Version 3 (V3) – DeepLabV3+ with Attention U-Net

Goal:

Improve spatial understanding and small-object focus in off-road semantic segmentation, after observing that loss engineering alone could not improve Mean IoU.

1. Motivation from V1 and V2

Baseline U-Net (V1) suffered from class imbalance and poor spatial precision.
Class-weighted Dice and Cross-Entropy loss (V2) stabilized training but did not significantly improve Mean IoU, indicating that the limitation was architectural.

2. Architectural Choice

V3 introduces a hybrid architecture combining:
DeepLabV3+ encoder with Atrous Spatial Pyramid Pooling (ASPP)
Attention U-Net style decoder

DeepLabV3+ captures multi-scale contextual information, while attention gates help the decoder focus on relevant spatial regions and suppress background noise.

3. Model Architecture Overview

Input Image (256×256×3)
ResNet50 Encoder (ImageNet pretrained)
ASPP (multi-scale context)
Upsampling Decoder
Attention-Gated Skip Connections
Final Softmax Output (10 classes)

4. Training Strategy

Encoder initialized with ImageNet weights
Encoder frozen during initial training phase
Dice + Categorical Cross-Entropy loss for stability
Adam optimizer with conservative learning rate
EarlyStopping and ModelCheckpoint callbacks

5. Evaluation Protocol

Pixel accuracy was not used due to class imbalance.
Evaluation focused on Mean Intersection-over-Union (Mean IoU)
Visual comparison of predictions vs ground truth

Qualitative assessment of boundaries and small objects

6. Observations and Results

The V3 model produced spatially coherent and smooth segmentations with correct global scene structure. Attention helped reduce noise and enforce regional consistency across large regions.

However, predictions were visually smoother than ground truth and failed to capture highly fragmented pixel-level annotations, resulting in moderate Mean IoU.

7. Key Insight

The discrepancy between qualitative improvements and numerical IoU highlights the impact of noisy and over-fragmented ground truth annotations. V3 demonstrated that architectural improvements enhance semantic understanding, but label structure remains a limiting factor.

8. Conclusion

V3 successfully addressed spatial focus and context modeling limitations observed in earlier versions. While Mean IoU improvements were modest, the model represents a significant qualitative improvement and provides a strong foundation for V4.