challenging task due to factors such as varying illumination, occlusions, complex terrain, and the high resolution of the data. Traditional image segmentation techniques often fail to handle these complexities effectively. There is a need for a more powerful and precise method that can manage both local details and global context in aerial imagery.

Motivation:

Aerial imagery is crucial for applications such as urban planning, agriculture, and disaster management. With the availability of high-resolution images, better tools are needed to automatically and accurately extract useful information. Deep learning models like UNet excel at capturing local features, while Vision Transformers (ViT) are great at understanding global context. Combining these strengths into a hybrid model could significantly improve the performance and accuracy of aerial image segmentation, making it more reliable for real-world use.

Existing System

The existing systems for aerial image segmentation mainly use either UNet or Vision Transformer (ViT) models individually. UNet is good at capturing local details through its encoder-decoder structure, making it effective for pixel-wise segmentation. However, it struggles with understanding the overall context of large, complex scenes. In contrast, ViT excels at capturing global patterns by processing image patches with self-attention, but it lacks precision in detecting fine details. Both models have limitations when used alone—UNet misses global context, and ViT misses local accuracy—making them less effective for challenging aerial images with varied lighting, terrain, and occlusions.

Both UNet and ViT have their strengths, but when used alone, they face limitations in handling the high variability, occlusions, and complex terrain found in aerial imagery. UNet may miss large-scale structures, while ViT may overlook fine edges or small objects. As a result, these existing systems may produce suboptimal results, especially in real-world applications where precise and reliable segmentation is crucial.

Objectives

- Accurately segment aerial images into land cover types like buildings, roads, vegetation, and water bodies.
- ✅ Develop a hybrid model combining UNet and Vision Transformer to capture both fine details and global context.
- ✅ Improve segmentation performance under challenging conditions such as varying lighting, shadows, occlusions, and complex terrains.
- ✅ Apply data augmentation (e.g., flipping, rotation, brightness changes) to enhance the model’s generalization to different environments.
- ✅ Compare the hybrid model's performance with existing models (UNet and ViT alone) using metrics like IoU, Dice Score, and Pixel Accuracy
- ✅ Build a system that can support real-world applications in urban planning, agriculture, and disaster response.

Abstract

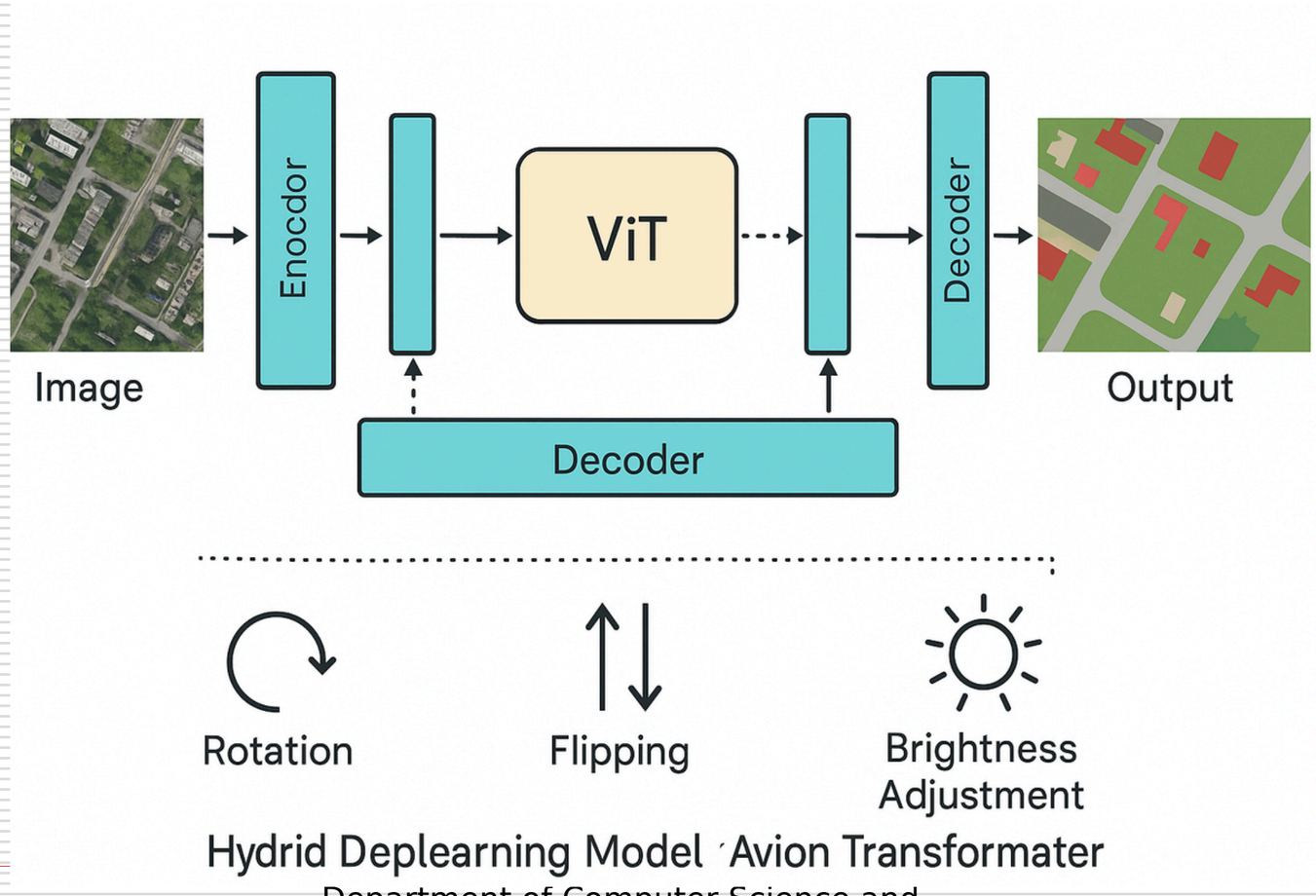
- Aerial images are very useful for things like city planning, farming, and tracking environmental changes. But these images are often complex and hard to analyze using traditional methods. This project presents a new approach that combines two powerful deep learning models—UNet and Vision Transformer (ViT)—to better segment (divide) these images into meaningful parts like roads, buildings, and greenery.

- UNet helps identify small, detailed features in the image, while ViT understands the overall context and large patterns. Together, they form a strong model that gives more accurate results. The method is tested on standard datasets and shows better performance than using UNet or ViT alone. It also works well even when the images have shadows, different lighting, or difficult terrain. This model can be used in real-world tasks like mapping land use, monitoring crops, and helping in disaster response.

Proposed System

- The proposed system is a hybrid deep learning model that combines two powerful techniques: UNet and Vision Transformer (ViT). UNet helps the model focus on small details and local features in the image, like edges and tiny structures. On the other hand, the ViT looks at the overall image to understand the bigger picture and context. By combining both, the model can accurately segment aerial images, identifying different areas like buildings, roads, and vegetation. The system also uses data augmentation methods like rotating, flipping, and adjusting brightness to make the model more adaptable to different conditions. After testing with standard datasets and evaluation methods, the system outperforms existing models, making it useful for tasks like urban planning, agriculture, and disaster response.

System Architecture



List of Modules

Core

- torch
- torchvision

Models

- timm (for Vision Transformer)
- segmentation-models-pytorch (for U-Net)

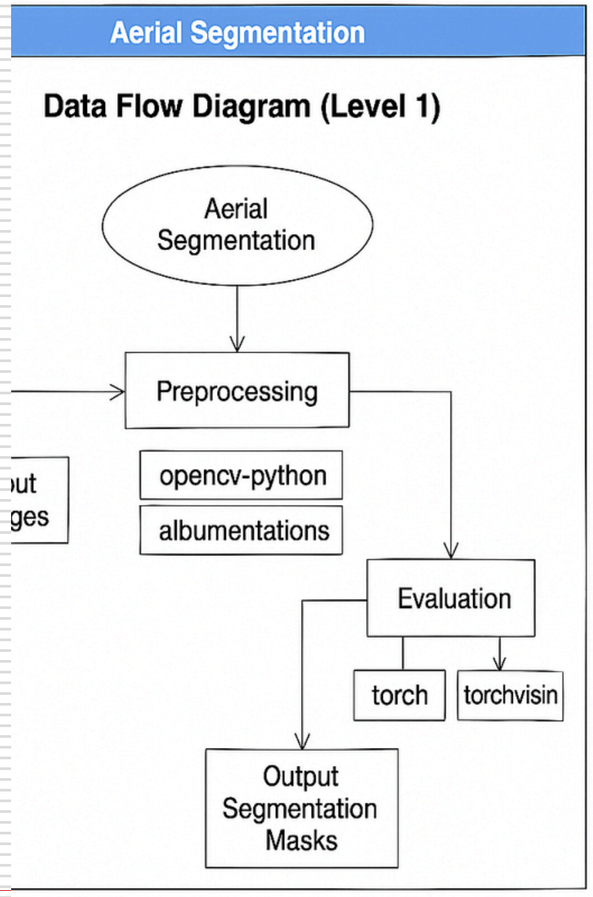
Image Handling

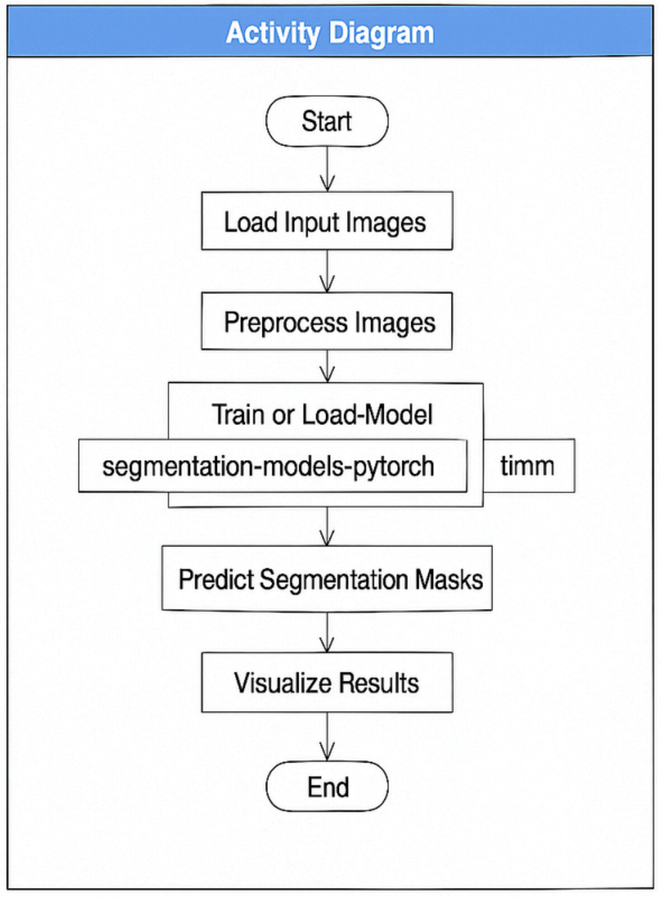
- opencv-python
- albumentations

Utilities.

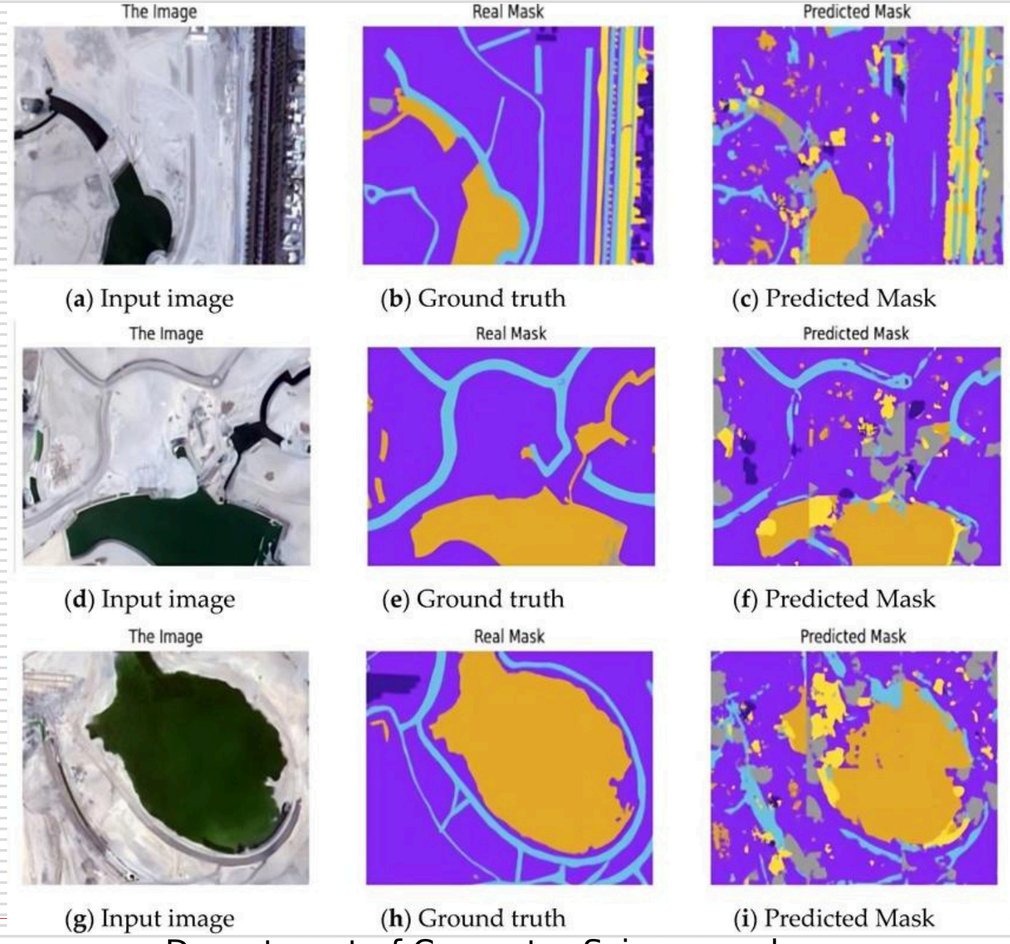
- matplotlib
- tqdm
- einops (for ViT tensor handling)

Functional Description for each modules with DFD and Activity Diagram





Implementation & Results of Module



Conclusion & Future Work

Conclusion:
This project successfully applied U-Net and Vision Transformers for segmenting aerial images. U-Net provided accurate results with fewer resources, while Vision Transformers captured long-range spatial features better. Both models showed strong performance in identifying objects from aerial views.

Future Work:

- Combine U-Net and Vision Transformer in a hybrid model for improved accuracy.
- Train on larger and more diverse datasets to boost generalization.
- Use post-processing techniques to refine segmentation boundaries.
- Deploy the model for real-time aerial analysis (e.g., drones or satellites).

References

1. U-Net: Convolutional Networks for Biomedical Image Segmentation Olaf Ronneberger, Philipp Fischer, Thomas Brox

2. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby (2021))

3. A Review of Deep Learning Approaches for Semantic Segmentation in Satellite Images by Heng Zhang et al. (2019)

Thank You