DesertVision++

CNN vs Transformer Architectures for Robust Off-Road Semantic Segmentation

Team: XDBoost

---

1. Executive Summary

Autonomous off-road systems require reliable scene understanding across varying terrain, lighting conditions, and environmental complexity. In this project, we developed and evaluated deep learning models for pixel-level semantic segmentation in desert environments.

We implemented and compared two architectures:

1. **U-Net++ (ResNet-50 backbone)** – CNN-based
2. **SegFormer-B2 (Transformer-based)**

Key results:

Model	Validation IoU	Test IoU
U-Net++	**0.7965**	0.3041
SegFormer-B2	—	**0.3823**

Although U-Net++ achieved high validation performance, it suffered from domain shift on the test distribution. SegFormer demonstrated better generalization, improving test IoU by **+7.82%**.

This study highlights the importance of architectural robustness under domain shift.

---

2. Problem Overview (Accessible Explanation)

The task is **semantic segmentation**: assigning a class label to every pixel in an image.

Each image contains 10 classes:

- Trees

- Lush Bushes
- Dry Grass
- Dry Bushes
- Ground Clutter
- Flowers
- Logs
- Rocks
- Landscape
- Sky

The challenge is not simply recognizing objects — but recognizing them under:

- Different lighting conditions
- Texture variations
- Rare object occurrences
- Distribution shifts between training and testing data

The real-world difficulty lies in **generalization**, not memorization.

---

3. Dataset and Experimental Setup

- 10 semantic classes
- Synthetic training distribution
- Validation split from same distribution
- Test distribution likely shifted

Images resized to:

```
512 × 512
```

Training performed on GPU Training time: ~2–3 hours per model

---

4. Model 1 — U-Net++ (CNN-Based)

Architecture

- Encoder: ResNet-50 (ImageNet pretrained)
- Decoder: Nested skip connections (U-Net++)
- Output: 10-class segmentation map

Why U-Net++?

- Strong multi-scale feature fusion
- Good small-object handling
- Proven performance in segmentation tasks

Loss Function

Hybrid Loss:

```
0.5 × Dice Loss + 0.5 × Focal Loss
```

- Dice optimizes region overlap
- Focal handles hard pixels and imbalance

Optimizer

- AdamW
- Learning rate: 2e-4
- Cosine Annealing scheduler
- Batch size: 4

---

5. U-Net++ Results

Validation (In-Distribution)

Mean IoU: **0.7965**

This indicates strong learning capacity within the synthetic domain.

Test (Distribution Shifted)

Mean IoU: **0.3041**

Per-class performance shows:

- Sky: 0.9863
- Landscape: 0.6872
- Trees: 0.4003
- Small objects (Flowers, Logs, Ground Clutter): ≈ 0

Interpretation

The model performs well on dominant, visually consistent classes but collapses on:

- Rare classes

- Small objects
- Texture-sensitive regions

This suggests strong **texture bias**, common in CNN architectures.

---

6. Model 2 — SegFormer-B2 (Transformer-Based)

Architecture

SegFormer is a hierarchical vision transformer:

- Backbone: MiT-B2
- Multi-scale attention mechanism
- Lightweight decoder
- Outputs at reduced resolution (upsampled during evaluation)

Unlike CNNs, transformers:

- Capture global context
- Are less reliant on local textures
- Better model structural relationships

Training Setup

- Pretrained backbone
- AdamW optimizer
- Learning rate: 6e-5
- Batch size: 4
- Mixed precision training

---

7. SegFormer Results

Test Mean IoU: 0.3823

Per-class highlights:

- Sky: 0.9819
- Landscape: 0.6822
- Dry Grass: 0.4201
- Trees: 0.2557

Small classes still underperform, but overall robustness improves.

Improvement over U-Net++:

+7.82% absolute mean IoU gain.

---

8. Comparative Analysis

Aspect	U-Net++	SegFormer
In-domain performance	Excellent	Not evaluated
Domain robustness	Weak	Better
Texture sensitivity	High	Lower
Global context modeling	Limited	Strong
Small object handling	Moderate	Limited (due to 1/4 resolution output)

Key Insight

CNN-based models tend to overfit to synthetic textures. Transformer-based models demonstrate improved structural reasoning under domain shift.

---

9. Domain Shift Analysis

The major gap (0.7965 → 0.3041) reveals:

1. Validation data shares distribution with training.
2. Test data likely differs in:
    - Lighting
    - Texture distribution
    - Object frequency
3. Micro IoU metric favors dominant classes.

Small classes collapse due to:

- Severe imbalance
- Low pixel coverage
- Limited augmentation diversity

---

10. Failure Case Observations

Observed issues include:

- Logs confused with ground clutter
- Flowers merged into dry grass
- Bushes merged with trees
- Rocks misclassified as landscape

These errors suggest:

- Feature confusion between visually similar categories
- Insufficient representation of rare categories

---

11. Strengths of This Study

- Two distinct architectures evaluated
- Hybrid loss experimentation
- Clear metric reporting
- Proper evaluation pipeline
- Honest domain shift analysis
- Visual qualitative comparisons saved

---

12. Limitations

- No heavy domain randomization
- No class-weighted transformer loss
- Small batch size (GPU constraint)
- No multi-scale training
- SegFormer output resolution limited (1/4 scale)

---

13. Proposed Improvements

1. Stronger Augmentation

- Gaussian noise
- Blur
- Perspective transforms
- Weather simulation
- Random cropping

2. Class Rebalancing

- Weighted cross-entropy
- Oversampling rare classes

3. Multi-Scale Supervision

Improve small-object segmentation.

4. Domain Adaptation

- Self-training with pseudo-labels
- Feature alignment techniques
- Synthetic-to-real transfer methods

---

14. Conclusion

DesertVision++ demonstrates that:

- High validation accuracy does not guarantee real-world robustness.
- Domain shift significantly impacts CNN-based segmentation models.
- Transformer architectures improve generalization under distribution changes.
- Rare class imbalance remains a critical challenge.

This project emphasizes the importance of architectural selection and domain-aware training strategies in autonomous off-road perception systems.