Dissertation Proposal

Sight in the Sea: Physically-Based Methods to Advance Underwater Vision for Robotic Systems

by
Katherine A. Skinner
Robotics Institute
University of Michigan

Committee:

Matthew Johnson-Roberson, Chair
Assistant Professor, Department of Naval Architecture and Marine Engineering
University of Michigan

Ram Vasudevan
Assistant Professor, Department of Mechanical Engineering
University of Michigan

Ryan M. Eustice
Associate Professor, Department of Naval Architecture and Marine Engineering
University of Michigan

Emily Mower Provost
Assistant Professor, Department of Computer Science and Engineering
University of Michigan

Brian Hopkinson
Associate Professor, Department of Marine Sciences
University of Georgia

April 18, 2018

Abstract

In recent years, advances in computer vision and deep learning have led to breakthroughs in perceptual capabilities of robotic systems. Autonomous cars can now drive thousands of miles without intervention using real-time localization and mapping. Robotic platforms with manipulators can pick and stow objects on cluttered shelves based on object and pose recognition. Yet there remain challenges to extending these advances to all domains, and to deploying these systems in the real world. Supervised learning methods require large amounts of training data with ground truth labels, which is difficult and time consuming to collect. Building vision systems that are robust to varying environmental conditions is also an open problem. These challenges are exacerbated in field robotics, where robots are deployed in natural, unstructured environments that have highly dynamic environmental conditions. For these applications, training data is expensive to collect and ground truth is difficult or impossible to gather.

This thesis focuses on robotic perception in field robotics, specifically in underwater environments. The underwater domain presents unique environmental conditions to robotic systems that exacerbate the challenges in perception for field robotics. One of these factors is the physical model of underwater light propagation. As a photon of light travels through an aqueous medium, it interacts with particulate matter in the water column, which can cause it to scatter or become absorbed completely. Absorption and scattering cause range- and wavelength-dependent attenuation of the signal that finally reaches the image sensor, which leads to exponential decay of color at different rates for different wavelengths. Scattering reduces the effective resolution of the image and produces a haze effect across the scene. As a result of these effects, assumptions commonly employed in computer vision algorithms for terrestrial applications, such as the brightness constancy constraint, are not valid in the underwater domain. Additionally, real-time depth sensors that have been a boon to terrestrial robotics applications are hindered underwater. Another factor that exacerbates these challenge is the difficulty in gathering training data and ground truth labels in subsea environments. Marine operations are expensive and difficult to carry-out, in many cases requiring expert knowledge and experience of operational challenges that arise at sea. Ground truth for true color of natural underwater scenes is impossible to ascertain. These factors make it difficult to apply computer vision and deep learning techniques developed for terrestrial applications to field robotics applications in the underwater domain.

This thesis focuses on developing unsupervised learning approaches to address the challenge of gathering ground truth labels in subsea environments. The proposed approaches explicitly incorporate the physical model of underwater light propagation to account for these range-dependent effects in a physically realistic way. Although this work specifically focuses on the underwater domain, its aim is to present novel frameworks for using physical models of environmental conditions in computer vision and unsupervised learning contexts to work towards robust perception systems for deploying robotic platforms in natural and unstructured environments.

CONTENTS

1 Introduction **1**

1.1 Motivation . . . 1
1.2 Underwater Image Formation . . . 3
1.3 Computer Vision Challenges . . . 5
1.4 Operational Challenges . . . 6
1.5 Problem Statement . . . 7
1.6 Contributions . . . 7
1.7 Proposed Schedule . . . 8

2 Unsupervised Generative Network to Enable Real-time Color Correction of Monocular Underwater Images **9**

2.1 Introduction . . . 9
2.2 Background . . . 10
2.3 Methodology . . . 11
2.3.1 Generating Realistic Underwater Images . . . 11
2.3.2 Underwater Image Restoration Network . . . 14
2.4 Experiments & Results . . . 16
2.4.1 Experimental setup . . . 16
2.4.2 Artificial Testbed . . . 16
2.4.3 Field Tests . . . 16
2.4.4 Network Training . . . 16
2.5 Results and Discussion . . . 17
2.6 Discussion & Conclusion . . . 20

3 Proposed Research in Learning for Simultaneous Color Correction and Depth Estimation of Stereo Underwater Imagery **22**

3.1 Introduction . . . 22
3.2 Background . . . 22
3.3 Methodology . . . 22
3.4 Experiments & Results . . . 22
3.5 Discussion & Conclusion . . . 22

4 Proposed Research in Real-time Underwater 3D Reconstruction **23**
4.1 Introduction . . . 23
4.2 Background . . . 23
4.2.1 Real-time Dense 3D Reconstruction . . . 23
4.2.2 Underwater 3D Reconstruction . . . 23
4.3 Methodology . . . 24
4.4 Experiments & Results . . . 24
4.5 Discussion & Conclusion . . . 24

5 Conclusions **25**

Bibliography **27**

LIST OF FIGURES

1.1 Map of tracklines from the National Centers for Environmental Information (NCEI) Marine Trackline Geophysical database showing coverage of ocean expeditions to collect geophysical observations of the seafloor between 1939-2018 (present). 1

1.2 Map of the Campeche Escarpment in the Gulf of Mexico. The bottom layer is the best bathymetry map available for the area pre-2013. The top layer shows a multibeam survey gathered during a 2013 expedition from R/V Falkor. The depths of the multibeam data range from $400m$ (red) to $3700m$ (blue). 2

1.3 Robotic platforms can be equipped with high resolution, color cameras to conduct systematic imaging surveys of the seafloor. 3

1.4 Abstraction of several water column effects, including scattering and attenuation, which are modeled as a function of the wavelength of light and range to the camera. 3

1.5 Sample underwater images from various test sites. 4

1.6 Point $\mathbf{P}(j)$ is observed in each of the stereo camera pairs k, both in the **L**eft and **R**ight cameras. 6

1.7 Timeline for proposed research. 8

2.1 Flowchart displaying both the WaterGAN and color correction networks. WaterGAN takes input in-air RGB-D and a sample set of underwater images and outputs synthetic underwater images aligned with the in-air RGB-D. The color correction network uses this aligned data for training. For testing, a real monocular underwater image is input and a corrected image and relative depth map are output. 9

2.2 WaterGAN: The GAN for generating realistic underwater images with similar image formation properties to those of unlabeled underwater data taken in the field. 12

2.3 Network architecture for color estimation. The first stage of the network takes a synthetic (training) or real (testing) underwater image and learns a relative depth map. The image and depth map are then used as input for the second stage to output a restored color image as it would appear in air. 14

2.4 (a) An artificial rock platform and (b) a diving color board are used to provide ground truth for controlled imaging tests in (c) a pure water tank to gather the MHL dataset. 16
2.5 Results showing color correction on the MHL, Lizard Island, and Port Royal datasets (from top to bottom). Each column shows (a) raw underwater images, and corrected images using (b) histogram equalization, (c) normalization with the gray world assumption, (d) a modified Jaffe-McGlamery model (Eqn. 3) with ideal attenuation coefficients, (e) Shin et al.'s deep learning approach, and (f) WaterGAN. 18
2.6 Zoomed-in comparison of color correction results of an image with and without skipping layers. 20

LIST OF TABLES

2.1 Color correction accuracy based on Euclidean distance of intensity-normalized color in RGB-space for each method compared to the ground truth in-air color board. 19
2.2 Variance of intensity-normalized color of single scene points imaged from different viewpoints. 19
2.3 Validation error in pixel value is given in RMSE in RGB-space. Validation error in depth is given in RMSE (m). 19

CHAPTER 1

INTRODUCTION

1.1 Motivation

Humans have sailed the sea and exploited its resources for thousands of years, but the first scientific expedition to methodically explore the oceans began in 1872 aboard the HMS *Challenger*. Over the course of 1000 days, the *Challenger* sailed around the world, collecting observations of marine life, measurements of water properties, and samples of seafloor sediment. Since then, scientific research vessels have covered much more of our oceans and waterways, which account for 70% of the Earth's surface (Fig. 1.1).

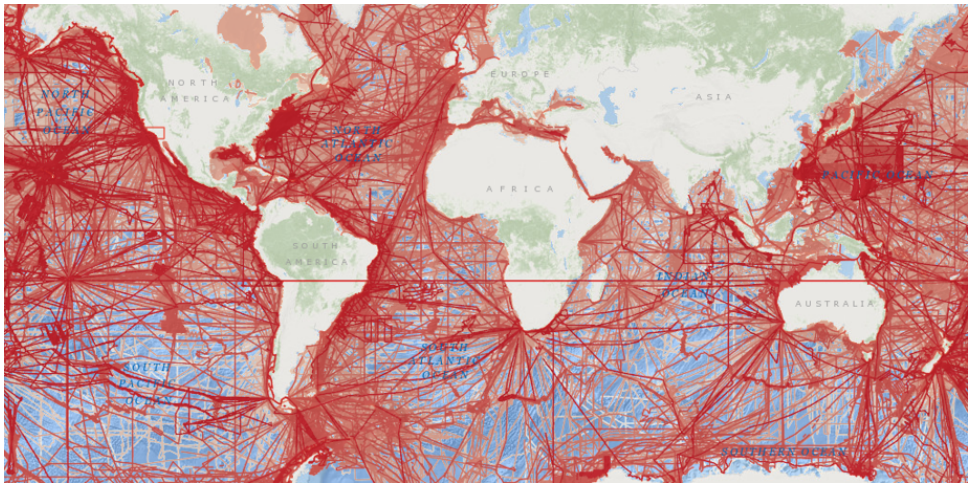


Figure 1.1: Map of tracklines from the National Centers for Environmental Information (NCEI) Marine Trackline Geophysical database showing coverage of ocean expeditions to collect geophysical observations of the seafloor between 1939-2018 (present).

Methods and sensors for collecting oceanographic data have drastically improved. Figure 1.2 shows two bathymetric maps of the Campeche Escarpment in the Gulf of Mexico. The base map was compiled from hydrographic charts and satellite altimetry and, until 2013, was the highest resolution bathymetric map available for the site. The top map was collected with a multibeam sonar during a survey from the R/V *Falkor*. The depths recorded range from $400m$ (red) to $3700m$ (blue).

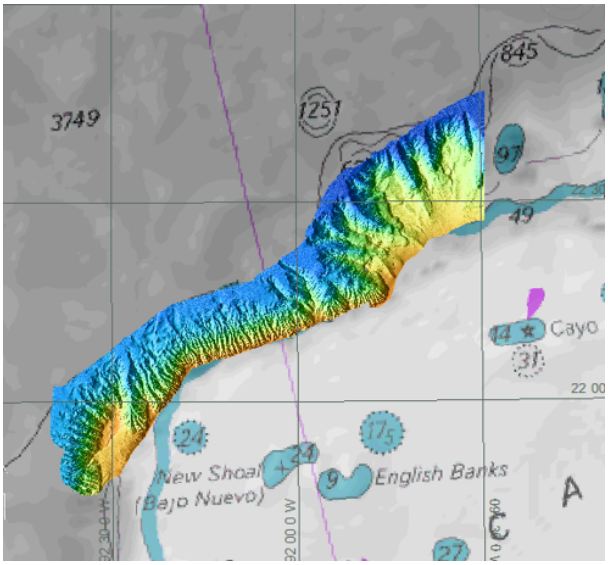


Figure 1.2: Map of the Campeche Escarpment in the Gulf of Mexico. The bottom layer is the best bathymetry map available for the area pre-2013. The top layer shows a multibeam survey gathered during a 2013 expedition from R/V Falkor. The depths of the multibeam data range from $400m$ (red) to $3700m$ (blue).

Acoustic sensors such as multibeam echosounders are commonly used for mapping bathymetry. Due to their long range underwater, sonar instruments can be mounted on ships to conduct surveys from the sea surface. The resolution of sonar systems is determined by the frequency and range to the scene. Side scan sonars operationg at higher frequencies can achieve higher resolution, but their operational range is limited due to attenuation through the water column. Synthetic aperture sonars (SAS) use state-of-the-art post-processing techniques to achieve up to $4cm$ resolution at a range of over $100m$. However, these systems must be used from a moving platform over a static scene, and their current cost inhibits widespread application. Additionally, sonar cannot gather color information of a scene, which is important for marine science applications.

Alternatively, optical sensors can be used to obtain high resolution color imagery of subsea environments. Optical sensing can be achieved even at beyond the photic zone using artificial lighting. However, optical sensors are subject to the attenuation of electromagnetic signals through the water column, which limits their ideal operational range. Generally, maximum visibility range is less than $25m$, with the world record range recorded at $79m$. In practice, ideal range for imaging systems is approximately $2-4m$. Thus, imaging surveys of the seafloor cannot be carried out from surface vessels in deeper waters. Furthermore, aperture limited by practical size, which also limits the field-of-view, or the observable area per frame.

Instead, many fields in marine science and engineering rely on underwater robotic platforms equipped with imaging sensors to provide high resolution, colored views of the seafloor. At depths less than $40m$, divers can carry diver rigs, or platforms equipped with cameras and navigation sensors to gather imagery of the seafloor (Fig. 1.3a). Remotely operated vehicles (ROVs) and autonomous underwater vehicles (AUVs) can be deployed in depths up to $11000m$ (Fig. 1.3b). Robotic systems are capable of carrying out large-scale surveys efficiently and effectively. Using computer vision techniques, images can then be integrated to create a map

or 3D model of the surveyed area.

(a) Diver rig survey at Hog Reef off the coast of Bermuda, conducted as a collaboration between Woods Hole Oceanographic Institution (WHOI), University of Michigan (UM), and University of Georgia (UGA).

(b) Deployment of the DROP Lab's Iver autonomous underwater vehicle (AUV) during a research cruise aboard the R/V *Falkor*.

Figure 1.3: Robotic platforms can be equipped with high resolution, color cameras to conduct systematic imaging surveys of the seafloor.

1.2 Underwater Image Formation

While recent decades have seen great advancements in vision capabilities of underwater platforms, the subsea environment presents unique challenges to optical imaging that are not present on land.

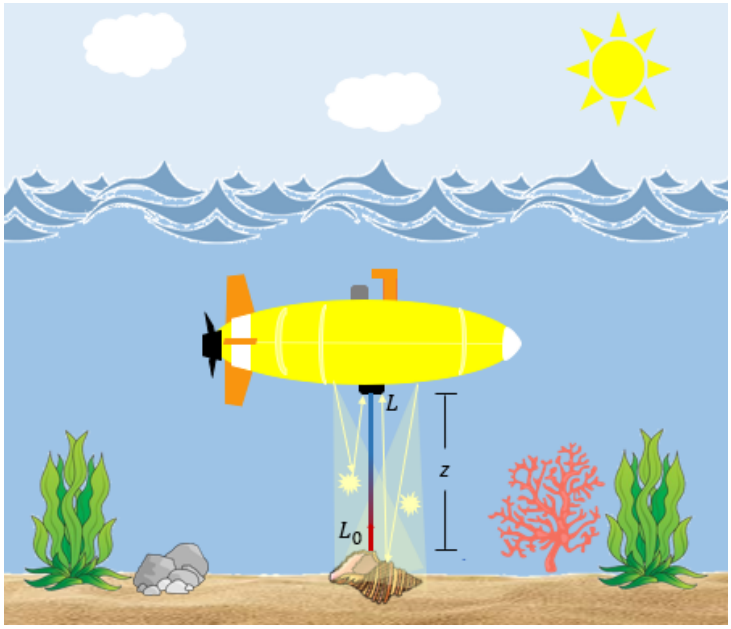
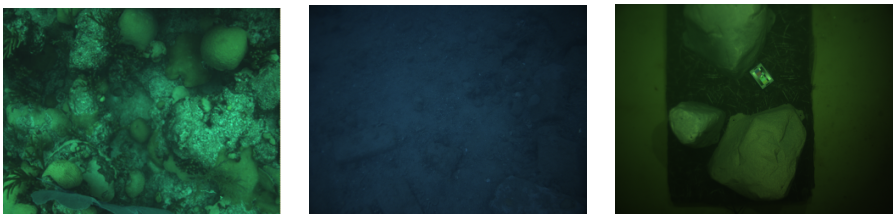


Figure 1.4: Abstraction of several water column effects, including scattering and attenuation, which are modeled as a function of the wavelength of light and range to the camera.

Figure 1.4 shows an abstraction of underwater light propagation between a camera system and a target scene. As a photon of light travels from the camera to the scene and back,

it interacts with surrounding water molecules and particulate matter in the water column, and can be either scattered or completely absorbed. Absorption is wavelength-dependent and significantly contributes to wavelength-dependent attenuation of light underwater. The rate of attenuation also depends on the properties of the water; it varies between fresh and salt water, coastal and ocean water, and even seasonally. Backscattering occurs when the light is scattered back towards the camera before it reaches the scene. This causes a haze effect, often referred to as veiling light. The backscattered signal is the main contributor to image degradation of underwater images. Forward scattering also occurs when light that is scattered away from the camera gets rescattered along the line of sight, leading to a blurring of the scene. However, its contribution to attenuation and image degradation is negligible compared to absorption and backscattering, so it is frequently omitted. Figure 1.5 provides examples of raw underwater images in different bodies of water to demonstrate the range of colors and quality that result from water column effects on underwater light propagation.



(a) Hog Reef, Bermuda (b) Port Royal, Jamaica (c) MHL Tank

Figure 1.5: Sample underwater images from various test sites.

Light propagation through a scattering media can be described by the radiance transfer equation (RTE). For a homogeneous water column, assuming no inelastic scattering or emission, a compact form of the RTE is given by:

$$L(d;\xi;\lambda) = L_0(d_0;\xi;\lambda)e^{-\beta(\lambda)z} + \frac{L_*(d;\xi;\lambda)e^{-K_d(\lambda)z\cos\theta}}{\beta(\lambda) - K_d(\lambda)\cos\theta}[1 - e^{-[\beta(\lambda)-K_d(\lambda)\cos\theta]z}] \tag{1.1}$$

L_0 is true radiance of the target scene, and L is the received radiance subject to water column effects. ξ is the 3D ray direction, λ is wavelength, z is range between the camera and the scene, d is depth (vertical range only), K_d is the attenuation coefficient due to diffuse downwelling, β is the beam attenuation coefficient representing loss of photons due to absorption and scattering, and θ denotes viewing direction.

More intuitively, the above equation describes the received radiance, L, as the combination of directly transmitted light, D, subject to attenuation, and backscattered light, B, which carries no information about the scene:

$$L = D + B \tag{1.2}$$

This model for underwater light propagation and its and application to perception of underwater robotic systems will be further explored throughout this work.

1.3 Computer Vision Challenges

The complex, nonlinear process of underwater image formation presents several challenges for the field of computer vision. On land, real-time depth sensing has been an incredible boon to perception of mobile robots [1], with advances in sensor modalities such as stereo cameras, light detection and ranging (LIDAR), and more recently color and depth (RGB-D) sensors [36]. Unfortunately, the latter two – which include time-of-flight range sensing and pattern projection structured light approaches – are still quite limited in their success underwater due to attenuation of electromagnetic signals through the aqueous medium [9].

In recent decades, stereo cameras have been popular sensing systems for underwater robots. With calibrated stereo pairs, high resolution images can be aligned with depth information to compute large-scale photomosaic maps or metrically accurate 3D reconstructions [13]. However, degradation of images due to range-dependent underwater lighting effects can hinder these approaches. The first step in traditional stereo vision pipelines is to detect distinct feature patches within an image. Common feature descriptors, including SIFT, SURF, and ORB, rely on image contrast, which is reduced in underwater images due to attenuation of light. Severe haze reduces visibility and can lead to total loss of information about image features. These factors make it challenging to detect distinct features in raw underwater imagery.

Once features are detected, feature matching is attempted across stereo image pairs (e.g. left-right images). With matched features and calibration, it is then possible to compute disparity, or depth maps, for each view. As more stereo views are gathered across a vehicle trajectory, feature matches can also be made across multiple viewpoints, such as when the vehicle crosses back over a feature it has seen before. These long-range matches can provide strong constraints on the relative pose of the vehicle throughout its trajectory, as well as on the geometry of the scene. One assumption that enables long-range feature matching on land is that one feature imaged from different viewpoints will appear with the same intensity across each image. This is known as the brightness constancy constraint (BCC). Similarly, state-of-the-art methods for real-time 3D reconstruction also rely on photometric consistency as a cue for tracking and fusing image patches from view-to-view.

These assumptions do not hold underwater. Figure 1.6 shows an illustration of a case where assumptions of photometric consistency fail: If the same scene point $P(j)$ is imaged from different viewpoints k and $(k+1)$ with a large range disparity between viewpoints, the same feature will appear with different radiance in the resulting images, $I_{k,L}$ and $I_{k+1,L}$ due to range-dependent attenuation.

One approach to restore underwater images is to use image processing techniques, such as histogram equalization, to stretch the effective contrast of the image. This can improve feature detection and matching, and often results in visually appealing images. However, image processing techniques have no knowledge of the physical process of underwater image formation; thus they do not account for range-dependent effects to restore photometric consistency across different viewpoints.

Instead, this thesis proposes novel methods for underwater image restoration that incorporate the physical model of underwater light propagation to improve accuracy and consistency of corrected underwater images. Still, this is an ill-posed problem. Scene geometry and object ra-

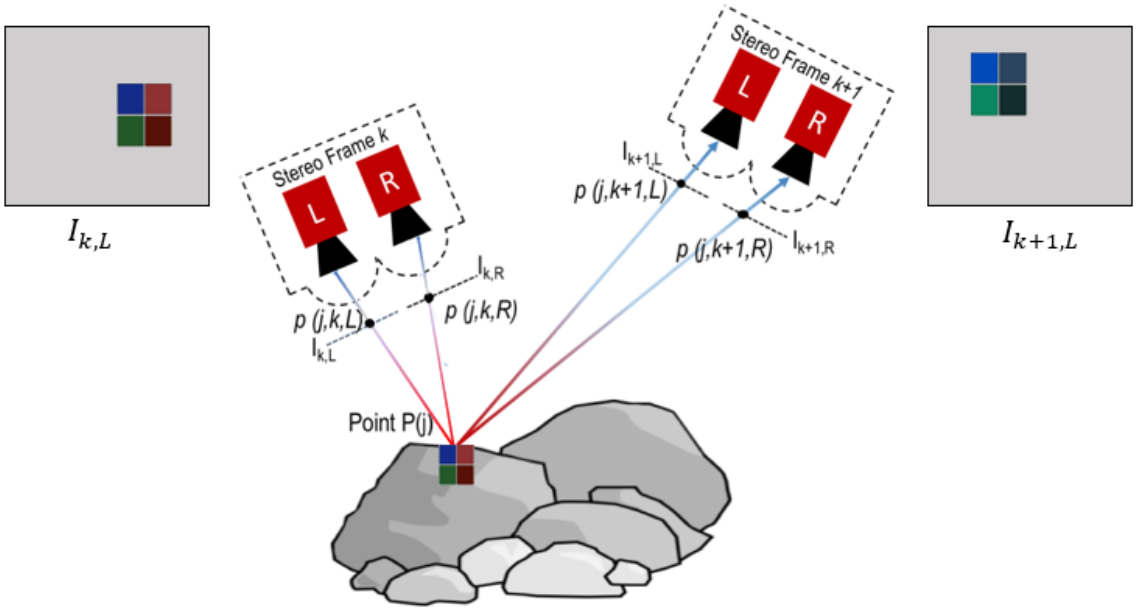


Figure 1.6: Point $\mathbf{P}(j)$ is observed in each of the stereo camera pairs k, both in the **L**eft and **R**ight cameras.

diance are interdependent, and automating solutions for these parameters is further confounded by other factors such as the imaging sensor and water properties.

1.4 Operational Challenges

It is also important to acknowledge challenges that arise in practice when working in marine environments. Marine operations are expensive, time-consuming, and sometimes dangerous to carry out. When using diver rigs, diver safety is a critical consideration for system design and mission planning. For boat operations, hiring a boat and crew with all necessary equipment can be expensive. Deep ocean expeditions can require being at sea for weeks at a time. Unpredictable weather can make conditions unsafe for sailing, or vehicle deployment. Wildlife may also interfere with divers or vehicles. In practice, these challenges leave a limited window of opportunity for data collection, ultimately limiting the amount of data that can be collected.

Another factor to consider is that WiFi and global positioning systems (GPS) do not operate underwater due to attenuation of electromagnetic signals. GPS is commonly used for localization on land as it provides highly accurate absolute position. Underwater robots must rely on other sensors – such as Doppler Velocity Logs (DVLs), inertial measurement units (IMUs), or camera systems – to determine relative position and orientation. Furthermore, ground truth position is only available when the vehicle is on the surface, where it is possible to get GPS data.

In general, ground truth data for other measurements is difficult to gather in subsea environments. Ground truth is important for validating developed methods or for supervised training for deep learning approaches. As mentioned previously, high resolution depth sensors used on land do not operate well underwater. Thus it is difficult to gather metrically accurate ground truth of depth or structure geometry. Ground truth color of submerged structures is also impractical to gather, unless the structure can be removed from the marine environment

for evaluation on land.

1.5 Problem Statement

Summarizing the above sections, this thesis proposal seeks solutions to the following challenges to perception in underwater robotics:

1. Assumptions used in state-of-the-art computer vision for terrestrial applications break down underwater due to water column effects.

2. The physical model for underwater image formation has been well-studied but there are challenges to implementing a generalizable, automated solution to account for water column effects in computer vision pipelines. The restoration of underwater images involves reversing effects of a complex physical process with prior knowledge of water column characteristics for a specific survey site. Parameters of this physical process are wavelength- and range-dependent.

3. Operational challenges in underwater robotics limit the amount and quality of data that can be collected, in practice. Additionally, ground truth of both geometry and color are particularly difficult to gather.

1.6 Contributions

The main objective of this proposed thesis is to improve perception for underwater robotic systems. To achieve this objective, this thesis proposal presents and proposes the following contributions:

1. Develop a deep learning framework that learns the underlying physical model of underwater light propagation in order to simulate and correct for water column effects on color in monocular underwater imagery. (Chapter 2)

2. (Proposed) Develop a method to learn both structure and color simultaneously for stereo underwater imagery to perform online depth estimation and color correction. (Chapter 3)

3. (Proposed) Develop a perception system that performs state-of-the-art real-time 3D reconstruction from an underwater robotic platform. (Chapter 4)

Work presented throughout this thesis proposal has been published in the following publications:

Jie Li*, **Katherine A. Skinner***, Ryan Eustice and Matthew Johnson-Roberson, "WaterGAN: Unsupervised generative network to enable real-time color correction of monocular underwater images." In IEEE Robotics and Automation Letters, 2017. *The authors contributed equally to this work.

Eduardo Iscar, **Katherine A. Skinner** and Matthew Johnson-Roberson, "Multi-view 3D reconstruction in underwater environments: evaluation and benchmark." In Proceedings of the

IEEE/MTS OCEANS Conference and Exhibition, Anchorage, USA, September 2017.

Katherine A. Skinner and Matthew Johnson-Roberson, "Underwater Image Dehazing with a Light Field Camera." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition – Workshops, 2017.

Katherine A. Skinner, Eduardo Iscar Ruland and Matthew Johnson-Roberson, "Automatic color correction for 3D reconstruction of underwater scenes." In Proceedings of the IEEE International Conference on Robotics and Automation, Singapore, 2017.

Katherine A. Skinner and Matthew Johnson-Roberson, "Towards real-time underwater 3D reconstruction with plenoptic cameras." In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Daejeon, Korea, 2016.

1.7 Proposed Schedule

Figure 1.7 shows a proposed schedule for research and publications.

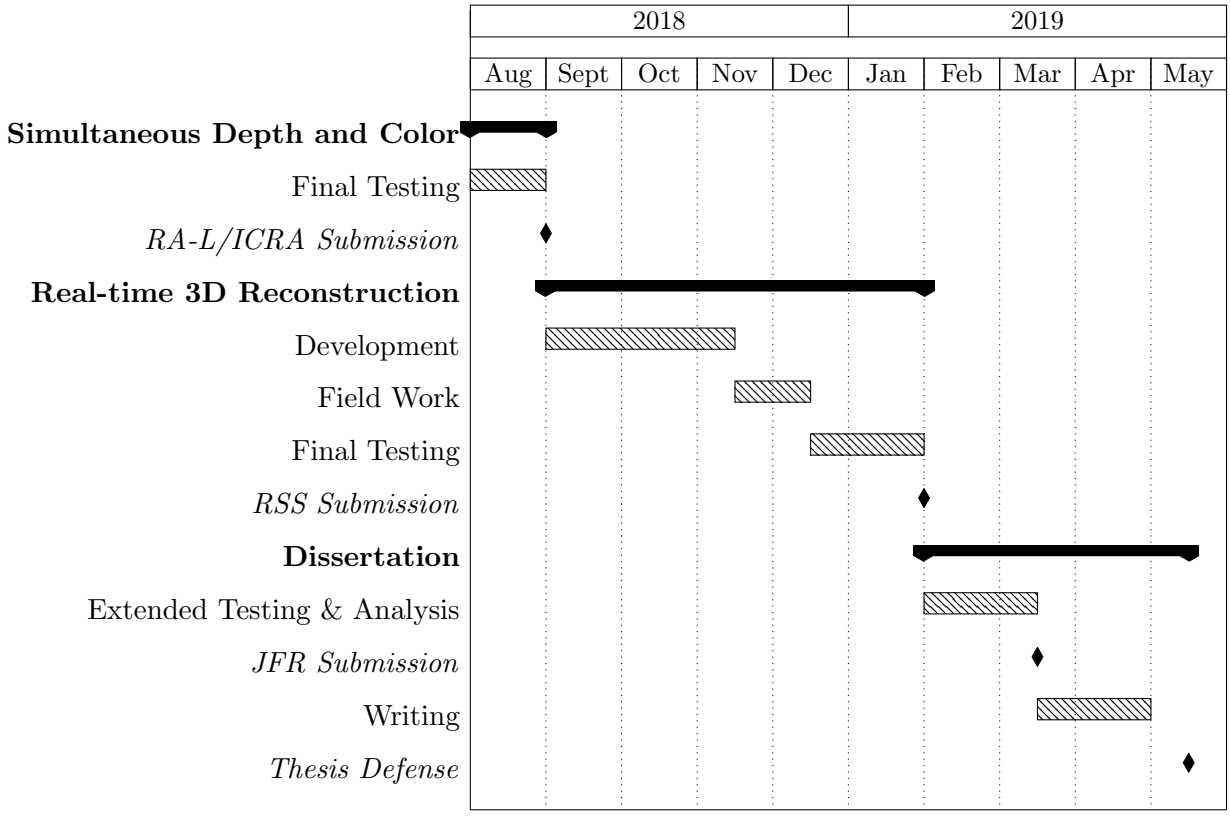


Figure 1.7: Timeline for proposed research.

CHAPTER 2

UNSUPERVISED GENERATIVE NETWORK TO ENABLE REAL-TIME COLOR CORRECTION OF MONOCULAR UNDERWATER IMAGES

2.1 Introduction

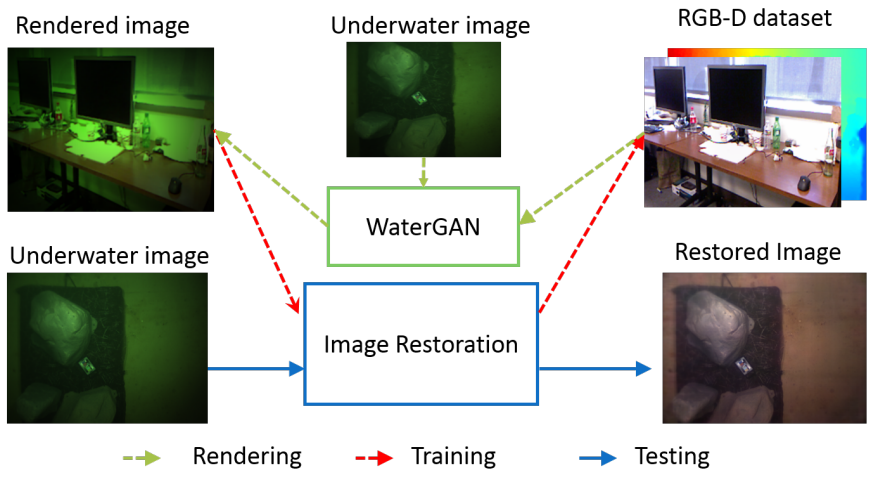


Figure 2.1: Flowchart displaying both the WaterGAN and color correction networks. WaterGAN takes input in-air RGB-D and a sample set of underwater images and outputs synthetic underwater images aligned with the in-air RGB-D. The color correction network uses this aligned data for training. For testing, a real monocular underwater image is input and a corrected image and relative depth map are output.

The restoration of underwater images involves reversing effects of a complex physical process with prior knowledge of water column characteristics for a specific survey site. Additionally, image restoration efforts must account for range- and wavelength-dependencies of water column effects in order to restore photometric constistency and brightness constancy. In recent years, advances in neural networks have enabled end-to-end modeling of complex nonlinear systems. Yet deep learning has not become as commonplace subsea as it has for terrestrial applications. One challenge is that many deep learning structures require large amounts of training data,

typically paired with labels or corresponding ground truth sensor measurements. Gathering large sets of underwater data with depth information is challenging in deep sea environments; obtaining ground truth of the true color of a natural subsea scene is also an open problem.

Rather than gathering training data, this chapter presents a novel approach, WaterGAN, a generative adversarial network (GAN) [7] that uses real unlabeled underwater images to learn a realistic representation of water column properties of a particular survey site. WaterGAN takes in-air images and depth maps as input and generates corresponding synthetic underwater images as output. This dataset with corresponding depth data, in-air color, and synthetic underwater color can then supplant the need for real ground truth depth and color in the training of a color correction network. As an application, this chapter also presents a novel network to perform monocular color correction based on generated data. The color correction network takes as input raw unlabeled underwater images and outputs restored images that appear as if they were taken in air.

This chapter is organized as follows: §2.2 presents relevant prior work; §2.3 gives a detailed description of the technical approach; §2.4.1 presents an experimental setup to validate our proposed approach; §2.5 provides results and a discussion of these results; lastly, §2.6 concludes the chapter.

2.2 Background

Prior work on compensating for effects of underwater image formation has focused on explicitly modeling this physical process to restore underwater images to their true color. Jordt et al. used a modified Jaffe-McGlamery model with parameters obtained through prior experiments [15] [8]. However, attenuation parameters vary for each survey site depending on water composition and quality. Bryson et al. used an optimization approach to estimate water column and lighting parameters of an underwater survey to restore the true color of underwater scenes [4]. However, this method requires detailed knowledge of vehicle configuration and the camera pose relative to the scene. The method developed in this chapter instead learns to model these effects using a deep learning framework without explicitly encoding vehicle configuration parameters.

Approaches that make use of the gray world assumption [14] or histogram equalization are common preprocessing steps for underwater images and may result in improved image quality and appearance. However, as such methods have no knowledge of range-dependent effects, resulting images of the same object viewed from different viewpoints may appear with different colors. Work has been done to enforce the consistency of restored images across a scene [3], but these methods require dense depth maps. Preliminary work aimed to relax this requirement using an underwater bundle adjustment formulation to estimate the parameters of a fixed attenuation model and the 3D structure simultaneously [33], but such approaches require a fixed image formation model and handle unmodeled effects poorly. The approach presented here can perform restoration with individual monocular images as input, and learns the relative structure of the scene as it corrects for the effects of range-dependent attenuation.

Several methods have addressed range-dependent image dehazing by estimating depth through developed or statistical priors on attenuation effects [5, 6, 18]. More recent work has focused on leveraging the success of deep learning techniques to estimate parameters of the complex

physical model. Shin et al. [28] developed a deep learning pipeline that achieves state-of-the-art performance in underwater image dehazing using simulated data with a regression network structure to estimate parameters for a fixed restoration model. The method developed in this chapter incorporates real field data in a generative network to learn a realistic representation of environmental conditions for raw underwater images of a specific survey site.

WaterGAN, is structured as a generative adversarial network (GAN). GANs have shown success in generating realistic images in an unsupervised pipeline that only relies on an unlabeled set of images of a desired representation [7]. A standard GAN generator receives a noise vector as input and generates a synthetic image from this noise through a series of convolutional and deconvolutional layers [27]. Recent work has shown improved results by providing an input image to the generator network, rather than just a noise vector. Shrivastava et al. provided a simulated image as input to their network, SimGAN, and then used a refiner network to generate a more realistic image from this simulated input [30]. To extend this idea to the domain of underwater image restoration, WaterGAN also incorporates easy-to-gather in-air RGB-D data into the generator network since underwater image formation is range-dependent. Sixt et al. proposed a related approach in RenderGAN, a framework for generating training data for the task of tag recognition in cluttered images [32]. RenderGAN uses an augmented generator structure with augment functions modeling known characteristics of their desired images, including blur and lighting effects. RenderGAN focuses on a finite set of tags and classification as opposed to a generalizable transmission function and image-to-image mapping.

2.3 Methodology

This chapter presents a two-part technical approach to produce a pipeline for image restoration of monocular underwater images. Figure 2.1 shows an overview of the full pipeline. WaterGAN is the first component of this pipeline, taking as input in-air RGB-D images and a sample set of underwater images to train a generative network adversarially. This training procedure uses unlabeled raw underwater images of a specific survey site, assuming that water column effects are mostly uniform within a local area. This process produces rendered underwater images from in-air RGB-D images that conform to the characteristics of the real underwater data at that site. These synthetic underwater images can then be used to train the second component of our system, a novel color correction network that can compensate for water column effects in a specific location in real-time.

2.3.1 Generating Realistic Underwater Images

WaterGAN is structured as a generative adversarial network, which has two networks training simultaneously: a generator, G, and a discriminator, D (Fig. 2.2). In a standard GAN [7] [27] the generator input is a noise vector z, which is projected, reshaped, and propagated through a series of convolution and deconvolution layers. The output is a synthetic image, $G(z)$. The discriminator receives as input the synthetic images and a separate dataset of real images, x, and classifies each sample as real (1) or synthetic (0). The goal of the generator is to output synthetic images that the discriminator classifies as real. Thus in optimizing G, we seek to

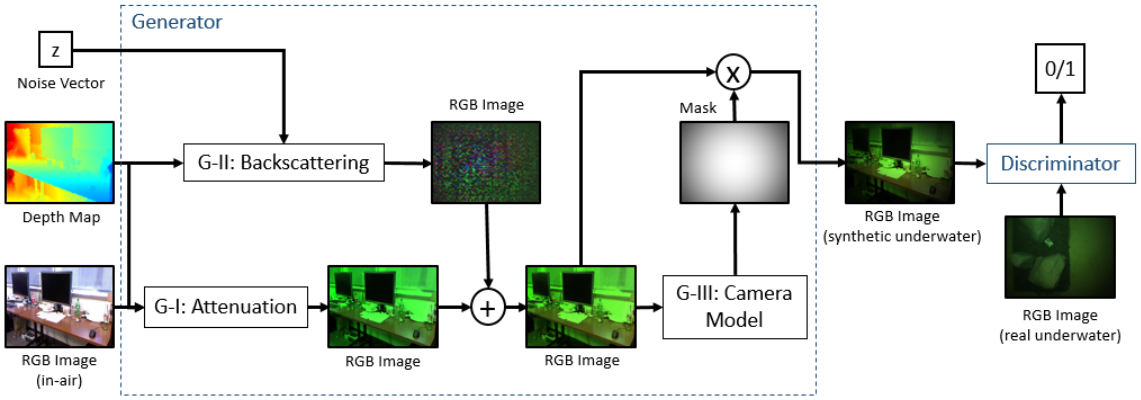


Figure 2.2: WaterGAN: The GAN for generating realistic underwater images with similar image formation properties to those of unlabeled underwater data taken in the field.

maximize

$$log(D(G(z)). \tag{2.1}$$

The goal of the discriminator is to achieve high accuracy in classification, minimizing the above function, and maximizing $D(x)$ for a total value function of

$$log(D(x)) + log(1 - D(G(z))). \tag{2.2}$$

The generator of WaterGAN features three main stages, each modeled after a component of underwater image formation: attenuation (G-I), backscattering (G-II), and the camera model (G-III). The purpose of this structure is to ensure that generated images align with the RGB-D input, such that each stage does not alter the underlying structure of the scene itself, only its relative color and intensity. Additionally, this formulation ensures that the network is using depth information in a realistic manner. This is necessary as the discriminator does not have direct knowledge of the depth of the scene. The remainder of this section describes each stage in detail.

G-I: Attenuation

The first stage of the generator, G-I, accounts for range-dependent attenuation of light. The attenuation model is a simplified formulation of the Jaffe-McGlamery model [8] [24],

$$G_1 = I_{air} e^{-\eta(\lambda) r_c}, \tag{2.3}$$

where I_{air} is the input in-air image, or the initial irradiance before propagation through the water column, r_c is the range from the camera to the scene, and η is the wavelength-dependent attenuation coefficient estimated by the network. The wavelength, λ, is discretized into three color channels. G_1 is the final output of G-I, the final irradiance subject to attenuation in the water column. Note that the attenuation coefficient is dependent on water composition

and quality, and varies across survey sites. To ensure that this stage only attenuates light, as opposed to adding light, and that the coefficient stays within physical bounds, η is constrained to be greater than 0. All input depth maps and images have dimensions of 48×64 for training model parameters. This training resolution is sufficient for the size of our parameter space and preserves the aspect ratio of the full-size images. Note that the generator can still achieve full resolution output for final data generation, as explained below. Depth maps for in-air training data are normalized to the maximum underwater survey altitude expected. Given the limitation of optical sensors underwater, it is reasonable to assume that this value is available.

G-II: Scattering

As a photon of light travels through the water column, it is also subjected to scattering back towards the image sensor. This creates a characteristic haze effect in underwater images and is modeled by

$$B = \beta(\lambda)(1 - e^{-\eta(\lambda)r_c}), \tag{2.4}$$

where β is a scalar parameter dependent on wavelength. Stage G-II accounts for scattering through a shallow convolutional network. To capture range-dependency, a 48×64 depth map is input into the generator, along with a 100-length noise vector. The noise vector is projected, reshaped, and concatenated to the depth map as a single channel 48×64 mask. To capture wavelength-dependent effects, this input is copied for three independent convolution layers with kernel size 5×5. This output is batch normalized and put through a final leaky rectified linear unit (LReLU) with a leak rate of 0.2. Each of the three outputs of the distinct convolution layers are concatenated together to create a $48 \times 64 \times 3$ dimension mask. Since backscattering adds light back to the image, and to ensure that the underlying structure of the imaged scene is not distorted from the RGB-D input, this mask, M_2, is added to the output of G-I:

$$G_2 = G_1 + M_2. \tag{2.5}$$

G-III: Camera Model

Lastly, the generator models vignetting and the sensor response function. Vignetting produces a shading pattern around the borders of an image due to effects from the lens, and it can be modelled by [22]:

$$V = 1 + ar^2 + br^4 + cr^6, \tag{2.6}$$

where r is the normalized radius per pixel from the center of the image, such that $r = 0$ in the center of the image and $r = 1$ at the boundaries. The constants a, b, and c are model parameters estimated by the network. The output mask has dimensions of the input images, and G_2 is multiplied by $M_3 = \frac{1}{V}$ to produce a vignetted image G_3,

$$G_3 = M_3 G_2. \tag{2.7}$$

As described in [22], the parameters can be constrained by:

$$(c \geq 0) \wedge (4b^2 - 12ac < 0). \tag{2.8}$$

Finally, I assume a linear sensor response function, which has a single scaling parameter k [4], with the final output given by

$$G_{out} = kG_3. \tag{2.9}$$

Discriminator

For the discriminator of WaterGAN, I adopt the convolutional network structure used in [27]. The discriminator takes an input image $48 \times 64 \times 3$, real or synthetic. This image is propagated through four convolutional layers with kernel size 5×5 with the image dimension downsampled by a factor of two, and the channel dimension doubled. Each convolutional layer is followed by LReLUs with a leak rate of 0.2. The final layer is a sigmoid function and the discriminator returns a classification label of (0) for synthetic or (1) for a real image.

Generating Image Samples

After training is complete, I use the learned model to generate image samples. For image generation, I input in-air RGB-D data at a resolution of 480×640 and output synthetic underwater images at the same resolution. To maintain resolution and preserve the aspect ratio, the vignetting mask and scattering image are upsampled using bicubic interpolation before applying them to the image. The attenuation model is not specific to the resolution.

2.3.2 Underwater Image Restoration Network

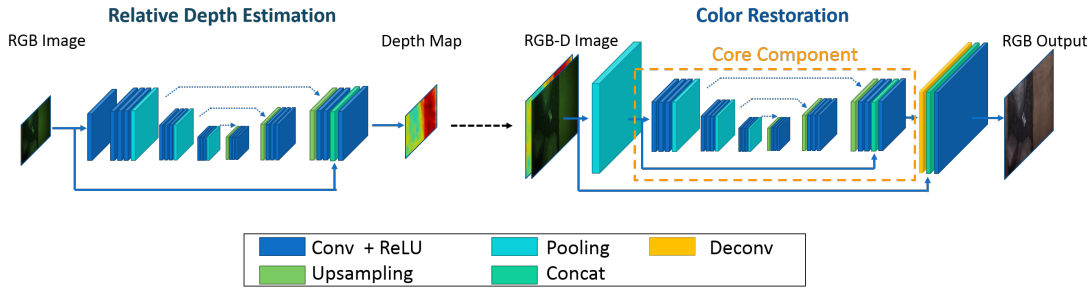


Figure 2.3: Network architecture for color estimation. The first stage of the network takes a synthetic (training) or real (testing) underwater image and learns a relative depth map. The image and depth map are then used as input for the second stage to output a restored color image as it would appear in air.

To achieve real-time monocular image color restoration, we propose a two-stage algorithm using two fully convolutional networks that train on the in-air RGB-D data and corresponding

rendered underwater images generated by WaterGAN. The architecture of the model is depicted in Fig. 2.3. A depth estimation network first reconstructs a coarse relative depth map from the downsampled synthetic underwater image. Then a color restoration network conducts restoration from the input of both the underwater image and its estimated relative depth map.

We propose the basic architecture of both network modules based on a state-of-the-art fully convolutional encoder-decoder architecture for pixel-wise dense learning, SegNet [2]. A new type of non-parametric upsampling layer is proposed in SegNet that directly uses the index information from corresponding max-pooling layers in the encoder. The resulting encoder-decoder network structure has been shown to be more efficient in terms of training time and memory compared to benchmark architectures that achieve similar performance. SegNet was designed for scene segmentation, so preserving high frequency information of the input image is not a required property. In our application of image restoration, however, it is important to preserve the texture level information for the output so that the corrected image can still be processed or utilized in other applications such as 3D reconstruction or object detection. Inspired by recent work on image restoration and denoising using neural networks [23][10], we incorporate skipping layers on the basic encoder-decoder structure to compensate for the loss in high frequency components through the network. The skipping layers are able to increase the convergence speed in network training and to improve the fine scale quality of the restored image, as shown in Fig. 2.6. More discussion will be given in §2.5.

As shown in Fig. 2.3, in the depth estimation network, the encoder consists of 10 convolution layers and three levels of downsampling. The decoder is symmetric to the encoder, using non-parametric upsampling layers. Before the final convolution layer, we concatenate the input layer with the feature layers to provide high resolution information to the last convolution layer. The network takes a downsampled underwater image of $56 \times 56 \times 3$ as input and outputs a relative depth map of $56 \times 56 \times 1$. This map is then upsampled to 480×480 and serves as part of the input to the second stage for color correction.

The color correction network module is similar to the depth estimation network. It takes an input RGB-D image at the resolution of 480×480, padded to 512×512 to avoid edge effects. Although the network module is a fully convolutional network and changing the input resolution does not affect the model size itself, increasing input resolution demands larger computational memory to process the intermediate forward and backward propagation between layers. A resolution of 256×256 would reach the upper bound of such an encoder-decoder network trained on a $12GB$ GPU. To increase the output resolution of our proposed network, we keep the basic network architecture used in the depth estimation stage as the core processing component of our color restoration net, as depicted in Fig. 2.3. Then we wrap the core component with an extra downsampling and upsampling stage. The input image is downsampled using an averaging pooling layer to a resolution of 128×128 and passed through the core process component. At the end of the core component, the output is then upsampled to 512×512 using a deconvolution layer initialized by a bilinear interpolation filter. Two skipping layers are concatenated to preserve high resolution features. In this way, the main intermediate computation is still done in relatively low resolution. We were able to use a batch size of 15 to train the network on a $12GB$ GPU with this resolution. For both the depth estimation and color correction networks, a Euclidean loss function is used. The pixel values in the images are normalized between 0 to

1.

2.4 Experiments & Results

2.4.1 Experimental setup

I evaluate the proposed method using datasets gathered in both a controlled pure water test tank and from real scientific surveys in the field. As input in-air RGB-D for all experiments, I compile four indoor Kinect datasets (B3DO [11], UW RGB-D Object [21], NYU Depth [31] and Microsoft 7-scenes [29]) for a total of 15000 RGB-D images.

2.4.2 Artificial Testbed

The first survey is done using a 4 ft $\times$ 7 ft man-made rock platform submerged in a pure water test tank at University of Michigan's Marine Hydrodynamics Laboratory (MHL). A color board is attached to the platform for reference (Fig. 2.4). A total of over 7000 underwater images are compiled from this survey.

(a) Rock platform (b) Color board (c) MHL test tank

Figure 2.4: (a) An artificial rock platform and (b) a diving color board are used to provide ground truth for controlled imaging tests in (c) a pure water tank to gather the MHL dataset.

2.4.3 Field Tests

One field dataset was collected in Port Royal, Jamaica, at the site of a submerged city containing both natural and man-made structure. These images were collected with a hand-held diver rig. For the experiments, I compile a dataset consisting of 6500 images from a single dive. The maximum depth from the seafloor is approximately $1.5m$. Another field dataset was collected at a coral reef system near Lizard Island, Australia [26]. The data was gathered with the same diver rig and I assumed a maximum depth of $2.0m$ from the seafloor. I compile a total number of 6083 images from the multi-dive survey within a local area.

2.4.4 Network Training

For each dataset, I train the WaterGAN network to model a realistic representation of raw underwater images from a specific survey site. The real samples are input to WaterGAN's discriminator network during training, with an equal number of in-air RGB-D pairings input

to the generator network. I train WaterGAN on a Titan X (Pascal) with a batch size of 64 images and a learning rate of 0.0002. Through experiments, I found 10 epochs to be sufficient to render realistic images for input to the color correction network for the Port Royal and Lizard Island datasets. I trained for 25 epochs for the MHL dataset. Once a model is trained, it can generate an arbitrary amount of synthetic data. For the experiments, we generate a total of 15000 rendered underwater images for each model (MHL, Port Royal, and Lizard Island), which corresponds to the total size of the compiled RGB-D dataset.

Next, this data is used to train the color correction network with the generated images and corresponding in-air RGB-D images. This set is split into a training set with 12000 images and a validation set with 3000 images. The networks are trained from scratch for both the depth estimation network and image restoration network on a Titan X (Pascal) GPU. The depth estimation network is trained for 20 epochs with a batch size of 50, a base learning rate of $1e^{-6}$, and a momentum of 0.9. The color correction network is trained using a two-level training strategy. For the first level, the core component is trained with an input resolution of 128×128, a batch size of 20, and a base learning rate of $1e^{-6}$ for 20 epochs. Then the whole network is trained at a full resolution of 512×512, with the parameters in core components initialized from the first training step. The full resolution model is trained for 10 epochs with a batch size of 15 and a base learning rate of $1e^{-7}$. Results are discussed in §2.5 for all three datasets.

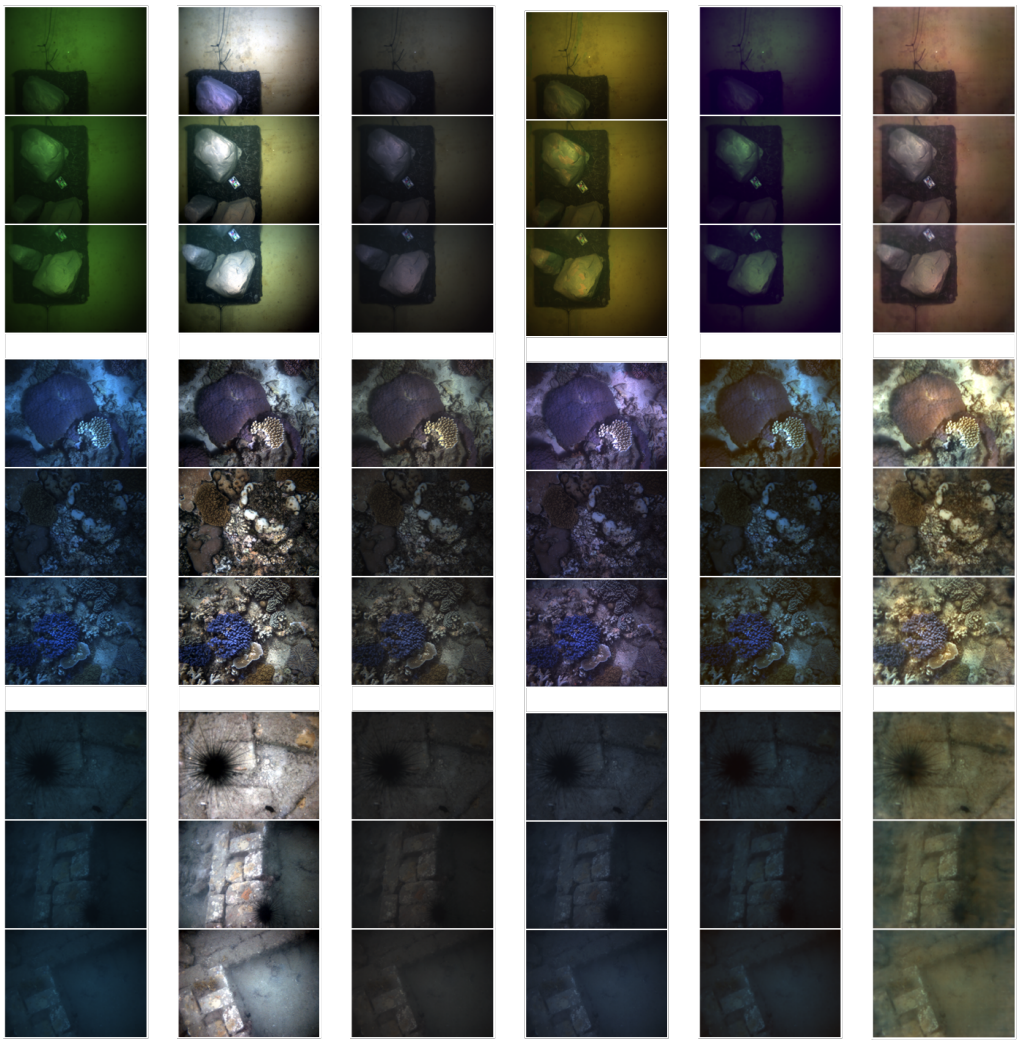
2.5 Results and Discussion

To evaluate the image restoration performance in real underwater data, I present both qualitative and quantitative analysis for each dataset. I compare the developed approach to image processing approaches that are not range-dependent, including histogram equalization and normalization with the gray world assumption. I also compare the results to a range-dependent approach based on a physical model, the modified Jaffe-McGlamery model (Eqn. 3) with ideal attenuation coefficients [15]. Lastly, I compare the proposed method to Shin et al.'s deep learning approach [28], which implicitly models range-dependent information in estimating a transmission map.

Qualitative results are given in Figure 2.5. Histogram equalization looks visually appealing, but it has no knowledge of range-dependent effects so corrected color of the same object viewed from different viewpoints appears with different colors. The proposed method shows more consistent color across varying views, with reduced effects of vignetting and attenuation compared to the other methods. I demonstrate these findings across the full datasets in the following quantitative evaluation.

I present two quantitative metrics for evaluating the performance of our color correction: color accuracy and color consistency. For accuracy, I refer to the color board attached to the submerged rock platform in the MHL dataset. Table 2.1 shows the Euclidean distance of intensity-normalized color in RGB-space for each color patch on the color board compared to an image of the color board in air. These results show that my method has the lowest error for blue, red, and magenta. Histogram equalization has the lowest error for cyan, yellow and green recovery, but my method still outperforms the remaining methods for cyan and yellow.

To analyze color consistency quantitatively, I compute the variance of intensity-normalized



(a) Raw (b) HE (c) GW (d) JM (e) [28] (f) WG

Figure 2.5: Results showing color correction on the MHL, Lizard Island, and Port Royal datasets (from top to bottom). Each column shows (a) raw underwater images, and corrected images using (b) histogram equalization, (c) normalization with the gray world assumption, (d) a modified Jaffe-McGlamery model (Eqn. 3) with ideal attenuation coefficients, (e) Shin et al.'s deep learning approach, and (f) WaterGAN.

pixel color for each scene point that is viewed across multiple images. Table 2.2 shows the mean variance of these points. WaterGAN shows the lowest variance across each color channel. This consistency can also be seen qualitatively in Fig. 2.5.

Table 2.1 Color correction accuracy based on Euclidean distance of intensity-normalized color in RGB-space for each method compared to the ground truth in-air color board.

	Raw	**Hist. Eq.**	**Gray World**	**Mod. J-M**	**Shin[28]**	**Prop. Meth.**
Blue	0.3349	0.2247	0.2678	0.2748	0.1933	**0.1431**
Red	0.2812	0.0695	0.1657	0.2249	0.1946	**0.0484**
Mag.	0.3475	0.1140	0.2020	0.2980	0.1579	**0.0580**
Green	0.3332	**0.1158**	0.1836	0.2209	0.2013	0.2132
Cyan	0.3808	**0.0096**	0.1488	0.3340	0.2216	0.0743
Yellow	0.3599	**0.0431**	0.1102	0.2265	0.2323	0.1033

Table 2.2 Variance of intensity-normalized color of single scene points imaged from different viewpoints.

	Raw	**Hist. Eq.**	**Gray World**	**Mod. J-M**	**Shin[28]**	**Prop. Meth.**
Red	0.0073	0.0029	0.0039	0.0014	0.0019	**0.0005**
Green	0.0011	0.0021	0.0053	0.0019	0.0170	**0.0007**
Blue	0.0093	0.0051	0.0042	0.0027	0.0038	**0.0006**

The trained network is also validated on the testing set of synthetic data and the validation results are given in Table 2.3. RMSE is used as the error metric for both color and depth. These results show that the trained network is able to invert the model encoded by the generator.

Table 2.3 Validation error in pixel value is given in RMSE in RGB-space. Validation error in depth is given in RMSE (m).

Dataset	**Red**	**Green**	**Blue**	**Depth RMSE**
Synth. MHL	0.052	0.033	0.055	0.127
Synth. Port Royal	0.060	0.041	0.031	0.122
Synth. Lizard	0.068	0.045	0.035	0.103

In terms of the computational efficiency, the forward propagation for depth estimation takes $0.007s$ on average and the color correction module takes $0.06s$ on average, which is efficient for real-time applications.

It is important to note that the depth estimation network recovers accurate relative depth, not necessarily absolute depth. This is due to the scale ambiguity inherent to the monocular depth estimation problem. To evaluate the depth estimation in real underwater images, the estimated depth is compared to the depth reconstructed from stereo images available for the MHL dataset in a normalized manner, ignoring the pixels where no depth is recovered from stereo reconstruction due to lack of overlap or feature sparsity. The RMSE of normalized estimated depth and the normalized stereo reconstructed depth is $0.11m$.

To evaluate the improvement in image quality due to skipping layers in the color correction network, the network is trained at the same resolution with and without skipping layers. For the first pass of core component training, the network without skipping layers takes around 30 epochs to reach a stable loss, while the proposed network with skipping layers takes around 15 epochs. The same trend holds for full model training, taking 10 and 5 epochs, respectively. Figure 2.6 shows a comparison of image patches recovered from both versions of the network. This demonstrates that using skipping layers helps to preserve high frequency information from the input image.

(a) Raw image patch

(b) Restored image without skipping layers

(c) Final output

Figure 2.6: Zoomed-in comparison of color correction results of an image with and without skipping layers.

One limitation of the model is in the parameterization of the vignetting model, which assumes a centered vignetting pattern. This is not a valid assumption for the MHL dataset, so the restored images still show some vignetting though it is partially corrected. These results could be improved by adding a parameter that adjusts the center position of the vignetting pattern over the image. This demonstrates a limitation of augmented generators, more generally. Since they are limited by the choice of augmentation functions, augmented generators may not fully capture all aspects of a complex nonlinear model [32]. I introduce a convolutional layer into the augmented generator that is meant to capture scattering, but I would like to experiment with adding additional layers to this stage for capturing more complex effects, such as lighting patterns from sunlight in shallow water surveys. To further increase the network robustness and enable the generalization to more application scenarios, I would also like to train the network across more datasets covering a larger variety of environmental conditions including differing illumination and turbidity.

Source code, sample datasets, and pretrained models are available at https://github.com/kskin/WaterGAN.

2.6 Discussion & Conclusion

This chapter proposed WaterGAN, a generative network for modeling underwater images from RGB-D in air. I showed a novel generator network structure that incorporates the process of underwater image formation to generate high resolution output images. I then presented a dense pixel-wise model learning pipeline for the task of color correction of monocular underwater images trained on RGB-D pairs and corresponding generated images. I evaluated the method on both controlled and field data to show qualitatively and quantitatively that the output is accurate and consistent across varying viewpoints. There are several promising directions for

future work to extend this network. Here I train WaterGAN and the color correction network separately to simplify initial development of our methods. Combining these networks into a single network to allow joint training would be a more elegant approach. Additionally, this would allow the output of the color correction network to directly influence the WaterGAN network, perhaps enabling development of a more descriptive loss function for the specific application of image restoration.

CHAPTER 3

PROPOSED RESEARCH IN LEARNING FOR SIMULTANEOUS COLOR CORRECTION AND DEPTH ESTIMATION OF STEREO UNDERWATER IMAGERY

3.1 Introduction

3.2 Background

3.3 Methodology

3.4 Experiments & Results

3.5 Discussion & Conclusion

CHAPTER 4

PROPOSED RESEARCH IN REAL-TIME UNDERWATER 3D RECONSTRUCTION

4.1 Introduction

4.2 Background

4.2.1 Real-time Dense 3D Reconstruction

Real-time dense 3D reconstruction is a critical perception task for fully autonomous robotic systems. Recently developed RGB-D simultaneous localization and mapping (SLAM) systems have demonstrated the ability to perform this task with a high degree of success for land-based applications [**?**] [20]. The reconstruction framework in this paper is adopted from fusion-based methods, specifically ElasticFusion [35]. Fusion-based methods perform online alignment of overlapping color images and depth maps to maintain and update a single 3D model, while tracking an estimated pose to enable global loop closures. Other fusion-based methods include KinectFusion [25] and Kintinuous, a spatially extended KinectFusion [34]. Each of these methods exploits RGB-D sensors capable of returning high-resolution color images with corresponding depth or range measurements to sense the surrounding environment.

4.2.2 Underwater 3D Reconstruction

Unfortunately, the aqueous medium presents unique challenges to transferring these terrestrial approaches to underwater environments. Complex effects on light propagation limit the effectiveness of active RGB-D sensors subsea and lead to a violation of the brightness constancy constraint (BCC) [19], each of which are commonly used to achieve real-time mapping in terrestrial applications. There have been recent advances in methods that use passive optical sensors for underwater applications. Several offline dense 3D reconstruction techniques have been developed and tested underwater with stereo camera data [12] [13]. Jordt-Sedlazeck et al. developed an underwater light propagation model to account for the effects of refraction through a flat port camera housing in both camera calibration and structure-from-motion [16] [17]. Additionally, Bryson et al. developed methods to perform color correction of images captured in underwater

environments for color consistent reconstructions [4]. While these methods provide a basis for overcoming several fundamental limitations of underwater vision, they are all offline techniques. The goal of this work is to advance towards the online dense 3D reconstruction required for many autonomous navigation and intervention tasks.

4.3 Methodology

4.4 Experiments & Results

4.5 Discussion & Conclusion

CHAPTER 5

CONCLUSIONS

BIBLIOGRAPHY

[1] Darpa urban challenge, 2007. URL http://archive.darpa.mil/grandchallenge/.

[2] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for scene segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):1–1, 2017. ISSN 0162-8828. doi: 10.1109/TPAMI.2016.2644615.

[3] M. Bryson, M. Johnson-Roberson, O. Pizarro, and S. Williams. Colour-Consistent Structure-from-Motion Models using Underwater Imagery. In *Rob: Sci. and Syst.*, pages 33–40, 2012.

[4] M. Bryson, M. Johnson-Roberson, O. Pizarro, and S. B. Williams. True color correction of autonomous underwater vehicle imagery. *J. Field Robotics*, pages 853–874, 2015. ISSN 1556-4967. doi: 10.1002/rob.21638.

[5] N. Carlevaris-Bianco, A. Mohan, and R. M. Eustice. Initial results in underwater single image dehazing. In *Proc. IEEE/MTS OCEANS*, pages 1–8, Seattle, WA, USA, September 2010.

[6] P. Drews, Jr., E. do Nascimento, F. Moraes, S. Botelho, and M. Campos. Transmission estimation in underwater single images. In *Proc. IEEE Int. Conf. on Comp. Vision Workshops*, pages 825–830, June 2013.

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Adv. in Neural Info. Proc. Syst.*, pages 2672–2680. Curran Associates, Inc., 2014. URL http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.

[8] J. S. Jaffe. Computer modeling and the design of optimal underwater imaging systems. *IEEE J. Oceanic Engin.*, 15(2):101–111, 1990. ISSN 0364-9059. doi: 10.1109/48.50695.

[9] J. S. Jaffe. Development of a laser line scan LIDAR imaging system for AUV use. Technical report, University of California, San Diego, 2011.

[10] V. Jain, J. F. Murray, F. Roth, S. Turaga, V. Zhigulin, K. L. Briggman, M. N. Helmstaedter, W. Denk, and H. S. Seung. Supervised learning of image restoration with convolutional networks. In *Proc. IEEE Int. Conf. Comp. Vision*, pages 1–8. IEEE, 2007.

[11] A. Janoch, S. Karayev, Y. Jia, J. T. Barron, M. Fritz, K. Saenko, and T. Darrell. A category-level 3-d object dataset: Putting the kinect to work. In *Proc. IEEE Int. Conf. on Comp. Vision Workshops*, pages 1168–1174, Nov 2011. doi: 10.1109/ICCVW.2011.6130382.

[12] M. Johnson-Roberson, O. Pizarro, S. B. Williams, and I. Mahon. Generation and Visualization of Large-scale Three-dimensional Reconstructions from Underwater Robotic Surveys. *Journal of Field Robotics*, 27(1):21–51, 2010.

[13] M. Johnson-Roberson, M. Bryson, B. Douillard, O. Pizarro, and S. B. Williams. Out-of-core efficient blending for underwater georeferenced textured 3d maps. In *IEEE Int. Conf. Comp. for Geospat. Res. and App.*, pages 8–15, 2013.

[14] M. Johnson-Roberson, M. Bryson, A. Friedman, O. Pizarro, G. Troni, P. Ozog, and J. C. Henderson. High-resolution underwater robotic vision-based mapping and 3d reconstruction for archaeology. *J. Field Robotics*, pages 625–643, 2016.

[15] A. Jordt. *Underwater 3D Reconstruction Based on Physical Models for Refraction and Underwater Light Propagation*. PhD thesis, Kiel University, 2013.

[16] A. Jordt-Sedlazeck and R. Koch. Refractive calibration of underwater cameras. In *Computer Vision, ECCV 2012*, volume 7576, pages 846–859, 2012.

[17] A. Jordt-Sedlazeck and R. Koch. Refractive structure-from-motion on underwater images. In *IEEE International Conference on Computer Vision*, pages 57–64, 2013.

[18] P. D. Jr., E. R. Nascimento, S. S. C. Botelho, and M. F. M. Campos. Underwater depth estimation and image restoration based on single images. *IEEE CG&A*, 36(2):24–35, 2016. doi: 10.1109/MCG.2016.26. URL http://dx.doi.org/10.1109/MCG.2016.26.

[19] P. H. K. Berthold and B. G. Schunck. Determining optical flow. *Artificial Intelligence*, 17.1-3:185–203, 1980.

[20] M. Keller, D. Lefloch, M. Lambers, S. Izadi, T. Weyrich, and A. Kolb. Real-time 3d reconstruction in dynamic scenes using point-based fusion. In *3D Vision-3DV 2013, 2013 International Conference on*, pages 1–8. IEEE, 2013.

[21] K. Lai, L. Bo, and D. Fox. Unsupervised feature learning for 3d scene labeling. In *Proc. IEEE Int. Conf. on Robot. and Automation*, pages 3050–3057, May 2014. doi: 10.1109/ICRA.2014.6907298.

[22] L. Lopez-Fuentes, G. Oliver, and S. Massanet. Revisiting image vignetting correction by constrained minimization of log-intensity entropy. In *Proc. Adv. in Comp. Intell.*, pages 450–463. Springer International Publishing, June 2015.

[23] X. Mao, C. Shen, and Y.-B. Yang. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. In *Advances in Neural Information Processing Systems 29*, pages 2802–2810. 2016. URL http://papers.nips.cc/paper/6172-image-restoration-using-very-deep-convolutional-encoder-decoder-networks-with-symm pdf.

[24] B. L. McGlamery. Computer analysis and simulation of underwater camera system performance. Technical report, UC San Diego, 1975.

[25] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohi, J. Shotton, S. Hodges, and A. Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *Mixed and augmented reality (ISMAR), 2011 10th IEEE international symposium on*, pages 127–136. IEEE, 2011.

[26] O. Pizarro, A. Friedman, M. Bryson, S. B. Williams, and J. Madin. A simple, fast, and repeatable survey method for underwater visual 3d benthic mapping and monitoring. *Ecology and Evolution*, 7(6):1770–1782, 2017. ISSN 2045-7758. doi: 10.1002/ece3.2701. URL http://dx.doi.org/10.1002/ece3.2701.

[27] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[28] Y.-S. Shin, Y. Cho, G. Pandey, and A. Kim. Estimation of ambient light and transmission map with common convolutional architecture. In *Proc. IEEE/MTS OCEANS*, pages 1–7, Monterey, CA, Sept. 2016.

[29] J. Shotton, B. Glocker, C. Zach, S. Izadi, A. Criminisi, and A. Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images. In *Proc. IEEE Conf. on Comp. Vision and Pattern Recognition*, pages 2930–2937, 2013.

[30] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb. Learning from simulated and unsupervised images through adversarial training. *CoRR*, abs/1612.07828, 2016. URL http://arxiv.org/abs/1612.07828.

[31] N. Silberman and R. Fergus. Indoor scene segmentation using a structured light sensor. In *Proc. IEEE Int. Conf. Comp. Vision Workshops*, pages 601–608, Nov 2011. doi: 10.1109/ICCVW.2011.6130298.

[32] L. Sixt, B. Wild, and T. Landgraf. Rendergan: Generating realistic labeled data. *CoRR*, abs/1611.01331, 2016. URL http://arxiv.org/abs/1611.01331.

[33] K. A. Skinner, E. Iscar, and M. Johnson-Roberson. Automatic color correction for 3D reconstruction of underwater scenes. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5140–5147, May 2017. doi: 10.1109/ICRA.2017.7989601.

[34] T. Whelan, M. Kaess, M. Fallon, H. Johannsson, J. Leonard, and J. McDonald. Kintinuous: Spatially extended kinectfusion. 2012.

[35] T. Whelan, S. Leutenegger, R. F. Salas-Moreno, B. Glocker, and A. J. Davison. Elasticfusion: Dense slam without a pose graph. In *Proceedings of Robotics: Science and Systems (RSS)*, 2015.

[36] Z. Zhang. Microsoft kinect sensor and its effect. *MultiMedia, IEEE*, 19(2):4–10, 2012.