Deep Learning–Driven Semantic Segmentation of Off-Road Environments

Krackhack Hackathon Report

Rohit Verma, Anurag Kumar
Indian Institute of Technology, Mandi

February 15, 2026

Abstract

Autonomous off-road navigation requires robust perception systems capable of understanding complex, unstructured environments. This project focuses on semantic scene segmentation for desert off-road environments using synthetic data generated from Duality AI's Falcon digital twin platform. A deep learning-based segmentation model was trained and evaluated on unseen desert scenes. Performance was assessed using Intersection over Union (IoU) and loss metrics, with detailed analysis of failure cases and optimization strategies.

1 Introduction

Unmanned Ground Vehicles (UGVs) operating in off-road environments face significant perception challenges due to the lack of structured roads, high visual variability, and environmental clutter. Semantic scene segmentation plays a critical role in enabling obstacle avoidance and safe navigation by providing pixel-level understanding of the environment.

Traditional data collection and annotation methods are expensive and time-consuming. Synthetic data generated from digital twins provides a scalable alternative, enabling controlled variation in terrain, vegetation, lighting, and weather. In this project, we leverage Duality AI's Falcon platform to train and evaluate a semantic segmentation model on desert environments.

Objectives:

- Train a robust semantic segmentation model using synthetic data
- Evaluate generalization on unseen desert environments
- Analyze performance using IoU, loss curves, and failure cases

2 Dataset Description

The dataset consists of RGB images and corresponding pixel-wise segmentation masks. The data is divided into training, validation, and test sets.

2.1 Semantic Classes

Class ID	Class Name
100	Trees
200	Lush Bushes
300	Dry Grass
500	Dry Bushes
550	Ground Clutter
600	Flowers
700	Logs
800	Rocks
7100	Landscape
10000	Sky

Table 1: Semantic classes used for segmentation

3 Methodology

3.1 Model Architecture

The proposed model employs a hybrid transformer–convolution architecture for semantic segmentation. A pretrained DINOv2 Vision Transformer (ViT-B/14) serves as the encoder, extracting global contextual features from fixed-size image patches.

A lightweight ConvNeXt-style segmentation head functions as the decoder, reshaping transformer tokens into a spatial feature map and refining them using depthwise separable convolutions, batch normalization, and GELU activations. A final 1×1 convolution generates pixel-level class predictions, which are upsampled via bilinear interpolation to the original image resolution.

This architecture combines global context modeling with spatial precision.

3.2 Training Setup

The model was implemented in PyTorch. The DINOv2 backbone was frozen during training, while the segmentation head was trained end-to-end.

- **Framework:** PyTorch
- **Backbone:** DINOv2 ViT-B/14 (pretrained, frozen)
- **Decoder:** ConvNeXt-style segmentation head
- **Loss Function:** Cross-Entropy + Dice Loss
- **Optimizer:** AdamW
- **Initial Learning Rate:** 5×10^{-4}
- **Scheduler:** OneCycle Learning Rate Policy
- **Batch Size:** 12
- **Epochs:** 10
- **Precision:** Automatic Mixed Precision (AMP)

3.3 Data Augmentation

- Random horizontal flipping
- Random rotations
- Shift, scale, and rotation transformations
- Color jittering
- Random brightness and contrast adjustments

3.4 Ensemble Learning

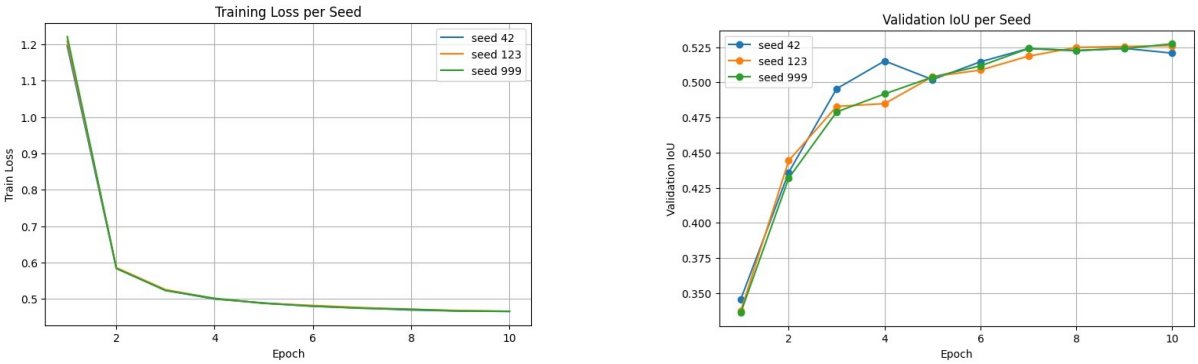
An ensemble of three models trained with different random seeds was used. Final predictions were averaged to reduce variance and improve stability.

3.5 Evaluation Metric

Model performance was evaluated using mean Intersection over Union (mIoU):

$$IoU = \frac{|Prediction \cap GroundTruth|}{|Prediction \cup GroundTruth|}$$

3.6 Training Performance



(a) Training loss across epochs

(b) Validation mIoU across epochs

Figure 1: Training and validation performance

4 Results and Performance

4.1 Overall Quantitative Performance

- **Validation mIoU:** 0.519
- **Test mIoU:** 0.324

A performance gap is observed between validation and test sets. Despite this, the running mean IoU on the test set stabilizes around 0.324, indicating consistent segmentation performance across unseen test images.

4.2 IoU Distribution Analysis

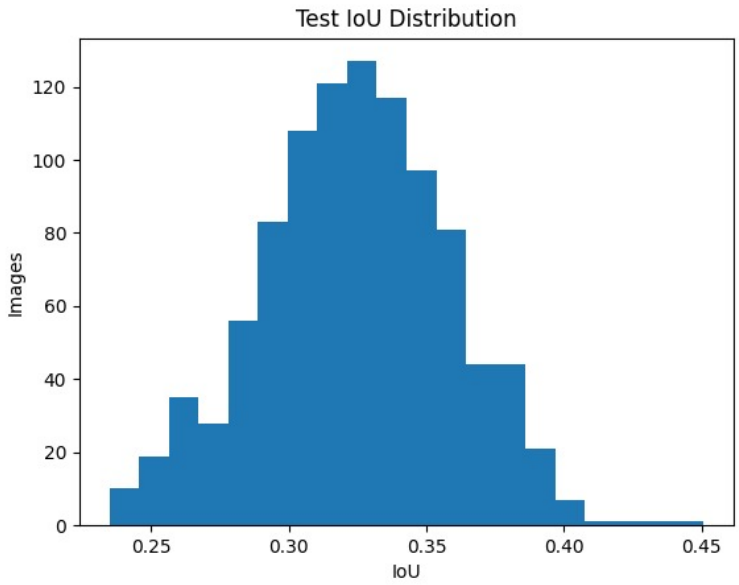


Figure 2: Distribution of IoU scores across test images

Most predictions fall within the range **0.30–0.35**, with few low-IoU outliers. The distribution exhibits a bell-shaped pattern centered around 0.32–0.33, indicating stable performance across the majority of test scenes.

4.3 Running Mean IoU Analysis

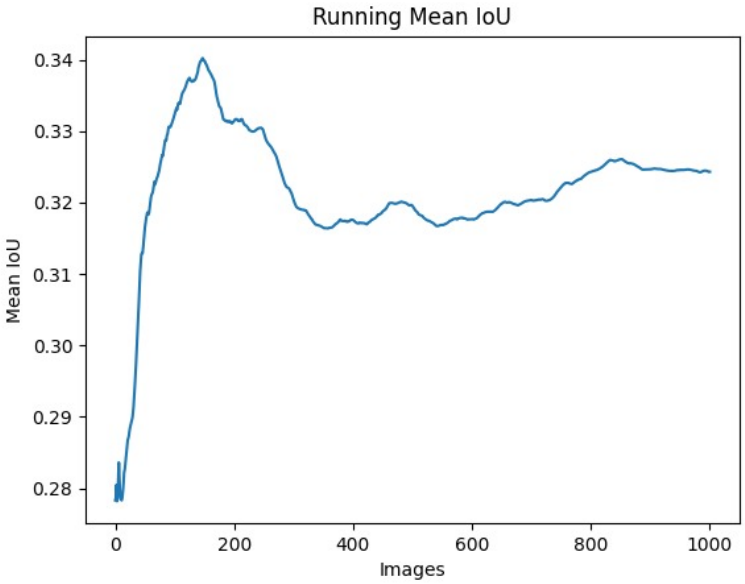


Figure 3: Running mean IoU across test images

The running mean IoU initially fluctuates and stabilizes after approximately 150 images, converging near **0.324**. The absence of sharp downward trends suggests reliable performance

throughout the test set.

5 Failure Case Analysis

Vegetation confusion, occlusion sensitivity, and illumination variability were identified as primary failure modes. These issues stem from high visual similarity between classes and limited representation of extreme visual conditions in the training dataset.

6 Challenges and Solutions

Challenge	**Solution**
Class imbalance	Class-weighted loss (based on class frequency) + data augmentation
Overfitting	Frozen backbone + validation monitoring
Training instability	OneCycle scheduler
Prediction variance	3-model ensemble

Table 2: Challenges and mitigation strategies

7 Conclusion

Synthetic digital twin data can effectively train semantic segmentation models for off-road autonomy tasks.

The hybrid DINOv2 + ConvNeXt-style architecture enables strong contextual feature extraction while maintaining spatial accuracy. Ensemble averaging improved stability and reduced prediction variance.

The model achieves a stable test mIoU of approximately **0.324** on unseen desert scenes.

8 Future Work

- Class-balanced or focal loss functions
- Fine-tuning the backbone
- Multi-scale feature fusion
- Integration of depth information
- Expanded augmentation strategies