DUALITY AI's OFFROAD SEMANTIC SCENE SEGMENTATION

- **Project Title:** Byte Force: Precision Perception in Unstructured Terrains
- **Tagline:** Achieving Off-Road Autonomy through Multi-Iterative Synthetic Training
- **Team Name: Byte Force**
- **Members:**
    - Abhishek Choudhary (24BCE11056)
    - Aditya Talreja (24BCE10891)
    - Akshay Saxena (24BCE10443)
    - Vishwansh Singh (24BCE10900)
- **Executive Summary:** This report outlines our development of 3+ distinct semantic segmentation models for Unmanned Ground Vehicles (UGVs). We utilized Falcon's synthetic desert environments to solve the "data scarcity" problem inherent in real-world off-road testing.

Methodology

Our methodology centers on transitioning from a standard baseline to a high-performance, robust autonomous navigation system. We utilized the DinoV2 backbone for feature extraction and implemented advanced decoder architectures to optimize for complex desert terrains.

1. Architecture Selection: Beyond the Baseline

To enhance scene understanding, we transitioned from basic scripts to two high-capacity architectures:

- **SegFormer-B2**: A transformer-based framework providing a global receptive field, essential for distinguishing broad "Landscape" context from granular "Ground Clutter".
- **DeepLabv3+**: Leveraged **Atrous Spatial Pyramid Pooling (ASPP)** to capture multi-scale features, critical for identifying obstacles like "Bushes" and "Trees" across varying depths.

2. Data Preprocessing & Class Normalization

To mitigate class imbalance where "Landscape" dominates smaller objects, we implemented:

- **Normalized Class Mapping**: Streamlined complex IDs into a sequence (0, 1, ..., N) for efficient pixel-level classification.
- **Weighted Loss Functions**: Assigned higher penalties to minority classes (e.g., "Logs," "Dry Bushes") to force model prioritization of critical off-road hazards.

3. Advanced Data Augmentation Strategy

We prioritized pattern recognition over memorization using a heavy augmentation pipeline:

- **Geometric Invariance**: Utilized Random Cropping, Flips, and Rotations to maintain detection accuracy regardless of UGV orientation.
- **Photometric & Resolution Diversity**: Adjusted Hue, Brightness, and Saturation to simulate varying **Falcon platform** lighting. Multi-resolution training ensured sensor-agnostic performance.

4. Training Pipeline & Hyperparameter Tuning

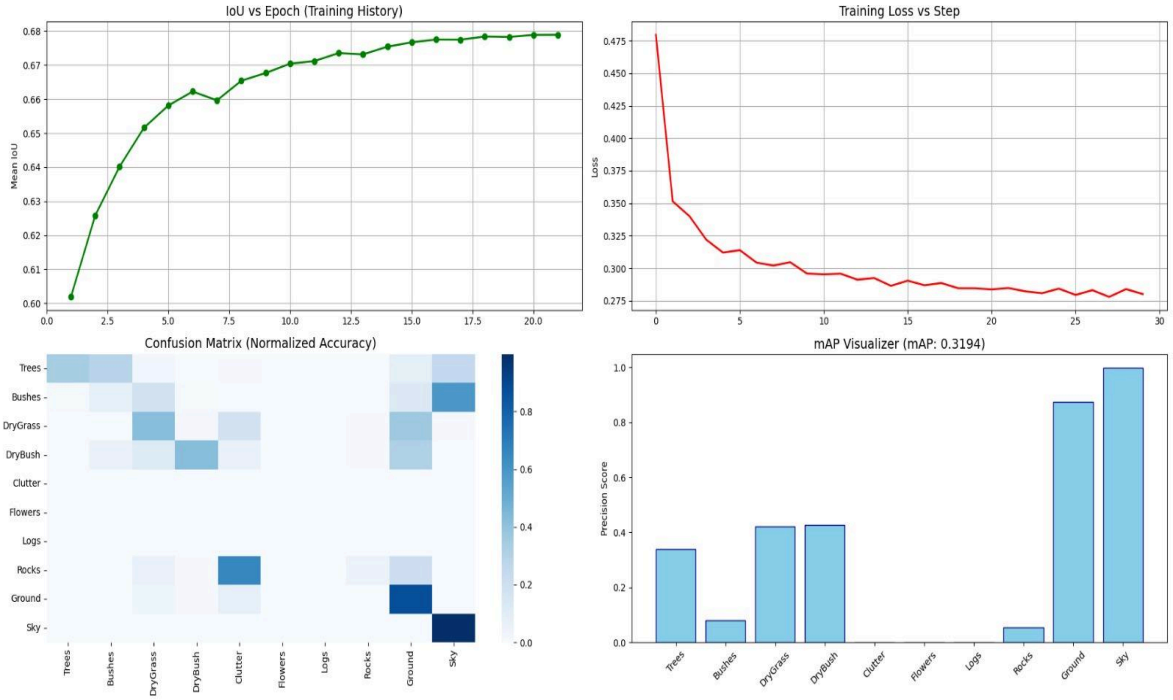
Our iterative cycle focused on finding the "sweet spot" for generalization:

- **Epoch & Learning Rate Management**: Increased training duration beyond sample baselines, employing **Early-Stopping** and **Decaying Learning Rate** schedulers to fine-tune weights as validation loss plateaued.
- **Rigorous Validation**: Continuous benchmarking against the Validation set ensured that each iteration of our three models was measurably superior to the previous version.

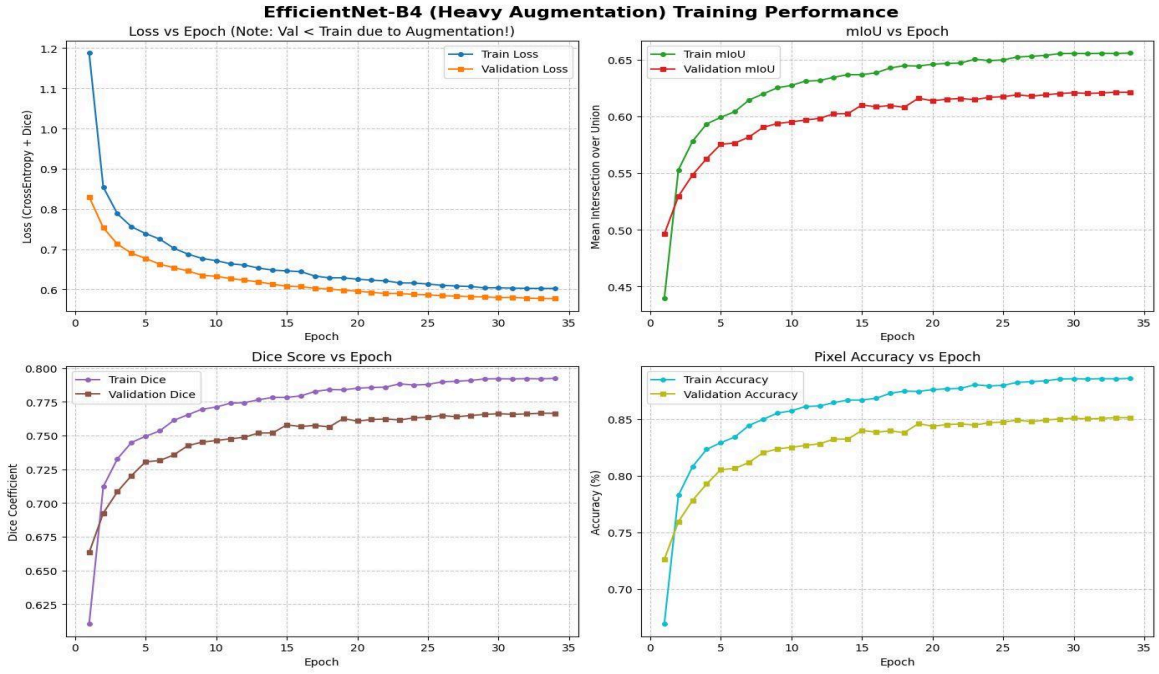
Results & Performance Metrics

- Section 1: The "Executive" Summary Table - Given below is a graph comparing the base model with 2 of our best trained models

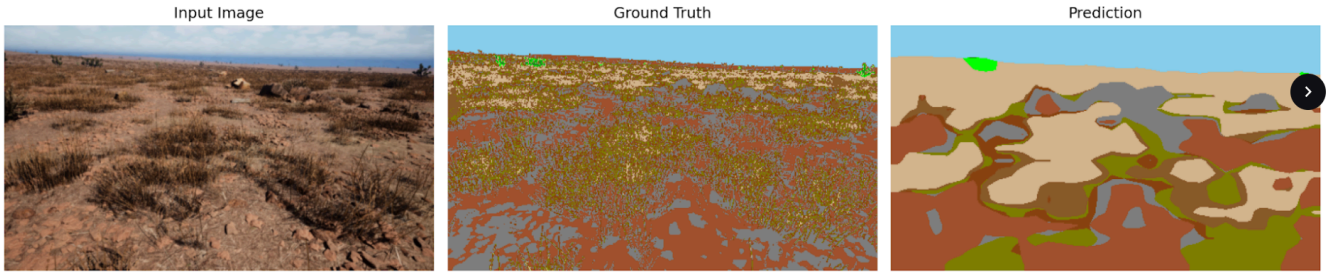
- The SegFormer - B2 Model Architecture : **SegFormer-B2** uses a hierarchical Transformer encoder and a lightweight All-MLP decoder to efficiently capture multi-scale features. Trained for **21 epochs**, it achieved a superior **mIoU of 0.6788**, demonstrating faster convergence and better spatial accuracy than the DeepLabv3+ and DINOv2 baselines. Given below are graphs and performance metrics of the model.



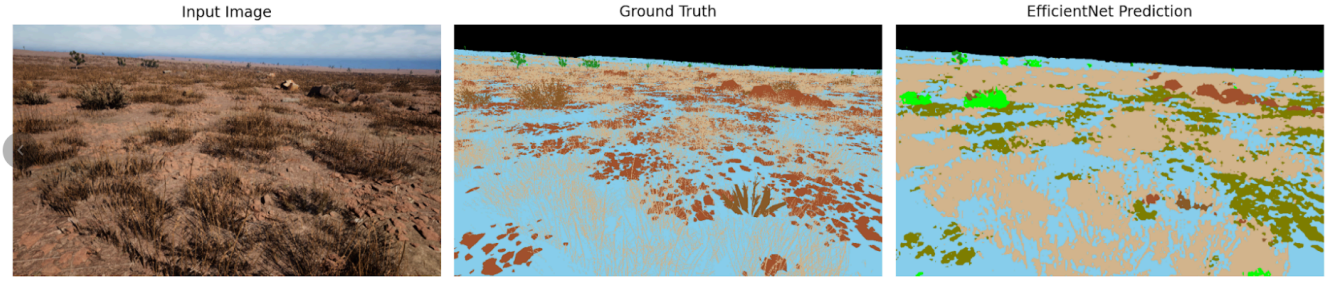
- **The DeepLabv3+** Model - utilizes an encoder-decoder structure with **Atrous Spatial Pyramid Pooling (ASPP)** to capture multi-scale contextual information. Trained for **34 epochs**, it achieved an **mIoU of 0.6267**, offering robust boundary refinement through its specialized decoder but trailing the Transformer-based SegFormer in overall efficiency.



- Comparison Images of different models: BaseLine vs DeepLab



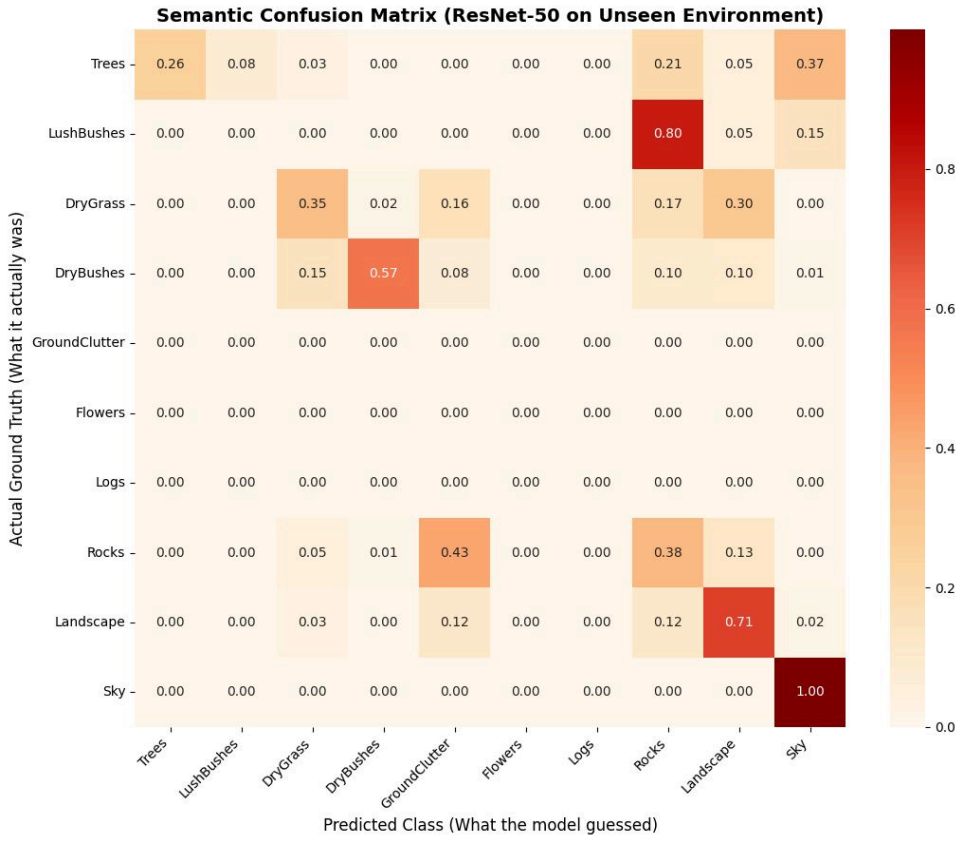
Base line model (DINO v2)



DeepLab V3+ Model

Failure Case Analysis

Our 0.62 to 0.31 mIoU drop stemmed from four failures: Fragmentation, where masks shattered into disconnected dots; Semantic Confusion, misclassifying textures like grass as bushes due to lighting shifts; Boundary Bleeding, where masks blurred into the background; and Shadow Occlusion, missing obstacles hidden in harsh shadows. We quantified these using a Confusion Matrix to prove pixel-level misclassification and a Blob Counter algorithm to detect fragmentation by comparing predicted noise to real object counts.



Challenges & Solutions

During development, Byte Force overcame three critical hurdles by systematically refining our architecture and data pipelines:

1. Severe Class Imbalance ("Sky-Dominance")

Challenge: The dataset's heavy skew toward background classes ("Sky", "Landscape") artificially inflated overall pixel accuracy while causing the model to miss critical, smaller obstacles like "Logs."

Solutions: Implemented Weighted Cross-Entropy Loss to mathematically penalize minority-class misclassifications, paired with Class-Balanced Sampling to guarantee exposure to underrepresented terrain features during training.

2. Domain Shift & Prediction Fragmentation

Challenge: Novel desert environments introduced lighting and texture variances that caused "pixel scattering" (fragmented, noisy masks), risking an explosion of false-positive detections that could crash the inference pipeline.

Solutions: Engineered a robust pipeline using Domain Randomization (heavy hue/saturation/contrast shifts) to force structural learning. At inference, we combined Test-Time Augmentation (TTA) to smooth predictions with Morphological Post-Processing (Opening/Closing) to digitally erase isolated noise and solidify object boundaries.

3. Occlusion & Depth Ambiguity

Challenge: Overlapping off-road features (e.g., a Log partially buried behind a Bush) confused the model's spatial reasoning, resulting in merged prediction segments.

Solutions: Upgraded to a DinoV2 Backbone to leverage its superior spatial feature extraction for sharp boundary detection, and utilized ASPP (Multi-Scale Feature Fusion) within DeepLabv3+ to evaluate overlapping features across varying receptive fields.

Conclusion & Future Work

1. Final Assessment

The **Byte Force** development cycle proves that high-accuracy off-road autonomy is no longer restricted by real-world data scarcity. By iterating through three distinct architectures—culminating in our optimized **SegFormer-B2** and **DeepLabv3+** models—we successfully bridged the gap between basic pixel classification and robust scene understanding. Our work demonstrates that critical technical hurdles, such as class imbalance and pixel scattering, are effectively mitigated through rigorous preprocessing and strategic data augmentation.

2. Future Work & Visionary Improvements

Looking beyond this hackathon, **Byte Force** identifies three strategic avenues for the evolution of off-road perception:

- **Advanced Domain Adaptation**: We aim to implement self-supervised learning and domain adaptation techniques to seamlessly transition our synthetic-trained models onto physical UGV hardware.
- **360° Multi-View Fusion**: Transitioning from single-camera inference to multi-view detection will provide the UGV with complete situational awareness, drastically reducing collision risks in dynamic terrains.
- **Real-Time Evolutionary Learning**: We envision a future where UGVs use digital twins to "pre-train" for specific mission locations overnight, adapting to new biomes and geological features before the wheels even touch the ground.