Boundaries-Informed Transformer for Terrain Segmentation with Enhanced Contour Detection: An exploration Study

Yuchuan Dong*
yuchuan@chalmers.se

Zinan Ma *
zinanm@chalmers.se

ABSTRACT

Off-road terrain detection is crucial for chassis tuning and autonomous navigation. One challenge that makes this task hard to resolve is that the terrain type is unstructured and highly varied. Current terrain detection techniques primarily rely on data from vehicle sensors and cameras, with neural networks used for classification. Current mainstream methods are CNN-based and transformer-based models. We focus on semantic segmentation, employing the Rellis-3D dataset, and simplify the task by reducing the number of terrain categories. Initially, we reproduce the transformer-based SegFormer network. Then, we enhance this model by incorporating a boundary detection module to provide structural guidance during segmentation, aiding the network in better delineating terrain types. During evaluation, our model, which did not use transfer learning, showed suboptimal performance. However, the integration of boundary detection resulted in noticeable improvements in some images, and an improvement of 5% accuracy on the test dataset. This work provides inspiration for future challenges, including how to effectively combine SegForm networks with boundary detection, and which boundary-informed segmentation strategies are most suitable. Further research will focus on the effectiveness and interpretability of these approaches.

Keywords

Terrain detection; SegFrom; Boundary information; Off-Road

1. INTRODUCTION

Accurate terrain segmentation is vital for autonomous navigation in **off-road** environments, where diverse and unstructured terrains pose significant challenges. **convolutional neural networks** (CNNs) have shown success in structured environments like urban scenes (e.g., Cityscapes and KITTI), they encounter difficulties in unstructured settings where terrain features are highly variable, where boundaries between different terrain types, such as rocks, vegetation, and water, are irregular and difficult to define. **Transformers**, with their ability to capture long-range dependencies, have emerged as a promising solution for segmentation tasks, but they still face challenges in delineating fine contours of objects and terrain.

In this exploratory study, we propose a boundaries-informed transformer network, inspired by Segformer, a novel positional-encoding-free and hierarchical Transformer encoder, to improve terrain segmentation by explicitly integrating boundary detection into the model. We argue accurate boundary detection is particularly important in off-road environments, where subtle transitions between terrain types (e.g., from gravel to rocks or water) can critically impact navigation decisions. Misclassifying these boundaries can lead to hazardous outcomes, such as mistaking rough or non-navigable surfaces for smooth, navigable terrain.

Due to resource constraints, we do not utilize pre-trained models. Instead, we train our model from scratch and ensure fair comparisons by maintaining the same training strategy across models—using identical input resolution, loss functions, and iterations. By combining the model's boundary detection capabilities, we aim to explore overall segmentation accuracy and robustness, particularly in challenging off-road environments.

We conduct our experiments on the RELLIS-3D dataset, which we coarse into six terrain categories, including smooth, rough, and bumpy navigable regions, non-navigable areas, and obstacles.

2. RELATED WORKS

2.1 Terrain Classification

There is a strong demand for real-time terrain recognition in various fields of off-road driving. For instance, in human-driven off-road vehicles, terrain-aware adjustments of the vehicle's configuration can prevent it from getting stuck. Similarly, robots navigating complex terrains for mining or construction require real-time terrain detection for autonomous navigation. There are several neural network approaches to implementing terrain classification. For example, Yun et al. [1] proposed a method that integrates point cloud data from radar transmission with camera data, while Shon et al. [2] utilized data collected from a vehicle's Controller Area Network (CAN) bus. The proposed method in this work focuses on the input data from cameras, using a computer vision-based approach for terrain detection. More specifically, this method addresses terrain classification as a semantic classification task.

2.2 Dataset

RELLIS-3D [3] and RUGD [4] are the latest datasets specif-

*Chalmers University of Technology, Rännvägen 6B, 41324 Gothenburg, Sweden.

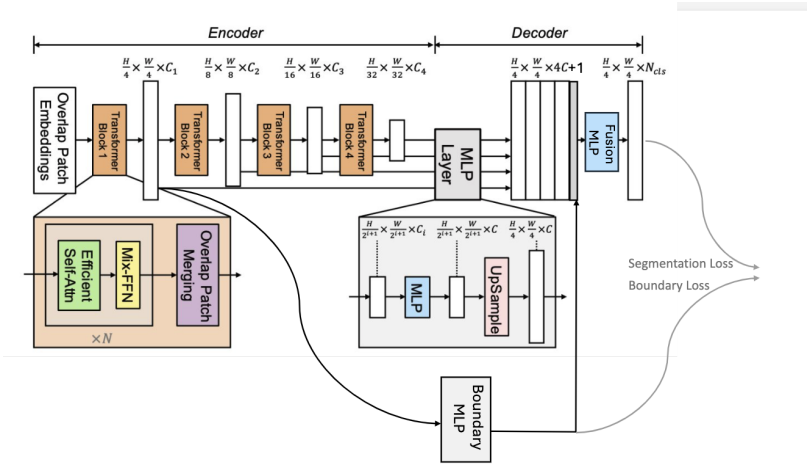

Figure 1: Proposed Network adapted from SegForm

ically designed for terrain detection tasks. RELLIS-3D is derived from RUGD but includes additional categories that RUGD does not have. Compared to RUGD, RELLIS-3D offers higher precision in its class annotations, making it more suitable for detailed and complex terrain classification tasks. In the proposed method, RELLIS-3D is employed,

2.3 Unstructured Semantic Segmentation

Unstructured semantic segmentation refers to the task in computer vision where images without a fixed structure or pattern are segmented semantically. In this task, the goal is to assign each pixel in the image to a specific category, although these categories may not have clear boundaries or regular arrangements.

The Pyramid Vision Transformer (PVT) [5] is a transformer-based architecture that incorporates a pyramid structure to capture multi-scale features, similar to CNNs. By downsampling features at different stages, PVT efficiently handles high-resolution images, making it ideal for tasks like object detection and segmentation. It combines the global context of transformers with the hierarchical feature extraction of CNNs for enhanced performance in vision tasks.

SegFormer[6] is an efficient design that uses a transformer for semantic segmentation. This architecture efficiently generates multi-level features from input images(pyramid structure), preserving both high-resolution and coarse-grain features. The model uses small 4x4 patches to ensure detailed feature extraction, especially beneficial for dense prediction tasks like semantic segmentation.

Except for the pixel-wise segmentation methods introduced before, boundary-informed segmentation has been studied in quite a few recent efforts. InverseForm[7] introduces a method to compensate for the capture in boundary transformation in cross entropy loss function. The proposed loss function in InverseForm measures distance of output boundary and annotated boundary instead of solely using cross entropy.

3. PROPOSED SOLUTION

Based on Segformer, the proposed method integrate a boundary-informed network and a specialized boundary-informed loss function. The boundary-informed approach is designed to improve the accuracy of object contours and terrain boundaries, which are critical for off-road navigation. Figure 1 illustrate the adapted construction from SegForm that proposed in this report. We describe how this boundary detection network is embedded into the overall architecture.

3.1 Boundary detection module

The Boundary Detection Module is a critical part of our segmentation work, specifically addressing a notable limitation of traditional segmentation approaches, including SegFormer we talk above. Even SegFormer utilize a hierarchical Transformer architecture with attention mechanisms to capture information across various image scales, it may overlooks the **explicit** representation of **boundary information**. The attention mechanism, while powerful in aggregating information from different regions, may not sufficiently highlight the transitions between objects, which are essential for accurate segmentation.

Our approach inspired from the way children color within the lines of a drawing. When provided with **clear boundaries**, kids can more easily fill in colors without straying outside the lines. Similarly, by using boundary information as a guiding mechanism, allowing the network to focus on identifying relevant regions and enhancing the overall quality of the output.

3.2 Boundary MLP

Our Boundary Detection Module identifies edges and transitions within the input images, serving as a crucial guide for the segmentation process. This module is applied exclusively to the first stage of the SegFormer model, where it can capture finer details and local features effectively. By providing **early guidance** on boundary locations, we enable the segmentation model to make more informed decisions throughout the subsequent layers.

The Boundary MLP utilizes a series of convolutional layers to extract **boundary information** from the input features. The process can be represented as follows:

1. The first convolution layer applies weights W_1 and bias b_1 to the input feature map F, resulting in:

$$B_1 = \text{ReLU}\left(W_1 * F + b_1\right)$$

2. A dilation convolution layer further processes:

$$B_2 = \text{ReLU}\left(W_2 * B1 + b_2\right)$$

3. Finally, the boundary output is computed as:

$$boundary\_out = W_3 * B_2 + b_3$$

3.2.1 *Fusion Layer*

The integration of boundary maps with segmentation outputs is achieved through a lightweight fusion process, combining these two critical information sources to generate the final segmentation mask. [1] Here the concrete process:

1. The outputs are concatenated::

$$C = \text{Concat}(\text{seg\_out}, \text{boundary\_out})$$

$$F_{\text{final}} = \text{ReLU}(W_f * C + b_f)$$

$$\text{final\_out} = W_{\text{seg}} * F_{\text{final}} + b_{\text{seg}}$$

2. A dilation convolution layer further processes:

$$B_2 = \text{ReLU}\left(W_2 * B1 + b_2\right)$$

3. Finally, the boundary output is computed as:

$$boundary\_out = W_3 * B_2 + b_3$$

3.3 Boundary-Informed Segmentation Loss

In the proposed approach, we use a combined loss function that integrates both boundary detection and segmentation, as boundaries can serve as a strong cue to improve segmentation. The goal is to leverage boundary information to provide better structural details to the segmentation output. There are more advanced Loss function such as Dice-Loss, Focal Loss, but we aim to keep model simple and illustrate how is boundary information works.

Segmentation Loss: For the segmentation loss, we use a weighted cross-entropy loss, which accounts for the class imbalance typically found in our terrain segmentation task. This ensures that underrepresented classes, such as smaller regions or less frequent terrain types, contribute equally to the training process, preventing the network from being biased toward the more common classes. The segmentation loss is defined as:

$$L_{\text{seg}} = \text{WeightedCrossEntropy}\left(P_{\text{seg}}, Y_{\text{seg}}\right)$$

Where P_{seg} is the predicted segmentation map and Y_{seg} is the ground truth segmentation.

Boundary Loss: The boundary loss aims to ensure that the network accurately detects object boundaries. Given the sparsity of boundary pixels in an image, we apply a weighted binary cross-entropy loss to balance the network's attention to these rare but important pixels. The segmentation loss is defined as:

$$L_{\text{bdr}} = \text{WeightedBinaryCrossEntropy}\ \left(P_{\text{bdr}}, Y_{\text{bdr}}\right)$$

Where P_bbdr is the predicted boundary map and Y_{bdr} is the ground truth boundary got from annotation image.

[1]We will discuss why the fuse the segmentation output and boundary information together, instead of using the segmentation output directly.

Boundary-Informed Segmentation Loss: Besides a fusing layer, the loss of boundary and segmentation is added to produce the final loss value. A coefficient λ adjust the weight of boundary loss to final loss. In the experiment, we set λ the value 6.

$$L_{\text{total}} = L_{\text{segmentation}} + \lambda \times L_{\text{boundary}}$$

4. EXPERIMENTS AND EVALUATION

Our network supports different MiT (Mix Transformer) backbones (B0-B4) across encoding stages (C1-C4), each with varying parameters. Due to resource constraints, we trained the Mit_B0 and Mit_B1 models, along with their **boundary-informed versions** (Mit_B0BI and Mit_B1BI), using identical input resolutions, loss functions, and training iterations for fair comparison.

For experiment setup, all models were trained on 512x512 resolution images over 25 iterations without pre-trained weights. In the following sections, Mit_B0BI refers to the Mit_B0 model with boundary-informed enhancements.

4.1 Results

During testing, the boundary-informed segmentation results generally outperform the simple SegFrom network. This is especially true for some simpler tasks, where boundary information can pay more attention to edges that are difficult to recognize.

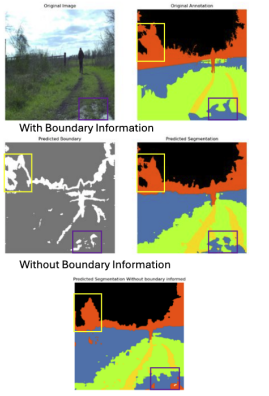

Figure 2: Boundary Information Guidance

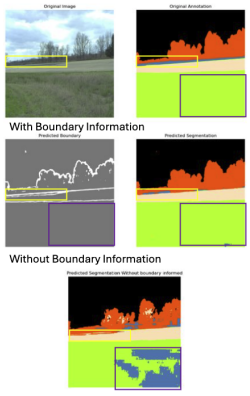

Figure 3: More Focused Attention

Figure 2 and Figure 3 presents two test cases, each illustrating the original image, ground truth, boundary detection, SegFormer segmentation result with boundary information (using model MiT_B0BI), and SegFormer segmentation result (using model MiT_B0). From the illustrated figures, we demonstrate that incorporating boundary information can enhance the focus of multi-head attention mechanisms on relevant regions, allowing the model to min- imize noise from other distinct areas.

The boundary map highlights edges within the scene, such as the contours of the tree and the pathway. Both figure show that by explicitly guiding the segmentation model with these boundary cues, Mit_B0BI produces sharper and more precise segmentation outputs. In the highlighted regions, the boundary-informed model (Mit_B0BI) more accurately delineates the tree's edges and better identifies the boundary between the muddy path and surrounding grassy areas.

Model	Params	BG	Smooth	Rough	Bumpy	Forbidden	Obstacle	MIoU
Mit_B0	7.7M	92.7	57.6	72.6	10.5	47.9	51.5	55.4
Mit_B0BI	8.4M	93.1	57.8	78.9	18.0	54.8	59.7	60.3
Mit_B1	30.7M	93.3	63.2	81.4	16.1	52.8	49.7	59.5
Mit_B1BI	33.4M	94.2	62.4	82.0	20.7	60.8	62.7	63.8

Table 1: Model performance comparison.(BI stand for Boundary-Informed model)

This boundary information guide the model to distinguish between adjacent terrain types, improving segmentation accuracy and detail, especially in complex, unstructured environments.

4.2 Evaluation

We evaluate our different model on the standard segmentation metrics: Intersection over Union (IoU), mean IoU (mIoU) and also model's parameters.

Intersection over Union (IoU) for class i:

$$IoU_i = \frac{\sum_I \sum_{x,y} 1(P(x,y) = i \text{ and } G(x,y) = i)}{\sum_I \sum_{x,y} 1(P(x,y) = i \text{ or } G(x,y) = i)}$$

Mean IoU (mIoU):

$$mIoU = \frac{\sum_i mIoU_i}{\sum_B 1}$$

MIoU (Mean Intersection over Union) is a metric that evaluates semantic segmentation performance by averaging the intersection over union between predictions and ground truth.

The tested models are Mit_B0, Mit_B0_BI, Mit_B1, Mit_B1_BI. Mit_B0 and Mit_B1 has $7.7M$ and $30.7M$ parameters respectively, and neither of them has been informed with boundary information.

Mit_B0_BI and Mit_B1_BI has $8.4M$ and $33.4M$ separately, both of them were informed with boundary information.

Table 1 shows the model evaluation, showing that models with boundary information have a better MIoU than the pure SegFormer models.

We even noticed that the model with boundary information and 8.4M parameters outperforms the model without boundary information, despite having more parameters (30.7M), achieving a higher MIoU.

4.3 Conclusion and Discussion

The model demonstrates improved performance with boundary-informed SegForm, showing about a 5% increase in MIoU, which confirms that incorporating boundary information enhances the original SegForm model. However, there is still significant room for further optimization.

Notably, the comparison section of the paper that proposed Gan-av[8] stated that the pure SegFormer can achieve an MIoU of 70 on this dataset. We hypothesize that the gap may be attributed to differences in pre-training, the refined loss function employed, or variations in the implementation of the SegFormer.

Given that we use weighted binary cross-entropy as the loss function, it is important to note a limitation highlighted in the InverseForm paper [7]: cross-entropy can be sensitive to slight image translations, sometimes resulting in a larger loss than even an incorrect segmentation. This suggests the need to explore alternative loss functions that are more suited to boundary detection.

Another challenge is how to effectively fuse SegForm and boundary-informed SegForm. Currently, we merge these two steps by concatenating their outputs and applying an additional convolutional layer. However, we believe there are more effective strategies to optimize the fusion process and improve the task performance.

5. REFERENCES

[1] H.-S. Yun, T.-H. Kim, and T.-H. Park, "Speed-bump detection for autonomous vehicles by lidar and camera," *Journal of Electrical Engineering & Technology*, vol. 14, pp. 2155–2162, 2019.

[2] H. Shon, S. Choi, and K. Huh, "Real-time terrain condition detection for off-road driving based on transformer," *IEEE Transactions on Intelligent Transportation Systems*, 2024.

[3] P. Jiang, P. Osteen, M. Wigness, and S. Saripalli, "Rellis-3d dataset: Data, benchmarks and analysis," 2020.

[4] M. Wigness, S. Eum, J. G. Rogers, D. Han, and H. Kwon, "A rugd dataset for autonomous navigation and visual perception in unstructured outdoor environments," in *International Conference on Intelligent Robots and Systems (IROS)*, 2019.

[5] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 568–578, 2021.

[6] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in neural information processing systems*, vol. 34, pp. 12077–12090, 2021.

[7] S. Borse, Y. Wang, Y. Zhang, and F. Porikli, "Inverseform: A loss function for structured boundary-aware segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5901–5911, 2021.

[8] T. Guan, D. Kothandaraman, R. Chandra, A. J. Sathyamoorthy, K. Weerakoon, and D. Manocha, "Ga-nav: Efficient terrain segmentation for robot navigation in unstructured outdoor environments," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 8138–8145, 2022.