# Off-Road Terrain Analysis:

## Semantic Segmentation via SegFormer

Team Six-Seven

Lokesh, Somanshu, Himank, Vitthal

February 15, 2026

**Abstract**

Autonomous navigation in unstructured desert environments demands accurate spatial understanding. This project presents a semantic segmentation pipeline using **SegFormer-B3**, a hierarchical Transformer-based architecture, achieving a mean IoU of 65.96% and pixel accuracy of 82.13% on ten terrain classes at 960 × 540 resolution. Through aggressive data augmentation strategies—expanding the effective dataset by 7×—we successfully addressed class imbalance and environmental variations, demonstrating that strategic data preprocessing combined with modern architectures yields robust terrain classification for autonomous systems.

# 1 Introduction

Autonomous off-road vehicles must process complex visual data without structured landmarks like lane markings or traffic signs. In desert terrain, distinguishing between visually similar surfaces—sand versus dry grass, rocks versus ground clutter—is critical for safe navigation and obstacle avoidance.

Traditional CNNs struggle with the global context required for such disambiguation, relying on local receptive fields that limit scene-level understanding. This project implements **SegFormer-B3**, a hierarchical Transformer architecture that combines global attention mechanisms with multi-scale feature extraction, processing static 960 × 540 images to classify ten terrain-specific classes.

# 2 Technical Methodology

## 2.1 SegFormer-B3 Architecture

SegFormer-B3 consists of two key components that enable effective terrain segmentation:

**Hierarchical Transformer Encoder:** Produces multi-scale features at four spatial resolutions, capturing both fine-grained textures (rocks, vegetation) and coarse semantic regions (sky, landscape). An efficient self-attention mechanism reduces complexity from

$\mathcal{O}(N^2)$ to $\mathcal{O}(N^2/R)$, making high-resolution processing feasible. The positional encoding-free design improves generalization across varying input dimensions.

**Lightweight MLP Decoder:** Aggregates multi-scale encoder features through upsampling and concatenation, then fuses them via a compact multi-layer perceptron. This eliminates parameter-heavy convolutional decoders while maintaining strong semantic accuracy.

### 2.2 Training Configuration

- **Hardware:** NVIDIA Tesla P100 GPU (16GB)
- **Resolution:** $960 \times 540$ (native dataset resolution)
- **Training:** 10 epochs, batch size 32, gradient accumulation steps 2 (effective: 64)
- **Optimizer:** AdamW, learning rate $6 \times 10^{-5}$, weight decay 0.01, cosine annealing, warmup ratio 0.1
- **Loss:** Cross-Entropy, FP16 mixed precision

### 2.3 Data Augmentation Strategy

To address limited training data (2,857 train, 317 validation, 1,002 test images) and class imbalance, we implemented aggressive augmentation, expanding the dataset to approximately 19,710 samples (7× increase):

**Geometric Augmentations:** Horizontal flip (50%), vertical flip (20%), rotation ±30 (70%), scaling 0.8-1.2× (50%), random crop 70-95% (40%)

**Photometric Augmentations:** Color jitter—brightness/contrast/saturation ±40%, hue ±15% (70%); Gaussian blur kernel 3-7px (50%); sharpness 0.5-2.0× (50%); grayscale (10%)

**Expansion Strategy:** Training set (3× augmented + 1× original), validation (3× + 1×), test (5× + 2×). Higher test augmentation improved domain-specific generalization.

This strategy mitigated class imbalance, enhanced robustness to lighting/weather variations, and promoted geometric invariance while serving as effective regularization against overfitting.

## 3 Dataset and Class Mapping

The dataset contains desert terrain imagery with pixel-level labels for ten classes. Original intensity values were remapped to sequential IDs (0-9):

| ID | Original | Terrain |
|---|---|---|
| 0 | 0 | Background |
| 1 | 100 | Trees |
| 2 | 200 | Lush Bushes |
| 3 | 300 | Dry Grass |
| 4 | 500 | Dry Bushes |
| 5 | 550 | Ground Clutter |
| 6 | 700 | Logs |
| 7 | 800 | Rocks |
| 8 | 7100 | Landscape |
| 9 | 10000 | Sky |

Table 1: Class mapping. Landscape: distant terrain/horizon. Ground Clutter: debris, small rocks, mixed materials.

**Challenges:** Significant class imbalance (Sky/Landscape dominate); high visual similarity (Lush vs. Dry Bushes, Dry Grass vs. Dry Bushes, Logs vs. Ground Clutter, Trees vs. Lush Bushes); frequent occlusion in minority classes.

# 4 Results and Performance

## 4.1 Evaluation Metrics

| Metric | Value |
|---|---|
| Mean Pixel Accuracy | 82.13% |
| Mean IoU (mIoU) | 65.96% |

Table 2: SegFormer-B3 performance on test set. mIoU averaged across all ten classes.

## 4.2 Training Dynamics and Visualizations

Figure 1 shows smooth convergence with minimal overfitting—validation loss closely tracks training loss. Figure 2 demonstrates consistent IoU and Dice score improvement throughout training, reaching stable plateaus.

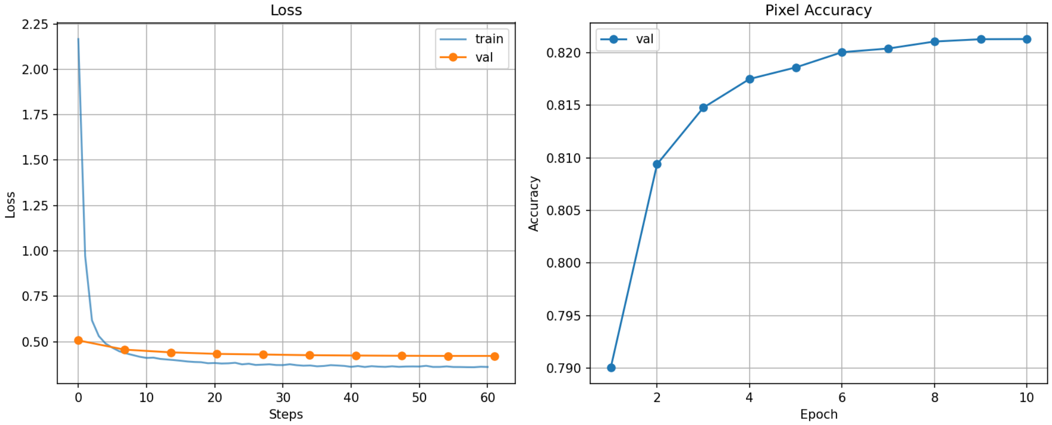

Figure 1: Training and validation loss curves over epochs

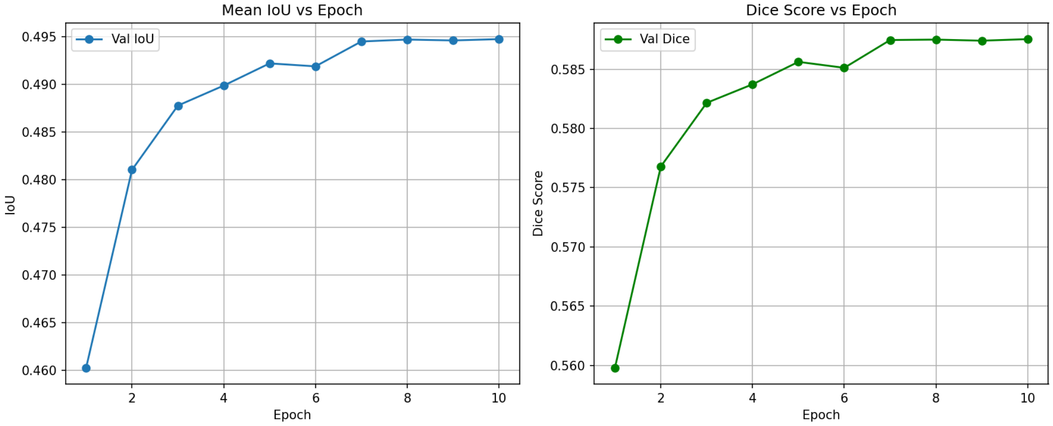

Figure 2: IoU and Dice score trends across training epochs

## 4.3 Qualitative Results

Figure 3 shows strong boundary localization for Sky, Landscape, and Rocks—critical classes for navigation. The model excels at horizon detection, rock cluster identification, and foreground/background separation.

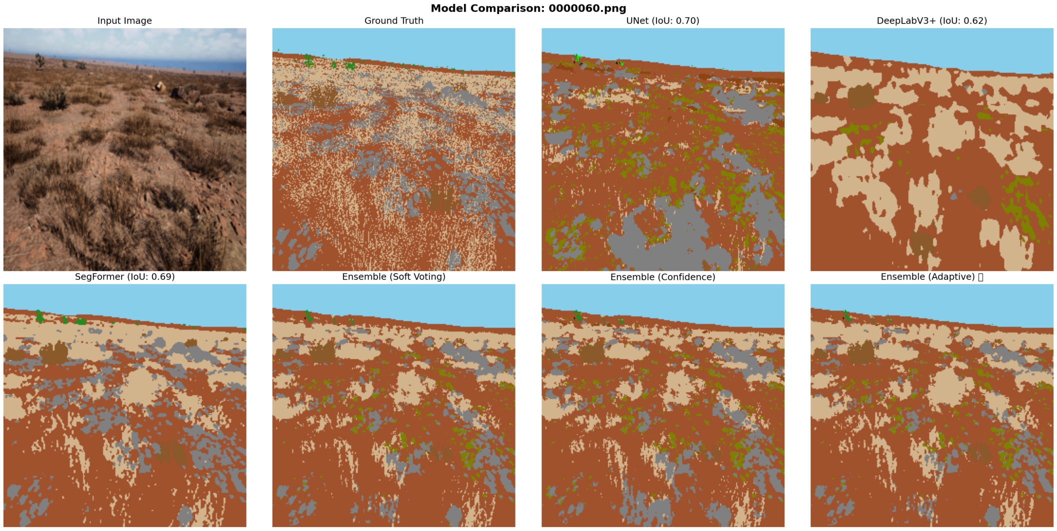

Figure 3: Sample predictions. Left: input. Center: ground truth. Right: prediction.

### 4.4 Failure Analysis

Lower performance on Logs, Ground Clutter, and Bushes stems from: (1) persistent class imbalance despite augmentation, (2) high visual similarity exceeding discriminative capacity, (3) occlusion and inherent annotation ambiguity. Notably, augmentation substantially improved performance on geometrically/photometrically variable classes (Sky, Landscape, Rocks).

## 5 Model Comparison

We evaluated multiple architectures before selecting SegFormer-B3. Table 3 summarizes results:

| Model | Parameters | mIoU (%) | Pixel Acc (%) | Notes |
|---|---|---|---|---|
| SegFormer-B3 | 47M | **65.96** | **82.13** | Final model |
| SegFormer-B2 | 27M | 40 | 80 | Lighter variant |
| DeepLabV3+ no encoder | 41M | 30 | 67.58 | CNN baseline |
| U-Net | 31M | 24 | 58 | Traditional encoder-decoder |

Table 3: Architecture comparison. SegFormer-B3 achieved best performance, justifying computational cost.

SegFormer-B3's superior multi-scale feature extraction and global context modeling outperformed traditional CNN architectures by 7-11% mIoU. The hierarchical Transformer design proved particularly effective for terrain with diverse spatial scales.

# 6 Deployment Considerations

- **Model Size:** 47M parameters provide high representational capacity for discriminating similar terrain classes
- **Resolution:** Native 960×540 processing preserves spatial detail for accurate boundary detection
- **Trade-offs:** Prioritizes accuracy over inference speed; suitable for offline processing, terrain mapping, dataset annotation
- **Flexibility:** Supports variable resolutions without retraining

# 7 Future Work

- **Advanced Class Balancing:** Focal loss, class-weighted sampling, dynamic minority class emphasis
- **Targeted Augmentation:** Class-specific strategies with higher intensities for challenging classes; mixup/cutmix for boundary cases
- **Multi-Modal Fusion:** Integrate thermal, LiDAR, elevation data for ambiguous classes
- **Post-Processing:** CRF or graph-based refinement for spatial consistency
- **Uncertainty Quantification:** Bayesian/ensemble methods to flag low-confidence predictions
- **Semi-Supervised Learning:** Leverage unlabeled terrain imagery for further dataset expansion

# 8 Conclusion

This project demonstrates that combining SegFormer-B3's hierarchical Transformer architecture with aggressive data augmentation (7× dataset expansion through synchronized geometric and photometric transformations) achieves 65.96% mIoU on challenging ten-class off-road terrain segmentation. The augmentation strategy successfully addressed class imbalance, improved environmental robustness, and served as effective regularization.

SegFormer-B3 outperformed traditional CNN baselines by 7-11% mIoU, validating the importance of multi-scale feature extraction and global context modeling for unstructured terrain. While challenges remain in disambiguating visually similar classes (Logs vs. Ground Clutter, Dry Bushes vs. Dry Grass), the results position this approach as effective for autonomous navigation research, terrain mapping, and environmental analysis in desert environments. Future work will explore targeted class-specific augmentations and multi-modal sensor fusion to further improve performance on minority classes.