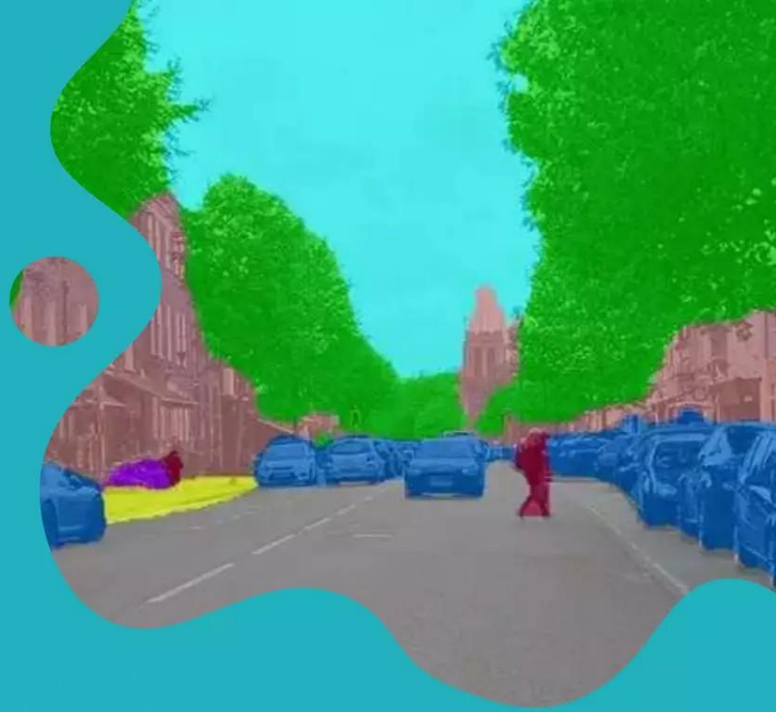
Semantic Segmentation Of Aerial Image



Introduction & Motivation

Aerial semantic segmentation is a computer vision task that involves classifying and labeling each pixel in an aerial image according to the type or category of the object or terrain it represents. This process goes beyond general object detection and aims to provide a detailed understanding of the visual content in aerial imagery.

Aerial semantic segmentation is crucial for various applications, such as urban planning, disaster management, agriculture, and autonomous vehicles. It enables the accurate labeling of objects and land cover in aerial imagery, leading to better decision-making, resource allocation, and improved safety and efficiency in these domains

Problem Statement

We aim to conduct a comparative analysis of various models developed through distinct methodologies to ascertain their respective efficacy under different scenarios. The selected models for concentrated evaluation include:

- UNet
- DeepLabV3+
- SegNet
- Swin

Literature Review

U-Net: Convolutional Networks for Biomedical Image Segmentation

- U-Net, a convolutional neural network architecture for biomedical image segmentation. The key aspect is its u-shaped architecture with a contracting path to capture context, and a symmetric expanding path enabling precise localization.
- The contracting path(Encoder) Captures context and features through convolution and pooling layers. The expansive path(Decoder) upsamples the features back to the original resolution using transposed convolutions, and concatenating high resolution features from the Encoder path. This allows the network to propagate context information to higher resolution layers.
- Key properties include:
  - End-to-end training from few images using aggressive data augmentation
  - Overlap-tile strategy to apply network to large images
  - Weighted loss function to separate fused cell segmentations
- The u-shaped architecture with concatenated contracting and expanding paths improves localization and makes efficient use of data.

Link for the Reference Paper: https://arxiv.org/pdf/1505.04597v1.pdf

Implementation

Data Set:

- We took a dataset from kaggle called Aerial Semantic Segmentation Drone Dataset.
- It has about 400 Original images taken from a drone, their labeled images, and a csv file consisting of the label names. There are a total of 23 labels in our dataset.
- We used the same dataset for all the models.
- The images are of size 4000x6000 in the dataset. We resized our images into different sizes (but mostly 512x512) for training the model.
- You can find our dataset at link: http://dronedataset.icg.tugraz.at/

Implementation

Unet:

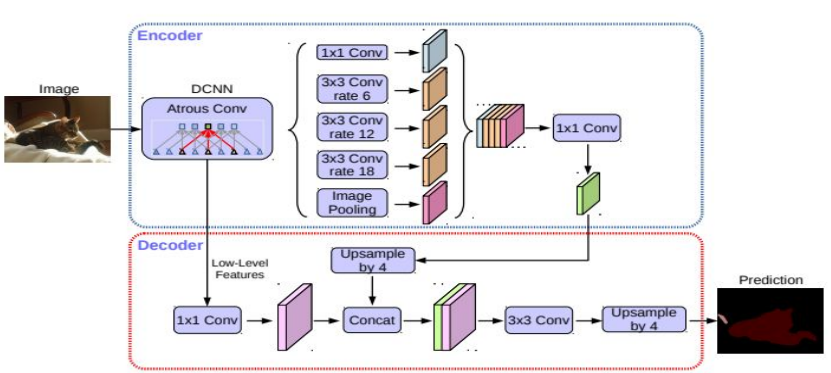
- Defined image, mask directory and loaded the data. Then splited the data into training set and test set(350:50). Then the data was preprocessed and images are resized. Final Image size is 512 * 512
- Then defined convolution block and upsampling block. Convolution block is composition of two conv layers and one maxpooling. Output is final output and skipped connection. Upsampling block is combination of transpose conv layer then merge then again 2 conv layer
- Taking these blocks as base, Unet model was build by using 5 convolution blocks and upsampling blocks and finally applying softmax.Total parameters are 34596311.
- Loss function used sparse categorical cross entropy function. Epochs = 150, batch size 16. Then the model was trained using resized trained dataset.
- Then model was evaluated using test dataset where accuracy obtained was approximately 79%.

Literature Review

DeepLabv3+: Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation

- DeepLabv3+ is a fully convolutional neural network (FCN) that uses atrous spatial pyramid pooling (ASPP) to capture context at multiple scales. ASPP consists of several parallel atrous convolutions with different rates, which allows the model to extract information from both fine-grained and coarse-grained features
- DeepLabv3+ also uses an encoder-decoder architecture to refine the segmentation results. The decoder upsamples the coarse-grained features produced by the encoder back to the original image size, and then uses these features to make more precise predictions.
- DeepLabv3+ delivers exceptional segmentation accuracy, effectively identifying and distinguishing different land cover classes in aerial imagery.

Implementation



DeepLabv3+:

- Git cloned the repository and divided our data into training, validation and test images and labels, defined functions of generating,reading and loading data with tensorflow
- Then defined the DeepLabv3+ network with ResNet-50 encoder (34 layers) and 2 ASPP modules, the decoder consisted of 4 bilinear upsampling layers. The total parameters turned out to be 11858007.
- The loss was calculated at each epoch while training and then plotted against epochs, defined inference and prediction functions to get the visualization of output segmented image
- The Hyperparameters are as follows: Image size of 512x512, batch size of 4, 50 training epochs, Adam optimizer with learning rate 0.001, Crossentropy loss function and achieved validation accuracy of 81.89

Literature Review

SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation

- SegNet consists of an encoder network, a decoder network, and a final pixel-wise classification layer.
- The key aspect is the decoder network, which uses pooling indices from the encoder layers to perform non-linear upsampling, mapping lower resolution encoder features to full input resolution features for pixel-wise classification.
- Advantages of this approach include improved boundary delineation, efficient use of memory by only storing pooling indices (vs encoder features), and ability to train end-to-end with stochastic gradient descent optimization.
- This paper also presents that:
    - Larger decoder networks improve accuracy, as does storing more encoder information
    - SegNet can match accuracy of other methods while requiring less memory

Link for the Reference Paper:https://arxiv.org/pdf/1511.00561v3.pdf

Implementation

SegNet:

- Cloned git repository to import all functions
- Defining Image and Mask Directories, Batch Size, Epochs, η, Determining Number of Classes, and Displaying Image Dimensions. Loading Image and Mask Paths, Creating a DataFrame, and Splitting into Training and Testing Sets.
- Created segnet model with 5 conv layers in encoder and 5 conv layers, 3 upsampling layers in decoder and zero padding, Batch normalization in both.Compiling the Model with Adam Optimizer. Training the Model using the Training Dataset and Validating on the Test Dataset.
- Visualizing Original Images, Predicted Masks, and Ground Truth Masks.
- The loss function used is categorical crossentropy.
- Hyperparameters are Adam Optimizer,batch size, input size of 512*512*3

Literature Review

Swin Transformer: Pure Transformer for Medical Image Segmentation:

- The Swin Transformer architecture introduces a hierarchical design that captures long-range dependencies efficiently. It divides the input image into non-overlapping patches and processes them through a series of hierarchical transformer layers. The key components include patch merging and patch partitioning, enabling the model to capture information at multiple scales.
- The hierarchical architecture is characterized by the following components Patch Embeddings,Transformer Layers,Patch Merging and Partitioning, Class Tokens
- The main advantages of this model is EfficientLong-RangeDependencies,Scalability,Competitive Performance.
- The main disadvantage of this model is Complexity,Interpretability.

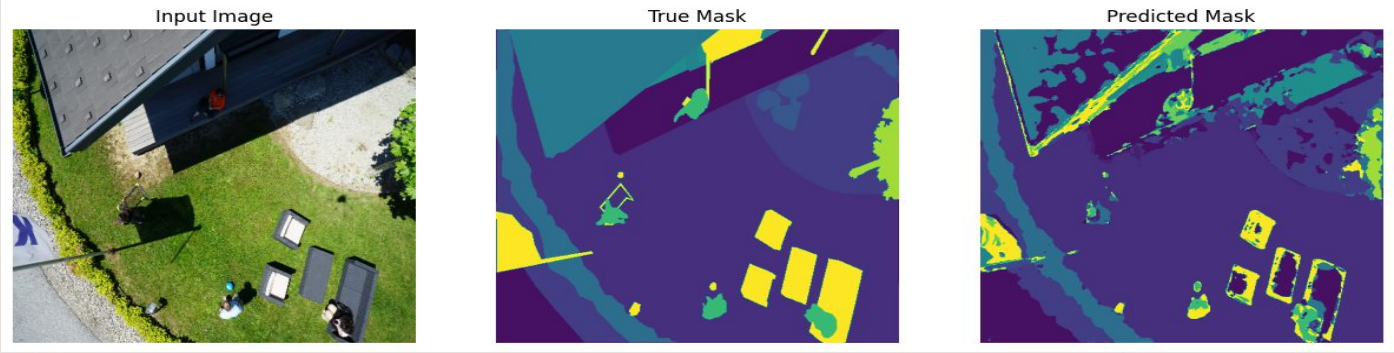
Link for the Reference Paper:https://arxiv.org/pdf/2105.05537.pdf

Implementation

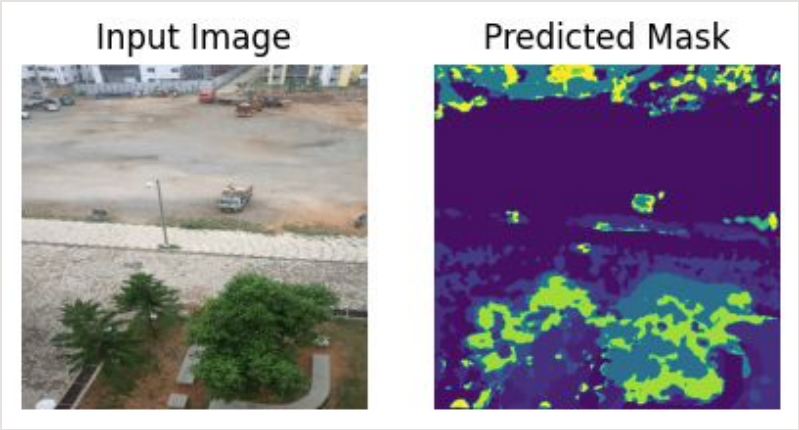
Swin Transformer:

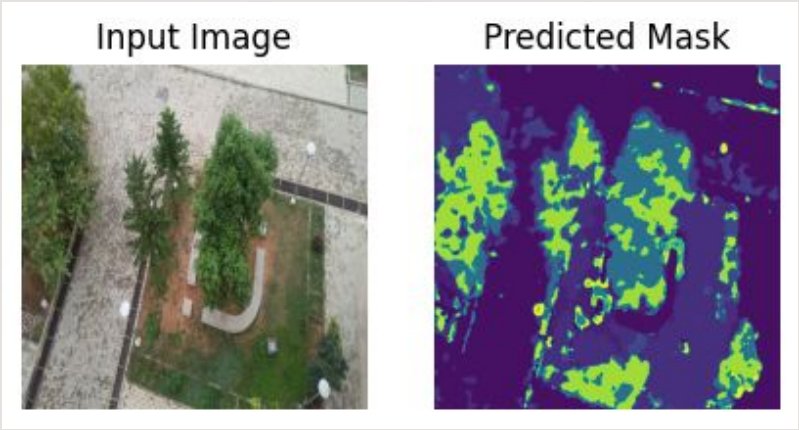
- Defining Image and Mask Directories, Batch Size, Epochs, Learning Rate, Height, and Width.Reading Sample Mask, Determining Number of Classes, and Displaying Image Dimensions. Loading Image and Mask Paths, Creating a DataFrame, and Splitting into Training and Testing Sets.
- Defining Functions to Read and Augment Images and Masks.Creating a Swin UNet Model using the keras_unet_collection library.Compiling the Model with Adam Optimizer, Focal Loss, and Metrics. Training the Model using the Training Dataset and Validating on the Test Dataset.
- Visualizing Original Images, Predicted Masks, and Ground Truth Masks.
- The loss function used is categorical crossentropy.
- Hyperparameters are Adam Optimizer,Droupout Rate,input size of 512*512*3, SoftMax is the activation function.

Unet Model predicted results

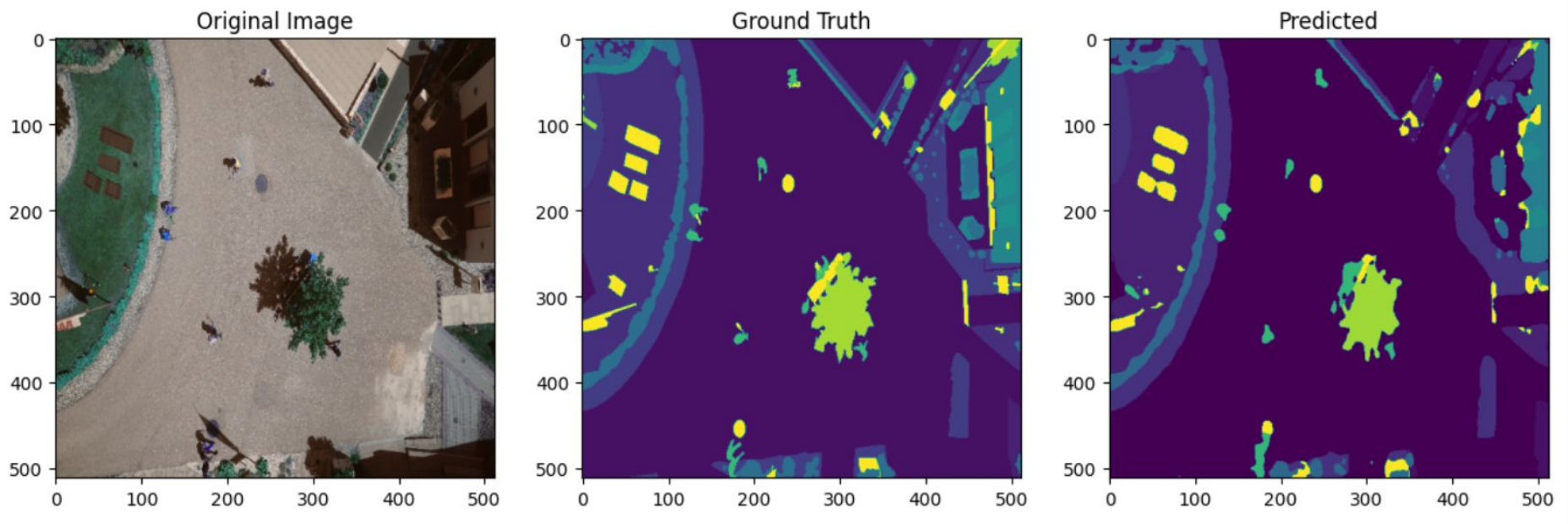


For Local Images:

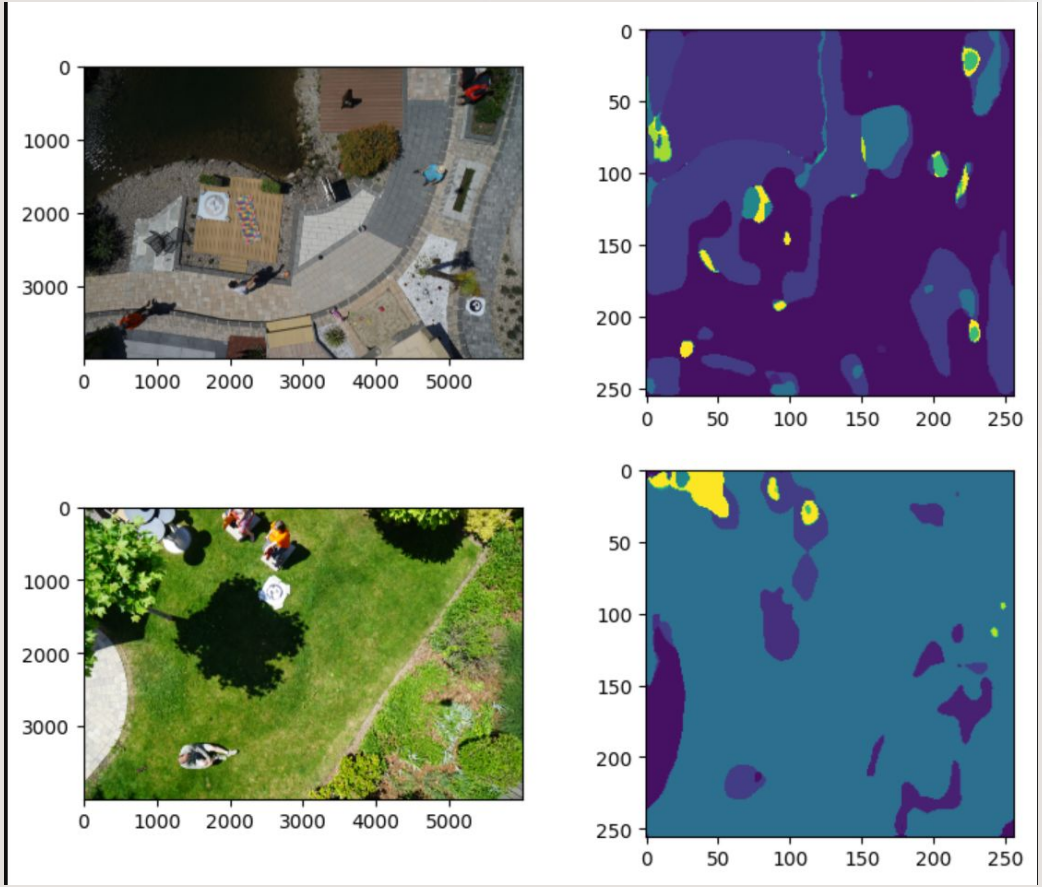




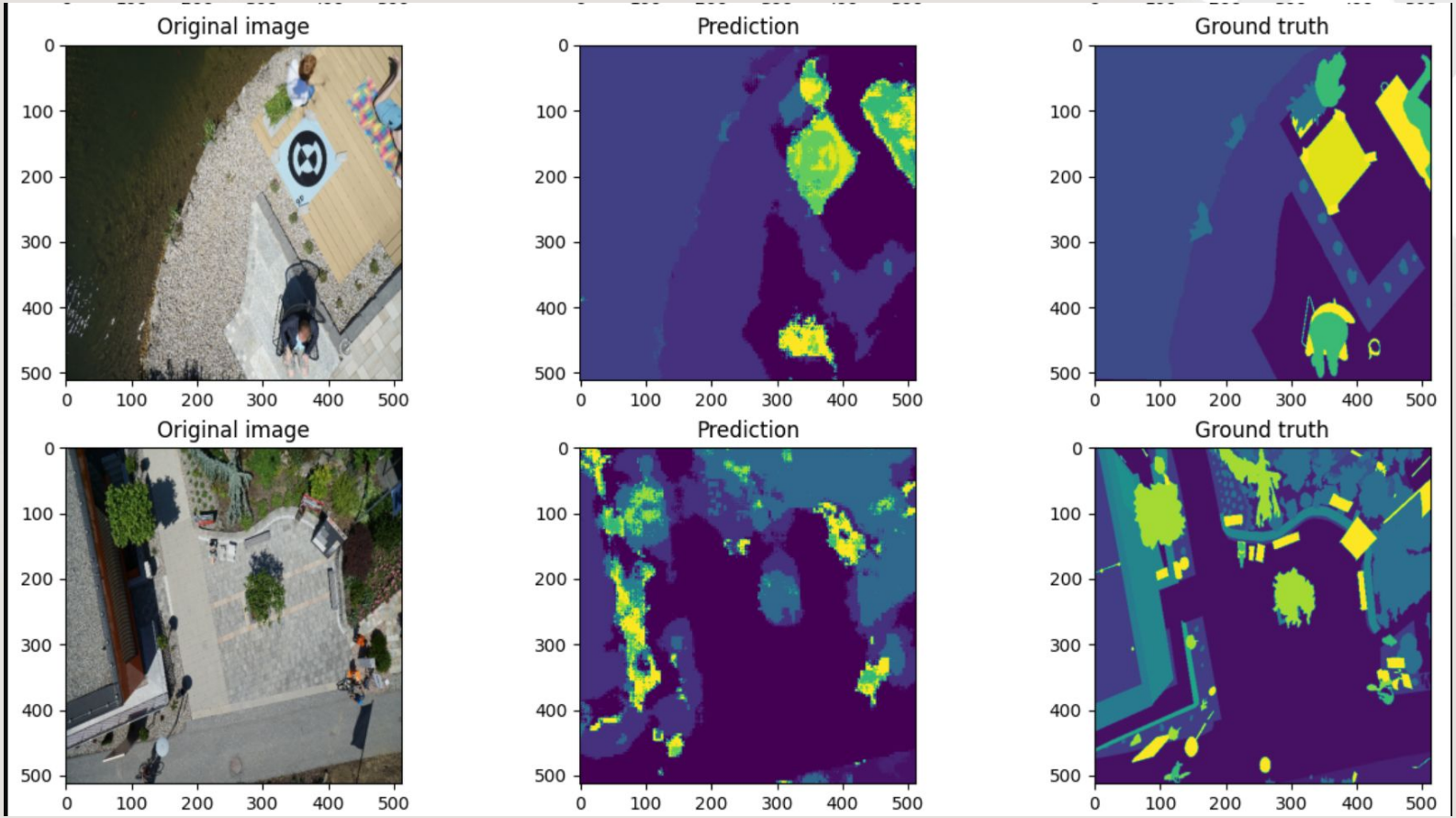
DeepLabV3+ Model predicted results



SegNet Model predicted results



Swin Transformer Model predicted results



Some Results from our codes

MODEL	EPOCHS	ACCURACY (TRAIN)	ACCURACY (VALID)	APPROX. TIME
UNET	150	79.34	78.82	120
DEEPLABV 3+	50	95.72	81.89	90
SEGNET	100	77.72	-	110
SWIN	10	73.63	70.18	120

Learnings

- Types of Semantic Segmentations (Semantic, Instance, Panoptic)

- Importance of GPUs in model trainings. Usage of Tensorflow, Pytorch

- U-net model is able to learn accurate and detailed segmentation even for small dataset. It has high accuracy for low resolution images.

- Deeplabv3+ is able to capture multi scale context information efficiently, which enables it to distinguish between objects with varying scale.

Learnings

- SegNet produces precise segmentation masks that are well-aligned with the input image.

- Swin transform maintain effective long range dependencies. It is scalable to handle input images of varying resolutions without significantly increasing computational complexity.

- Each model has there own advantage. We have to choose the optimal model considering factors such as computational resources, dataset size and nature of feature you need to capture.

Individual Contributions

N A POORNA CHANDRA (ee20btech11032): Swin Transformer

KSDS SIIDHHU (ee20btech11020) : UNet

VIBHUVAN M (ee20btech11055) : DeepLabV3+

ABHIROOP CHINTALAPUDI (ai20btech11005) : SegNet

THANK YOU