Spring 5-8-2023

Deep Learning of Semantic Image Labels on HDR Imagery in a Maritime Environment

Charles Montagnoli
montagnc@my.erau.edu

DEEP LEARNING OF SEMANTIC IMAGE LABELS ON HDR IMAGERY IN A MARITIME ENVIRONMENT

by

Charles T. Montagnoli

A Thesis Submitted to the College of Engineering Department of Mechanical Engineering
in Partial Fulfillment of the Requirements for the Degree of

Master of Science in Mechanical Engineering

Embry-Riddle Aeronautical University

Daytona Beach, Florida

Monday 24th April, 2023

DEEP LEARNING OF SEMANTIC IMAGE LABELS ON HDR IMAGERY IN A MARITIME ENVIRONMENT

by

Charles T. Montagnoli

This thesis was prepared under the direction of the candidate's Thesis Committee Chair, Dr. Eric Coyle, Professor, Daytona Beach Campus, and Thesis Committee Members Dr. Patrick Currier, Professor and Department Chair of Mechanical Engineering, Daytona Beach Campus, Dr. Jianhua Liu, Associate Professor, Daytona Beach Campus. This thesis was submitted to the Department of Mechanical Engineering in partial fulfillment of the requirements for the degree of Master of Science in Mechanical Engineering.

Thesis Review Committee:

Eric Coyle, Ph.D.
Committee Chair

Jianhua Liu, Ph.D.
Committee Member

Jean-Michel Dhainaut, Ph.D.
Graduate Program Chair,
Mechanical Engineering

Patrick Currier, Ph.D.
Committee Member and Department Chair,
Mechanical Engineering

Jim Gregory, Ph.D.
Dean, College of Engineering

Christopher Grant, Ph.D.
Associate Vice President of Academics

Date

Acknowledgements

I would like to begin by thanking my family and friends for their continued support throughout this process, especially my partner Jenn and my parents Brian and Carol. I'd also like to thank the continued support and guidance of my research advisor, Dr. Eric Coyle, and committee member Dr. Patrick Currier. Additionally I'd like to thank the members of the ERAU RobotX team and the RAAS research lab for their dedication to the pursuit of maritime autonomy. This thesis would not be possible without all of your support.

The research for this thesis was conducted in conjunction with the Deparment of the Navy, Office of Naval Research, in pursuit of works for two Naval Engineering Education Consortium Grants. I would like to thank the sponsoring facilities of these grants, NSWC Carderock and NUWC Keyport, for their continued support and guidance through the process of working on these grants, and in turn this thesis. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Office of Naval Research.

Abstract

Researcher:	Charles T. Montagnoli
Title:	Deep Learning of Semantic Image Labels on HDR Imagery in a Maritime Environment
Institution:	Embry-Riddle Aeronautical University
Degree:	Master of Science in Mechanical Engineering
Year:	2023

Situational awareness in the maritime environment can be extremely challenging. The maritime environment is highly dynamic and largely undefined, requiring the perception of many potential hazards in the shared maritime environment. One particular challenge is the effect of direct-sunlight exposure and specular reflection causing degradation of camera reliability. It is for this reason then, in this work, the use of High-Dynamic Range imagery for deep learning of semantic image labels is studied in a littoral environment. This study theorizes that the use HDR imagery may be extremely beneficial for the purpose of situational awareness in maritime environments due to the inherent advantages of the technology. This study creates labels for a multi-class semantic segmentation process, and performs well on water and horizon identification in the littoral zone. Additionally, this work contributes proof that water can be reasonably identified using HDR imagery with semantic networks, which is useful for determining the navigable regions for a vessel. This result is a basis on which to build further semantic segmentation work upon in this environment, and could be further improved upon in future works with the introduction of additional data for multi-class segmentation problems.

Table of Contents

List of Figures

Page

Figure

Chapter I

Introduction

Perception systems used for the situational awareness of a vehicle or vessel have become more prevalent due to advancements in mobile robotics and autonomy. The National Institute of Science and Technology defines situational awareness as ”Perception of elements in the system and/or environment and a comprehension of their meaning, which could include a projection of the future status of perceived elements and the uncertainty associated with that status.”[Ross, 2022]. By 2030, it is expected that the market for automotive sensors will nearly double from where it was in 2021 [Pre, 2022]. The increase in sensors and sensing modalities are driven by a desire for safer systems and a slow march towards autonomy. In the automotive market, these sensors can help with lane-keeping, adaptive cruise-control, assistive braking, blind-spot detection, and various other human assistance tasks, depending on the usability of the sensor for that specific purpose. [Vargas et al., 2021]

The marine environment poses an interesting problem in the sensing space due to a multitude of factors, including vessel size, use-case, operational speeds, and many other factors that vary from vessel to vessel. Safety systems must address different challenges compared to land vehicles, and the challenges of sensing in the marine environment are unique. Situational awareness for mobile platforms that operate in more standardized environments such as roadways or indoor spaces benefit from the simplicity of the environment. Roadways tend to have defined lane sizes, defined lane markings, and defined traffic signaling. By contrast, while there are channel markers and signage to define waterways in some locations, the maritime environment as a whole is vast and widely undefined. Inherent to the maritime environment are unique environmental hazards and difficulties such as water splash, unexpected vessel motions due to water conditions, and sunlight exposure, including specular reflection. While each of these issues are not strictly endemic to the maritime environment, their effects are still of importance to any study operating within the maritime environment.

Additionally, the study of autonomy in the maritime environment could have a large impact on the safety and efficiency of the worldwide freight industry. In land-based freight, companies looking to push the industry forward have been advancing the technologies for autonomy of the freight trucking industry [Flämig, 2016]. It is an obvious trend that autonomy is seen as a path towards improved safety and efficiency [Grote, 2020]. Worldwide, the United Nations Conference on Trade and Development reports that approximately 80% of all global trade by-volume is transported via maritime freight [UNCTAD, 2022]. Additionally, the volume of freight being transported via freight vessels in the maritime environment is constantly increasing year-over-year, meaning that congestion on the seas and in ports-of-call has become an increasing logistical issue. In 2021, the world saw just how catastrophic the failure of a single canal can be on the worldwide freight industry. The infamous Suez Canal blockage held up over 360 freight ships that were waiting to clear the canal to deliver goods [Russon, 2021]. By many estimates, the economic knock-on effect of this blockage was extremely high, with reports estimating the blockage to cost over 6.5 million USD for every minute that ships were not able to transit the canal [Russon, 2021]. While this is an extreme case in the maritime freight industry, it showcases just how important the accuracy and efficiency of this trade network is to global trade. Studies into the use of advancing technologies in the maritime environment, like this one, could help to prevent incidences in the future.

When discussing the freight industry, it's important to make the distinction about where this study would be of most importance. Most of freight vessel operation occurs in open-water environments, where the problem of object detection and avoidance is much simpler. This study instead aims to focus on situational awareness in the littoral zone. The littoral zone is a portion of rivers and seas that is close to the shore. While this definition is relatively vague, it's generally seen as all portions of water close to the shore, and in oceanic cases is considered to include the continental shelf. This zone is where some of the densest surface-based traffic will occur. Fully autonomous freight ships are not likely to take over the industry anytime soon, but the investigation of situational awareness within the maritime environment could prove important for the future of the industry. In particular, this study intends to focus on the detection and classification of objects portion of situational awareness.

Significance of the Study

A significant amount of research has been conducted in the fields of machine learning and mobile robotics, and with the growth of more intelligent systems entering the consumer market, this research and development only continues to grow. The interest in mobile robotics and autonomy is being driven heavily by the automotive industry. In fact, it is estimated that some 75 billion US dollars have been spent on investments into autonomous driving alone [Pre, 2022]. This sizable investment comes from the automotive industry's desire to push their industry forward into the future, and this funding drives a significant amount of research into mobile robotics and autonomy. While this investment is large, it focuses solely on the problem of mobile ground vehicles, particularly cars, and does little to expand into other spaces or modes of transportation. Of particular interest to this research is autonomy and situational awareness within the maritime environment. There has been some research into this topic, as will be discussed in the literature review, yet it is still in its infancy compared to research for autonomous vehicles. The investment in the automotive industry is not without its benefit to the general field though, as this new market of autonomous and intelligent vehicles has driven innovation in the sensor space for use on mobile robotic platforms [Yeong et al., 2021]. The advancing technology as well as the infancy of the maritime autonomous environment promotes the interest in performing this study.

Additionally, this study looks to further the research that is being done within the littoral zone. This area is of interest due to the high volume of boating traffic that operates within this zone, both commercial and recreational, and the unique challenges for situational awareness that it provides.

Semantic Segmentation

This study is aimed at the use of semantic segmentation deep learning methods. Semantic segmentation is a pixel-wise approach to machine learning problems, allowing for a heightened granularity in the final solution. For each pixel within an image, the learner will attempt to learn what class it belongs to and the broader context around why it belongs to a specific class. This allows for the learner to try and better understand what individual elements within an image relate to the classes of interest.

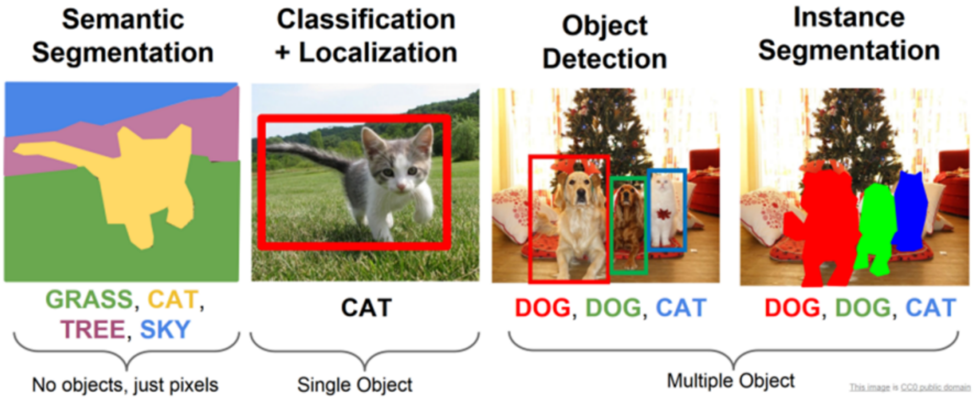


Figure 1.1: A Visual Representation of the Differences in Popular Image-Based Machine Learning Methods [Rieder and Verbeet, 2019]

Fig. 1.1 shows the differences in various types of image-based machine learning methods that are commonly used. Left-most is the type of machine learning method being applied in this study, semantic segmentation, and it can be seen from the image how each individual pixel in the image has been given a label that corresponds to a color. In general, this method of applying a colored pixel representation to each pixel based on it's class is called masking. Looking at the images in Fig. 1.1, it can be seen how the mask representation in the left-most image correlates to the real-world representation of the cat image that can be seen directly next to it. By labeling each pixel that correlates to a class of interest with a colored mask, the learner can use these colored masks to relate each pixel to it's class. It is anticipated that the clutter present within the littoral zone will benefit from the granularity provided by a semantic segmentation approach to deep learning.

Expected Contributions

This study is aimed at examining the use of semantic image labels on HDR imagery for deep learning in the maritime environment, particularly in the littoral zone. In pursuit of this study, there are particular outcomes that are expected as contributions to the furthering of research in this field. These contributions are as follows:

- Generating valuable image ground-truth labels for semantic segmentation within the operational environment and contribution to a dataset for future publication.
- Identifying and defining potential labeling conventions for littoral environment studies.
- Preliminary study of semantic deep learning networks on the data produced for this study.
- This is a preliminary study of semantic networks to identify navigable regions within the maritime environment.

Driving these desired contributions are short-comings in the field of research in the maritime environment. Robotic platforms in the maritime environment suffer from a lack of adequate data with semantic labels for use in machine learning, and the study of HDR data in the maritime environment for situation awareness is of particular interest.

Limitations

The data used in this study was collected prior to this beginning of this study. The team was not able to, in the timeline of this study, collect additional data which could improve upon these results. Additionally, due to the difficulty and time consumption of the process of preparing detailed semantic labels for imagery, the team was limited to producing 600 HDR images with highly-detailed semantic labels. Lastly, the hardware available to the team for training purposes was limited to a single desktop computer equipped with high-end CPU and GPU components, of particular note for deep learning is a limit of 24GB of VRAM.

Definition of Terms

Learning Rate
A tuning parameter in the optimization routine for a neural network to adjust how much weights and parameters are changed.

Littoral Zone
A portion of the maritime environment most often associated with the near-shore and a few hundred meters off-shore.

Maritime Environment
The operational zone for surface vessels, whether it be rivers, lakes, seas, or oceans.

Neural Network
A machine learning model designed to learn in a similar way to human learning, originally modeled after the human brain.

Perceptron
A neural network component that performs computations to detect features or extract information from an input.

List of Acronyms

ANN	Artificial Neural Network
ASV	Autonomous Surface Vessel
ASPP	Atrous Spatial Pyramid Pooling
CNN	Convolutional Neural Network
DNN	Deep Neural Network
DCNN	Deep Convolutional Neural Network
GPU	Graphical Compute Unit
GPS	Geographic Positioning Service
HDR	High Dynamic Range
IMU	Inertial Measurement Unit
LiDAR	Light Detection and Ranging
R-CNN	Regional-Convolutional Neural Network
SDR	Standard Dynamic Range

Chapter II

Literature Review

Machine learning is a topic of high relevance to the field of mobile robotics. Object detection and object tracking for mobile robots benefits greatly from the use of machine learning, particularly deep learning. Deep learning is extremely important to the world of mobile robotics and situational awareness because of its ability to handle complex data. Deep learning has been found to be extremely effective at learning from 3D point-cloud data and imagery data, which are most commonly the type of data obtained from the sensors like cameras, LiDAR, and Radar, used for autonomy and situational awareness purposes.

Machine Learning and Mobile Robotics

Mobile Robotics refers to automated or remotely operated machines which use sensors in conjunction with various other technologies to interpret and maneuver within their environment [Kaiser et al., 1995]. Mobile Robotics is a field with broad applications across a large diaspora of use-cases. One of the more common uses might be the use of autonomous mobile robots for maneuvering of inventory within warehouses [Ng et al., 2020],[Bogue, 2016].

Machine learning has been an integral part of mobile robotics for decades, and the benefits from machine learning are often what can make a mobile robotic application viable. According to [Kaiser et al., 1995]. ”Applying machine learning techniques can help mobile robots meet the need for increased safety and to adapt to the real-world operation demands” [Kaiser et al., 1995]. They go on to elaborate about how learned behaviors are capable of making mobile robots adaptable to their environment. Machines learning to perform a designated task, while being able to learn the variability that may come with performing this task, allows for them to learn adaptability.

[Chen et al., 2022] performs experiments to analyze how different machine learning methods can be used as a means for perception-based ”threat detection” in the maritime environment using perception systems. In this context, threat detection meaning anything that could be perceived as disruptive to the vessel’s movements or tasks. They analyze the machine learning methods by which behaviors of maritime threat vessels can be identified. They implement machine learning methods like Markov models and k-means clustering to identify and classify behaviors deemed to be inherent to a maritime vessel that is a threat. Their experiments show how machine learning perception methods can be used, as a means with high potential, for accurate detection of maritime threat vessels.

Deep Learning

In general, Deep Learning refers to a neural network architecture in which there are more than one hidden layers within the neural network [Long et al., 2014]. The number of layers and perceptrons per layer are dependent upon the problem at hand and the network architecture.

Deep Learning Neural Networks (DNNs) are especially successful in the task of image analysis, which leads to the use of deep learning for situational awareness of mobile robots. Cameras and other sensors allow for mobile robots to gather information about their surroundings, but deep learning is what allows the elevation of that data to provide context and awareness. Deep learners can be beneficial in situations with complex data such as imagery or spatial data from a mobile robot. Deep learning’s capability to abstract complex data allows for the learner to automatically extract optimal features. By extracting the optimal features for the learner, less human supervision is required, and the learner can even extract information that humans may not be able to perceive [Xie et al., 2017].

Deep Learning Computing

The process to train a deep learning network requires a computer with a significant amount of computational power. The most important factor for computation with these deep learning networks is the Graphical Processing Unit (GPU). This is due to the fact that performing the computation necessary requires a large amount of specialized parallel computations to be executed, which the GPU is often better at performing than the Central Processing Unit (CPU), the CPU by contrast is better at generalized repetitive computations.

Consider the generalized architecture of a convolutional neural network in Fig. 2.1. This simplified representation of a CNN for classification of imagery serves as an easy example for why parallel processing is necessary. From left to right, the image is processed in the input layer, which then feeds into the hidden Convolutional Layers, with pooling operations occurring between layers. These convolutional layers are made up of some number of perceptrons that perform mathematical operations on the input data. In each CNN the number of perceptrons per layer will vary, however in general it may be hundreds or thousands per layer producing an ever-growing number of computations. Through the use of parallel computing, these extremely complex neural networks with potentially millions of computations can be tackled much more efficiently. The parallel computing capabilities allow for the specialized calculations taking place in these convolutional layers to be executed in parallel - meaning that the computations happen at the same time as one another.

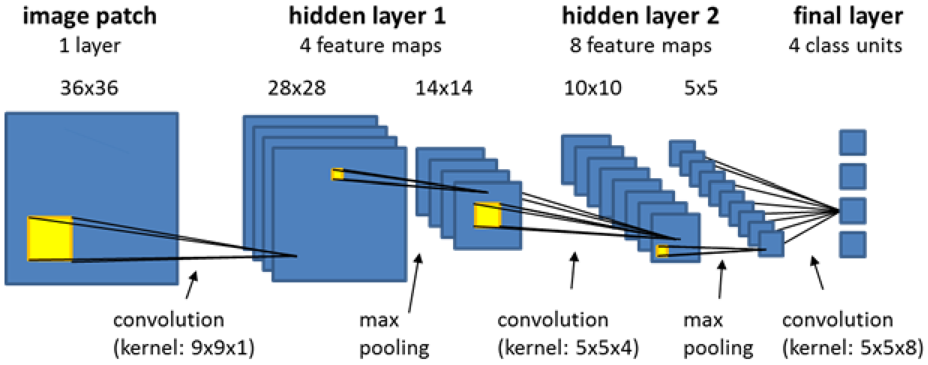


Figure 2.1: Generic CNN Example with Pooling and Convolution [Ecognition, 2021]

Convolutional operations have no dependency on one another, they're self-contained operations. This means that there would be no advantage to performing convolutions in a sequential process. Therefore, it is more efficient to perform convolutions in parallel with one another. When training a CNN with use of a GPU, the advantages of parallelization can be fully utilized.

Deep Learning Applications to Mobile Robotics

Deep learning has become a pivotal part of mobile robotics and autonomy in recent years. When operating a mobile robot that is leveraging sensors such as camera, LiDAR, or radar, deep learning has been shown to be an extremely integral component in situational awareness tasks. Deep learning neural networks are able to make use of the highly complex data derived from these sensors, and extract meaningful features and relationships that can be used to learn things about their operational environment. Deep learners are also extremely adept problem solver for non-trivial tasks such as navigation and control.

Deep Learning of Spatial Information Through the use of deep learning, mapping and localization can be attained without complete knowledge of the environment, as discussed above. This idea can permeate further into the use of strictly spatial information for mapping and navigation.

[Tsai et al., 2021] propose a Deep Learning method for the mapless navigation and control of a mobile robot through the use of LiDAR spatial data. Without a map of the environment, a target navigational way-point is given to the robot as a destination. Using a DCNN trained on the LiDAR data, they propose a method by which to perform navigational control based on LiDAR data and estimated robot position. They use the DCNN to predict the motion control commands needed to navigate the unknown environment.

As one of the main sensors used in the autonomous driving space, LiDAR based perception has been explored widely in the world of autonomous driving vehicles. Most autonomous driving vehicles will have LiDAR, Radar, and Cameras as their main sensors used for the purpose of situational awareness. [Li et al., 2020] performed a review of Deep Learning of LiDAR Point Clouds in Autonomous Driving, and identified many of the trends and issue within the field. They identified three general topics of Deep Learning with spatial LiDAR data:

1. 3D Point Cloud Semantic Segmentation
2. 3D Object Detection/Localization
3. 3D Object Classification/Recognition

3D Point Cloud Semantic Segmentation Semantic Segmentation is the pixel-wise approach to classification, generally used through pixel-wise classification of imagery data. In 3D Point Cloud Semantic Segmentation, the points of the points clouds are considered as the pixels of the semantic segmentation problem. Similar to typical semantic segmentation but in a 3-dimensional space, each point from a point cloud will have a class predicted from the list of classes the network is trained on. One of the main differentiating factors in 3D Point Cloud Semantic Segmentation methods tends to be the method or criterion by which points within a cloud are grouped [Li et al., 2020].

3D object detection/localization The general object detection/localization problem is simply to identify if an object is within the area of interest and where it may be within the area. With 3D point cloud data, detection and localization is one of the simpler problems to be solved. LiDAR are designed to give returns that can identify where within the space (relative to the sensor) the object is located and what area it occupies.

To achieve true detection and localization, techniques like clustering of points within the point cloud may be implemented as a means to filter false returns from data that represents a true object [Tsai et al., 2021]. Additionally, the localization of these objects could be achieved through the use of creating a centroid from the spatially clustered object information [Li et al., 2020].

3D object classification/recognition Object classification and recognition with 3D point cloud information is similar to typical object classification and recognition problem. It is desired to find what, if any, groups of point cloud information can be classified into the desired classes/tags. The problem can be defined as, given cluster(s) of 3D spatial point cloud data, determine which classes, if any, they belong to.

Semantic Deep Learning of Visual Information

Thanks to the advantages of deep learning with visual information as highlighted previously, deep learning with visual information may be one of the most explored topics in all of deep learning. There are many different neural network models used for image processing tasks as well as applications for their use. This section will focus specifically on the use of semantic segmentation models applied to imagery data.

Semantic segmentation networks have been applied to a multitude of applications and have been proven to show high accuracy results on imagery data for use in situational awareness tasks. Many different model architecture types have been generated and applied to the task of semantic segmentation and instance segmentation. Using an R-CNN approach for instance segmentation, [Ren et al., 2015a] proposed a Mask R-CNN model for object instance segmentation. This network uses essentially an ensemble of Faster R-CNN outputs to generate bounding box coordinates, object class identifications, and finally binary masks for segmentation. This model is a method in which instance segmentation of objects within an image can be performed off the back of a network designed to run at a relatively high speed.

The advancement of attention-based models in recent years have gone to show how extremely successful these networks can be at learning tasks and generating useful outputs. Attention-based models, talked about in-depth by [Vaswani et al., 2017], which use an attention metric to influence the network's hypothetical ”attention” to focus on the most important aspects for it's learning task(s). Extremely popular attention-based transformer models have produced impressive tools like GPT-3 [Brown et al., 2020], the model behind OpenAI's ChatGPT. The power of attention-based models is truly impressive, so applying them to the problem of computer vision machine learning is especially enticing. In a paper by [Zhang et al., 2019], an attention-based learner is proposed which uses attention to softly weight multi-scale features at each pixel location within an image. The result is an attention mechanism which focuses the network on the most 'important' portions of an image for the purpose of generating semantic segmentation masked outputs. The model is able to use the attention metric as a way to weight objects based on their scale within an image.

Dilated Convolutional Models, and the DeepLab family of models specifically, have proven to have extremely performance on complex tasks. These dilated convolutional models use dilated convolutions to expand the field-of-view of the network without further down-sampling imagery [Yu, 2016]. This provides a benefit by increasing the amount of context that can be gained by the network without performing convolutions on the image to the point of lost context. The networks that use dilated convolution for semantic segmentation show high performance on the standard tests, for example the initial DeepLab [Chen et al., 2016],[Chen et al., 2017] networks were capable of greater than 70% accuracy on the 2012 PASCAL VOC challenge. Further expanding upon these successful networks, DeepLabV3+ was created years later and improves upon this result achieving 89% accuracy on the 2012 PASCAL VOC challenge for segmentation[Chen et al., 2018],[Minaee et al., 2021]. This is the network architecture that will be used in this thesis research.

Deep Learning in the Maritime Environment

As stated, the deep learning space for situational awareness in the maritime environment has been less explored than land-based environments. An important distinction between the maritime environment and many land-based operational environments is the largely unbounded and undefined nature of the maritime environment. On-road camera situational awareness can benefit from the relatively well-defined operational environment, such as the existence of frequent and standardized road-signage, lane-markings, traffic signals, and standard rules for traffic patterns. While the maritime environment, particularly when operating within a channel, may have some signage and loosely defined travel lanes, the general operational environment is relatively vast and undefined.

In a master's thesis presented at Embry-Riddle Aeronautical University in 2017, Robert Goring presents a review entitled ”Feasibility of neural networks for maritime visual detection on a mobile platform”[Goring, 2017]. This study focuses on the use of two visual data sets available to the research group, one of camera imagery from an ASV and one from an ROV. For this study, [Goring, 2017] assessed the feasibility of Faster R-CNN on visual data. Faster R-CNN is a Regional Convolutional Neural Network architecture built on the back of the original R-CNN network that is, Faster R-CNN is designed as a more efficient version of R-CNN. The method which Faster R-CNN uses consists of a base CNN model which aims to provide region proposals for detection, these detections are then fed into the classifier[Ren et al., 2015b]. This study successfully uses this data to perform object detection and classification on their data, achieving a high mAP of greater than 90%. The data used in this study is relatively low resolution imagery based on older camera architecture. The data is obtained from standard definition digital cameras at a resolution of 1920x1200, which equates to a roughly 2.3MP image.

[Lin et al., 2022] propose a method for perception in the maritime environment based on the use of LiDAR as the primary sensor. In their approach, they supplement the use of LiDAR detections with the vessel's Automatic Identification System(AIS). AIS is a broadcast system which operates in the very-high-frequency (VHF) maritime band of frequencies to communicate with other vessels in the area [USCG, 2023]. By supplementing the LiDAR returns with the robust AIS system, this approach leverages the capabilities of a CNN to perform real-time object detection. This method projects the detections of objects in the 3D space down to a 2D overhead view to give a clearer picture of where objects are in the plane of the water. This method of deep learning in the maritime environment provides relatively high success of classification in high-density maritime environments, providing for 65.4% mAP in on-water testing. This study claims to be the first one published which uses a CNN for deep learning of LiDAR data in the maritime environment.

Chapter III

Methodology

With the focus of this study on deep learning in the maritime environment, this requires a significant amount of work to prepare the data and tune network performance for optimal results. As a result, the methods section will begin with a discussion of the data used, followed by the sensor suite used for data collection and the deep learning methods that will be applied.

Data Collection

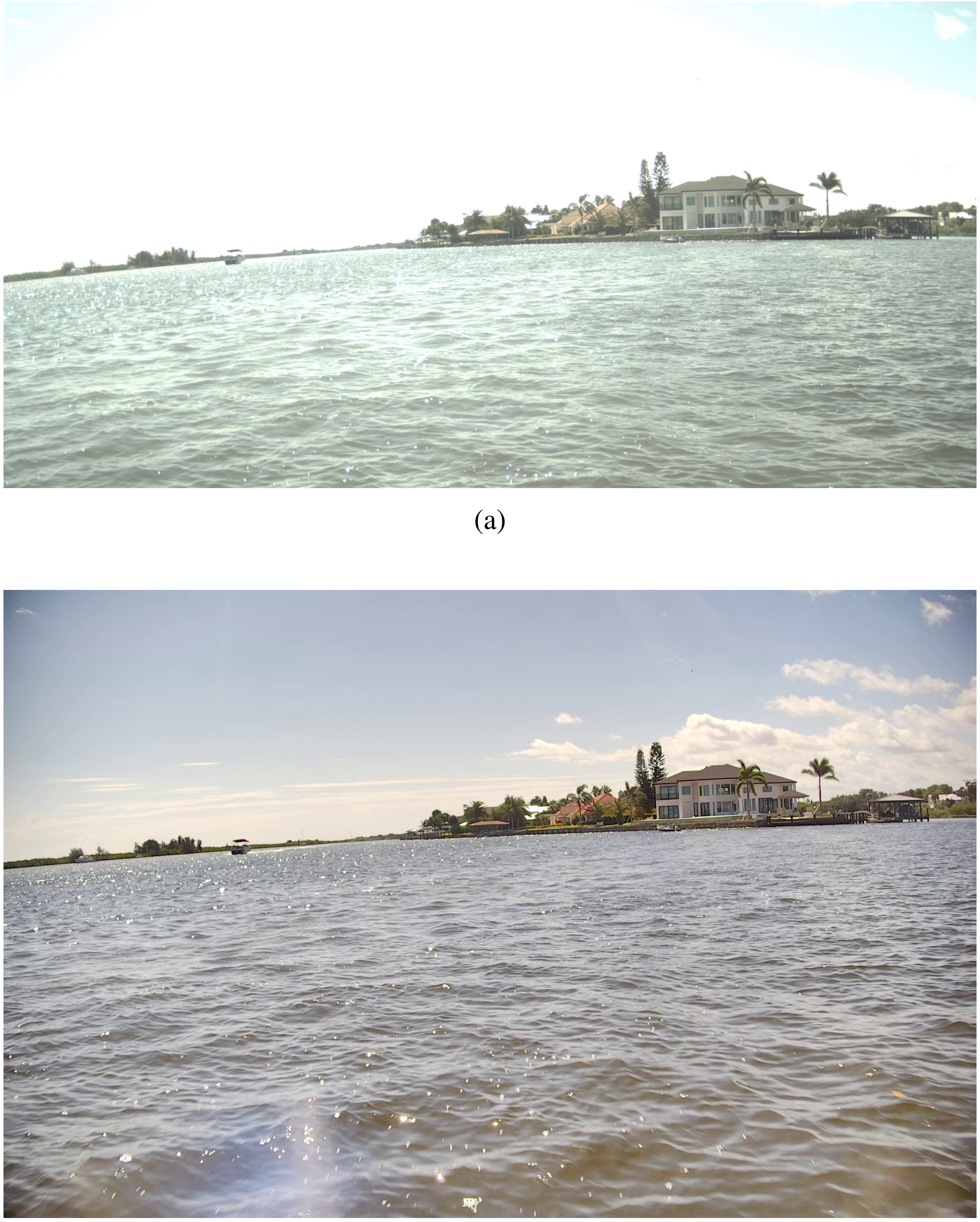
The data used in this study was collected by the research group using a custom built sensor suite. The custom built sensor suite contains multiple modalities of sensing across visual and spatial perception. These sensors allowed for the collection of multi-modal data for the maritime environment, as well as localization of where the data was collected. The data that was of the most importance to this study, however, was the HDR camera imagery.

The reason that exclusively the HDR imagery was chosen for this study is due to factors including the benefits gained from HDR imagery and the lack of other studies in this space that involve HDR imagery. The research team feels as through their access to a high definition 5MP HDR camera provides for the study of the usage of this camera technology in the maritime environment. For this study, it is expected that the greater image contrast and pixel-wise color information from the HDR camera is expected to be beneficial, or at least of no harm, to the task of pixel-wise semantic segmentation.

The Sensor Suite

There are 6 cameras in the sensor suite. Three of the cameras are 4k High-Definition visible light cameras, two of them are LWIR Thermal Cameras, and lastly there's one HDR, High Dynamic-Range, Camera. The HDR camera is a Leopard Imaging LI-IMX490 5MP camera that captures video in 2880x1860 pixel resolution. This camera covers the front-facing field-of-view of the sensor suite.

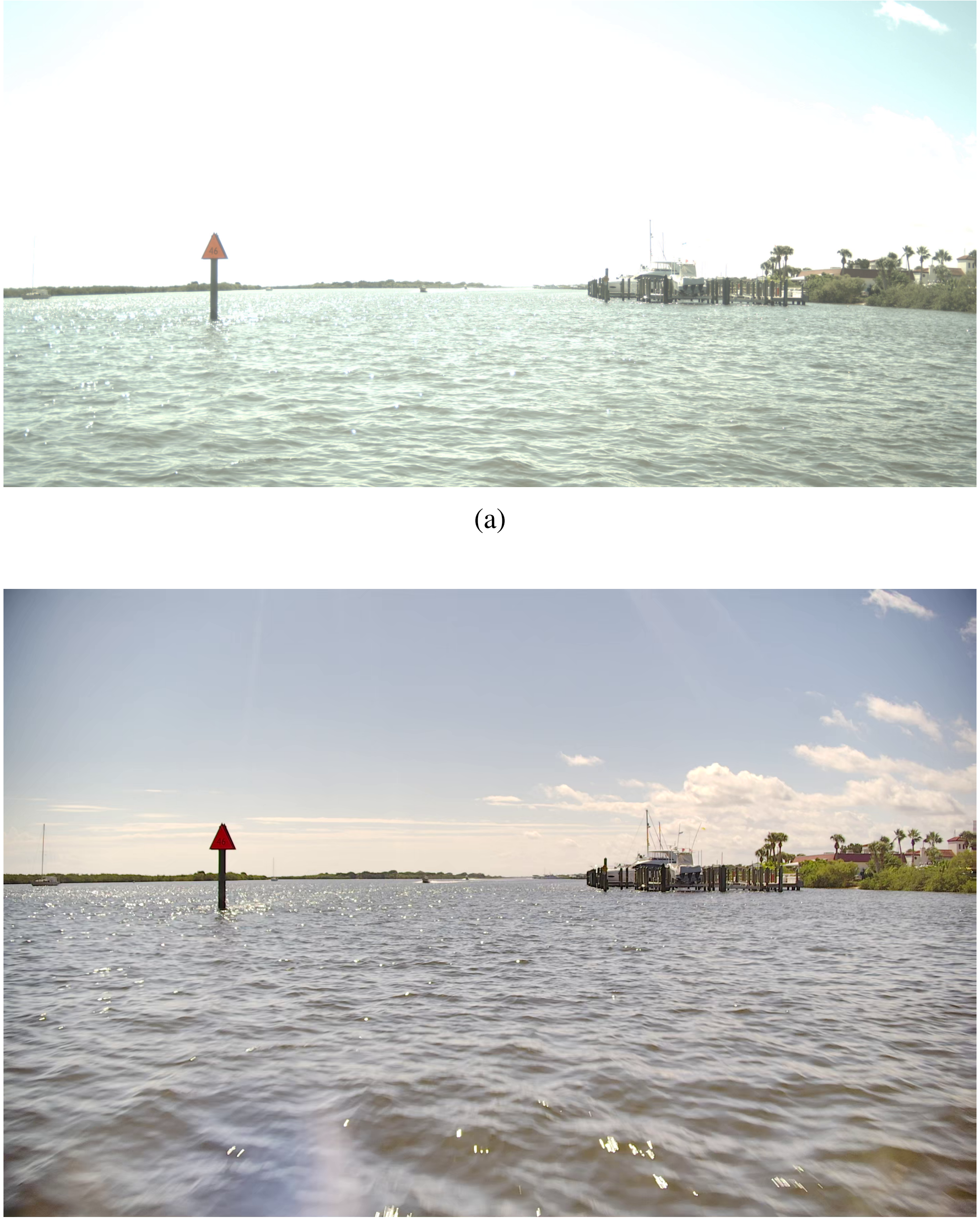
The HDR imagery is of particular interest to this study due to the significant advantages expected of HDR imagery in the maritime environment. These potential advantages are more evident in high-contrast situations. The maritime environment often has high-contrast scenes to be captured in imagery due to exposure from the sun. The brightness of the sun as well as the reflection of the sun off the water can wash-out, or over-saturate, the imagery captured by standard definition cameras. A washed out image is one in which the contrast between colors is low, meaning that the color information within the image to differentiate colors and objects from one another is lower [Vigier et al., 2016]. Wash-out can lead to loss of information within the image, and depending on the severity of the wash-out can even lead to an image with almost no discernible or useful information. As demonstrated in Fig. 3.1a / 3.1b and 3.2a / 3.2b, the HDR imagery is capable of capturing the context within the image without becoming washed-out by the exposure from direct sunlight. This attribute of HDR cameras makes their use extremely desirable when working in the maritime environment, particularly for the purpose of image classification and object detection. Even in indirect sunlight situations, showcased in Fig. 3.3a / 3.3b, the color information retained within the HDR image is much greater than in the 4k image. However, it should be studied if and how these perceived benefits can be leveraged for Deep Neural Network learning tasks.



(a)

(b)

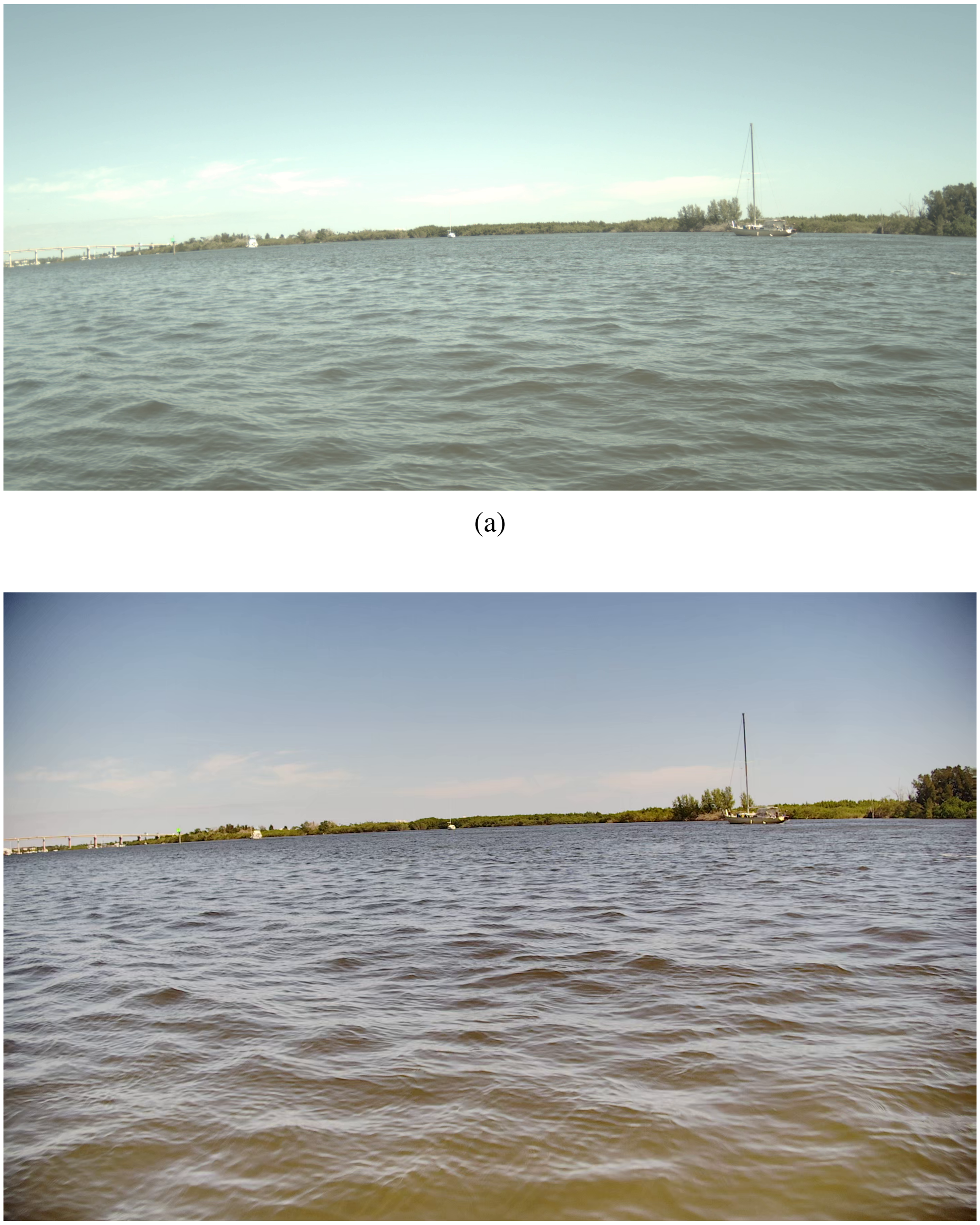
Figure 3.1: Samples from Dataset:
(a) Direct Sunlight Imagery from Sensor Suite's Center-Facing 4k Camera, Showing Loss of Pixel-Wise Contrast In the Presence of Direct Sunlight Conditions
(b) Direct Sunlight Imagery from Sensor Suite's Center-Facing HDR Camera, Showing Robust Color Information In the Presence of Direct Sunlight Conditions



(a)

(b)

Figure 3.2: Samples from Dataset:
(a) Direct Sunlight Imagery from Sensor Suite’s Center-Facing 4k Camera, Showing Loss of Pixel-Wise Contrast In the Presence of Direct Sunlight Conditions
(b) Direct Sunlight Imagery from Sensor Suite’s Center-Facing HDR Camera, Showing Robust Color Information In the Presence of Direct Sunlight Conditions



(a)

(b)

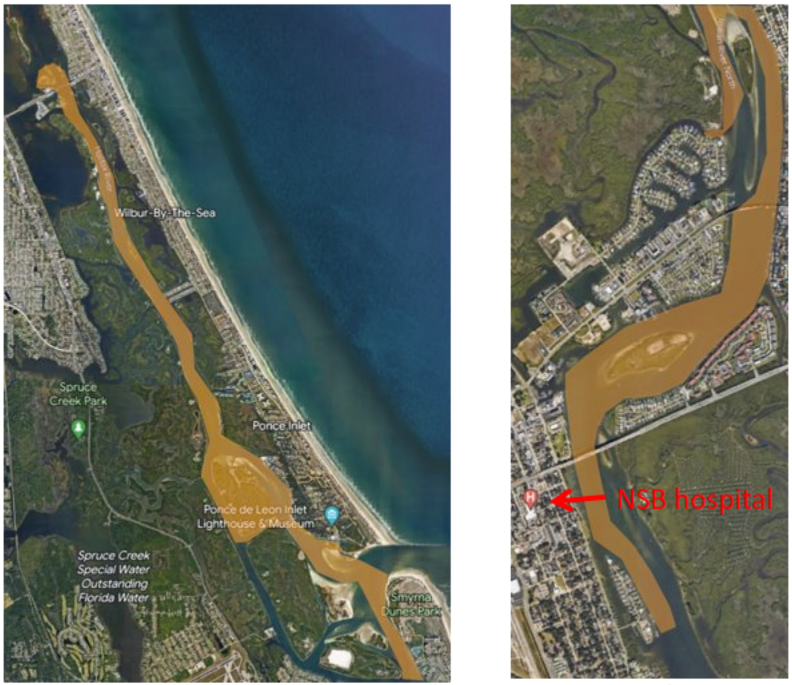
Figure 3.3: Samples from Dataset:
(a) Indirect Sunlight Imagery from Sensor Suite’s Center-Facing 4k Camera
(b) Indirect Sunlight Imagery from Sensor Suite’s Center-Facing HDR Camera

Additionally, the HDR camera that the team has access to, the 'LI-IMX490-GW5400-GMSL2' [LeopardImaging, 2023] from Leopard Imaging provides meets or exceeds the specifications desired for use by the research team in a maritime environment. The lens FoV is 65 degrees, providing adequate coverage of the front-facing FoV. The camera and its peripherals are also rated at IP67 Intrusion Protection, exceeding the team's design requirement of IP65 Intrusion Protection. The camera can accept 9-18VDC input power, allowing for use of 12VDC power for all cameras within the sensor suite. Lastly, and of most importance to this study, the camera produces images at 2880(H) x 1860(V), providing for images with greater than 5MP image quality. The image size obtained allows for high granularity in the labelling process, and though the imagery must be resized for use in the neural network, the use of a dilated convolution method aims to help retain some of this granular information.

One benefit brought on by the HDR camera that was not yet discussed was its capability to differentiate between sky and water. As discussed previously, High-Dynamic Range imagery benefits from creating images with greater contrast within the image. This contrast means that differently colored objects are more easily separable from one another in the image based on the color disparity between them. With this, an HDR camera could be used to provide horizon detections from the camera data alone.

Data Collection

The collection of this data was performed over the course of approximately one year in Florida, specifically on the Halifax River of Florida's Intracoastal Waterways. All of the data collection that was used in this study was performed through mounting of the sensor suite on a small personal fishing boat. Though the sensor suite is designed for use on ASVs, the nature of the data collection lends itself better to manned out-board motor vessels which can be more easily controlled to drive near areas of interest.


Wilbur-By-The-Sea
Spruce Creek Park
Ponce Inlet
Ponce de Leon Inlet Lighthouse & Museum
Spruce Creek Special Water Outstanding Florida Water
Smyrna Dunes Park
NSB hospital

Data path on October 16

Data path on September 25

Figure 3.4: Collection Area of Data

The dataset being used consists of 4 days of data collection across September and October of 2021. Fig. 3.4 shows the areas of operation for the sub-set of the data labeled for this study. The Florida Intracoastal Waterways are a littoral zone, and it can be seen from Fig. 3.4 just how much density of both in-water and near-water perceivable objects there could be.

Labeling of Data

The data collected from the sensor suite needed to be labeled for semantic segmentation. Semantic labels provide a class for each individual pixel within an image. To perform this labeling, the team used Supervisely as the service to host their data for labeling. The use of this service allows for the team to work cooperatively on the data for the different studies being performed.

The labeling of the data for semantic segmentation requires that each pixel within the image get a label - whether that is one of the labels of interest for the study or a background label. The set of labels that are of interest to the partnering facility at NSWC Keyport are much more expansive than what will be used in this study, but the narrowed down list used in this study is derived from a collaborative effort between ERAU and NSWC Keyport. The labels of interest for this study are:

- Water
- Moving Boats
- Stationary Boats
- Signage
- Buoys
- Piers & Docks
- People
- Vegetation
- Buildings
- Bridges

This reason that this set of labels was chosen is due to a multitude of factors, the main of which was from the outline of work desired by the sponsors of this research at NSWC Keyport as part of the Naval Engineering Education Consortium Grant that this work is in pursuit of. Additionally, an examination of the data collected confirmed that there seemed to be an adequate representation of these classes within the imagery for use in a machine learning model. Lastly, there is an interest in the use of a high-level tagging system to apply sets of tags to the data at a later date. These tags would be deterministic features of the object being observed.

- Tag 1: Land-based, Surface-based, Air-based, Subsurface-based, Undetermined
- Tag 2: Man-made, Natural, Undetermined
- Tag 3: Moving, Stationary, Undetermined

These sets of tags are a part of future works to be performed by the research group, however they play a role in the way that the data was labeled for this study. The data within this study was labeled to identify vessels that were in-motion or stationary, which can allow for the future work(s) to take advantage of this more highly detailed labelling scheme. These tags or a tagging system are not intended to be applied in this study.

Any pixels within the imagery that are not labeled will be automatically assigned to class 0: 'Background' which is a default background class. This background class is important because not all items within the image are discernible to either the researcher labeling the data or the learner, and additionally there may be some items within each image that have little to no relevance to the study.

Shown in Fig. 3.5 is an example piece of labeled data from the dataset. This example shows just how complex operation within the littoral zone can become based on the density of objects and varying scale of their presence. Many instances of classes at varying scales, interlaced in and among other classes represent the complexity of the problem and the granularity needed to differentiate the classes.

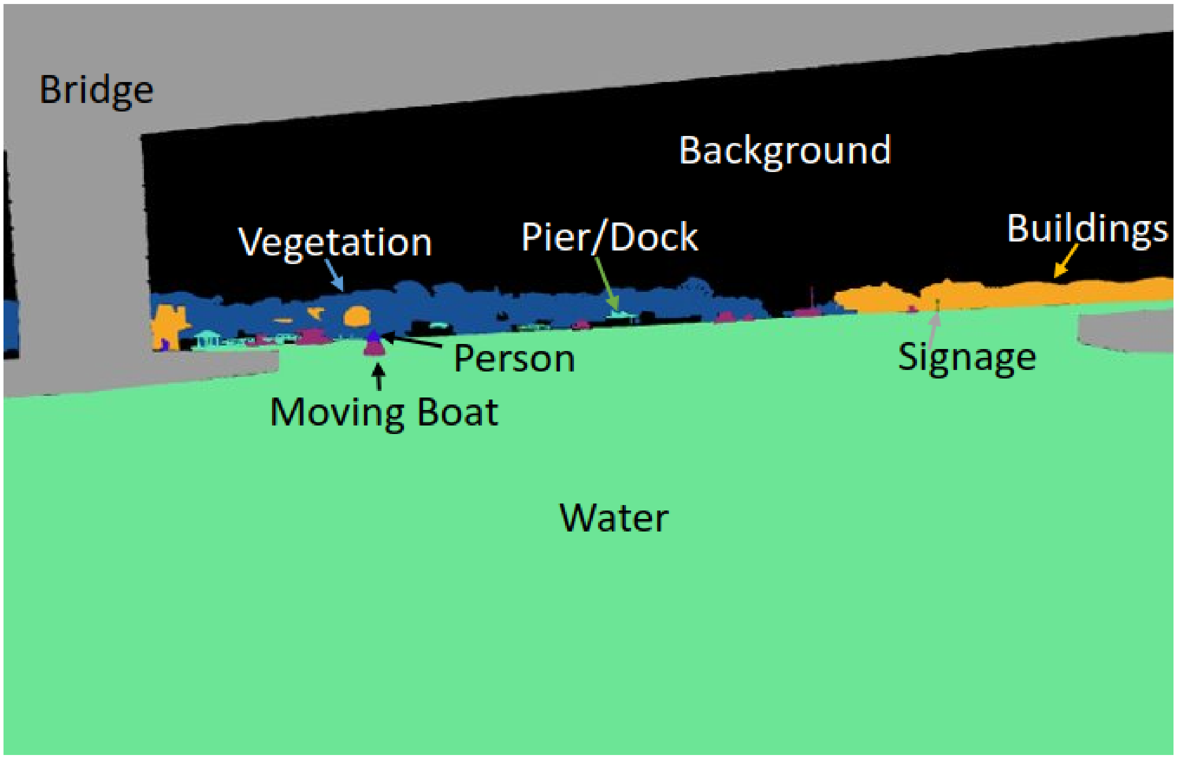
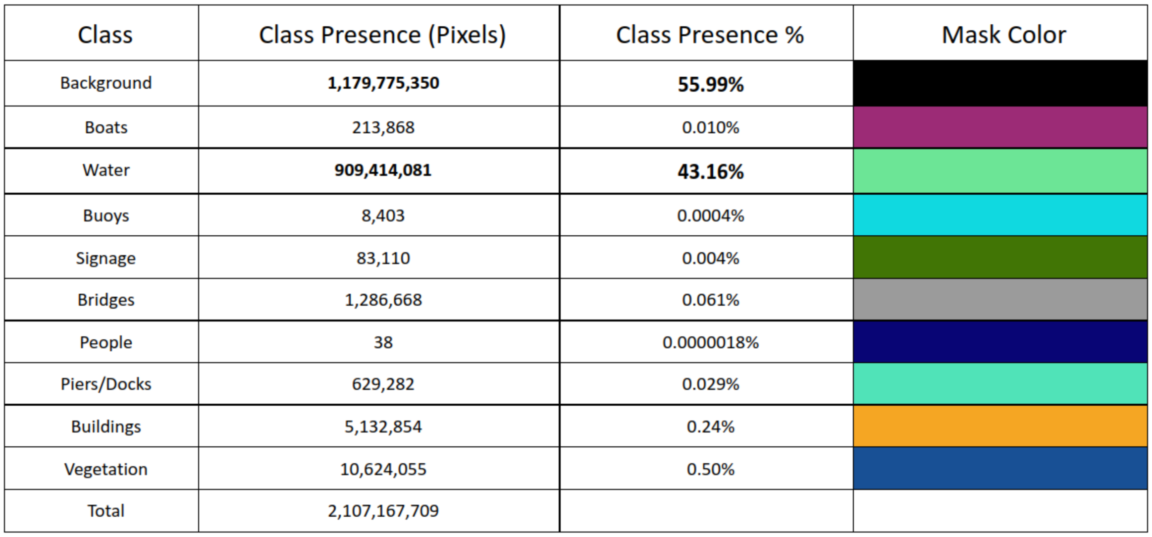


Figure 3.5: Labeled Masked Image, Displaying Object Density in the Littoral Zone and Varying Scale of Represented Classes



Class	Class Presence (Pixels)	Class Presence %	Mask Color
Background	**1,179,775,350**	**55.99%**	⬛
Boats	213,868	0.010%	🟪
Water	**909,414,081**	**43.16%**	🟩
Buoys	8,403	0.0004%	🟦
Signage	83,110	0.004%	🟩
Bridges	1,286,668	0.061%	⬜
People	38	0.0000018%	🟦
Piers/Docks	629,282	0.029%	🟩
Buildings	5,132,854	0.24%	🟧
Vegetation	10,624,055	0.50%	🟦
Total	2,107,167,709		

Figure 3.6: Class Representation within Dataset

The labeling is performed by applying bit-masks to each image of the appropriate label. The labels for each class consist of a specific color bit-mask that is placed onto the image. The process of placing colored bit-mask labels on the images is done through either painting pixels or applying a polygon operation to a group of pixels.

The DeepLabV3+ Network

The deep learning architecture chosen by the research group for this study is DeepLabV3+, a member of the DeepLab family of deep learning networks. DeepLabV3+ is an evolution of DeepLabV3[Chen et al., 2018], which in itself was an evolution on the previous iterations of the DeepLab network architecture. This section will give some in-depth background information on the network architecture being used, and some explanations as to why it is believed that this would be the best network for this use-case.

The original DeepLabV1 network architecture proposed a novel method for semantic image segmentation classification by combining two existing methods into one architecture. The first is DCNN, a Deep Convolutional Neural Network. Deep Convolutional

Neural Networks are CNN architectures which tend to have additional convolutional and pooling layers making them 'Deep'[Long et al., 2014]. The authors of the DeepLabV1 paper introduce an augmentation to the traditional DCNN through the addition of Atrous Convolutions[Chen et al., 2016]. In addition to the DCNN with Atrous Convolutions, the original paper applies something called Fully-Connected Conditional Random Fields iteratively on the output of the DCNN, progressively smoothing the output boundaries and segments on the image. Though this is integral to DeepLabV1, this process was removed in future iterations of DeepLab, including DeepLabV3+ used in this study.

Atrous Convolutions provide an alternative approach to performing convolutions. These convolutions use a dilation rate to extend the field of view of the convolutional layers. By dilating the kernel, the convolution operation is able to consider broader context within the image. By dilating the kernel, the computational cost is not increased since the kernel is still the same size. Standard convolution operations could be seen as Atrous Convolution where the atrous rate is r=1. In essence, atrous convolution is an efficient way to gain broader context without costing us any additional computations. Through increasing the atrous rate, the model's field of view increases and allows objects to be encoded at varying scales.

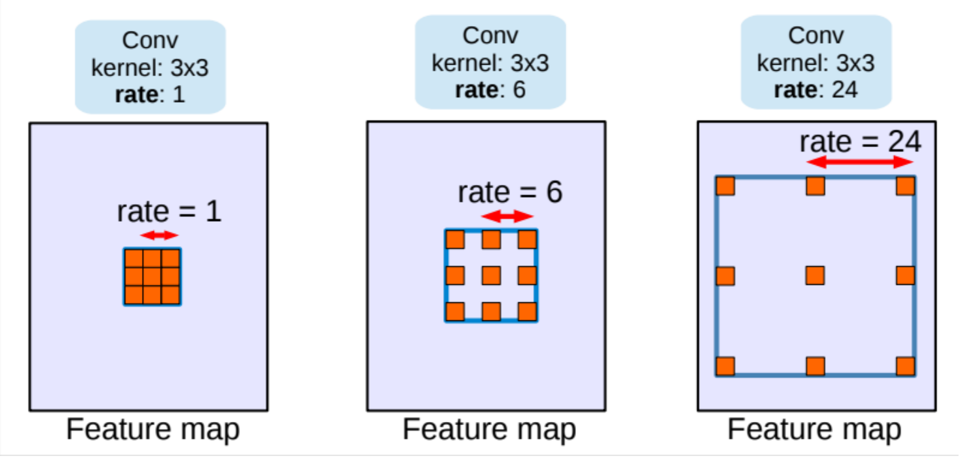


Figure 3.7: Atrous Convolution with a 3x3 Kernel Shown at Different Rates

DeepLabV2 builds on DeepLabV1 and introduces a new idea call Atrous Spatial Pyramid Pooling (ASPP)[Chen et al., 2016]. Atrous Spatial Pyramid Pooling is used to tackle the problem of objects being presented at different scales. Since Atrous Convolutions can encode information at different scales, ASPP is used in conjunctions with Atrous Convolutions to gain improved performance of segmentation results. DeepLabV2 still uses the Fully-Connected Conditional Random Fields approach which will be removed in DeepLabV3.

Atrous Spatial Pyramid Pooling is an important improvement made to the DeepLab architecture. ASPP takes advantage of the multi-scale information collected from Atrous Convolutions at different scales to apply simultaneous parallel filters for classification.

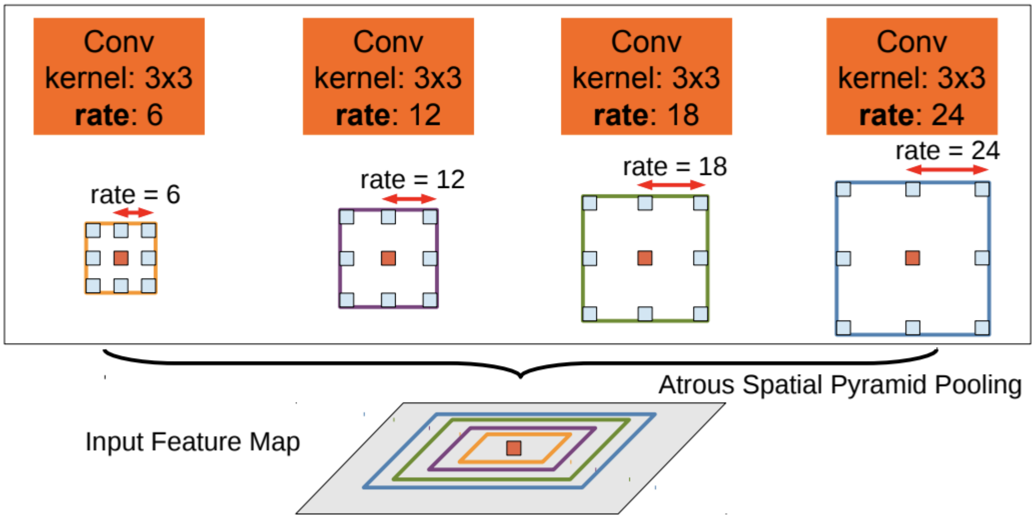


Figure 3.8: Atrous Spatial Pyramid Pooling [Chen et al., 2017]

As can be seen from Fig. 3.8, the ASPP method classifies pixels with the use of the multi-scale information from Atrous Convolutions. The orange pixel is being classified through the use of parallel filters with differently scaled context. These cascading filters of differing scale can be portrayed to look like a pyramid of atrous convolutional outputs - hence Atrous Spatial Pyramid Pooling.

DeepLabV3 builds on the previous iterations of the DeepLab family. The DCNN from the previous architectures remains, though with an improvement that includes cascading atrous convolutions[Chen et al., 2017]. By cascading atrous convolutions with varying rates, the learner can have additional convolutional layers. These cascading atrous convolutions allow for progressively more information to be extracted by the learner, and due to the setup of the cascading convolutions these layers can keep the output size the same. Meaning that this addition allows for additional context to be extracted without the negative effects usually associated with including these additional layers.

As mentioned previously, the Fully-Connected Conditional Random Fields portion of the learner was removed in this iteration. The CRF portion of the learner did not allow for ’end-to-end’ learning, since the CRF function was essentially a filter on the output. The removal of CRF allows the learner to be that ’end-to-end’ learner. The ASPP portion of the network was left the same.

DeepLabV3+ builds on the previous learners, but changes the architecture into an encoder-decoder network. The desire to use an encoder-decoder network is driven by the previous networks struggles with loss of object boundary clarity [Chen et al., 2018]. The proposed network fixes this problem and improves the network overall.

The encoder in DeepLabV3+ works much the same as previous iterations of DeepLab, however now the network makes use of something called ’Separable Atrous Convolution’. Separable Atrous Convolutions take the normal Atrous Convolution operation and separate it into two operations; Depthwise Atrous Convolutions and Pointwise Atrous Convolutions. Depthwise Atrous Convolutions involve performing operations on each input channel, while Pointwise Atrous Convolutions combine the outputs of the Depthwise Convolutions to each channel individually.

The decoder in DeepLabV3+ is used as the means to recollect the context about object boundaries which were absent in the previous architectures without an explicit decoder step. The decoder in DeepLabV3+ employs up-sampling operations to generate greater spatial resolution of the segmentation. All of this combined allows DeepLabV3+ to boast some of the As this network has become one of the most widely used deep learning architectures for semantic segmentation and effective [Chen et al., 2018], it has been chosen for this research study.

Application of DeepLabV3+

Through the use of a Docker container, TensorFlow release 22.07 was setup to train and run DeepLabV3+ on images with semantic masks. All of the scripts for setting up and running DeepLabV3+ were run in this docker container with Python 3. The research computer being used by the team was running on Linux with Ubunutu 20.04LTS. This operating system was selected due to the free and open source nature of the operating system, its native capability to support Docker and Python, as well as the research team's experience with the operating system. Additionally, Ubuntu has native support for the installation and use of CUDA drivers. CUDA is the parallel computing architecture used by NVIDIA, the manufacturer of the GPU used in this study, to accelerate computing on GPU demanding tasks such as machine learning [Dehal et al., 2018].

For this reason, the team assembled a custom computer which included the most capable machine learning GPU available at the time: The NVIDIA RTX 3090. The RTX 3090 GPU that the team has boasts 10,496 CUDA cores, 24 GB of VRAM, and is capable of training the HDR imagery at 512x512 with a batch size of up to 16. The network is setup to run on a 512x512 image in order to try and balance accuracy and speed. Even though this GPU is extremely capable, training speeds slow down exponentially if the image too large. Additionally, this series of GPUs are capable of using mixed precision floating-points. The advantage of mixed precision in training deep learning networks is increasing the speed on the most mathematically intense models [Micikevicius et al., 2017].

Loss Function

The loss function used to train DeepLabV3+ is Sigmoid Focal Cross Entropy Loss. This loss function was selected since it was known that class imbalance would be inherent to our data. When attempting to identify both the water and Sigmoid focal cross entropy loss is based on focal loss, which is designed for a classification problem in which the classes are highly imbalanced [Lin et al., 2017], as expected in this study. Focal loss is designed to down-weight the well-classified examples in order to focus on the harder examples. In other words, focal loss puts the impetus on learning the classes which have fewer annotations in the training data. This is an important focus for this study due to the nature of the data. Within the dataset of 600 images, the most represented classes are water and background, meaning that all other classes have less representation, and can be seen as harder to identify examples. Focal loss is capable of accounting for this class imbalance in order to give a greater chance at classifying these harder examples. In practice, the focal loss is applied using a balancing factor of α to tune the focal loss.

$$FocalLoss(p) = -\alpha(1-p)^{\gamma}log(p) \tag{III.1}$$

In Eq.III.1, α is a balancing factor, γ is a tunable focusing parameter, and p is the model's estimated probability of the class [Lin et al., 2017]. This form of the focal loss is meant to greatly emphasize the importance of correcting mis-classified examples for the learner. The reason that it is called sigmoid focal cross entropy is that the loss combines the sigmoid operation for computing p with the focal loss computation. This loss function is implemented using TensorFlow and Keras in Python alongside the DeepLabV3+ implementation.

Learning Rate Scheduling One of the ways in which neural networks can be tuned to ensure they're producing an appropriate result is through scheduling the learning rate [Kim et al., 2021] [Senior et al., 2013]. Neural networks perform gradient descent on the data to generate weights and predictions, and this gradient descent is, in essence, seeking out the optimal result of the network's training. The learning rate is one of the most important factors in a neural network as it determines if, and at what rate the gradient descent converges. With too low of a gradient descent there may not be enough exploration, too high can lead to jumping from one local maxima/minima to another. This is why it can be important to explore the space with the learning rate. One way that this can be done is through learning rate scheduling.

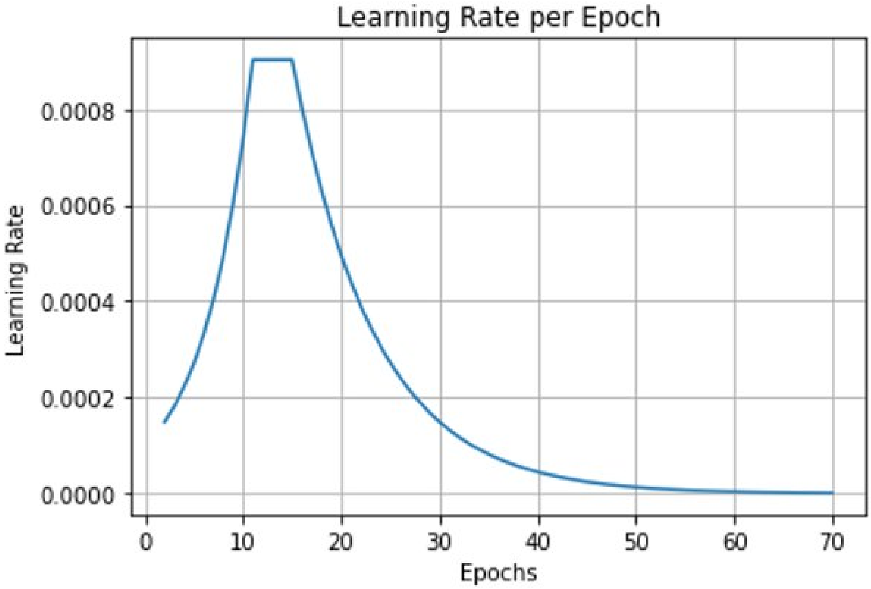


Figure 3.9: Example Learning Rate Scheduler Used for Training, Beginning at a Low Learning Rate and Exploring Before Decelerating the Learning Rate Back Down.

The learning rate scheduling approach that has been employed in this research involved a warm-up, exploration, and cool-down. This means that the learning rate starts low, is increased exponentially until plateauing for several epochs, and then is exponentially lowered until the training is complete. Through this scheduling scheme, the intention is to allow the learning rate to have the ability to explore the space without leaving it too large or too small for the entirety of the training. This combines two methods which are often used in the literature; learning rate warm-up and learning rate decay.

Network Optimizer The optimizer is an important part of any deep learning neural network, they're algorithms which are used to minimize the error and maximize efficiency [Postalcıoğlu, 2019]. They update and modify attributes, like the learning rate or some of the weights, of the network as it trains in an effort to reduce the overall loss and improve accuracy [Sun et al., 2020]. The optimizer being used in this training is the Adam optimizer. Adam optimization implements a stochastic gradient descent(SGD) method based on the adaptive estimation of first and second-order moments [Kingma and Ba, 2014]. The creators of the Adam optimizer, Kingam et al., boast Adam as being ”computationally efficient, has little memory requirement, invariant to diagonal re-scaling of gradients, and is well suited for problems that are large in terms of data/parameters”[Kingma and Ba, 2014]. Choosing an optimizer is a difficult task due to the number of available options, each with their own positives. Adam was chosen because most applications of DeepLabV3+ also use Adam.

Chapter IV
Results

This project aims to show that semantic segmentation network architectures with HDR imagery within the maritime environment are feasible for situational awareness tasks. The following results will show the successes of the prescribed methods for this research as well as discussing short-comings and ways to fix them in the future.

Metrics

The metrics by which a neural network can be assessed are extensive. Generally, studies will focus on a few of these metrics depending on what is most important to the study being done. For this study, Categorical Accuracy, Precision, and Recall will be the main metrics tracked from the network training. These metrics give insight into the successes and failures of the network being trained. Additionally, these metrics are among the standard metrics used in many similar studies on semantic networks.

Categorical Accuracy When discussing the results of a neural network, it is often important to discuss the accuracy of that network. After all, the accuracy can be the most telling of the success that a network is achieving at classifying the desired classes. That said, for this study categorical accuracy was chosen as the main performance metric. Categorical accuracy (CA) differs from standard accuracy by using the one-hot encoded labels of the data. Therefore, when the network calculates categorical accuracy, the accuracy is calculated by predicting the total frequency at which y_{pred}, the predicted class, matches y_{true}, the one-hot encoded label for that class [TensorFlow, 2023]. More simply, categorical accuracy is the number of True Positive (TP) class predictions divided by the number of class examples.

$$CA = \frac{TP}{y_{true}} \tag{IV.1}$$

Precision The precision of a neural network is the precision with which the network predicts with respect to the labels provided [Google, 2022]. When calculating Precision, the predictions are separated into True Positives (TP) and False Positives (FP). True positives are predictions that apply a class to the correct object according to the provided label, while False Positives are predictions which apply an incorrect class to an object according to the provided label. In other words, for every time that a prediction is made, how often is the prediction correct? This can be represented with the equation IV.2. Precision takes into account False Positive predictions which can allow us to gain further insight about how precisely the learner is predicting classes. A large number of False Positives, and therefore a lower precision, would suggest the learner is applying class predictions too often and imprecisely.

$$Precision = \frac{TP}{TP+FP} \tag{IV.2}$$

Recall The recall of a neural network is how often the network recalls a prediction correctly with respect to the labels provided [Google, 2022]. When calculating Recall, the predictions are separated into True Positives and False Negatives. True Positives are predictions that apply a class to the correct object according to the provided label, while False Negatives (FN) are predictions which incorrectly do not classify an object as having a class when compared to the provided label. In other words, for every time a class is seen, how often is the prediction correctly giving that class its label? This can be represented with the equation IV.3. Recall takes into account False Negatives, or simply how many times the learner missed the presence of a class predictor. This can allow for insight into what the learner is having a hard time recalling from training when performing predictions.

$$Recall = \frac{TP}{TP+FN} \tag{IV.3}$$

F1 Score The F1 score of a network is often used as a means to describe the overall performance of the network alongside accuracy. The F1 score is the harmonic mean of Recall and Precision, resulting in a composite score between the two. F1 score can be seen as more instructive to the network's overall performance since it wraps multiple metrics into one score. It is noted that F1 score, as well as the relationship between recall and precision incidentally, are effected by the class imbalance of the specific problem at hand [Khan et al., 2020]. This means that the relationships that exist between these scores will vary based on class imbalance, but their use is still a great indicator for model performance overall.

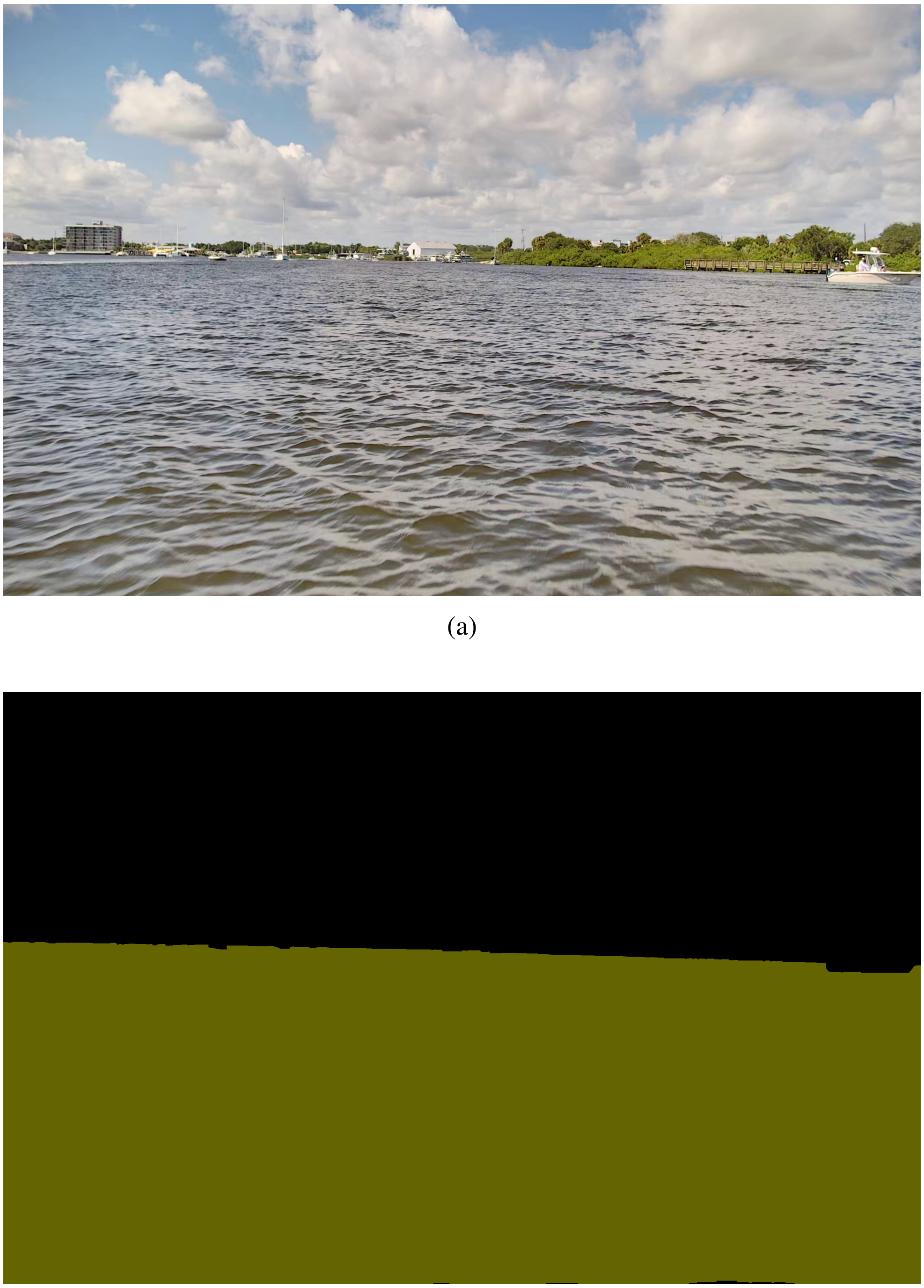
$$F1Score = 2 * \frac{Recall * Precision}{Recall + Precision} \tag{IV.4}$$

Network Results

The network being used, DeepLabV3+, was pre-trained on ImageNet-50 with frozen weights to take advantage of transfer learning. Transfer learning allows the use of pre-trained weights from a proven dataset to be frozen, meaning that the learner will have learned some amount of low-level image-feature information before any training has been done with the intended data. These frozen weights allow for small-dataset studies, like the one being performed here, to have a chance at success without the being majorly limited by an amount of data. Training a network fully from end-to-end requires significantly more data than training through transfer learning. For training, validation, and testing, the research team produced a set of 600 HDR images with highly-detailed semantic labels, as described in the methodology.

Network Results - Water Only Results

From this point, a selection of images was organized from their dataset to be labeled for training, validation, and testing. This subset of imagery consists of a 600 images that were selected from two of the three data collection days within the dataset. From these images, the next goal was to produce a result that proved the learner was setup properly to learn on the data at hand. To do this, time was spent labelling only the water within the images, resulting in a dataset that consisted of only background and water as the classes. Anything within an image that was not the water was left as background.



(a)

(b)

Figure 4.1: Water-Only Training Sample:
(a) Example Training Image
(b) Example Mask with Water Only Labels

The data shown in Fig. 4.3a / 4.3b was used to train a network for detecting what was water within the images, with all non-water sections being background. For a water-only result, the network is capable of reaching an accuracy of approximately 71.5%. At this accuracy, the network is capable of classifying the water, delineating the line where the water meets the line of the horizon, which can be seen in Fig. 4.2.

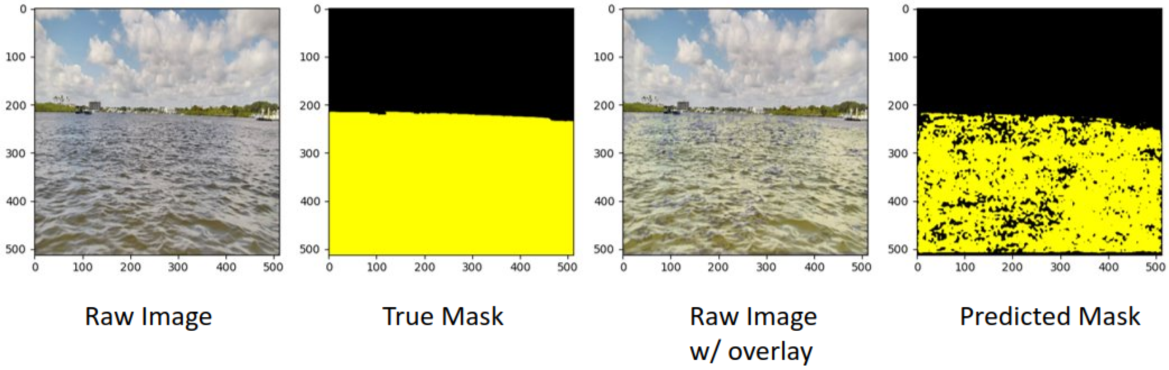
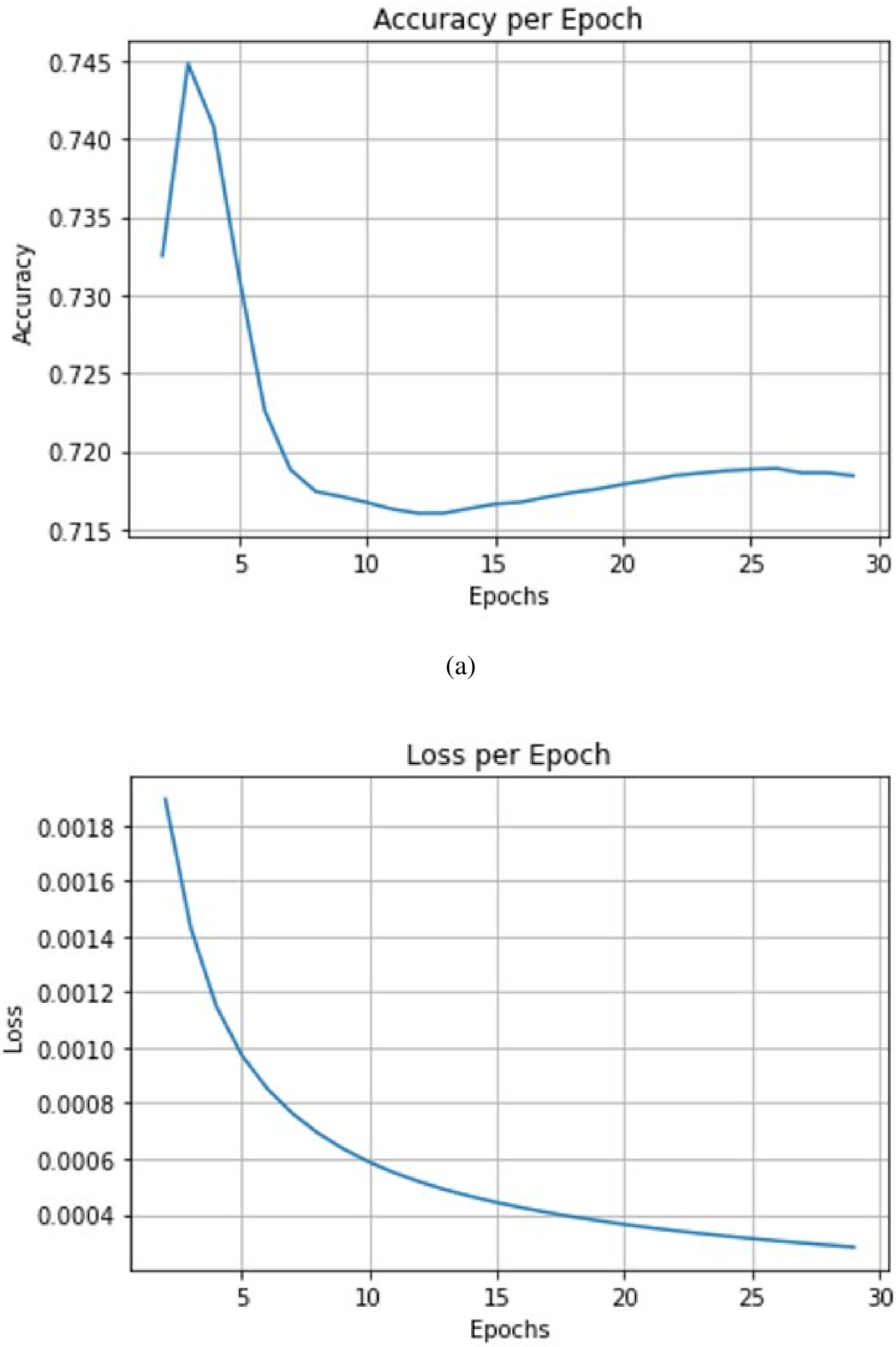


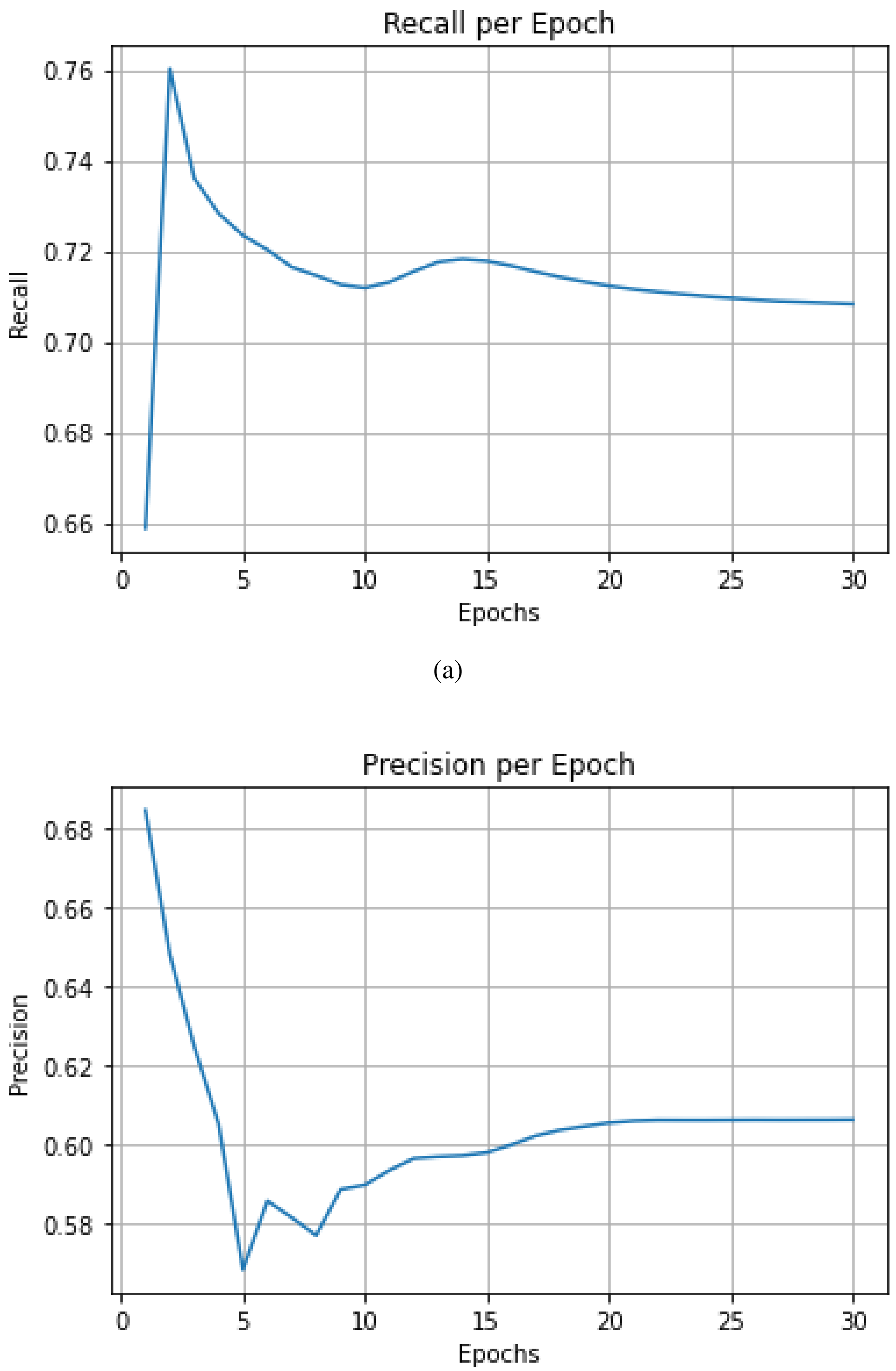
Figure 4.2: Water Only One-Hot Mask Prediction Example on 512x512 Test Image



(a)

(b)

Figure 4.3: Water-Only Metrics:
(a) Network Accuracy per Epoch, Demonstrating Convergence to Result Near 26 Epochs.
(b) Network Loss per Epoch, Demonstrating Proper Convergence Across Epochs.



(a)

(b)

Figure 4.4: Water-Only Metrics:
(a) Network Recall per Epoch, in Fig. 4.2 with Some False Positives can be Observed.
(b) Network Precision per Epoch, Demonstrating Acceptable Precision.

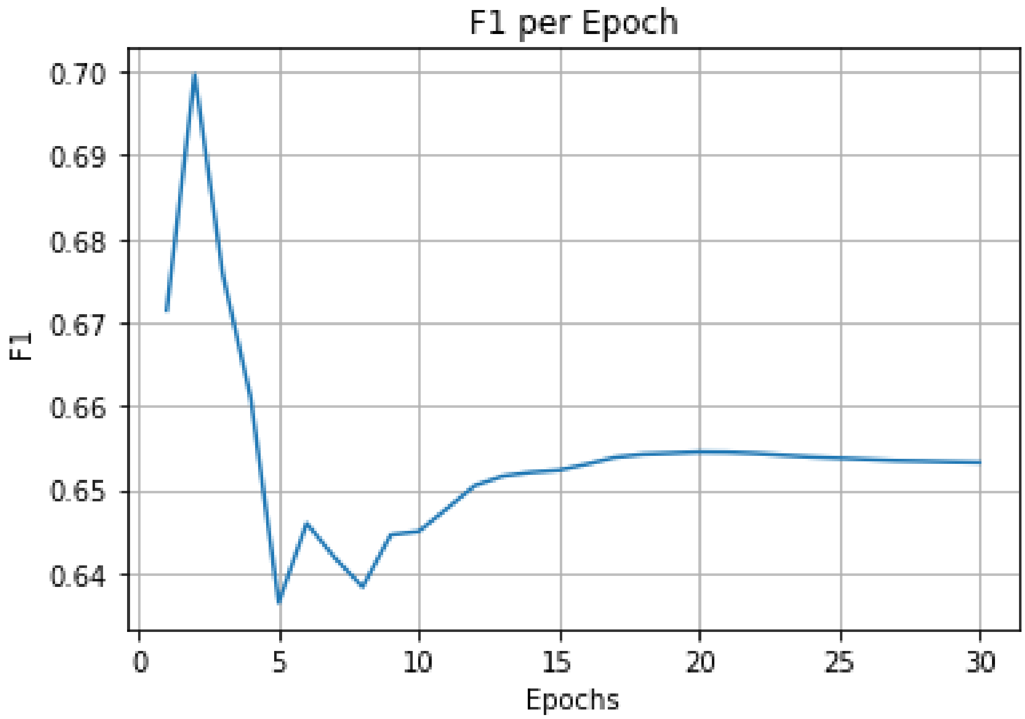


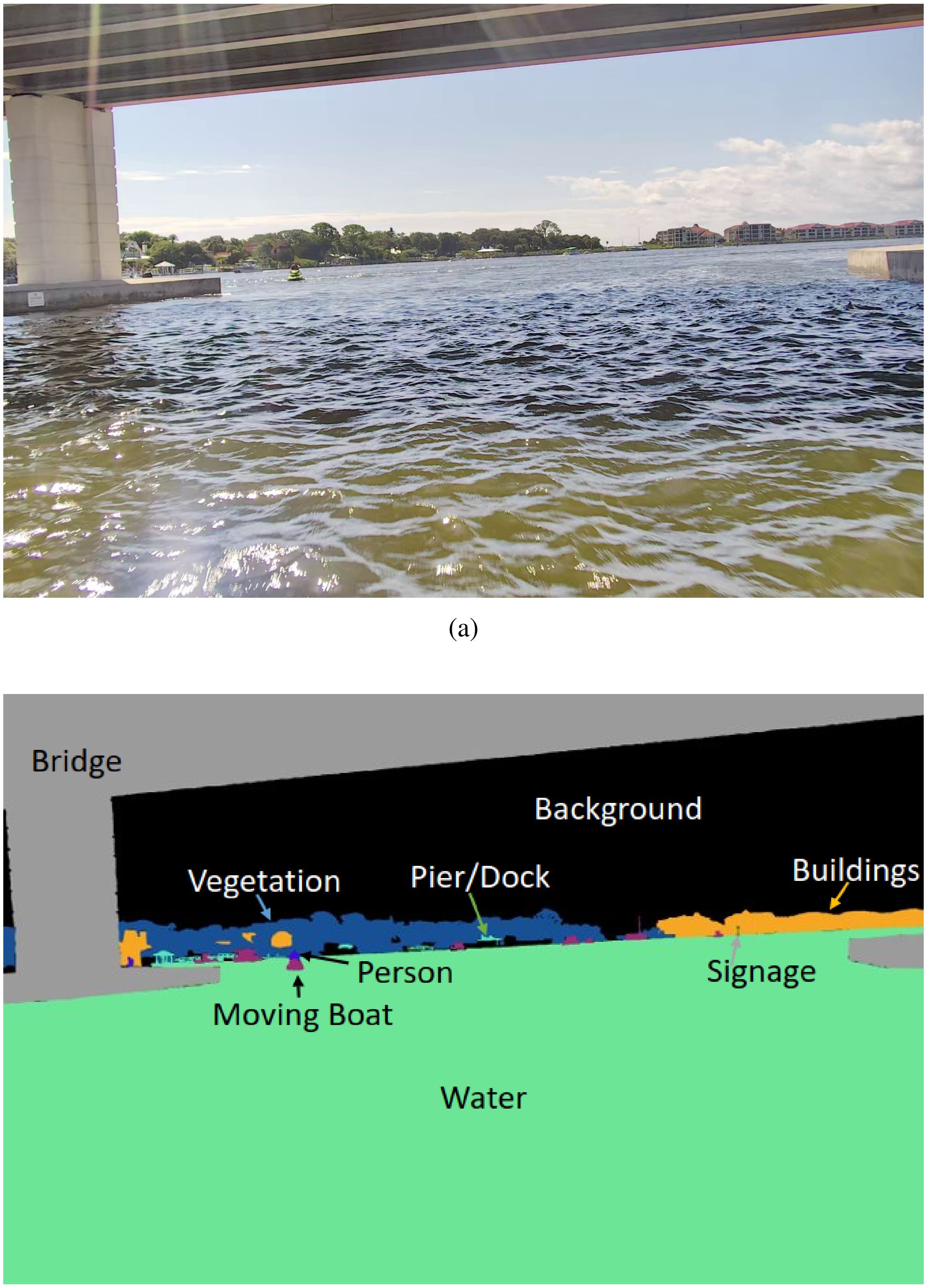
Figure 4.5: F1 Score, Combining the Outcomes from Recall and Precision

The metrics shown describe a learner that has converged onto the solution with relative success. The Recall and Precision are both greater than 60% across classes, with the F1 score landing above 65%. The overall Accuracy is just over 71% and the loss steadily decreased, signifying no significant issues in model convergence. The resultant model proves a successful semantic result for a water/non-water semantic segmentation deep neural network. This result proves that the approach to labelling this data can provide some useful benefit for the purposes of situational awareness in a maritime environment. There is some other research that has been done for the purpose of water surface and horizon detection with shown benefits for tasks such as obstacle detection [Sheikh and Afanasyev, 2018], [Hożyń and Zalewski, 2020], [Paccaud and Barry, 2018]. Additionally, this is the first of these results that was obtained using strictly HDR imagery data in the maritime environment.

Network Results - Multi-Class Results

While the results from the water-only network show promise for basic situational awareness tasks, there is a greater potential for this learner to do even more. To that end, it was attempted to train the network on a multi-class setup encompassing some of the classes outlined in the earlier section on data collection. Being able to have the learner identify additional classes of information is a desirable step knowing how to interact with a potential obstacle. While this is a desirable result, it should be noted that this increases the difficulty on the learner significantly. No longer must it only learn whether something is water or non-water object, it must now attempt to correctly identify what those non-water objects may be.

With the data that was prepared for this study, shown in Fig. 3.6, there were a total of 10 classes of interest overall which were labeled by the team. It was known from the outset that not all 10 would have enough data for this initial study, but for the completeness of the dataset all 10 were labeled regardless of their representation within the 600 image set. Fig. 3.6 shows just how large the class disparity was within this set of data. It is expected that with future efforts, these under-represented classes can be brought into the fold to meet the needs of future networks.



(a)

(b)

Figure 4.6: Multi-Class Sample:
(a) Example Training Image
(b) Example Mask with Multiple Class Labels

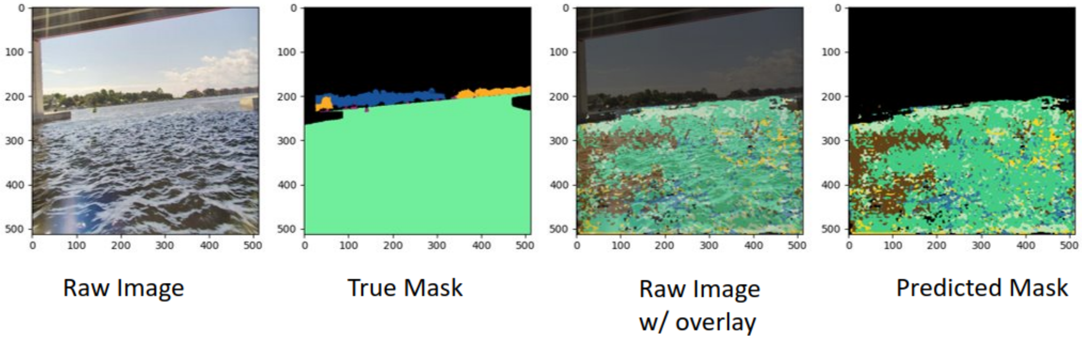
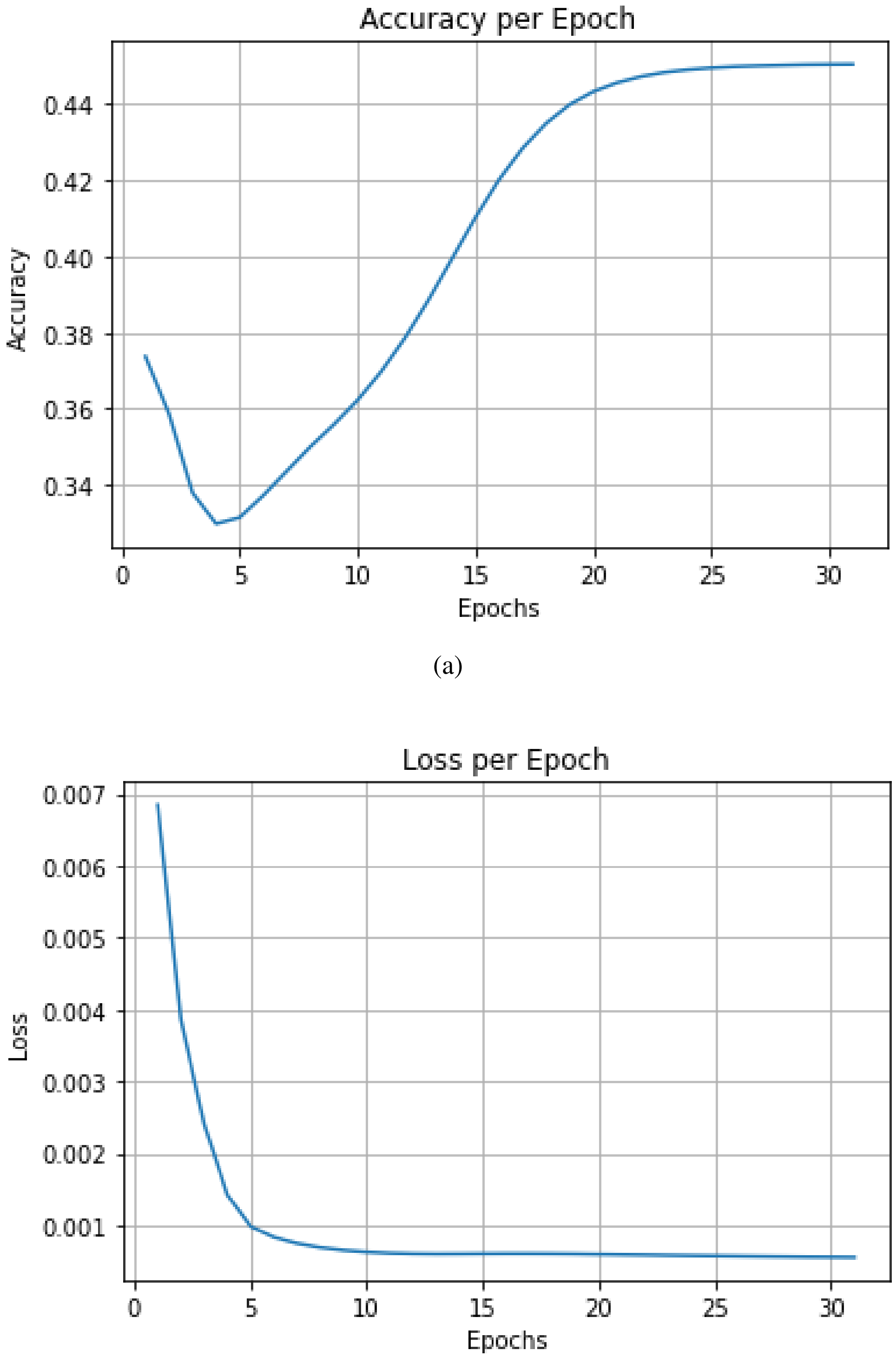


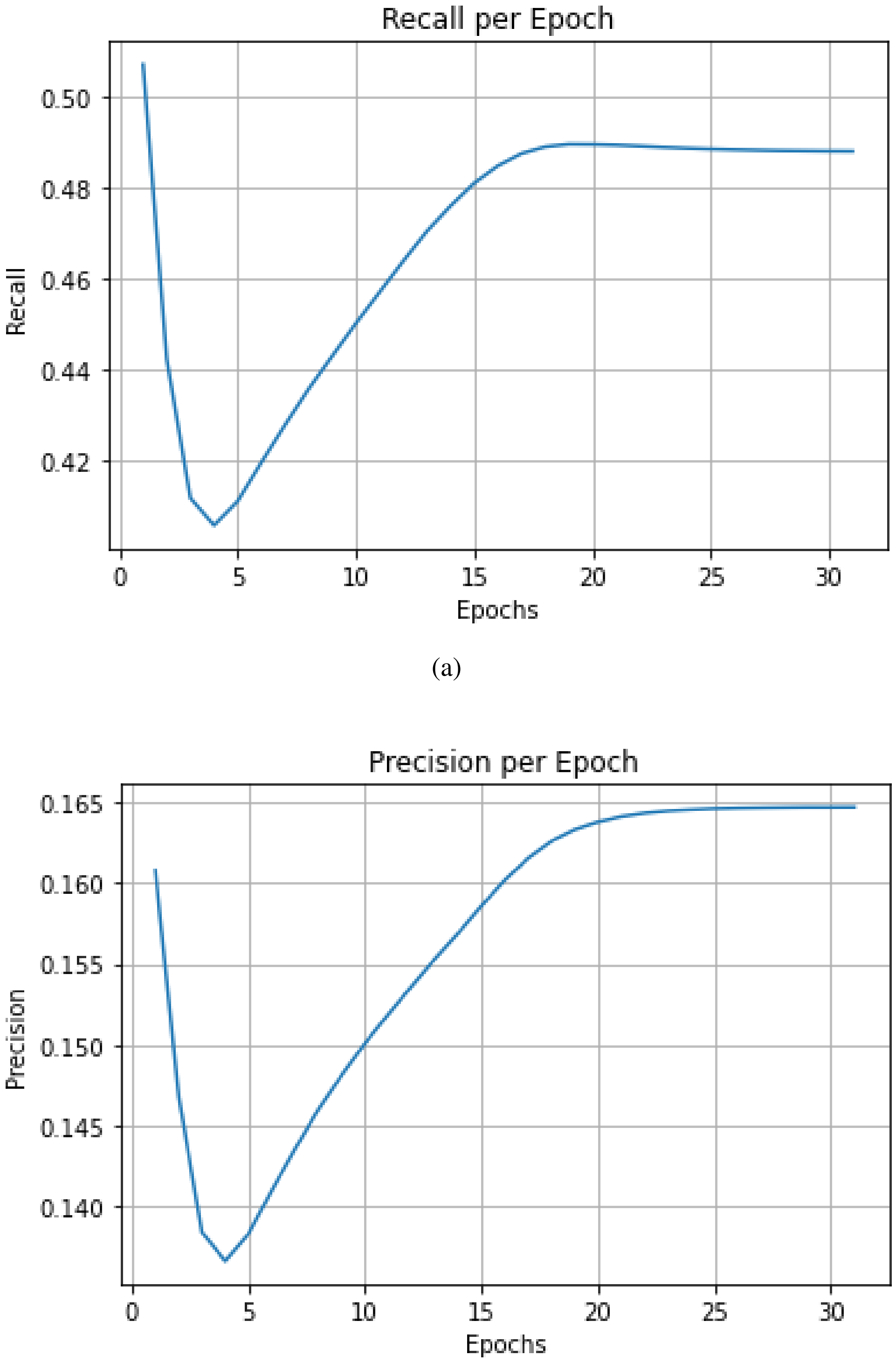
Figure 4.7: Multi-Class Training Result



(a)

(b)

Figure 4.8: Multi-Class Metrics:
(a) Multi-Class Accuracy, Showing a Low Accuracy, Worse than Acceptable.
(b) Multi-Class Loss, Showing Network Convergence on Data Provided.



(a)

(b)

Figure 4.9: Multi-Class Metrics:
(a) Multi-Class Recall, Showing a Low Recall, Suggesting Network Struggles to Learn all Classes.
(b) Multi-Class Precision, Showing very Low Precision, Suggesting Network Struggles to Identify Trained Classes.

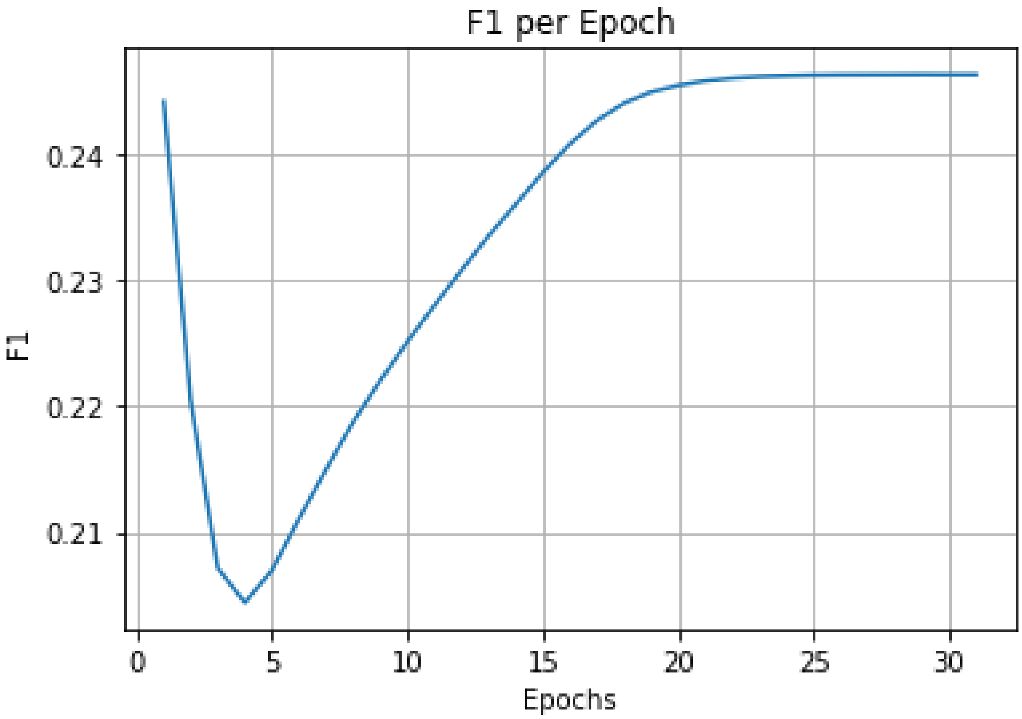


Figure 4.10: F1 Score, Combining the Outcome from Recall and Precision. Very Low F1 Score Suggests Learner is Struggling to Learn all Classes.

It is clear from the results of this network, that the learner has had a difficult time actually learning all of the classes that were asked of it. There is clearly the learned shape of the water, and even good portions of the water are correctly classified, but the learner is clearly confused when attempting to learn all of these classes. We can attribute this to multiple factors, including the class disparity and the lack of data for the network to train on. Looking at the class representation in Fig. 3.6, over 99% of the pixels represented in the data being used are either Water or Background. This great disparity means that the learner has less than 1% of all pixels to work with for learning any of the other classes. As was mentioned in Eq. III.1, the loss function applied with this learner was aimed at helping to learn harder to learn examples. Though we believe we've applied the correct loss function to help with the class disparity, there is still an inherent problem in the raw representation of each class. Looking at the data in Fig. 3.6 shows that classes that may be of major importance to situational awareness in the maritime environment, such as Boats, have more almost 4,000x less pixels than a class like water. Even a more represented class like Buildings, which we may consider to be less important for maritime situational awareness, has nearly 200x less pixels than water. All this is to say that irregardless of class disparity, there seems to be an issue in overall class representation. If the issue of insufficient data for learning the multi-class problem is addressed through data augmentation methods and introduction of additional new data, results similar to that of the water-only results are to be expected.

Training Time

The hardware used for training in this study, mentioned in Chapter III, was some of the best consumer-grade hardware that the team could have access to at the time of this research's inception. The training of the networks on this hardware took approximately 1 minute per epoch, and most instances of the network were run with between 25 and 75, depending on what was being tested. The final iterations of the networks that have been shown in this paper were run at between 35 and 50 epochs, meaning that these networks produced a result in approximately 35 to 50 minutes. This processing time is largely thanks to the power provided by the team's training PC. Times would be drastically greater with a weaker computer.

Chapter V

Discussion, Conclusions and Recommendations

Discussion

The work which was done in this thesis aims to provide a preliminary study for the feasibility of semantic segmentation on HDR imagery in the maritime environment for situational awareness purposes. The work which was done aimed to use the research group's HDR image data which is advantageous in the presence of environmental conditions like low-light and direct sun exposure.

Conclusions and Recommendation for Future Works

The work performed in this study created an small dataset of highly-detailed labeled HDR imagery for use in semantic segmentation in maritime environment, and tested the efficacy of this small dataset in performing object detection with the DeepLabV3+ network. This study was able to produce results with greater than 60% accuracy in multiple variations, however with the important caveat that it only performs well on two of the classes; Water and Background. This means that while the learner is capable of achieving this accuracy, it is only accurately identifying the water and the information that is non-water. From this study, we find that the amount of data present for use in learning was not sufficient for a multi-class problem.

It was found in this study that the network was able to accurately use the HDR imagery to identify the water in the maritime environment. This capability of the network allows for the accurate detection of the horizon using only the HDR camera, which can be used as a basic means of situational awareness. Additionally, this study performed multi-class semantic segmentation on the same HDR imagery with labels for up to 10 classes. It was found that for a significant number of the classes within the imagery labeled, there was simply not enough data to create a learner that was accurate enough to provide any use for the task of situational awareness. Though significant efforts were put forth to improve the results on the multi-class problem, there were no significant improvements made through hyper-parameter tuning and optimization efforts.

There are improvements that could be made to this study for the purpose of future studies. The main recommendation for future works would be to label more images for the network to train on. While the 600 images that the research group had labeled could be used for the task of water classification, there was no high level of success in classification for the multi-class problem. This problem stems from the lack of data available for the learner to use to learn more classes. It is recommended that for future studies with this data, a greater amount of time and resources are dedicated to building up the dataset for the learner. In this study, it was found that manually labelling the HDR imagery data with the detail required took a significant amount of time. For images with multiple instances of multiple classes, it could take upwards of 25-30 minutes to get the labels applied to that image. This time dedication to the task of labeling imagery means that the research group spent hundreds of hours between two researchers just to label the 600 images used in this study. Any future study which would expand on the work done here would need to dedicate the time and resources to build-up the labeled dataset so that the learner can have a greater chance of success in the multi-class problem.

Additionally, there are research studies which suggest some optimizations and changes which are standard to the industry aren't as ideal as many claim them to be. According to a 2017 study by Wilson et al., adaptive gradient optimization methods like Adam may actually prove to generalize worse on some deep learning problems when compared to the more generic gradient descent (GD) or stochastic gradient descent (SGD) [Wilson et al., 2017]. Though optimization methods like Adam tend to be commonplace in computer vision deep learning methods, there may be merit in exploring other methods as a means to improve the results of this study.

Lastly, there is research to suggest that a change in ideology about the way hyper-parameters are tuned may be beneficial. In a 2017 study by [Smith et al., 2017], they studied networks in which increased batch sizes lead to an equal or greater performance in classification when compared to a study performing learning rate decay methods [Smith et al., 2017]. This was studied on learners using SGD with momentum, Nesterov momentum, and Adam optimization routines. Their study suggests that with increased batch sizes alone, not only can learning rate scheduling be avoided entirely, but that comparable results can be achieved with little other hyper-parameter tuning. This comes with the caveat that greater batch sizes increase the amount of hardware memory required to train on image data.

References

[Pre, 2022] (2022). Automotive sensor market (by type: Pressure sensors, temperature sensors, speed sensors, motion sensors, and gas sensors; by application: Chassis, powertrain, safety amp; security, body electronics, and telematics) - global market size, trends analysis, segment forecasts, regional outlook 2022 - 2030.

[Bogue, 2016] Bogue, R. (2016). Growth in e-commerce boosts innovation in the warehouse robot market. *Industrial Robot: An International Journal*, 43(6):583–587.

[Brown et al., 2020] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

[Chen et al., 2022] Chen, I., Huang, L., Qiao, J., Tamir, D. E., and Rishe, N. (2022). Combining perception considerations with artificial intelligence in maritime threat detection systems. In *2022 17th Annual System of Systems Engineering Conference (SOSE)*. IEEE.

[Chen et al., 2016] Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2016). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs.

[Chen et al., 2017] Chen, L.-C., Papandreou, G., Schroff, F., and Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation.

[Chen et al., 2018] Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation.

[Dehal et al., 2018] Dehal, R. S., Munjal, C., Ansari, A. A., and Kushwaha, A. S. (2018). GPU computing revolution: CUDA. In *2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*. IEEE.

[Ecognition, 2021] Ecognition (2021). Using deep learning models / convolutional neural networks.

[Flämig, 2016] Flämig, H. (2016). Autonomous vehicles and autonomous driving in freight transport. *Autonomous Driving*, page 365–385.

[Google, 2022] Google (2022). Classification: Precision and recall nbsp;—nbsp; machine learning nbsp;—nbsp; google developers.

[Goring, 2017] Goring, R. (2017). Feasibility of neural networks for maritime visual detection on a mobile platform. In *Doctoral Dissertations and Master's Theses*.

[Grote, 2020] Grote, G. (2020). Safety and autonomy: A contradiction forever? *Safety Science*, 127:104709.

[Hożyń and Zalewski, 2020] Hożyń, S. and Zalewski, J. (2020). Shoreline detection and land segmentation for autonomous surface vehicle navigation with the use of an optical system. *Sensors*, 20(10):2799.

[Kaiser et al., 1995] Kaiser, M., Klingspor, V., del R. Millan, J., Accame, M., Wallner, F., and Dillmann, R. (1995). Using machine learning techniques in real-world mobile robots. *IEEE Expert*, 10(2):37–45.

[Khan et al., 2020] Khan, A. A., Akbar, R., and Sattar, T. (2020). Comparison of supervised learning models for bankruptcy prediction: A case study of textile industry in pakistan. In *Intelligent Information and Database Systems*, pages 61–71. Springer.

[Kim et al., 2021] Kim, C., Kim, S., Kim, J., Lee, D., and Kim, S. (2021). Automated learning rate scheduler for large-batch training.

[Kingma and Ba, 2014] Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization.

[LeopardImaging, 2023] LeopardImaging (2023). Li-usb30-imx490-gw5400-gmsl2-065h.

[Li et al., 2020] Li, Y., Ma, L., Zhong, Z., Liu, F., Cao, D., Li, J., and Chapman, M. A. (2020). Deep learning for lidar point clouds in autonomous driving: A review.

[Lin et al., 2022] Lin, J., Diekmann, P., Framing, C.-E., Zweigel, R., and Abel, D. (2022). Maritime environment perception based on deep learning. *IEEE Transactions on Intelligent Transportation Systems*, 23(9):15487–15497.

[Lin et al., 2017] Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). Focal loss for dense object detection.

[Long et al., 2014] Long, J., Shelhamer, E., and Darrell, T. (2014). Fully convolutional networks for semantic segmentation.

[Micikevicius et al., 2017] Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G., and Wu, H. (2017). Mixed precision training.

[Minaee et al., 2021] Minaee, S., Boykov, Y. Y., Porikli, F., Plaza, A. J., Kehtarnavaz, N., and Terzopoulos, D. (2021). Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1–1.

[Ng et al., 2020] Ng, M.-K., Chong, Y.-W., Ko, K.-m., Park, Y.-H., and Leau, Y.-B. (2020). Adaptive path finding algorithm in dynamic environment for warehouse robot. *Neural Computing and Applications*, 32(17):13155–13171.

[Paccaud and Barry, 2018] Paccaud, P. and Barry, D. A. (2018). Obstacle detection for lake-deployed autonomous surface vehicles using RGB imagery. *PLOS ONE*, 13(10):e0205319.

[Postalcıoğlu, 2019] Postalcıoğlu, S. (2019). Performance analysis of different optimizers for deep learning-based image recognition. *International Journal of Pattern Recognition and Artificial Intelligence*, 34(02):2051003.

[Ren et al., 2015a] Ren, S., He, K., Girshick, R., and Sun, J. (2015a). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, pages 91–99.

[Ren et al., 2015b] Ren, S., He, K., Girshick, R., and Sun, J. (2015b). Faster r-cnn: Towards real-time object detection with region proposal networks.

[Rieder and Verbeet, 2019] Rieder, M. and Verbeet, R. (2019). Robot-human-learning for robotic picking processes.

[Ross, 2022] Ross, R. S. (2022). Engineering trustworthy secure systems. Technical report.

[Russon, 2021] Russon, M.-A. (2021). The cost of the suez canal blockage.

[Senior et al., 2013] Senior, A., Heigold, G., Ranzato, M., and Yang, K. (2013). An empirical study of learning rates in deep neural networks for speech recognition. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE.

[Sheikh and Afanasyev, 2018] Sheikh, T. S. and Afanasyev, I. M. (2018). Stereo vision-based optimal path planning with stochastic maps for mobile robot navigation. In *Intelligent Autonomous Systems 15*, pages 40–55. Springer International Publishing.

[Smith et al., 2017] Smith, S. L., Kindermans, P.-J., Ying, C., and Le, Q. V. (2017). Don't decay the learning rate, increase the batch size.

[Sun et al., 2020] Sun, S., Cao, Z., Zhu, H., and Zhao, J. (2020). A survey of optimization methods from a machine learning perspective. *IEEE Transactions on Cybernetics*, 50(8):3668–3681.

[TensorFlow, 2023] TensorFlow (2023). Tf.keras.metrics.categoricalaccuracy nbsp;: nbsp; tensorflow v2.12.0.

[Tsai et al., 2021] Tsai, C.-Y., Nisar, H., and Hu, Y.-C. (2021). Mapless LiDAR navigation control of wheeled mobile robots based on deep imitation learning. *IEEE Access*, 9:117527–117541.

[UNCTAD, 2022] UNCTAD (2022). Review of maritime transport 2022.

[USCG, 2023] USCG (2023). Automatic identification system (ais) overview.

[Vargas et al., 2021] Vargas, J., Alsweiss, S., Toker, O., Razdan, R., and Santos, J. (2021). An overview of autonomous vehicles sensors and their vulnerability to weather conditions. *Sensors*, 21(16):5397.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

[Vigier et al., 2016] Vigier, T., Krasula, L., Milliat, A., Da Silva, M. P., and Le Callet, P. (2016). Performance and robustness of hdr objective quality metrics in the context of recent compression scenarios. In *2016 Digital Media Industry Academic Forum (DMIAF)*, pages 59–64.

[Wilson et al., 2017] Wilson, A. C., Roelofs, R., Stern, M., Srebro, N., and Recht, B. (2017). The marginal value of adaptive gradient methods in machine learning.

[Xie et al., 2017] Xie, D., Zhang, L., and Bai, L. (2017). Deep learning in visual computing and signal processing. *Applied Computational Intelligence and Soft Computing*, 2017:1–13.

[Yeong et al., 2021] Yeong, D. J., Velasco-Hernandez, G., Barry, J., and Walsh, J. (2021). Sensor and sensor fusion technology in autonomous vehicles: A review. *Sensors*, 21(6):2140.

[Yu, 2016] Yu, F. (2016). Multi-scale context aggregation by dilated convolutions. In *Proceedings of the 4th International Conference on Learning Representations*.

[Zhang et al., 2019] Zhang, L., Guo, X., Zhang, J., Cai, J., and Wei, Y. (2019). Attention to scale: Scale-aware semantic image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3640–3649.