COMP9517 Group Project – Semantic Segmentation

An innovative approach to 2D WildScene Natural Environments

Liam Chen   
email: z3278107@ad.unsw.edu.au  
student id: z3278107

Jennifer Wang   
email: z5311246@ad.unsw.edu.austudent id: z5311246Shree Baskar  
email: z5444481@ad.unsw.edu.austudent id: z5444481

Amulya Lokeshwari Sri Raja Kalidindi  
email: z5520321@ad.unsw.edu.austudent id: z5520321The Duy Nguyen   
email: z5532839@ad.unsw.edu.austudent id: z5532839

*Abstract*—TODO. The output module is generated using a computer with GPU memory NVIDIA GeForce RTX 3060 6G, programming with python 3.8.19. A python requirements.txt can be found in the submission

Keywords—semantic segmentation, deep learning, image classification, convolutional networks, Jaccard similarity, Intersection over Union (IoU), U-Net, SegNet, VGG, Deep CNN, ensemble learning, semantic scene understanding

# Introduction

Semantic segmentation (per-pixel class labelling) is a critical task in computer vision (Szeliski, 2021, p.413) and it has gained a lot of attention in the research field recently. This task is significant in applications such as autonomous driving, environmental monitoring, search and rescue, and automation, where understanding of surrounding environments is required to navigate and perform tasks effectively. (Vidanapathirana et al. 2023). While there are many notable structured 2D and 3D urban scene datasets to study the urban environment semantic segmentation (Vidanapathirana et al. 2023), natural scene semantic segmentation is less common and therefore warrants increased focus.

To address this gap, the WildScenes dataset was created (Vidanapathirana et al., 2023). It contains RGB images of traversals over six months across two locations: Venman National Park and Karawatha Forest Park in Brisbane, Australia. A total of 21.28 km was travelled and 9306 images were captured, each with a resolution of 2016 x 1512. Each image was annotated manually using a coarse-to-fine approach with multiple rounds of auditing to overcome the challenges in labelling natural environments. These challenges include ambiguity of object boundaries for elements with similar features (i.e., leaves and shrubs, or dirt and grass), and differentiating similar regions.

This project aims to implement natural scene semantic segmentation using state-of-the-art deep learning models such as U-Net, SegNet and VGG. We will leverage powerful architectures to achieve high accuracy in segmenting various natural elements that are above the benchmarks published in the original research paper. Additionally, techniques like superpixels and Conditional Random Fields (CRFs) will be incorporated to refine the segmentation boundaries, enhancing the overall performance of the model.

# Literature Review

Semantic segmentation is the process of classifying each pixel in an image as belonging to one of a set of predefined classes. Many techniques have been developed over the years, ranging from early image-processing techniques to convolutional neural networks (CNNs).

The work of Shelhamer et al. (2015) introduced Fully Convolutional Networks for semantic segmentation. These replaced the fully connected layers in traditional CNNs with a decoding, or upsampling, layer to enable pixel-wise classification on the original image scale. This paved the way for further developments in deep learning such as U-Net (Ronneberger et al., 2015), DeepLab (Chen et al., 2018), and SegNet (Badrinarayanan et al., 2017).

Despite these recent advancements, natural scene segmentation has received little attention in research. It poses additional challenges due to its less structured and highly variable elements (Vidanapathirana et al., 2023).

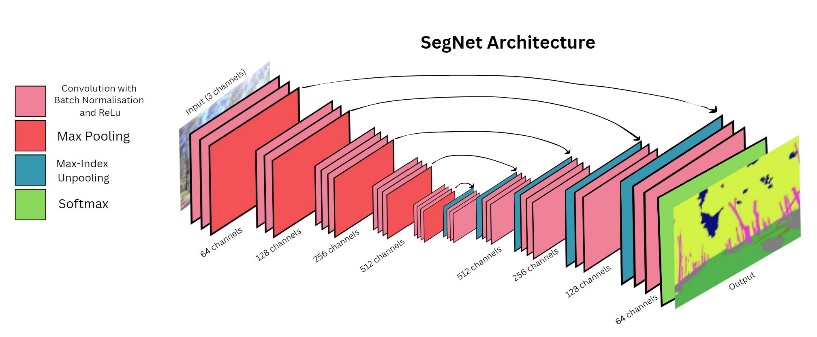
## Fully Convolutional Neural Networks for Semnatic Segmentation (Shelhamer et al., 2015)

This method, proposed by Shelhamer et al., addresses the problem of applying traditional CNNs to a pixel-wise classification problem. Instead of using a final, fully connected classification layer, they use a fully convolutional encoder-decoder architecture to produce an output with the same spatial resolution of the input and compute pixel-wise labels.

Each layer in the FCN is a 3D array with dimensions height, width, and number of channels. The first layer is always the original input image. The Encoder layers are similar to traditional CNNs, involving convolutional layers followed by activation functions and downsampling to extract features in a fine-to-coarse manner (low level features to high-level features). Then, the spatial resolution of the output is increased in the Decoder layers using learned deconvolution and upsampling layers. Skip-connections are used to connect the coarse outputs back to the original pixels for pixel-wise classification. It was found that utilising more low-level information improved the quality of the segmentation as it was able to preserve boundary information. Upscaling from the final encoder layer directly produced a very coarse output with little boundary definition.

This method could be utilised for the WildScenes dataset as it is proven to be effective for image segmentation. However, given the ambiguous and detailed boundaries present in natural environments, additional care should be taken to ensure the boundaries are well-defined in the segmentation. Some methods to address this challenged are mentioned below.

## SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation (Badrinarayanan et al., 2015)

The SegNet architecture has an encoder, decoder and a pixel-wise classification layer. This differs from FCN in its upsampling method. The decoder upscales the feature maps by storing indices used in max pooling by the encoder in place of skip connections. This save storage space and increases the efficiency of the algorithm as compared to other FCNs which utilise the entire feature map during upsampling. The decoder is followed by a final, pixel-wise classification layer which employs a softmax activation function to produce class probabilities. This model was found to be effective for semantic segmentation and slightly more efficient.

Areas with limited variation were classified more accurately. A weakness of SegNet is chaotic/cluttered scenes, which was attributed to the presence of infrequent classes. Training on a larger dataset with more balanced class distributions was proposed as a solution.

This model could work for the WildScenes dataset as it is proven to be effective for the task of semantic segmentation. However, some natural scenes have high variability, so we might run into the same issue as above, where the model may find classifying infrequent classes difficult.

## Superpixels and Conditional Random Fields (Zhao et al., 2018)

This paper presents a method for semantic segmentation utilising superpixels and conditional random fields (CRFs). Traditional methods that have performed well often struggle with producing well-defined and accurate boundaries. The proposed method aims to overcome this limitation.

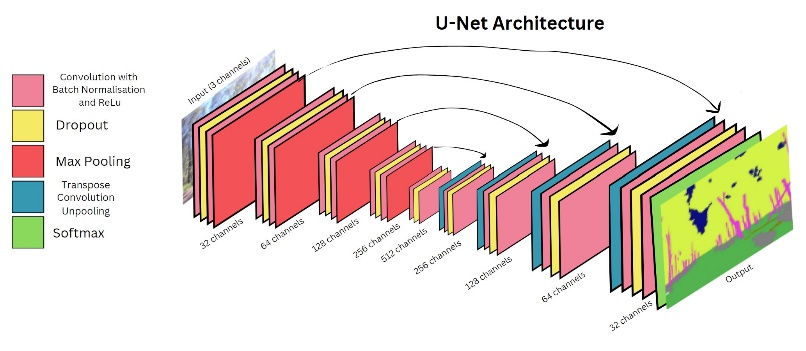
The method generates superpixels using Simple Linear Iterative Clustering. It extracts features from these superpixels. A CRF model is then created, where superpixels are the nodes and the relationships between them are edges. The segmentation problem is the formulated as an energy minimisation problem which can be solved using graph cuts.

This was particularly effective in improving the edge definition of the FCN, which loses some low-level information to upsampling. First, the FCN was used to define a baseline segmentation. Then, the output of the CRF was fused with the segmentation to refine boundary definition.

## Learning Hierarchical Features for Scene Labelling (Farabet et al., 2012)

A challenge faced in semantic segmentation is learning objects at different scales. Sometimes, the label class might depend on long-range information that is not captured through a convolutional filter. This paper proposes a multi-scale approach to learn features hierarchically. The network processes the input at multiple scales using image pyramids to capture local and global context. The CNN produces an initial pixel-wise classification. Then, the image is decomposed into superpixels producing a second, over-segmentation. This helps refine the original segmentation using graph-based optimisation on the superpixels, utilising CRFs for smoothing of boundaries. This was found to be effective especially when images had objects of various sizes. It achieved a similar boundary definition as the previous method, having used a similar boundary optimisation method.

## **TODO – make the above more concise. Maybe discuss the pros and cons of each. Add U-Net and SegFormer.**

**Figure 1:** SegNet architecture diagram. There are 13 convolutional layers in both the encoder and the decoder. Each set of convolutional layers in the encoder is followed by max pooling, and each set in the decoder is followed by max-index unpooling. The final layer is a softmax classification, which assigns a label to each pixel.

**Figure 2**: U-Net architecture diagram. There are 13 convolutional layers in the encoder and decoder. Each set of convolutional layers in the encoder is followed by max pooling, and each set in the decoder is followed by unpooling using transpose convolution and skip-connection. The final classification layer uses a soft-max function, and in between convolutional layers, dropout is used to prevent overfitting.

# Methods

The WildScenes dataset was used for this task. It contains 9306 RGB images with a resolution of 2016 x 1512, annotated manually using a coarse-to-fine approach.

The dataset was split using the methods from the original paper (Vidanapathirana et al., 2023). The dataset was split into training, validation, and tests sets to ensure a good class distribution and avoid geographical overlap. A minimum distance of 45 meters was kept between samples in different sets, grouped based on their (x, y) coordinates using k-means clustering. One thousand candidate splits were generated, evaluated, and the optimal split was chosen. We used the entire dataset to train our models, and manually verified the class distribution in the training, validation, and testing datasets.

## **Data Split**

## **Model Building**

### **U-Net Baseline**

### **SegNet Baseline**

### **SegFormer Baseline**

### **Superpixels and Conditional Random Fields**

### **Traditional Segmentation Methods**

## **Evaluation Criteria**

Table Type Styles

| Table Head | Table Column Head | | |
| --- | --- | --- | --- |
| Table column subhead | Subhead | Subhead |
| copy | More table copya |  |  |

1. Sample of a Table footnote. (*Table footnote*)
2. Example of a figure caption. (*figure caption*)

# Experimental Results

# Discussion

# Conclusion

# References

CSIRO, “WildScenes Dataset.” Version 2, Jun. 12, 2024. [Online]. Available: <https://doi.org/10.25919/5hzc-5p73>

K. Vidanapathirana, M. Raison, S. Denman, M. J. Milford, and S. Anavatti, “WildScenes: A Benchmark for 2D and 3D Semantic Segmentation in Large-scale Natural Environments,” arXiv:2312.15364, 2023. [Online]. Available: <https://arxiv.org/abs/2312.15364>

CSIRO, “WildScenes: A Benchmark for 2D and 3D Semantic Segmentation in Natural Environments,” 2023. [Online]. Available: <https://csiro-robotics.github.io/WildScenes/>

V. Badrinarayanan, A. Kendall, and R. Cipolla, “SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 12, pp 2481-2495, Dec. 2017. doi: 10.1109/TPAMI.2016.2644615.

J. Long, E. Shelhamer, and T. Darrell, “Fully Convolutional Networks for Semantic Segmentation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, June 2015, pp.3431-3440. doi: 10.1109/CVPR.2015.7298965.

R. Szeliski, Computer Vision: Algorithms and Applications, 2nd ed. New York, NY, USA: Springer, 2021, p. 413.

O. Ronneberger, P. Fischer, and T. Brox. “U-Net: Convolutional Networks for Biomedical Image Segmentation,” in Proceedings of the International Conference on Medical Computing and Computer-Assisted Intervention (MICCAI), Munich, Germany, Oct. 2015, pp. 234-241. doi: 10.1007/978-3-319-24574-4\_28.

L.-C. Chen, G. Papanderou, F. Schroff, and H. Adam, “Rethinking Atrous Convolution for Semantic Image Segmentation,” arXiv preprint arXiv:1706.05587, 2017. [Online]. Available: <https://arXiv.org/abs/1706.05587>

C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning Hierarchical Features for Scene Labeling," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 8, pp. 1915-1929, Aug. 2013. doi: 10.1109/TPAMI.2012.231.

H. Zhao, Y. Shi, and S. Zhang, "Superpixels and Conditional Random Fields for Image Segmentation: A Comprehensive Study," IEEE Transactions on Image Processing, vol. 27, no. 8, pp. 3727-3741, Aug. 2018. doi: 10.1109/TIP.2018.2825123.

**TODO: change references to the below format (in-text references using [1], [2], etc) after all references are finished**

1. G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*
2. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
3. I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.