EXPLORING UNCHARTED TERRAIN: OFF-ROAD DRIVABLE AREA SEGMENTATION WITH CYCLEGAN

Prakadeeswaran Manivannan, Harish Gnana Sekaran, Supreeth Gadamsetty & Amitabha Deb
CSCI 5527: Deep Learning Project Progress Report
University of Minnesota, Twin Cities

1 INTRODUCTION

Making off-road environments accessible to autonomous vehicles presents an exhilarating yet formidable challenge. Off-road terrains are characterized by their diversity, featuring rocky paths, mud, sand, vegetation, and more, each demanding distinct navigation strategies. The scarcity of labelled data and the limitations of sensors in these challenging conditions make it demanding to develop robust perception and control systems. Real-time decision-making is essential, ensuring the vehicle can adapt swiftly to the continuously shifting off-road conditions, while also addressing safety concerns that arise in the absence of conventional road infrastructure. Interacting with dynamic elements like moving obstacles and changing terrain adds complexity, further testing the adaptability of autonomous systems. This endeavour is intriguing due to the variety of problems it presents, its real-world impact on applications like search and rescue, and the need for technological innovation across sensor technology, computer vision, machine learning, and control systems. It also offers opportunities for environmental sustainability and the enhancement of adventure and recreational activities, making it a field ripe with potential and excitement.

2 PROBLEM DEFINITION

Producing highly accurate ground truth masks for off-road image segmentation poses a labour-intensive challenge, which is exacerbated by the limited availability of off-road datasets. This scarcity becomes particularly apparent when compared to the abundance of urban datasets like cityscapes.

3 LITERATURE REVIEW

Image segmentation is a vital computer vision technique that enables the precise identification and delineation of objects of regions of interest within digital images. Mask R-CNNHe et al. (2017) is a state-of-the-art deep learning architecture used for instance segmentation, which is the task of not only detecting objects in an image but also segmenting them into precise pixel-level masks. It builds upon the Faster R-CNNRen et al. (2015) architecture, which is a widely used framework for object detection. At a high level, Mask R-CNN consists of several modules. Backbone is a standard convolution neural network (typically, ResNet50 or ResNet101)He et al. (2015) that serves as a feature extractor. Feature Pyramid Network (FPN) improves the standard feature extraction pyramid by adding a second pyramid that takes the high-level features from the first pyramid and passes them down to the lower layer.

The U-Net paper Ronneberger et al. (2015) introduces a convolutional neural network architecture designed for biomedical image segmentation tasks. Notable for its U-shaped architecture, the model combines contracting and expansive paths to capture contextual information and refine segmentation. Skip connections aid in preserving spatial information during the upsampling process. U-Net has proven effective in various medical image segmentation applications, providing accurate delineation of structures in biomedical images. The architecture's versatility and success in segmentation tasks have made it a pivotal contribution to the field of deep learning in medical imaging.

In the paper 'Segment Anything'Kirillov et al. (2023), a new task, model, and dataset for image segmentation has been introduced. Segment Anything Model (SAM) has learned a general notion of what objects are – this understanding enables zero-shot generalization to unfamiliar objects

and images without requiring additional training. Segment Anything Model (SAM) uses vision transformer-based image encoder to extract image features and compute an image embedding, and prompt encoder to embed prompts and incorporate user interactions. Then extracted information from two encoders is combined into a lightweight mask decoder to generate segmentation results based on the image embedding, prompt embedding, and output token

The "Yamaha-CMU-Off-Road" dataset (YCOR) https://theairlab.org/yamaha-offroad-dataset/ consists of 1,076 images captured across different seasons in Western Pennsylvania and Ohio. It features eight distinct classes, including "sky," "rough trail," "smooth trail," "traversable grass," "high vegetation," "non-traversable low vegetation," and "obstacle." The labelling process involved a polygon-based interface, and labels were refined using a Dense CRF for label density. Manual inspection and correction ensured label accuracy.

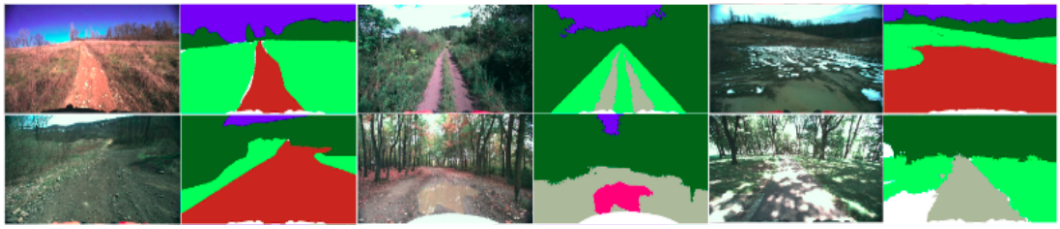

Figure 1: Yamaha Dataset(Maturana et al. (2018))

YCOR is considered more diverse and challenging than a comparable dataset, "DeepScene"http://deepscene.cs.uni-freiburg.de/. It exhibits a higher degree of variability and complexity, making it valuable for computer vision and autonomous navigation research. DeepScene, in contrast, has a left-right bias and a more predictable structure. Despite its challenges, YCOR is a valuable resource for researchers, although its size may be limiting for some applications.

The paper "Image-to-Image Translation with Conditional Adversarial Nets"Isola et al. (2016) introduces conditional GANs (cGANs) for image-to-image translation tasks. It proposes a framework that pairs a generator with a discriminator to generate images from labelled data, allowing various transformations like turning satellite images into maps. cGANs prove effective in tasks such as style transfer, object transfiguration, and more. The authors demonstrate their approach's success across diverse applications, making it a fundamental paper in the field of generative adversarial networks for controlled image transformation.

The paper "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks"Zhu et al. (2020) introduces CycleGAN, a novel approach for unpaired image translation tasks. It leverages cycle consistency to learn mappings between two domains, enabling the transformation of images without paired data. This framework has broad applications, including style transfer, art generation, and image enhancement. By employing adversarial networks and cycle constraints, the model ensures that the translated images retain their original content and style. CycleGAN has become a cornerstone in the field of unsupervised image translation, enabling creative and practical applications in various domains without the need for paired training data.

4 Our Goal

In our project, we plan to employ CycleGAN as a pivotal tool for generating ground truth labels due to its unique ability to address the challenges associated with off-road data. CycleGAN facilitates the creation of synthetic off-road images and corresponding segmentation masks, effectively expanding the available dataset without the need for extensive manual labelling. By leveraging CycleGAN, we not only alleviate data scarcity issues but also enhance domain adaptation, making the segmentation model more robust to variations in off-road environments. This approach contributes to more cost-effective and efficient data generation. It allows us to tackle the scarcity of labelled off-road data while ensuring the model's capacity to perform effectively in a wide range of off-road scenarios.

5 METHODOLOGY

During the initial phase of our project, we focused on data exploration and understanding, as well as a comprehensive study of CycleGAN to ensure we have a solid grasp of its functioning and capabilities. This phase allowed us to identify specific dataset nuances and requirements unique to off-road scenarios. The dataset used is the "Yamaha Off-Road Dataset" which has a lot of variability.

In the later stages of the project, we shifted our attention to assessing the practical utility of the generated off-road images and segmentation masks. This evaluation involved rigorous testing of the model's segmentation capabilities on an entirely unseen dataset, mimicking real-world scenarios that autonomous vehicles encounter. The goal is to measure the model's generalization and ability to handle diverse off-road environments effectively. By carrying out these comprehensive testing procedures, we aim to validate the model's robustness, thereby ensuring that it can reliably assist in off-road autonomous vehicle applications and contribute to the successful navigation of complex terrains.

5.1 SEGMENT ANYTHING

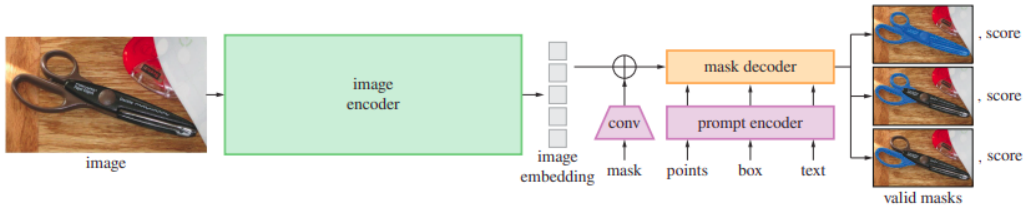

Figure 2: Segment Anything Model (SAM) overview(adapted from Kirillov et al. (2023)).

Segment Anything Model (SAM), has three main components: an image encoder, a flexible prompt encoder, and a fast mask decoder.

The image encoder uses a Vision Transformer (ViT) that has been pre-trained using the MAE method. This allows it to efficiently process high-resolution images. The prompt encoder handles two types of prompts: sparse prompts like points, boxes, and text, which are embedded using positional encodings and CLIP; and dense prompts like masks, which use convolutions.

The mask decoder maps the image embedding, prompt embeddings, and an output token to a predicted mask. It uses a modified Transformer decoder with prompt self-attention and cross-attention to update the embeddings. After two decoder blocks, the image embedding is upsampled and fed to an MLP which outputs a mask probability at each pixel location.

To handle ambiguous prompts corresponding to multiple objects, SAM can predict multiple masks per prompt, up to 3 masks. It also predicts a confidence score for each mask to rank them. The model is trained using a mix of geometric prompts, simulating an interactive segmentation setup.

The overall model design focuses on efficiency - with a precomputed image embedding, the prompt encoder and mask decoder can run on CPU in about 50ms. This enables real-time interactive prompting and segmentation.

The mask prediction is supervised using a combination of focal and dice loss. The model is trained using prompts randomly sampled over 11 rounds per mask, allowing it to be easily integrated into the data collection pipeline.

5.2 CYCLEGAN

CycleGAN, short for Cycle-Consistent Generative Adversarial Network, is a type of generative adversarial network (GAN) architecture introduced for unpaired image-to-image translation. It was proposed by Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros in their 2017 paper "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks."

The key idea behind CycleGAN is to learn mappings between two domains (e.g., photos and paintings) without requiring paired examples during training. Traditional supervised approaches for image translation often rely on datasets where each image in one domain is paired with a corresponding image in the other domain. However, obtaining such paired datasets can be challenging and expensive.

CycleGAN addresses this limitation by introducing a cycle-consistency loss, which enforces that the mapping from one domain to another and back should return the original image. This helps the model learn meaningful mappings even without paired data. The architecture consists of two generators and two discriminators.

The training process involves adversarial training, where the generators try to fool the discriminators, and the discriminators aim to distinguish real from generated images. Additionally, cycle-consistency loss is employed to ensure that the reconstructed images are close to the originals. The overall objective function combines adversarial losses and cycle-consistency losses.

CycleGAN has been successfully applied to various tasks, such as style transfer, object transfiguration, and other image-to-image translation problems, demonstrating its ability to learn effective mappings between diverse image domains.

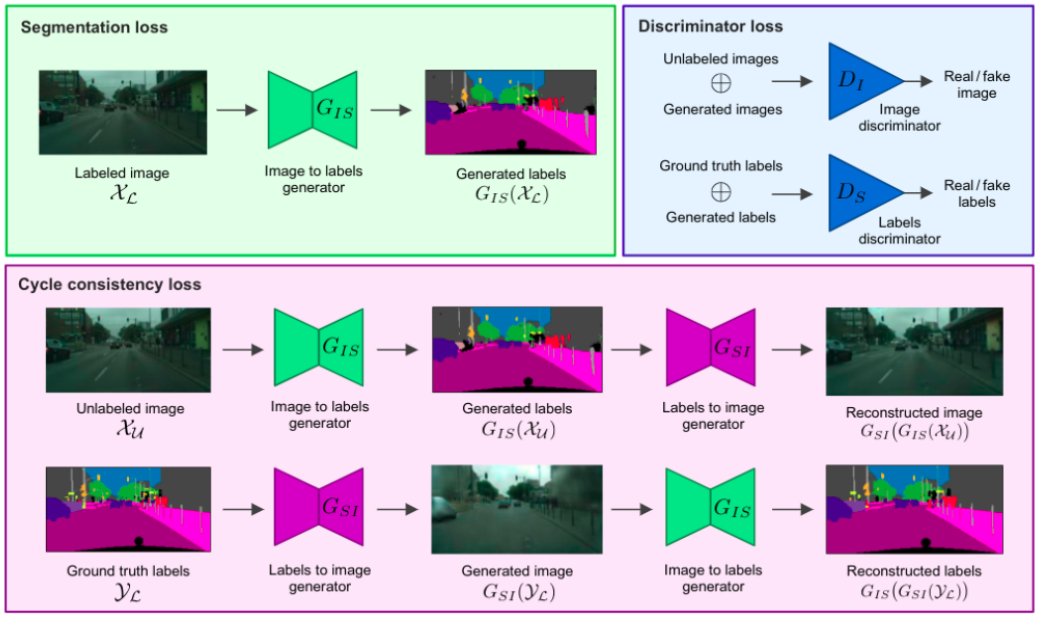

Figure 3: CycleGAN Architecture (adapted from Arnab Kumar Mondal (2019)).

6 Experiments and Results

6.1 Segment Anything

We used a pretrained SAM model to predict some of the masks for the images in the dataset. Considering zero shot learning, SAM performs quite well but it is unable to detect the changes in elevations while the textures remain same on the image. Finetuning SAM on the dataset might improve its performance on this specific task.

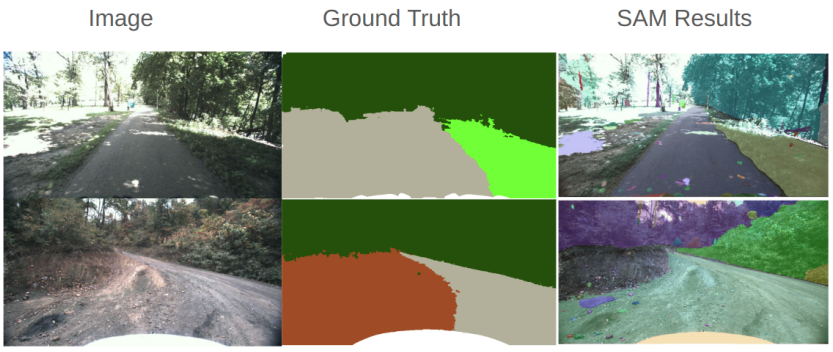

Figure 4: Results using pretrained SAM

6.2 CycleGAN

Subsequently, we implemented the CycleGAN model using popular deep-learning framework PyTorch. Our CycleGAN implementation trains two generators and two discriminators to perform image translation between images and segmentation labels. The generator architecture consists of convolutional layers with instance normalization, employing residual blocks for effective feature preservation. The discriminator utilizes convolutional layers with leaky ReLU activation. Hyperparameters include the use of 6 residual blocks in the generators, Leaky ReLU activation with a negative slope of 0.2, and instance normalization. The training loop spans 50 epochs, involving adversarial and cycle consistency losses. Identity loss is also incorporated to maintain content identity in translated images. The model utilizes the Adam optimizer with a learning rate of 2e-4 and a lambda value of 10 for loss weighting.Additionally, learning rate scheduling is implemented for gradual decay. The weight initialization is performed based on the module's type, such that for convolutional and linear layers, the weights are initialized using a normal distribution with a mean of 0 and the specified gain. If the module has bias terms, they are initialized to a constant value of 0. For Batch Normalization layers, the weight parameters are initialized to 1, and the bias terms are set to 0

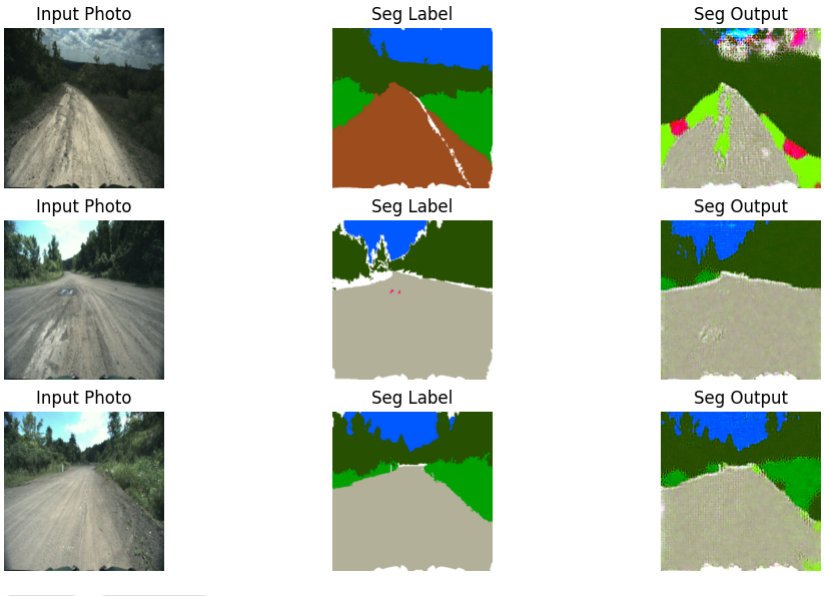

Figure 5: Input to labels - CycleGAN output

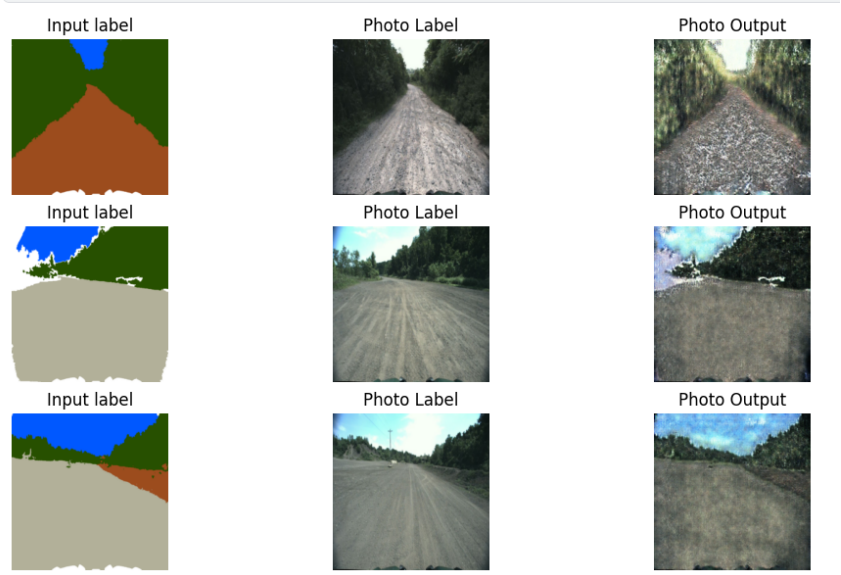

Figure 6: Lables to input - CycleGAN output

6.3 Evaluation

6.3.1 Quantitative

We have used Dice scores and IoU Medium as the primary metrics for quantitative evaluation. Dice coefficient is calculated by 2*intersection divided by the total number of pixel in both images while IoU is the area of the intersection over union of the predicted segmentation and the ground truth. Dice coefficient and IoU are the most commonly used metrics for semantic segmentation because both metrics penalize false positives, which is a common factor in highly class imbalanced datasets

On using the valid folder in the YCOR for testing purposes we get the following scores -
Dice score for generated masks - 0.563
Dice score for generated photos - 0.618
IoU for generated masks - 0.404
IoU for generated photos - 0.574

The code to reproduce our results can be found at `https://github.com/SupreethGadam/Deep_Learning_Project-Segmentation_with_CycleGAN-Off_Road_Driving_Area.git`

6.3.2 Contributuion

The primary contribution of this project was the initially reimplementing Cycle-GAN in PyTorch and then to generate segmentation labels for the Yamaha Dataset, a dataset focused on off-road scenes. The overarching goal was to enhance the availability of labeled data for off-road environments, addressing challenges associated with obtaining accurate and diverse annotations. The evaluation of the model's performance involved both qualitative and quantitative assessments, with a focus on metrics such as the Dice score and Intersection over Union (IoU) score. The qualitative evaluation included a visual assessment of the generated segmentation labels to ensure their coherence and relevance to the characteristics of off-road scenes in the Yamaha Dataset. In addition, the project systematically explored the impact of different hyperparameters, including variations in learning rates, batch sizes, and optimization strategies, with the objective of identifying configurations that optimized results for the off-road segmentation task. The investigation extended to the influence of various weight initialization strategies on the training process and the resulting segmentation quality. Multiple experiments were conducted using different initialization methods, such as Xavier or He

initialization, to observe how these choices affected convergence and the model's ability to capture complex features in off-road scenes. Furthermore, the project delved into the effects of training for longer epochs, recognizing the potential benefits of extended training durations in capturing more intricate patterns and achieving better model convergence. The combination of these contributions, encompassing Cycle-GAN-based label generation, systematic hyperparameter exploration, varied weight initialization, and extended training, constituted a comprehensive effort to advance the state-of-the-art in off-road scene segmentation.

7 CONCLUSION

The report explores using CycleGAN for off-road drivable area segmentation to address the scarcity of labeled off-road data. CycleGAN can generate synthetic images and segmentation masks, effectively expanding the dataset without extensive manual labeling. This helps tackle data scarcity issues and enhance domain adaptation.

The Yamaha off-road dataset used has high variability across seasons and terrains like trails, grass, and obstacles. Methods explored include Segment Anything Model (SAM) for interactive segmentation and CycleGAN for unpaired image-to-image translation.

Experiments involved training a CycleGAN model in PyTorch for image translation between photos and segmentation masks. Both quantitative evaluation using metrics like Dice score and IoU and qualitative visual assessments were done.

The project systematically studied the effects of hyperparameters, weight initialization strategies, and extended training on improving segmentation quality. Multiple configurations were tested to optimize for the off-road environment.

In conclusion, the project contributed towards advancing off-road scene segmentation by reimplementing CycleGAN for synthetic label generation, thorough hyperparameter evaluation, and comprehensive qualitative and quantitative analysis.

ACKNOWLEDGEMENT

We would like to extend our gratitude to Professor Ju Sun and Teaching Assistants Tiancong Chen, Jiandong Chen for their support throughout the semester.

REFERENCES

Aniket Agarwal Arnab Kumar Mondal. Revisiting cyclegan for semi-supervised segmentation. *arXiv preprint arXiv:1908.11569*, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL http://arxiv.org/abs/1512.03385.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. *CoRR*, abs/1703.06870, 2017. URL http://arxiv.org/abs/1703.06870.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. *CoRR*, abs/1611.07004, 2016. URL http://arxiv.org/abs/1611.07004.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023.

Daniel Maturana, Po-Wei Chou, Masashi Uenoyama, and Sebastian Scherer. Real-time semantic mapping for autonomous off-road navigation. In *Field and Service Robotics*, pp. 335–350. Springer, 2018.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR*, abs/1506.01497, 2015. URL http://arxiv.org/abs/1506.01497.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 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, 2015.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks, 2020.