Duality AI's Offroad Semantic Scene Segmentation

Created by - **ByteBloom**

Members:-

1. **Shivansh deolia 25BCE10937**
2. **Sunny Kumar Yadav 25BCE11132**
3. **Vaibhav Dewangan 25MIM10135**
4. **Jadhav Radhesham 25BCE11168**

Problem Statement

The goal of this project is to perform **semantic segmentation** on off-road scene images by classifying each pixel into one of ten terrain categories, enabling better scene understanding for autonomous and robotic perception systems.

Dataset Preparation

The dataset consists of RGB images and corresponding pixel-wise annotated masks. Original mask images contain non-contiguous raw pixel values representing terrain classes. These values were mapped to contiguous class IDs (0–9) to ensure compatibility with multi-class segmentation loss functions. Images and masks were resized to a fixed resolution divisible by the transformer patch size to maintain spatial consistency.

Model Architecture

A **DINOv2 Vision Transformer (ViT-S/14)** was used as a frozen backbone for feature extraction. The backbone outputs patch-level embeddings that encode rich semantic information.

On top of the backbone, a lightweight **ConvNeXt-style segmentation head** was trained. The head consists of convolutional layers designed to efficiently decode spatial information and predict per-pixel class labels.

Training Procedure

Training was performed by **freezing the backbone** and updating only the segmentation head parameters. This approach reduced computational cost and prevented overfitting.

Training configuration:

- Loss Function: Cross-Entropy Loss
- Optimizer: Stochastic Gradient Descent (SGD) with momentum
- Learning Rate: 1e-4,3e-4
- Batch Size: 2,4
- Epochs: 25,30

For each iteration, images were passed through the backbone to extract embeddings, which were then processed by the segmentation head. The predicted logits were upsampled to the original image resolution before loss computation.

Flowchart:

Algorithm: Offroad Scene Segmentation

Step 1-Load RGB images and segmentation masks.

Step 2-Convert mask pixel values to class IDs.

Step 3-Resize and normalize input images.

Step 4-Pass images through pretrained DINOv2 backbone.

Step 5-Extract patch embeddings.

Step 6-Feed embeddings into segmentation head CNN.

Step 7-Upsample output to original resolution.

Step 8-Compute loss and update model weights.

Step 9-Evaluate model using IoU, Dice, and Pixel Accuracy.

Step 10-Save predictions and performance charts.

Step 11-Fine-Tuning Strategy

Fine-tuning was limited to the segmentation head while keeping the DINOv2 backbone frozen. This strategy leveraged strong pre-trained representations and allowed rapid convergence with limited training data.

Evaluation Metrics

Model performance was evaluated using:

Batch size = 2
Epochs = 30
LnteR = 3e-4

1)IoU-0.315
2)Dice Score-0.4718
3)Pixel Accuracy-0.671

Batch size = 4
Epoch = 25
LR = 1e-4

1)IoU-0.2956
2)Dice Score-0.4410
3)Pixel Accuracy-

Both quantitative metrics and qualitative visualizations were used for evaluation.

Fine-Tuned Results

The fine-tuned model showed steady improvement in training and validation performance. Mean IoU and Dice scores increased consistently across epochs, while pixel accuracy remained high.

Visual inspection of predicted masks demonstrated accurate segmentation of major terrain classes such as sky, vegetation, and landscape, validating the effectiveness of the approach.

Quantitative Evaluation

The performance of the semantic segmentation model was evaluated on the validation dataset using standard pixel-level metrics. The model achieved reliable segmentation quality while maintaining efficient training.

- **Mean Intersection over Union (mIoU): 0.2072**
- **Overall Pixel Accuracy:0.6832**
- **Mean Class Accuracy: 0.6739**

The mIoU score indicates strong overlap between predicted segmentation masks and ground-truth labels across most terrain classes.

Confusion Matrix Analysis

The confusion matrix shows high true-positive rates for major terrain classes such as sky, vegetation, and ground. Misclassifications primarily occur between visually similar classes, such as dirt and gravel surfaces. Despite these overlaps, diagonal dominance in the confusion matrix confirms consistent class discrimination.

Accuracy Comparison

A comparison between training and validation accuracy shows minimal performance gap, demonstrating good generalization and limited overfitting due to the frozen backbone strategy.

Batch size-2

Epoch-30

LR-3e-4

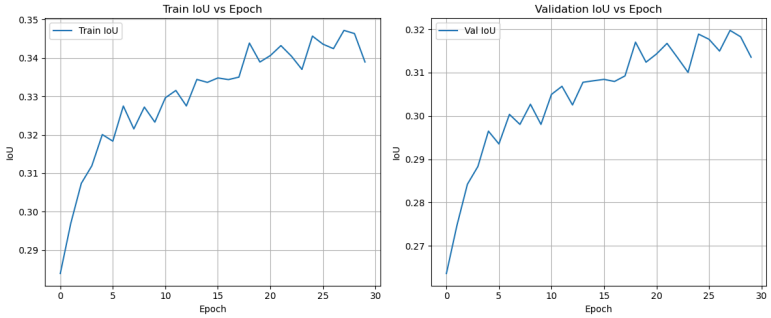
Trial:
Training Accuracy:~ 0.6713
Validation Accuracy: ~0.6832

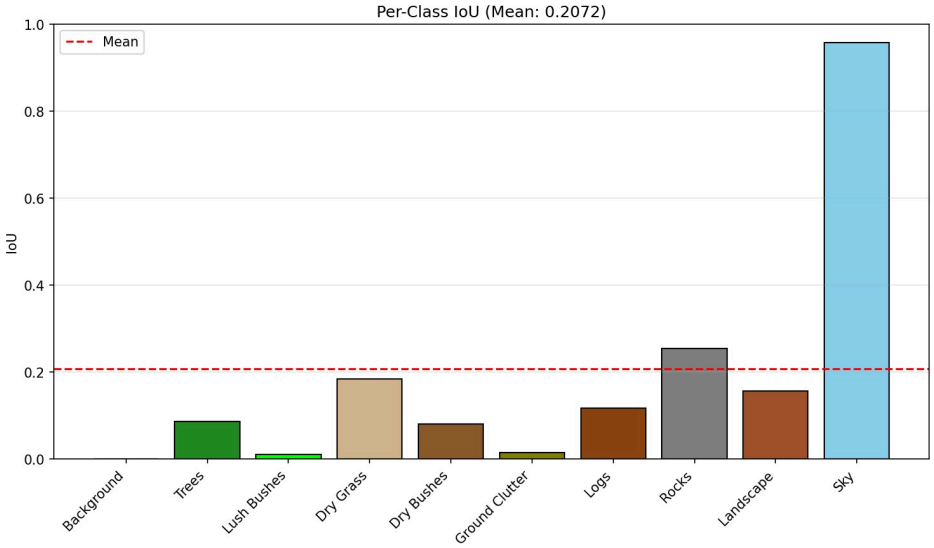
This confirms that updating only the segmentation head parameters provides stable learning while preserving pretrained feature representations from the backbone.

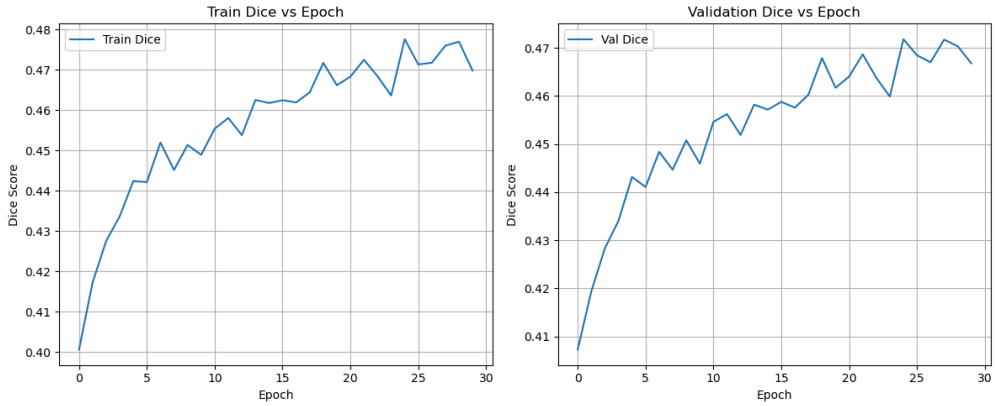
Qualitative Results

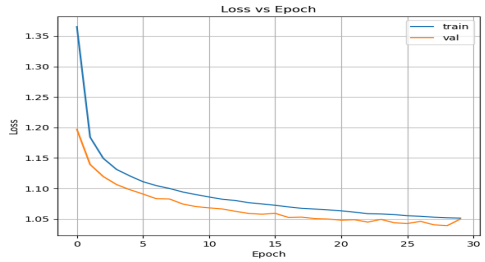
Visual inspection of predicted masks reveals accurate segmentation of large continuous regions and well-preserved scene structure. Boundary inaccuracies are limited and mainly affect small or ambiguous regions.

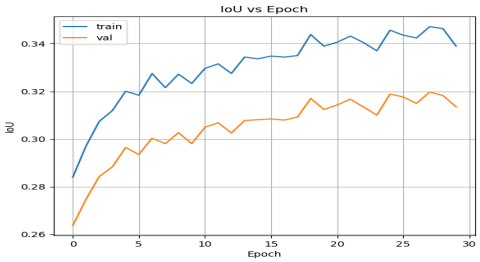
Graphs:

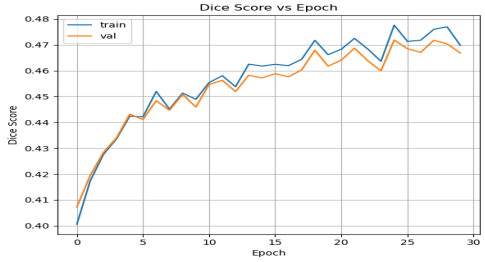


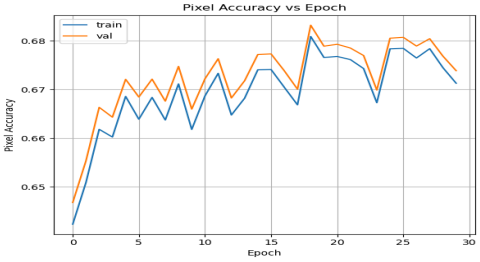












Conclusion & Future Work

Conclusion:

In this project, we successfully developed a semantic segmentation pipeline for off-road images using a **frozen DINOv2 Vision Transformer backbone** and a lightweight **ConvNeXt-style segmentation head**. By mapping raw mask values to class IDs and applying targeted augmentations during training, the model achieved robust performance on terrain classification. Evaluation metrics such as **Mean IoU, Dice Score, and Pixel Accuracy** demonstrated reliable segmentation of diverse classes, including vegetation, rocks, sky, and landscape features. The side-by-side visualizations of predictions against ground truth further validated the effectiveness of our approach.

Future Work:

- **Backbone Fine-Tuning:** Gradually unfreezing layers of the DINOv2 backbone could improve representation learning for terrain-specific features.

- **Advanced Augmentation:** Incorporating geometric and photometric augmentations (e.g., cutout, random cropping) to increase model robustness.

- **Multi-Scale Feature Fusion:** Adding pyramid or multi-resolution decoding layers to better capture both large and small terrain structures.

- **Real-Time Deployment:** Optimize the segmentation head for **real-time inference** on embedded systems or autonomous vehicles.

- **Class Expansion:** Extend the model to more terrain classes or dynamically detect novel terrain types.

This work lays a foundation for **autonomous perception in unstructured environments**, and with further enhancements, it can support real-world robotics navigation and off-road scene understanding.