Attention Guided Off-Road Semantic Segmentation

Omkar Sargar
M.S. in Robotics
Northeastern University
Boston, MA 02130
sargar.o@northeastern.edu

Ronak Bhanushali
M.S. in Robotics
Northeastern University
Boston, MA 02130
bhanushali.r@northeastern.edu

1 Introduction

1.1 Problem Statement

The problem of navigating in off-road environments is one of the most challenging tasks faced by autonomous vehicles. Unlike well-structured urban environments, off-road environments are characterized by their unpredictable and noisy nature, with a wide range of obstacles and terrain types that can impede vehicle movement. Traditional path planning and computer vision techniques are often insufficient to cope with the complexity of these environments, leading to suboptimal performance and safety risks. To address these challenges, there is a pressing need to develop robust and lightweight methods that can be deployed onboard autonomous vehicles. These methods should be capable of accurately identifying and classifying different terrains, obstacles, and objects to facilitate safe and efficient navigation. This problem is especially critical for applications such as search and rescue missions, precision agriculture, and autonomous driving, where reliable and accurate perception capabilities are essential.
Github

1.2 Motivation

Semantic Segmentation models show promising potential in this area as they provide a detailed and context oriented description of a given situation. By dividing the scene into different segments and labelling each segment, it is possible to to identify different elements of the environment, including terrain types, obstacles, and objects. These segments can be further used to create a costmap of the area. Given the potential of semantic segmentation models for addressing the challenges of off-road navigation, there is a strong motivation to develop more efficient and effective methods. These methods should be capable of providing accurate and reliable perception capabilities in real-time, while also being lightweight and robust enough to operate under challenging environmental conditions. The development of such methods has significant implications for the field of autonomous vehicles, enabling safer and more efficient navigation in off-road environments and contributing to the development of more advanced and sophisticated robotic systems.

2 Literature Survey

Semantic segmentation is a fundamental task in computer vision, and has been widely researched and studied in the literature. Over the years, several methods have been proposed for semantic segmentation, ranging from traditional approaches based on random forests [1], [2] to deep learning-based methods such as Fully Convolutional Networks FCN [3], SegNet [4], and U-Net [5]. Deep learning-based methods have gained significant attention due to their ability to learn high-level features and effectively handle semantic segmentation problems.

Traditional deep learning models struggle to perform well on off-road datasets due to unclear boundaries and ambiguous class features. Prior work in unstructured and uneven terrain navigation include [6] which uses a proprioceptive-based model to classify terrains using vibration data. [7], [8] present appearance-based models that leverage data from sensors like cameras and radar.

Attention modules have been discussed in [9] where the authors used attention modules to aggregate multi-scale context information. More recently, the transformer architecture, which uses self-attention to weigh the importance of different parts of the input, has been shown to outperform traditional convolutional neural networks (CNNs) on several datasets. [10] proposes GA-Nav, a novel group-wise attention mechanism to distinguish between navigability levels of different terrains, which can improve different backbone designs. [11] proposes a novel hierarchically structured Transformer encoder that outputs multiscale features and avoids complex decoders.

3 Methods

We successfully implemented three deep-learning models for semantic segmentation. U-Net was taken as the baseline model for this problem statement as it is a relatively lightweight model which has been recorded to perform well even on smaller datasets.

3.1 U-Net

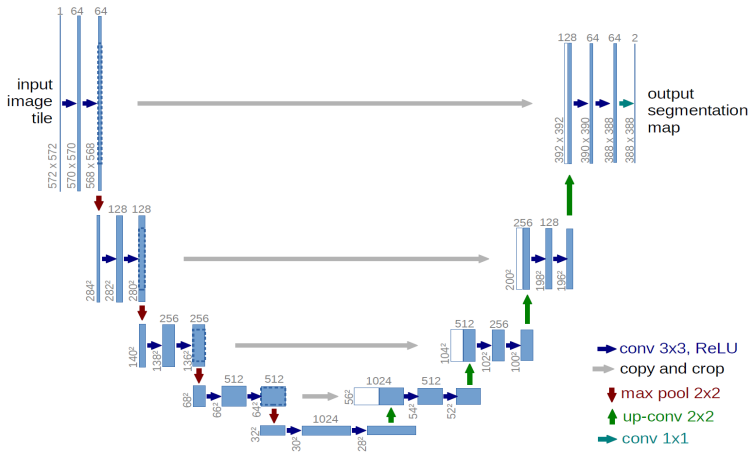

Figure 1: U-Net architecture

U-Net is a popular and effective deep-learning architecture for semantic segmentation tasks. It was proposed by Ronneberger et al. in 2015 and has since been widely adopted in the computer vision community. The architecture consists of an encoder and a decoder network, where the encoder network progressively reduces the spatial resolution of the input, while the decoder network upsamples the feature maps to the original resolution. The encoder consists of several convolutional layers, each followed by a rectified linear unit (ReLU) activation and a max-pooling operation. These operations progressively reduce the spatial resolution of the input and increase the number of channels. The decoder network, consists of several convolutional layers and up-sampling operations, each followed by a ReLU activation. The up-sampling operations increase the spatial resolution of the feature maps and decrease the number of channels. Skip connections are used to connect corresponding feature maps from the encoder and decoder networks, allowing the model to combine low-level and high-level features. The skip connections provide the decoder network with access to both fine-grained and high-level information, enabling accurate segmentation of objects at different scales. The final layer of the decoder network is a 1×1 convolutional layer followed by a softmax activation, which outputs a probability distribution over the classes for each pixel in the input image.

U-Net's ability to capture both local and global context makes it a suitable choice for handling the challenges presented by unstructured environments. Additionally, U-Net is a relatively lightweight and efficient model architecture, making it a practical choice for real-world robotic applications. The number of parameters in U-Net is relatively low, which reduces the risk of overfitting and allows the model to be trained with limited data. The model architecture is also quite simple compared to some other deep learning methods, making it easier to implement and train. These factors made U-Net an attractive baseline model for our task.

3.2 Attention U-Net

Attention U-Net is a variant of the original U-Net architecture that incorporates attention mechanisms to improve the accuracy of semantic segmentation. [12] proposed a novel attention gate (AG) model for medical imaging that automatically learns to focus on target structures of varying shapes and sizes. Models trained with AGs implicitly learn to suppress irrelevant regions in an input image while highlighting salient features useful for a specific task. The proposed Attention Gates are incorporated into the standard U-Net architecture to highlight salient features that are passed through the skip connections, see Figure 2. Information extracted from the coarse scale is used in gating to disambiguate irrelevant and noisy responses in skip connections. This is performed right before the concatenation operation to merge only relevant activations.

Attention mechanisms help the model to focus on the most informative parts of the input image, allowing it to allocate more resources to important regions and suppress irrelevant information. These attention blocks allow the model to capture more precise spatial relationships between objects, making it more effective at segmenting objects with complex shapes and fine-grained details. The attention blocks also allow the model to better handle variations in lighting conditions, occlusions, and other challenges commonly encountered in unstructured environments. These features made Attention U-Net a promising solution for a model that could potentially surpass the performance of the baseline model.

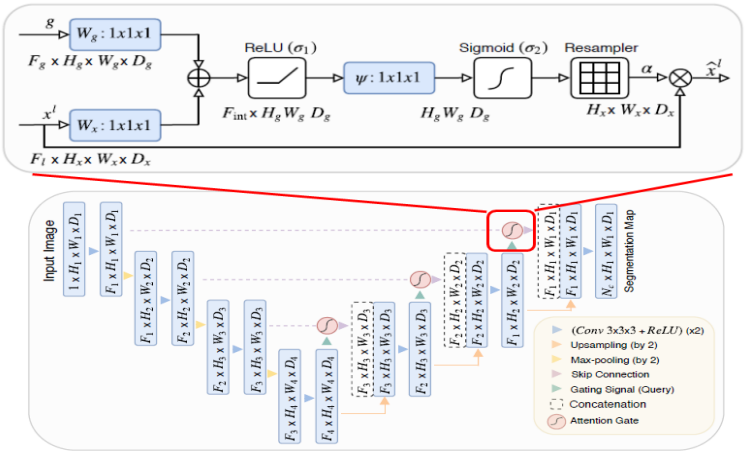

Figure 2: Attention U-Net architecture

3.3 Segformer

Segformer is a recently proposed architecture for semantic segmentation that combines the advantages of transformer architecture and convolutional neural networks (CNNs). The transformer architecture has been widely used in natural language processing tasks and has recently been applied to computer vision tasks with promising results. Unlike CNNs, transformers operate on sequences of vectors, allowing them to model global relationships between different parts of the input. In the Segformer architecture, the input image is first divided into non-overlapping patches, which are processed by hierarchical transformer encoder. The transformer layers learn to capture global dependencies

between different parts of the input, allowing the model to better understand the context of each patch. The decoder consists of a lightweight All-MLP block to fuse these multi-level features to produce the final semantic segmentation mask. Segformer has been shown to achieve state-of-the-art performance on various semantic segmentation benchmarks, especially in scenarios where objects are small and densely packed.

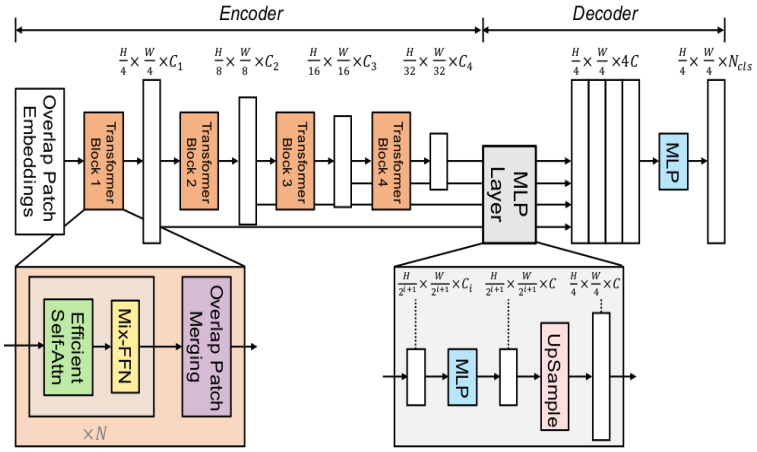

Figure 3: Segformer architecture

Compared to U-Net and Attention U-Net, Segformer has several advantages. First, Segformer can better capture global dependencies between different parts of the input, allowing it to better model long-range interactions between objects. This makes it particularly effective in scenarios where objects are densely packed or where contextual information is important for accurate segmentation. Additionally, the recorded performance of Segformer on other datasets also made Segformer a promising choice for semantic segmentation in unstructured environments, especially in scenarios where objects are small and densely packed.

3.4 Using Weighted Cross Entropy Loss

Another method that could help in increasing the accuracy of the model is using an updated loss function that accounts for the frequency of classes in each label and accordingly assigns weights to all classes. This loss function is inspired from loss for Long Tailed Learning [https://arxiv.org/pdf/2110.04596.pdf]. This loss function prioritizes the classes with less data points in loss function rather than the high frequency classes. This method could improve accuracy lost due to the class imbalance problem and help detect the low occuring classes in a given frame. Class weightage is determined as follows

$$\alpha_i = \frac{1-\beta}{1-\beta^{\mathrm{n}_i}}$$

where:

α_i: Weight of class i for loss calculation

β: Hyperparameter that determines the weight given to the classes

n_i: Count of samples for class i in the current batch

4 Experiments

4.1 Experimental Settings

The idea is to train the models individually on each dataset. The evaluation metrics that we plan to use are mIoU, Pixel Accuracy, F1 score and Recall. We look forward to investigate the following things -

1. How does the addition of an attention mechanism to the baseline model affect performance?

2. How does the use of Weighted Cross-Entropy affect the model performance?

4.1.1 Datasets

We used two publicly available datasets: RUGD and RELLIS-3D, in our experiments. The RUGD dataset contains over 7,400 high-resolution annotated images captured from a variety of urban and rural environments, such as forests, mountains, and deserts, with diverse lighting and weather conditions. Each image in the RUGD dataset is annotated with pixel-level semantic labels for 24 different classes. For experimenting with U-Net and Attention U-Net the RUGD dataset was spilt to consist of 5934 images for training and 742 images each for validation and testing. Due to the high computational resource requirement for training Segformer, a smaller dataset of 26 images for training and 8 images, each for validation and testing was used in the experiments with Segormer on RUGD. The RELLIS3D dataset contains over 13,556 LiDAR scans and 6,235 annotated images captured from various off-road environments, such as fields, forests, and trails, with diverse terrain and vegetation conditions. Each scan in the RELLIS3D dataset is annotated with voxel-level semantic labels for 20 different classes. For experimenting with U-Net and Attention U-Net the RELLIS-3D dataset was spilt to consist of 3303 images for training, 982 images for validation, and 1674 images for testing.

4.1.2 Implementation Details

Following model parameters were used to train the models:

Table 1: Parameters for training on RUGD

Method	Epochs	Learning Rate	Optimiser	Batch Size	Momentum
U-Net	12	0.001	Adam	16	0.9
Attention U-Net	13	0.001	Adam	16	0.9
Segformer	25	0.00006	AdamW	2	-

For training on RUGD the ReduceLROnPlateau scheduler was used with the following parameters: factor=0.1, patience=3.

Table 2: Parameters for training on RELLIS-3D

Method	Epochs	Learning Rate	Optimiser	Batch Size	Momentum
U-Net	30	0.008	SGD+Momentum	16	0.95
Attention U-Net	30	0.006	SGD+Momentum	16	0.95

For training on RELLIS-3D the ReduceLROnPlateau scheduler was used with the following parameters: factor=0.1, patience=3.

4.2 Evaluation of U-Net, Attention U-Net and Segformer on RUGD

The performance of three different semantic segmentation models, U-Net, Attention U-Net, and Segformer, was evaluated on the RUGD dataset. The results show that U-Net achieved a mean Intersection over Union (mIoU) score of 0.1320, a Pixel Accuracy of 0.8226, an F1 Score of 0.1488,

and a Recall of 0.1600. Attention U-Net achieved a slightly lower mIoU score of 0.1262, but had a slightly higher Pixel Accuracy of 0.8242, F1 Score of 0.1418, and Recall of 0.1529.

Table 3: Evaluation of U-Net, Attention U-Net and Segformer on RUGD

Method	mIoU	Pixel Accuarcy	F1 Score	Recall
U-Net	0.1320	0.8226	0.1488	0.1600
Attention U-Net	0.1262	0.8242	0.1418	0.1529
Segformer*	0.2732	0.3331		

*Segformer was trained and tested on a smaller dataset

The high pixel accuracy in both U-Net and Attention U-Net suggest that the model is correctly classifying most of the pixels, but the spatial overlap between the predicted and ground-truth masks is low. This is due to the fact that there exists a huge class imbalance in the RUGD dataset and the model has not been able to learn to segment the classes in the long tail of the dataset.

Even though Segformer was trained on a substantially smaller dataset, it achieved the highest mIoU score of 0.2732, and had a reasonable corresponding pixel accuracy of 0.3331 compared to U-Net and Attention U-Net. This result may be attributed to the fact that Segformer is designed to capture long-range dependencies between different parts of the input, making it particularly more effective in scenarios where the objects are densely packed or where contextual information is important for accurate segmentation. The closeness in the mIoU and pixel accuracy for Segformer suggests that it was not affected as severely as the other models by class imbalance.

4.3 Evaluation of U-Net and Attention U-Net on RELLIS-3D

Table 4: Evaluation of U-Net and Attention U-Net on RELLIS-3D

Method	mIoU	Pixel Accuarcy	F1 Score	Recall
U-Net	0.1632	0.8943	0.1811	0.8943
Attention U-Net	0.1866	0.9422	0.1997	0.2021
Attention U-Net with weighted cross entropy	0.0055	0.0357	0.0098	0.0226

The performance of two different semantic segmentation models, U-Net, and Attention U-Net was evaluated on the Rellis-3D dataset. Similar to the results of RUGD, the mIoU is low but the pixel accuracy is quite high because it fails to detect the less occurring classes. However in case of Rellis-3D the mIoU, Pixel Accuracy, F1 Score and recall all go up after adding attention. As evident from the **??** certain classes such as "barrier" were accurately segmented after adding the attention module. Its evident that the weighted cross entropy loss seems to add a lot of weightage to the low occurring classes which leads to the mask containing majority of those classes as seen in figure 4. The model is very sensitive to the hyper parameter β. The results seen are with β set to 0.5.

5 Conclusion and Future Work

5.1 Discussion

In our project, we compared three different models: U-Net, Attention U-Net, and Segformer for performing semantic segmentation in challenging off-road environments. Our hypothesis that adding a visual attention module will help in increasing the performance of baseline models was proved true through our testing. The Segformer model performed the best in terms of mIoU, followed by Attention U-Net and then the vanilla U-Net. Although the performance of Attention U-Net is better than U-Net the difference between their performance is quite smaller than we expected. This is something we plan to investigate, along with some other avenues.

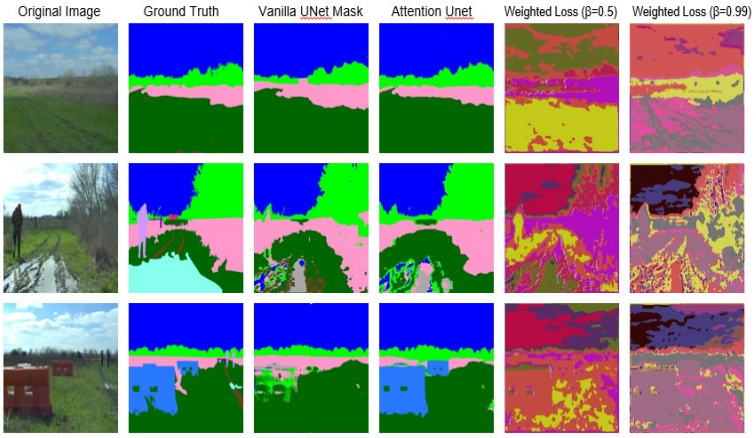

Figure 4: Inference output for different methodologies

5.2 Future Work

The future scope of the project involves the following.

Investigating other attention mechanisms: While the attention gate used in this project has shown to be effective, there are other attention mechanisms such as self-attention and channel attention that could be explored for further improvement.

Incorporating other sensor modalities: To improve the robustness of the model, integrating other sensor modalities such as LIDAR which is readily available in the Rellis 3D dataset could be explored. This could provide the model with more comprehensive information about the environment, which could improve its ability to segment the scene accurately.

Incorporating distributional robustness loss: As seen, class imbalance can really affect the way certain classes are detected. One potential solution is to incorporate distributional robustness loss into the training process. This would encourage the model to learn features that are more robust to distributional shift, improving its generalization ability in new and unseen environments. Exploring the effectiveness of this approach on the proposed model could be a valuable future direction.

Transfer learning: Transfer learning has shown to be an effective method for training deep learning models on small datasets. Using pretrained weights for a similar Attention U-Net model or a vision transformer model followed by transfer learning has potential to significantly improve performance.

Deployment on a real-world robot: The ultimate goal of this research is to deploy the model on an autonomous vehicle for off-road navigation. Therefore, testing the model on a real-world robot in a variety of off-road environments would be an important next step.

Exploration of other segmentation architectures: In addition to UNet and Segformer, other segmentation architectures such as DeepLab v3, PSPNet, and BiSeNet v2 could be explored to compare their performance with the proposed method.

References

[1] Zhensong Wang, Lifang Wei, Li Wang, Yaozong Gao, Wufan Chen, & Dinggang Shen. (2018, February). Hierarchical vertex regression-based segmentation of head and neck CT images for radiotherapy planning. IEEE transactions on image processing : a publication of the IEEE Signal Processing Society. Retrieved April 27, 2023, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5954838/

[2] Gao, Y., Shao, Y., Lian, J., Wang, A. H., Chen, R. C., Shen, D. (2016). Accurate Segmentation of CT Male Pelvic Organs via Regression-Based Deformable Models and Multi-Task Random Forests. IEEE Transactions on Medical Imaging, 35(6), 1532–1543. https://doi.org/10.1109/tmi.2016.2519264

[3] Long, J. (2014, November 14). Fully Convolutional Networks for Semantic Segmentation. arXiv.org. https://arxiv.org/abs/1411.4038

[4] Badrinarayanan, V. (2015, November 2). SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. arXiv.org. https://arxiv.org/abs/1511.00561

[5] Ronneberger, O. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. https://lmb.informatik.uni-freiburg.de/Publications/2015/RFB15a/

[6] Wilson, G., Ramirez-Serrano, A. (2014). Terrain Roughness Identification for High-Speed UGVs. Journal of Automation and Control Research. https://doi.org/10.11159/jacr.2014.002

[7] Matsuzaki, S., Yamazaki, K., Hara, Y., Tsubouchi, T. (2018). Traversable Region Estimation for Mobile Robots in an Outdoor Image. Journal of Intelligent and Robotic Systems, 92(3–4), 453–463. https://doi.org/10.1007/s10846-017-0760-x

[8] Reina, G., Milella, A., Rouveure, R. (2015). Traversability analysis for off-road vehicles using stereo and radar data. In HAL (Le Centre pour la Communication Scientifique Directe). Le Centre pour la Communication Scientifique Directe. https://doi.org/10.1109/icit.2015.7125155

[9] W. Wang, Y. Gao, H. Zhang, Y. Liu, and X. Zhang, "LAANet: Lightweight attention-guided asymmetric network for real-time semantic segmentation," in 2021 IEEE International Conference on Multimedia Expo Workshops (ICMEW), Jul. 2021, pp. 1-6. doi: 10.1109/ICMEW53674.2021.9439127.

[10] Geometric Algorithms for Modeling, M. (n.d.). GANAV: Efficient Terrain Segmentation for robot navigation in unstructured outdoor environments. GAMMA. Retrieved April 27, 2023, from https://gamma.umd.edu/researchdirections/autonomousdriving/offroad/

[11] Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., amp; Luo, P. (2021, October 28). Segformer: Simple and efficient design for semantic segmentation with Transformers. arXiv.org. Retrieved April 27, 2023, from https://arxiv.org/abs/2105.15203

[12] Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., Kainz, B., Glocker, B., amp; Rueckert, D. (2018, May 20). Attention U-net: Learning where to look for the pancreas. arXiv.org. Retrieved April 27, 2023, from https://arxiv.org/abs/1804.03999