Semantic Segmentation Using DINO ViT and U-Net

Semantic segmentation assigns a class label to every pixel in an image. In this project, I focus on segmenting aerial drone images from the AeroScapes dataset into 11 different classes such as roads, buildings, cars , etc.

Dataset

The AeroScapes dataset contains aerial images and corresponding pixel-level segmentation masks. It includes 11 semantic classes.

- Images are resized to **518x518** pixels to align with the Vision Transformer (ViT) patch size (patch size 14 to 518/14 . Results in approx 37 tokens).

- Masks are resized using nearest neighbor interpolation to maintain discrete labels.

Data set link - https://www.kaggle.com/datasets/kooaslansefat/uav-segmentation-aeroscapes/data

Model Architecture

The model is a U-Net variant built with:

- **Encoder:** A pre trained DINO Vision Transformer (vit_base_patch14_dinov2), which outputs a final feature map of shape **[B, 768, 37, 37]** . This backbone is frozen during training.

- **Decoder:** Four upsampling blocks progressively increase spatial resolution through transposed convolutions:

    - up1: 768 → 512 channels, 37×37 → 74×74

    - up2: 512 → 256 channels, 74×74 → 148×148

    - up3: 256 → 128 channels, 148×148 → 296×296

    - up4: 128 → 64 channels, 296×296 → 592×592 (larger than original 518)

- To match the target image size, a **center crop** is applied to the 592×592 output feature map, cropping it to **512×512**. This ensures consistent output size for final prediction.

- **Final layer:** A 1×1 convolution reduces 64 channels to the number of classes (11), producing per-pixel class scores.

Training Details

- **Loss:** CrossEntropyLoss, to measure the difference between predicted class scores and ground truth pixel labels.
- **Optimizer:** Adam optimizer with learning rate 1e-4 for stable and efficient updates.
- **Training:** The ViT backbone is frozen to leverage pretrained features; only decoder layers are trained.
- **Epochs:** 11 epochs total.

Evaluation Metrics

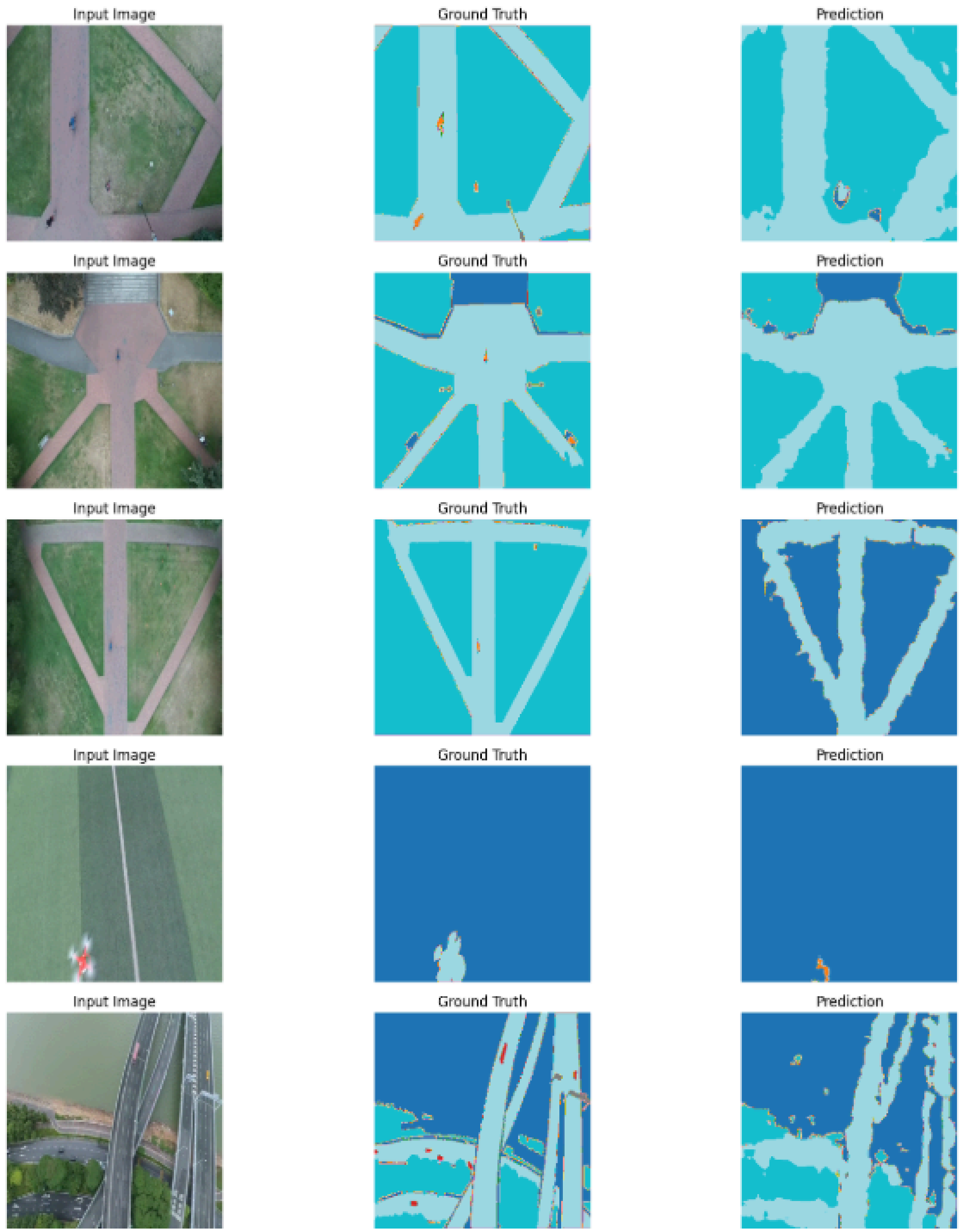
- **Mean Intersection over Union (mIoU):** Average overlap between predicted and true masks, calculated over all classes. It is a standard segmentation quality measure.
- **Pixel Accuracy:** Percentage of pixels classified correctly over the entire image.

Results

Epoch No.	Training Loss	Validation mIoU	Pixel-Wise Accuracy
1	1.1374	0.199	0.6408
2	0.4045	0.322	0.859
4	0.2148	0.354	0.8802
6	0.175	0.3627	0.8837
7	0.1617	0.3594	0.8853
8	0.1523	0.3683	0.8846
9	0.1464	0.3618	0.887
10	0.1372	0.3745	0.8824
11	0.1329	0.3823	0.889

The steady increase in both metrics shows that the model is learning useful segmentation patterns.

Some visual representations of the Input , Ground Truth and Predicted Images are shown below ,


Input Image
Ground Truth
Prediction
Input Image
Ground Truth
Prediction
Input Image
Ground Truth
Prediction
Input Image
Ground Truth
Prediction
Input Image
Ground Truth
Prediction

Challenges and Scope of Improvement

- Currently, the frozen backbone limits adaptation to the segmentation task; fine-tuning could improve results.
- Decoder lacks skip connections, which could help recover fine-grained spatial details.
- Alternative loss functions (Dice, Focal) could better handle class imbalance.
- Longer training with learning rate scheduling may boost performance.

Presentation Link - https://youtu.be/h-K6XyY-x-w?si=bvMmiE8Rc742R1_K