Semantic Segmentation for Off-Road Traversal

E. Marquez, A. Niemczura, M. Ellis, R. Chen, P. Eyabi, M. Faykus, A. Pickeral, M. Smith
Holcombe Department of Electrical and Computer Engineering, Clemson University

Abstract

For Off-Road Unmanned Ground Vehicles (UGVs) to safely traverse terrain, they require detailed understandings of their surrounding environment. RGB cameras are typically used to capture a vehicle's surroundings. To translate camera data into meaningful insights, we use semantic segmentation.

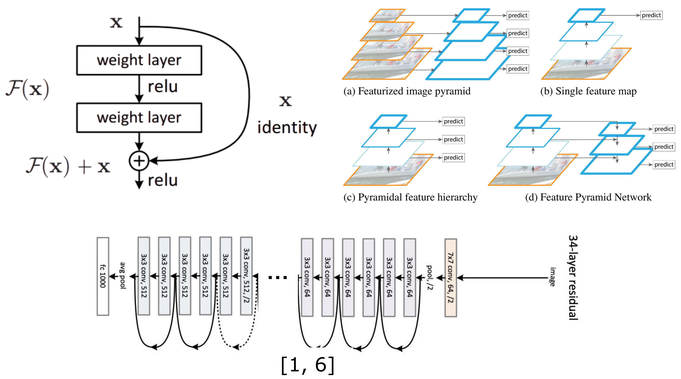
Semantic segmentation classifies every pixel in an image, providing fine-grained detail about an image scene. This makes it a strong task for determining traversable terrain. However, semantic segmentation typically has a high computational cost, which can be detrimental to real-time systems. Ensuring high accuracy in segmentation with low computational cost is paramount.

This research aims to evaluate different types of deep learning architectures in predicting traversable terrain using semantic segmentation for safe autonomous driving while maintaining fast inference speeds.

Background

FPN-ResNet

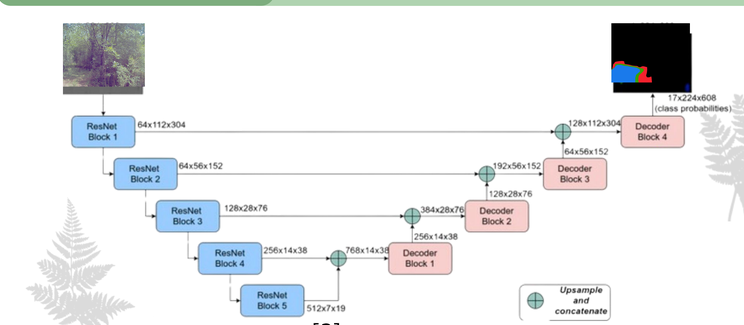
Features Residuals (skip connections) that allow for very deep networks



[1, 6]

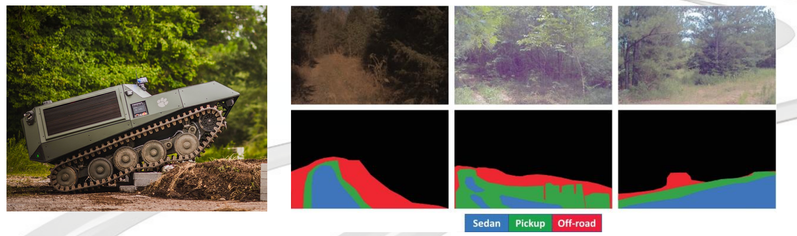
UNet-ResNet

Unet w/ ResNet as the encoder for the residual connection strength



[3]

Methodology



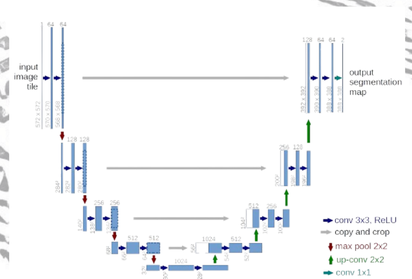
CaT Off-Road Dataset [5]

$$\mathrm{IoU}_{\mathrm{class}} = \frac{\mathrm{Prediction}_{\mathrm{class}} \cap \mathrm{GroundTruth}_{\mathrm{class}}}{\mathrm{Prediction}_{\mathrm{class}} \cup \mathrm{GroundTruth}_{\mathrm{class}}}$$

$$m\mathrm{IoU} = \frac{\sum \mathrm{IoU}_{\mathrm{class}}}{n_{\mathrm{classes}}}$$

UNet

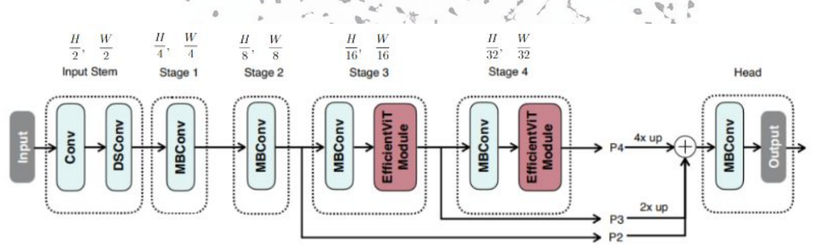
Uses a contraction and expansion path to combine fine-grained and coarse details



[2]

EfficientViT

Merges convolutional layers with lightweight attention blocks (EfficientViT Module)



[4]

Results

Inference and Memory Metrics (NVIDIA V100)

Model	Inference Speed	Parameters	Model Size	Inference Memory Usage
UNet	20.17 ms	30 M	118 MB	523 MB
FPN-RN34	35.09 ms	22 M	91 MB	978 MB
UNet-RN34	16.98 ms	24 M	93.4 MB	**371 MB**
EfficientViT	**14.62 ms**	**0.7 M**	**2.9 MB**	402 MB

Individual Class and Mean IoU Accuracy (%)

Model	Background	Sedan	Pickup	Off-Road	mIoU
UNet	90.40	88.25	60.28	77.13	78.94
FPN-RN34	90.34	88.01	60.24	77.12	70.01
UNet-RN34	**98.87**	94.88	86.84	85.78	90.84
EfficientViT	98.16	**98.22**	**92.01**	**93.09**	**95.37**
PSPNet*	_	91.64	69.08	81.00	80.57

*CaT baseline state of the art results [5]

Conclusion

Ultimately, EfficientViT shows strong results for perception in off-road terrain. The combination of convolutional layers and lightweight attention mechanisms leads to strong accuracy compared to previous baseline state-of-the-art results while maintaining a fast inference speed. Additionally, low model size and inference memory usage make EfficientViT ideal for edge devices with lower memory and computational power capabilities. Our results prove this model is suitable for real-world deployment on a UGV.

Acknowledgements

Thank you to the College of Engineering, Computing and Applied Sciences and Clemson's Undergraduate Research Program for their support in this research. Clemson University is also acknowledged for their generous allotment of compute time on the Palmetto Cluster.

Sources

[1] T. -Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan and S. Belongie, "Feature Pyramid Networks for Object Detection," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 936-944, doi: 10.1109/CVPR.2017.106.
[2] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18* (pp. 234-241). Springer International Publishing.
[3] Neven, Robby, and Toon Goedemé. 2021. "A Multi-Branch U-Net for Steel Surface Defect Type and Severity Segmentation" *Metals* 11, no. 6: 870. https://doi.org/10.3390/met11060870
[4] H. Cai, J. Li, M. Hu, C. Gan, and S. Han, "EfficientViT: Multi-Scale Linear Attention for High-Resolution Dense Prediction," arXiv e-prints, p. arXiv:2205.14756, May 2022.
[5] Sharma, S., Dabbiru, L., Hannis, T., Mason, G., Carruth, D. W., Doude, M., ... & Tang, B. (2022). Cat: Cavs traversability dataset for off-road autonomous driving. *IEEE Access*, *10*, 24759-24768.
[6] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).