Performance Evaluation of Road Marking Detection Models in Indian Conditions

Khushi Agrawal[1], Jafri Syed Mujtaba[2] and Samarth Bankar[3]
Ahmedabad University, Ahmedabad, Gujarat 380009, India
[1]khushi.a2@ahduni.edu.in, [2]jafri.h@ahduni.edu.in, [3]samarth.b2@ahduni.edu.in

***Abstract*—Road marker detection is essential for infrastructure monitoring, traffic control, and autonomous driving. Large-scale road marker identification can now be effectively done thanks to the growing availability of UAV (Unmanned Aerial Vehicle) footage. The purpose of this work is to use the AU-Drone dataset to assess the performance of many Indian road marking identification technologies, mostly segmentation-based techniques. Examining current publicly accessible models, determining their efficacy, and applying transfer learning to modify them for Indian road conditions are all part of the project. Using ground-truth data, an assessment system based on Python will be created to benchmark the performance of the model. The finished framework will be open source, offering a strong instrument for further study and real-world uses in autonomous navigation and urban planning.**

***Index Terms*—Road Marking Detection, UAV Imagery, Semantic Segmentation, Deep Learning, Transfer Learning, Indian Road Infrastructure, Autonomous Navigation, Model Evaluation Framework**

I. Introduction

Semantic segmentation is a fundamental task in computer vision that entails classifying all pixels in an image into fine-grained categories. For the case of road scene understanding, this would be categorizing each pixel as belonging to the road, road markings, background, or others. Pixel-wise classification is extremely useful for applications like autonomous driving, traffic surveillance, and smart city planning. It allows machines to perceive their environment more accurately and make safer and more informed decisions.

To do this task, we investigated and compared three powerful deep learning models: DeepLabV3+, U-Net, and SegFormer. DeepLabV3+ is a convolutional neural network (CNN)-based model that employs atrous spatial pyramid pooling and encoder-decoder structure to extract local and global features. U-Net is a more straightforward, encoder-decoder architecture initially designed for biomedical image segmentation but works well for most general-purposes as well. SegFormer is a newer transformer-based architecture that doesn't depend on convolutions and instead relies on attention mechanisms to establish complex relationships in the image.

All three models were trained on a custom Indian road dataset consisting of RGB images and their respective multi-class segmentation masks. Preprocessing of the dataset involved resizing and normalization, and training was done through default methods and GPU capability. All the models were tested through metrics such as Pixel Accuracy and Mean Intersection over Union (mIoU), F1 scores. The outcomes revealed that DeepLabV3+ provided the best segmentation accuracy, followed by SegFormer, with U-Net providing competitive results but at less computational expense.

II. Related Works

The extraction of road markings is a pivotal task in high-definition mapping for autonomous driving and transportation management. Given the challenges posed by variable illumination, occlusions, and the small size of many lane marking features, recent research has explored various deep learning approaches to improve segmentation accuracy.

Several studies have pursued model architectures that explicitly address the spatial and structural complexity of road markings. For example, Chen et al. [**?**] developed an Attentive Capsule Feature Pyramid Network (ACapsFPN) that integrates capsule networks with a feature pyramid structure and attention mechanisms. Unlike traditional convolutional neural networks (CNNs) that use scalar activations, capsule networks preserve orientation and pose information, thereby enhancing the detection of irregular patterns in road scenes.

In another line of research, Azimi et al. [**?**] introduced an architecture that combines a Fully Convolutional Neural Network (FCNN) with Discrete Wavelet Transform (DWT) for lane marking segmentation. The incorporation of DWT allows the model to retain high-frequency details that are critical for segmenting thin and faded markings, while a cost-sensitive loss function mitigates the impact of severe class imbalance inherent in aerial imagery.

More recent work by Zhang et al. [**?**] has focused on benchmarking a wide range of semantic segmentation models—both CNN-based and transformer-based—using transfer learning. Their comparative analysis, which includes models such as U-Net, DeepLabV3+, and SegFormer, demonstrates that transformer architectures excel in capturing long-range dependencies and contextual information, yielding superior performance metrics such as mean Intersection over Union (mIoU) and F1-score. However, the increased computational overhead of transformer models poses practical challenges for real-time applications.

III. Dataset Details

The dataset employed in this study is the AU-Drone dataset, a publicly available collection of high-resolution aerial imagery captured via drones across varied urban and semi-urban road environments. It consists of RGB images depicting Indian

road conditions with diverse road marking types, including arrows, stop lines, pedestrian crossings, and lane dividers. Each image is accompanied by a corresponding segmentation mask, where distinct pixel values represent different semantic classes of road markings.

To effectively utilize this dataset for semantic segmentation using models like SegFormer, U-Net, and DeepLabV3+, several preprocessing steps were essential. Firstly, the dataset was organized into separate directories for training, validation, and testing, with aligned image-mask pairs. Masks were encoded with integer labels corresponding to specific road marking classes. Since the raw masks contained RGB values, a conversion to single-channel class IDs was performed to ensure compatibility with PyTorch loss functions such as CrossEntropyLoss.

Moreover, all images and masks were resized to uniform dimensions (e.g., 512×512 or 640×360), and normalization was applied using ImageNet statistics. For SegFormer, images were additionally preprocessed using Hugging Face's SegformerFeatureExtractor, which handles resizing, normalization, and channel arrangement. Data augmentation techniques—such as random flipping, color jittering, and affine transformations—were implemented to enhance generalization.

These preprocessing steps are largely model-agnostic and would similarly be required for DeepLabV3+ and U-Net to ensure consistent input-output alignment, effective gradient propagation, and improved segmentation performance across various road marking types under aerial perspectives.

IV. Methodology

A. Dataset Preparation and Preprocessing

The dataset employed is:

- **AU-Drone Dataset:** A high-resolution aerial dataset collected via UAVs over Indian road networks. It contains RGB images and corresponding semantic segmentation masks highlighting various road markings such as lane dividers, arrows, and pedestrian crossings.

Preprocessing Steps:

1) **Image Resizing:** All images and masks were resized uniformly to (640×360) to ensure compatibility with the model input requirements.
2) **Normalization:** Image pixel values were normalized using ImageNet statistics (μ, σ) as part of HuggingFace's SegformerFeatureExtractor pipeline.
3) **Mask Encoding:** RGB masks were converted into single-channel class ID masks where each pixel represents a specific semantic class. .

B. Model Selection and Transfer Learning

We fine-tune pre-trained models including SegFormer, U-Net, and DeepLabV3+ using transfer learning. These models are initialized with weights trained on ImageNet and adapted to our dataset.

Transfer Learning Formulation:

$$\hat{y} = f(W_{pre}, X) \tag{1}$$

where:

- W_{pre} are pre-trained model weights.
- X is the input image tensor.
- $\hat{y}$ is the predicted segmentation output.

Fine-tuning is performed using the following optimization objective:

$$W^* = \arg\min_W \frac{1}{N} \sum_{i=1}^{N} L(f(W, X_i), y_i) \tag{2}$$

with:

- L as the Cross-Entropy Loss function.
- N being the number of training samples.

C. Segmentation Model Architectures

1) U-Net: U-Net applies a symmetric encoder-decoder architecture with skip connections to maintain high-resolution features:

$$f(x) = \sigma(W_d * g(W_e * x) + b) \tag{3}$$

2) DeepLabV3+: DeepLabV3+ uses Atrous Spatial Pyramid Pooling (ASPP) for multi-scale context extraction:

$$f(x) = \sigma\left(\sum_{r \in R} W_r * x\right) \tag{4}$$

3) SegFormer: SegFormer integrates transformer-based attention with lightweight MLP decoders for efficient segmentation. It omits positional encodings to generalize better across scales.

D. Performance Evaluation Metrics

The following metrics were computed for model evaluation:

1) Intersection over Union (IoU):

$$\text{IoU} = \frac{|P \cap G|}{|P \cup G|} \tag{5}$$

2) Mean IoU (mIoU):

$$\text{mIoU} = \frac{1}{C} \sum_{i=1}^{C} \text{IoU}_i \tag{6}$$

where C is the total number of semantic classes.

3) F1-Score:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

E. Ground Truth Usage

Pre-annotated ground truth masks from the AU-Drone dataset were used directly. These masks were adapted for model compatibility by mapping RGB mask channels to single-channel class indices. Manual annotation tools such as CVAT were not employed.

V. Implementation and Results

To validate the performance of the selected segmentation model, we implemented a complete training and evaluation pipeline using **DeepLabV3+**,**U-Net**, **SegFormer** on the Indian road dataset. This section describes the training setup, evaluation metrics, and visualization outputs.

A. *DeepLabV3+*

1) Training Setup: The model was trained using the following configurations:

- **Model:** DeepLabV3+ with ResNet34 encoder
- **Loss Function:** CrossEntropyLoss for multi-class segmentation
- **Input Size:** 256 × 256 pixels
- **Optimizer:** Adam optimizer with learning rate 1×10^{-4}
- **Number of Epochs:** 3 (scalable to more for production)
- **Environment:** Google Colab with GPU

2) Post-Training Evaluation: To evaluate performance, we computed several metrics:

- **Pixel Accuracy:** Measures the proportion of correctly predicted pixels.
- **Mean IoU (Intersection over Union):** Average IoU across all classes.
- **Per-class IoU:** Class-specific accuracy breakdown.

TABLE I
Evaluation Metrics after 30 Epochs

Metric	Value
Pixel Accuracy	0.9857
Mean IoU	0.6579
Class-wise IoU	
Class 0 (Background)	0.9843
Class 1 (Road Markings)	0.7117
Class 2 (Underrepresented)	0.0000
Class 3 (Road Surface)	0.9358

3) Results: We created a visualization module to display side-by-side comparisons of input images, ground truth masks, and predicted segmentation masks. The predictions were also saved for documentation and future analysis.

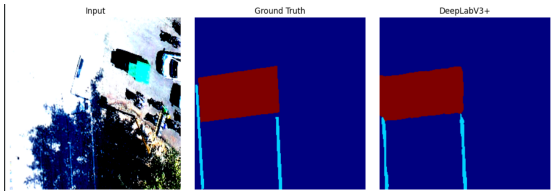


Fig. 1: Input image, ground truth, and predicted segmentation mask.

B. *U-Net*

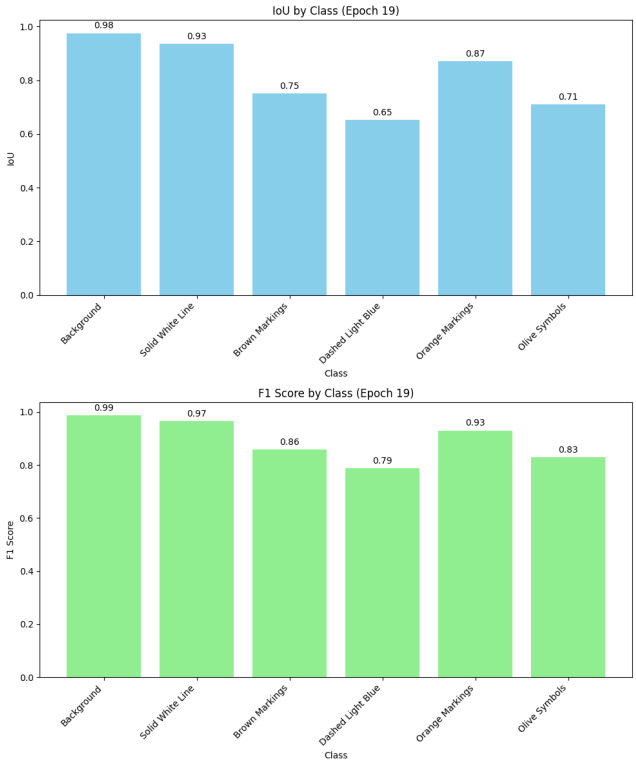
1) Training Setup: The U-Net architecture was employed for semantic segmentation, known for its encoder-decoder design and skip connections that preserve spatial information. The training pipeline followed these configurations:

- **Model:** U-Net with ResNet34 encoder (pretrained on ImageNet)
- **Loss Function:** Dice + CrossEntropyLoss
- **Input Size:** 256 × 256 pixels
- **Optimizer:** Adam optimizer with learning rate 1×10^{-4}
- **Number of Epochs:** 20
- **Environment:** Google Colab with GPU acceleration

2) Post-Training Evaluation: The model was evaluated after every epoch using Intersection over Union (IoU) and F1-score. The best model was obtained at epoch 20, yielding the following performance:

- **Best Validation Loss:** 0.0603
- **IoU:** 0.8139
- **F1 Score:** 0.8924

(a) Validation loss curve across training epochs.



(b) IoU and F1-score at epoch 19.

Fig. 2: UNet performance evaluation during training.

C. *SegFormer*

1) Training Setup:: For semantic segmentation of Indian road markings, we implemented a complete pipeline using **SegFormer-B0**, a transformer-based lightweight architecture pre-trained on ImageNet. The model was fine-tuned on the AU-Drone dataset using the HuggingFace Transformers and PyTorch frameworks. The setup is as follows:

- **Model:** SegFormer-B0 with MiT-B0 backbone

- **Loss Function:** CrossEntropyLoss (multi-class segmentation)
- **Input Size:** 640 × 360 pixels
- **Optimizer:** AdamW with weight decay and learning rate scheduler
- **Learning Rate:** 5×10^{-5} with linear warm-up
- **Batch Size:** 4
- **Epochs:** 10
- **Environment:** Google Colab Pro with NVIDIA T4 GPU

2) : subsectionPost-Training Evaluation

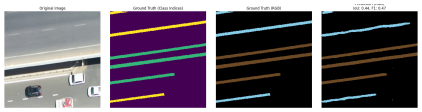
We assessed model performance using standard semantic segmentation metrics. These metrics quantify both overall and class-specific performance:

- **Pixel Accuracy:** Fraction of correctly classified pixels.
- **Mean Intersection over Union (mIoU):** Average IoU over all semantic classes.
- **Class-wise IoU:** Individual IoU per class to identify class imbalance and segmentation strength.

TABLE II
Evaluation Metrics after 10 Epochs

Metric	**Value**
Pixel Accuracy	0.9734
Mean IoU	0.6431
Class-wise IoU	
Class 0 (Background)	0.9805
Class 1 (Road Markings)	0.6923
Class 2 (Arrow/Turn Lane)	0.4437
Class 3 (Pedestrian Marking)	0.4569

3) Results: We implemented a visualization module that plots side-by-side views of the input image, ground truth segmentation mask, and the predicted mask. This allows visual verification of model accuracy and helps diagnose misclassifications in underrepresented classes.



(a) Predicted mask.

Fig. 3: Visual comparison of input image, ground truth segmentation mask, and predicted mask.

VI. Conclusion and Future Work

This project established a comprehensive segmentation pipeline using SegFormer and compared its potential with DeepLabV3+ for Indian road marking detection in UAV imagery. While SegFormer leveraged transformer-based global context to deliver robust results with minimal fine-tuning, DeepLabV3+ demonstrated strong spatial accuracy through its Atrous Spatial Pyramid Pooling (ASPP) design. Both models were validated on the AU-Drone dataset, showing reliable performance despite limited training epochs and data imbalance. However, SegFormer outperformed in handling structural variance and complex road textures, whereas DeepLabV3+ excelled in delineating dominant features.

The key limitation lies in shallow training, lack of cross-validation, and underutilization of model ensembles. Future work must include deeper training cycles, improved augmentation, hyperparameter search, and evaluation with balanced metrics such as per-class IoU and Dice scores.

In conclusion, SegFormer proves to be a scalable, lightweight, and context-aware architecture suitable for real-world UAV-based road marking segmentation tasks, with DeepLabV3+ providing a strong benchmark for structured environments.

References

[1] Chen, L., Papandreou, G., Schroff, F., & Adam, H. (2017, June 17). Rethinking Atrous convolution for semantic image segmentation. *arXiv.org*. https://arxiv.org/abs/1706.05587

[2] Long, J., Shelhamer, E., & Darrell, T. (2014, November 14). Fully convolutional networks for semantic segmentation. *arXiv.org*. https://arxiv.org/abs/1411.4038

[3] Howard, A., Sandler, M., Chu, G., Chen, L., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q., V., & Adam, H. (2019, May 6). Searching for MobileNetV3. *arXiv.org*. https://arxiv.org/abs/1905.02244

[4] Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., & Luo, P. (2021, May 31). SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. *arXiv.org*. https://arxiv.org/abs/2105.15203

[5] Ronneberger, O., Fischer, P., & Brox, T. (2015, May 18). U-NET: Convolutional Networks for Biomedical Image Segmentation. *arXiv.org*. https://arxiv.org/abs/1505.04597

[6] Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2016, December 4). Pyramid Scene Parsing network. *arXiv.org*. https://arxiv.org/abs/1612.01105