A Comparison and Extension of Deep Learning Methods for Semantic Segmentation in the Context of Coral Reef Survey Imaging

by

Andrew King

(Under the Direction of Suchendra M. Bhandarkar)

Abstract

We examine two main classes of deep learning methods, patch-based convolutional neural network (CNN) architectures and fully convolutional neural network (FCNN) approaches, for semantic segmentation and object classification of coral reef survey images. Using image data collected from underwater video of marine environments, we compare five common CNN architectures and observe Resnet152 [1] to achieve the highest accuracy. For our comparison of FCNN approaches, we test three common architectures and one custom modified architecture and observe the best performance with Deeplab v2 [2]. We expand on our initial approaches by proposing a technique that utilizes the multi-view image data commonly extracted, yet often discarded, in video or remote sensing domains. We examine the use of stereoscopic image data for FCNN approaches and multi-view image data for patch-based CNN methods. Our proposed TwinNet architecture is the top performing FCNN. Among patch-based multi-view approaches, our proposed nViewNet-8 architecture yields the highest accuracy on this task.

INDEX WORDS: Deep Learning, Artificial Intelligence, Computer Vision, Machine Learning, Semantic Segmentation, Convolutional Neural Networks, Fully Convolutional Neural Networks, Image Classification, Multi-view Image Data, Coral Reef, Marine Science

A Comparison and Extension of Deep Learning Methods for Semantic Segmentation in the Context of Coral Reef Survey Imaging

by

Andrew King

B.A., Southern Virginia University, 2016

A Thesis Submitted to the Graduate Faculty
of The University of Georgia in Partial Fulfillment
of the
Requirements for the Degree

Master of Science

Athens, Georgia

2018

A Comparison and Extension of Deep Learning Methods for Semantic Segmentation in the Context of Coral Reef Survey Imaging

by

Andrew King

Approved:

Major Professor: Suchendra M. Bhandarkar

Committee: Brian M. Hopkinson
Frederick Maier

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
August 2018

ACKNOWLEDGMENTS

This research was funded in part by a Robotics Research Equipment Grant by the Faculty of Robotics and the Office of Vice President for Research, The University of Georgia, Athens, Georgia, to Dr. Bhandarkar and Dr. Hopkinson.

TABLE OF CONTENTS

List of Figures

List of Tables

CHAPTER 1

INTRODUCTION

In this thesis, we detail a variety of deep learning methods for effectively mapping coral reef ecosystems. Traditional approaches to this task have often been limited either due to the labor-intensive nature of manual mapping by human divers, or in the depth and scope achievable by methods such as aerial photography. Using underwater image data, it has become possible for experts to manually annotate images of the coral reef in order to achieve accurate classification. This process, however, is both time consuming and labor intensive.

It is important to attempt to overcome these barriers to coral reef mapping and classification in order to monitor the health of these marine ecosystems. There is currently a state of marine environmental crisis brought about by major declines in coral reef ecosystems [3]. By improving mapping and monitoring tools that can estimate the abundance of organisms in a given ecosystem, it is possible to track the changes in the health of coral reef environments.

To automate and streamline the annotation task, we first explore the use of convolutional neural network (CNN) architectures. We compare the accuracy of annotation completed using the following CNN architectures: VGG16 [4], InceptionV3 [5], InceptionResNetV2 [6], Resnet50 and Resnet152 [1]. Next, we compare the above CNN architectures to prior methods that have been used for semantic segmentation and object classification of underwater images of marine environments. Among the compared architectures Resnet152 [1] performs the best on this task with 90.03% accuracy, compared to the traditional SVM and texton dictionary approach, which performs with 84.80% accuracy.

We next detail the use of fully convolutional neural networks (FCNNs) on this task. FCNNs generate a class prediction for every pixel in a given image in order to simultaneously perform

object classification and semantic segmentation of the image. We compare FCNN models including FCN8s [7], Dilation8 [8], and DeepLab v2 [2] to DilationMod, a custom modification of the Dilation8 architecture that we designed specifically for this task. The best accuracy is observed from the DeepLab v2 architecture with 67.70% pixelwise accuracy.

The performance results of these models demonstrate that modern deep learning architectures can produce better results than traditional methods for the task of object classification and semantic segmentation in underwater coral reef images.

In the third chapter we describe how these deep learning methods can be extended by utilizing multiple viewpoints to make a more accurate prediction. Frequently, in capturing data for mapping tasks of underwater environments, images of subjects are taken from multiple points of view. We propose approaches that use this information in order to improve the accuracy of both FCNN and patch-based models.

We propose a method that uses stereoscopic image pairs to improve the accuracy of FCNNs on the task of semantic segmentation of coral reef images. Our method uses both left-perspective and right-perspective rectified images to generate a disparity map, which is then added as a fourth channel. Next, we propose the TwinNet architecture, which accepts stereo image pairs as inputs and uses a weight sharing scheme similar to those seen in Siamese networks [9]. Using the left-perspective and right-perspective images, the network is able to learn spatial features rather than relying on hand-engineered features that are provided to the network explicitly. Our TwinNet architecture is able to perform with 66.44% pixelwise accuracy on this task.

To improve the accuracy of patch-based approaches to semantic segmentation, we explore the use of multiple-viewpoint images for single-entity classification. To create a three-dimensional semantic segmentation, we first create a three-dimensional mesh and then perform a classification on each mesh face. To improve the overall accuracy of this classification, we propose different ensemble voting schemes. We also propose the nViewNet architecture, which can receive a variable number of images (with a specified maximum) as inputs and learn a combination of the inputs to

output a single-entity classification. NViewNet outperforms ResNet152 [1] with a top accuracy of 94.26%.

For both FCNN and patch-based approaches, we show that utilizing image data acquired from varying points of view can improve classification accuracy in the semantic segmentation task.

CHAPTER 2

A COMPARISON OF DEEP LEARNING METHODS FOR SEMANTIC SEGMENTATION OF CORAL REEF SURVEY IMAGES[1]

[1] A. King, S. M. Bhandarkar, and B. M. Hopkinson. Submitted to 2018 Computer Vision and Pattern Recognition Workshops, March 26, 2018

ABSTRACT

Two major deep learning methods for semantic segmentation, i.e., patch-based convolutional neural network (CNN) approaches and fully convolutional neural network (FCNN) models, are studied in the context of classification of regions in underwater images of coral reef ecosystems into biologically meaningful categories. For the patch-based CNN approaches, we use image data extracted from underwater video accompanied by individual point-wise ground truth annotations. We show that patch-based CNN methods can outperform a previously proposed approach that uses support vector machine (SVM)-based classifiers in conjunction with texture-based features. We compare the results of five different CNN architectures in our formulation of patch-based CNN methods. The Resnet152 CNN architecture is observed to perform the best on our annotated dataset of underwater coral reef images. We also examine and compare the results of four different FCNN models for semantic segmentation of coral reef images. We develop a tool for fast generation of segmentation maps to serve as ground truth segmentations for our FCNN models. The FCNN architecture Deeplab v2 is observed to yield the best results for semantic segmentation of underwater coral reef images.

2.1 INTRODUCTION

A fundamental issue limiting ecological studies in marine environments, such as coral reefs, is the difficulty of generating accurate and repeatable maps of the underlying ecosystems. Manual *in situ* mapping performed underwater by human divers is extremely time consuming, whereas aerial photography and satellite remote sensing are both severely limited by the fact that seawater absorbs light strongly, thereby limiting monitoring to very shallow marine ecosystems [10]. Acoustic methods are able to map the ocean floor at a large spatial scale, but are not suitable for mapping marine ecosystems at finer spatial scales.

This paper describes our ongoing work on the mapping and monitoring of coral reef ecosystems. Coral reefs provide habitat to a wide diversity of organisms and also substantial economic and

cultural benefits to the several million people who live in adjacent coastal communities [11]. However, coral reefs worldwide are being increasingly threatened by a variety of natural and anthropogenic stressors such as global climate change, ocean acidification, sea level rise, pollutant runoff, sedimentation, and overfishing [12, 13]. These stressors have caused coral reef ecosystems worldwide to suffer from massive, rapid declines over the past three decades, resulting in a state of marine environmental crisis [3]. Given their precarious state, improved mapping and monitoring tools are urgently needed to detect and quantify the changes in coral reef ecosystems at appropriate scales of temporal and spatial resolution.

Traditional reef surveys for mapping, classification, and enumeration of underwater taxa have been performed *in situ* by scuba divers trained in marine ecology. While accurate, *in situ* surveys are time consuming, expensive, and allow only limited coverage of the coral reef. With recent advances in autonomous underwater vehicles (AUVs) equipped with high-resolution cameras, *in situ* surveys are being increasingly replaced by image/video-based robotic surveys. In addition, computer vision, pattern recognition, and machine learning techniques are enabling the generation of detailed, large-scale maps of underwater environments [14]. AUVs traveling systematically through the coral reef environment are able to continuously acquire high-quality images of small portions of the coral reef ecosystem. Using computer vision algorithms, the individual images are then assembled into a large-scale, 3D reconstruction (or map) of the coral reef ecosystem accompanied by semantic classification of the various coral taxa, thereby permitting one to estimate the spatial distribution of these taxa on the coral reef. Figure 3.3 depicts the 3D reconstruction of a coral reef accompanied by the semantic classification of its constituent taxa.

Recent advances in the field of deep learning have resulted in significant progress in image object classification and, more recently, in semantic image segmentation. The advances in deep learning have given researchers in a variety of fields sufficient cause to reexamine traditional methods for image segmentation and object classification to determine if deep learning approaches can indeed improve performance. One such field is coral reef ecology, where several approaches to assessing the ecological state of coral reef ecosystems entail analysis of data on the spatial dis-

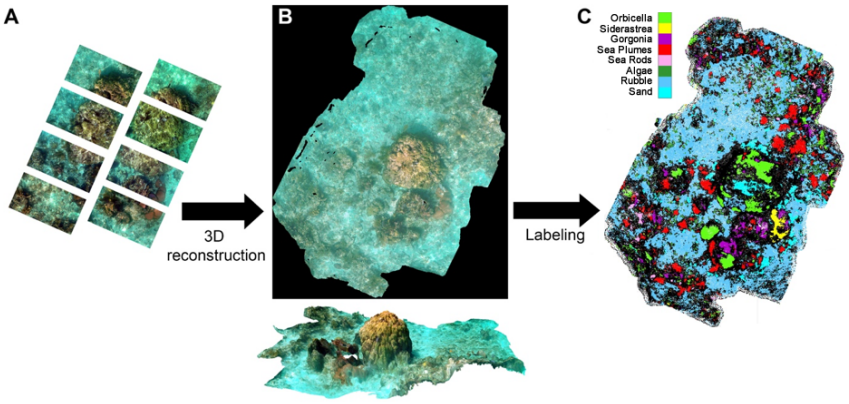


Figure 2.1: 3D reconstruction and annotation of a coral reef ecosystem.

tribution of sessile organisms, including hard corals, soft corals, and algae, and open space for settlement [15, 16]. This data is commonly obtained from underwater images acquired *in situ* by human divers or by autonomous or remotely operated underwater robotic vehicles.

Traditionally, overhead images of coral reef sections are manually annotated by domain experts. During the annotation process, experts are presented with pseudorandomly generated pixel positions in an image and are required to provide a classification label for each of these pixels. Once a large enough pixel sample is collected, it is possible to robustly estimate the abundance of each organism group in the coral ecosystem. A significant shortcoming of this process is that it is labor intensive, which in turn limits the scale and frequency of coral ecosystem assessment.

In this paper, we first examine the annotation task and show how it can be automated using known convolutional neural network (CNN) architectures. We compare the annotation accuracy of known CNN architectures such as VGG16 [4], InceptionV3 [5], InceptionResNetV2 [6], Resnet50 and Resnet152 [1]. We further compare these CNN architectures to previous work in the areas of semantic segmentation and object classification in the context of analysis of underwater coral reef images.

To localize the various coral taxa, we adopt a patch-based CNN approach, which first segments the coral reef images into uniform regions, often using well known algorithms such as simple linear iterative clustering (SLIC) [17] or graph cuts [18]. Patches from each region are then extracted and classified, resulting in a semantic segmentation map of the original image. The patch-based CNN approaches are typically limited by the corresponding segmentation algorithm used when trying to localize organisms within the coral reef.

We also examine fully convolutional neural network (FCNN) models, which are capable of performing simultaneous semantic segmentation and object classification by generating a class prediction for each pixel in an image. We compare the performance of the following FCNN models: FCN8s [7], Dilation8 [8], DeepLab v2 [2], and DilationMod, which is a custom modification of the Dilation8 architecture designed by us for the specific task of semantic segmentation of underwater coral reef images. We show that modern deep learning architectures are indeed capable of outperforming conventional methods for semantic segmentation and object classification in underwater coral reef images.

2.2 Background

2.2.1 Convolutional Neural Networks (CNNs)

Convolutional neural networks (CNNs) have seen enormous success in a wide range of classification tasks. The first CNN architecture that we consider for our implementation of a patch-based approach to semantic image segmentation and object classification is the VGG16 architecture [4]. This architecture was proposed in 2014 by Simonyan and Zisserman [4] of the Visual Geometry Group for the purpose of image classification. The VGG16 architecture represents a significant improvement over previous networks by its use of small 3×3 kernel filters instead of the larger kernel filters common at the time. The VGG16 CNN architecture is comprised of 13 convolutional layers and three fully connected (FC) layers for a total of 16 weight layers. We also consider the InceptionV3 architecture proposed by Szegedy et al. [5]. The InceptionV3 architecture

works to improve upon previous CNN architectures through its defining contribution – the inception module. The inception module tries to approximate an optimal sparse convolutional neural network, allowing the InceptionV3 architecture to deepen (i.e., add layers) while staying within common GPU memory constraints.

As the CNNs grow deeper, the gradient updates become vanishingly small in the upper layers of the network, presenting significant difficulties during the training process. This phenomenon, termed the *vanishing gradient problem*, is addressed by He et al. [1] in their formulation of the ResNet CNN architecture. ResNet makes use of residual blocks that attempt to estimate or fit a residual mapping as opposed to a direct mapping. The ResNet residual blocks make use of a skip connection that passes information directly from the first layer of the block to the last. The intermediate layers then learn a residual from the input layer. This allows the gradient to be preserved across several CNN layers. We consider both the 50-layer ResNet50 architecture and the 152-layer ResNet152 architecture in this paper [1]. Finally, we also consider the Inception-ResNetV2 architecture proposed by Szegedy et al. [6], which combines the Inception architecture with the ResNet residual block architecture.

2.2.2 Fully Convolutional Neural Network (FCNN) Architectures

Among the fully convolutional neural network (FCNN) models for simultaneous semantic image segmentation and object classification, we first consider the FCN8s architecture proposed by Shelhamer et al. [7]. The FCN8s architecture represents the first successful attempt to repurpose an existing CNN architecture designed for image classification for the task of semantic image segmentation. To repurpose a CNN-based classifier for semantic image segmentation, Shelhamer et al. [7] use the existing VGG16 classification architecture [4] as their base model. They eliminate the fully connected CNN layers in the VGG16 architecture, replacing them with 1-by-1 convolution layers with an overall depth equal to the number of classes. This results in an end-to-end trainable model for semantic image segmentation, eliminating the need for separate segmentation and patch-wise classification phases. The FCN8s architecture requires whole-image ground truth

segmentation maps for the purpose of training. The training loss is evaluated by comparing the network output against the ground truth segmentation map. The segmentation map that results from the FCN8s architecture is downsampled to 1/32 of the original size. Simple bilinear interpolation can be used to expand the image, but this results in poor segmentation localization. To address this problem Shelhamer et al. [7] propose a scheme to feed information from previous layers (where the feature maps are larger and hence of higher resolution) and use transposed convolution to upsample the final segmentation map.

Yu and Koltun [8] present a new FCNN architecture termed Dilation8. They base Dilation8 on the FCN8s architecture [7] and improve on its results. They contend that CNN models designed specifically for classification, such as VGG16, need to be rethought for the task of semantic segmentation. Dilation8 removes some of the max pooling layers in VGG16 in order to preserve spatial resolution. Rather than using iteratively larger kernels to maintain a large receptive field, they modify the convolution operator itself as shown in equation (2.1).

$$(F *_l k)(p) = \sum_{s+lt=p} F(s)k(t) \tag{2.1}$$

Yu and Koltun [8] modify the standard equation for discrete convolution where $*$ refers to the convolution operation, F represents a discrete function, and k represents a discrete kernel. Yu and Koltun [8] use parameter l to effectively dilate the convolution kernel by factor l. This means that a one-dilated convolution would be equivalent to standard convolution. The use of dilation allows the receptive field to grow while still maintaining the same number of parameters. Furthermore, Yu and Koltun [8] also implement a context module that is layered after the network. The context module supports an exponential expansion of the receptive field, allowing the network to exploit contextual information at multiple scales. The approach outlined by Yu and Koltun only downsamples the image to 1/8 of its original size, as opposed to 1/32 in the FCN8s architecture proposed by Shelhamer et al. [7].

The final FCNN model that we consider in this paper is Deeplab v2, proposed by Chen et al. [2]. Chen et al. refine previously proposed FCNN models by employing the ResNet [1] as their

base architecture instead of VGG16. Deeplab v2 uses dilated convolution instead of traditional convolution in its Resnet implementation, in a manner similar to Dilation8. Furthermore, Deeplab v2 adds a post-processing step based on a conditional random field (CRF) for refinement of the semantic segmentation map. We compare the performance of the aforementioned FCNN models including one based on a modification of Yu and Koltun's Dilation8 architecture [8] on our dataset of coral reef survey images.

2.2.3 Related Work

Beijbom et al. [3] investigated automated approaches to determine the spatial distribution of the various organisms in a coral reef ecosystem using survey images. They also outlined many of the obstacles unique to this task [3]. They noted the various challenges faced by coral reef image analysis on account of the extreme variations in the size, color, shape, and texture of each of the organism classes (i.e., taxa) and the organic and ambiguous nature of the class boundaries. Furthermore, dramatic changes in water turbidity between sites due to ocean currents and the presence of plankton and algal blooms could greatly alter the ambient lighting and image colors, making the task of automated image analysis even more difficult [3]. Beijbom et al. [3] employed a maximum response filter bank in conjunction with a multiscale patch and texton dictionary based approach to characterize the features in an underwater coral reef image [19]. These features were then input to a support vector machine (SVM) to classify the patches as belonging to the various organism classes.

Treibitz et al. [20] present a wide field-of-view fluorescence imaging system called FluorIS based on a consumer-grade RGB camera that is enhanced for greatly increased sensitivity to chlorophyll-a fluorescence. Images acquired using FluorIS are shown to exhibit high spectral correlation with *in situ* spectrometer measurements. FluorIS is shown to be capable of reliable image acquisition during day and night under varying ambient illumination conditions. In follow-up work, Alonso et al. [21] present a CNN-based scheme for end-to-end semantic segmentation of coral reef

images given sparsely or weakly labeled training data. In particular, they show how augmentation of RGB images with fluorescence data (as done by FluorIS) can be used to generate a dense semantic labeling by fine-tuning an existing encoder-decoder CNN model. However, their scheme is restricted to a binary labeling of images as coral or non-coral in contrast to our work, which entails fine-grained categorization of coral reef surfaces into multiple biological classes.

In this paper, we compare the performance of the approach of Beijbom et al. [3] with that of various deep learning approaches on our coral reef image dataset. We show the superiority of deep learning on coral reef survey images. Given the variance that can occur between different locations as well as over time, we propose that deep CNN-based approaches to semantic image segmentation and object classification are particularly well suited for tasks in this problem domain.

2.3 Evaluation of Patch-Based CNN Approaches

2.3.1 Data Collection

The coral reef underwater image dataset was collected from coral reefs off the Florida Keys by a team of swimmers/divers. An underwater stereo camera rig (GoPro Dual Hero system) was used to collect the underwater video data while swimming over sections of the reef. The rig was carried over the reef in a serpentine pattern in order to capture the entire seafloor for a given region of the coral reef. Images were extracted from the video data at a rate of two frames per second. A subset of the collected images were then annotated by experts to provide ground truth pixel classifications. During the annotation process, an individual pixel in an image is selected in a pseudorandom fashion. The pixel is shown along with its spatial context to an expert who then assigns it to one of the following 10 classes: (1) *Acropora palmata*, (2) *Orbicella spp.*, (3) *Siderastrea siderea*, (4) *Porites astreoides*, (5) *Gorgonia ventalina*, (6) sea plumes, (7) sea rods, (8) algae, (9) rubble, and (10) sand.

The first four classes, i.e., *A. palmata*, *Orbicella spp.*, *Siderastrea siderea*, and *P. astreoides*, represent the different species of coral commonly found on reefs in the Florida Keys. The remaining single-species class, i.e., *Gorgonia ventalina*, represents the common sea fan. The

Table 2.1: Results of the patch-based CNN architectures. SGD refers to the stochastic gradient descent algorithm.

Architecture	Accuracy	Optimizer	Batch Size
SVM and Texton Dict.	84.80		
VGG16	87.34	SGD	32
InceptionResNetV2	84.79	SGD	32
InceptionV3	84.69	SGD	32
Resnet50	88.10	SGD	32
Resnet152	**90.03**	SGD	16

remainder of the classes are multi-species classes or general classes. A total of 9,511 pixels were annotated among the collected 1,807 images. We extracted a square region centered around each annotated pixel to create a dataset of 9,511 classified images.

2.3.2 METHODS

We compare five commonly used CNN architectures known to perform well on patch classification tasks. We compare the performance of well known CNN architectures, such as VGG16 [4], InceptionResNetV2 [6], InceptionV3 [5], Resnet50 and Resnet152 [1], to that of the SVM-based and texton dictionary-based approach proposed by Beijbom et al. [3]. We initialize the aforementioned CNN models using pretrained weights on the Imagenet dataset. The top fully connected layers of the CNNs are removed and replaced with a customized layer, the output of which matches the number of classes under consideration.

We employ a bottleneck approach in which features from the convolutional layers of the network are saved and used to train the top layers of the CNN model before training the entire CNN model. Training the top layers of the CNN ensures that the pretrained weights are not significantly altered via large gradient updates. The newly created top layers have a fully connected layer with ReLU activation functions and dropout followed by a softmax activation layer with 10 units (the

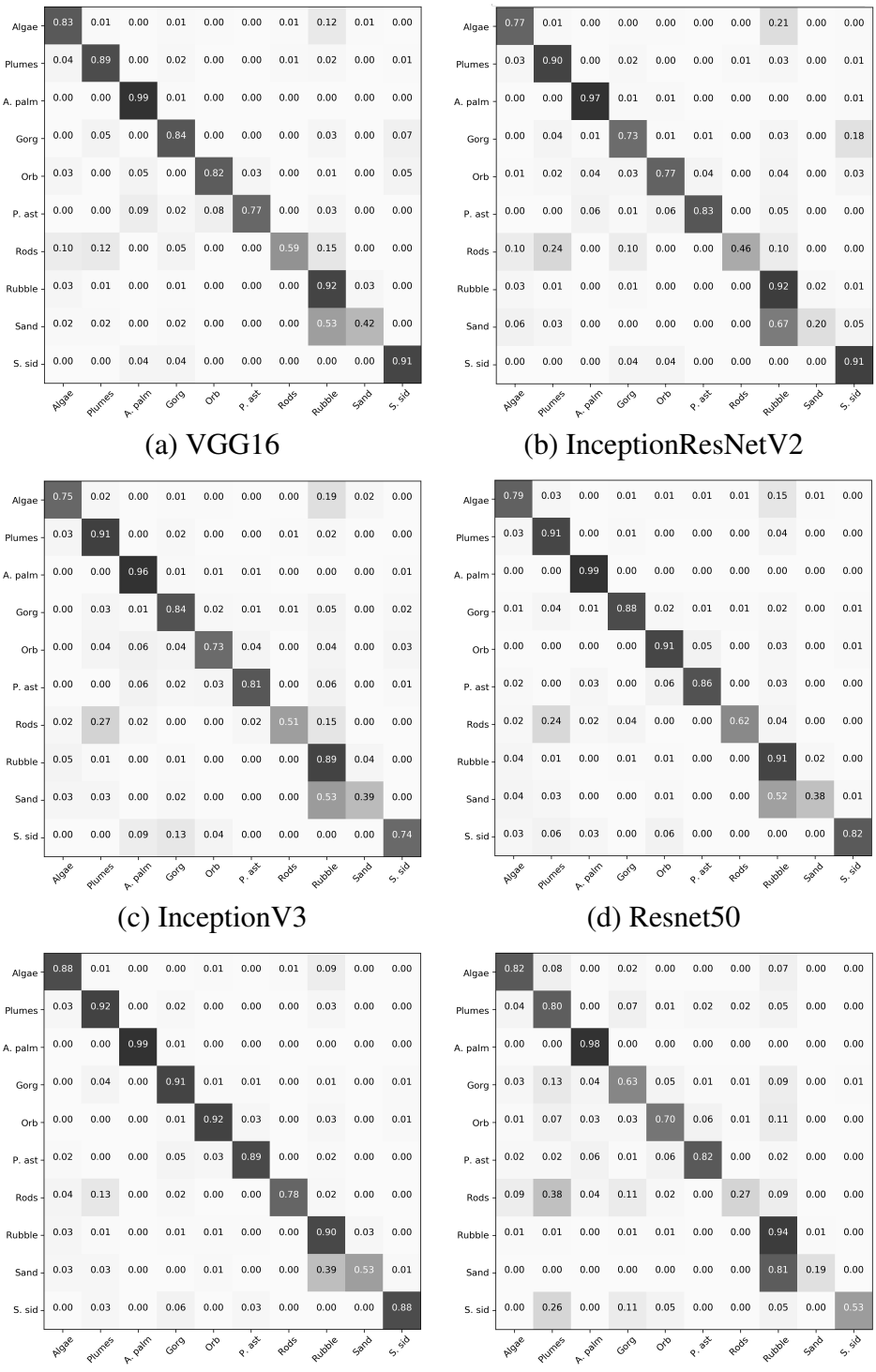
number of classes). We use a batch size of 32 for all of the CNN models except Resnet152 (which requires a smaller batch size of 16) in order to train them using an Nvidia GTX 1080 GPU card. All the CNN models are trained using stochastic gradient descent (SGD) to optimize the pretrained weights. The top layer of each CNN model is trained with a learning rate of 1×10^{-3} and a weight decay rate of 5×10^{-4}, after which the entire network is trained with a learning rate of 1×10^{-4} and weight decay rate of 1×10^{-6}. The networks are trained in increments of 50 epochs until the loss function is no longer observed to be steadily decreasing.

We also replicate the support vector machine (SVM)-based approach of Beijbom et al. [3] and test it on our dataset. We use grid search to optimize the SVM hyperparameters. To ensure experimental validity, we separate our dataset into two sets, a testing set and a training set. We train our models with the training set and then report the model performance on the unseen testing set. The overall accuracy across all classes is reported.

2.3.3 Performance of the CNN Architectures

Table 2.1 summarizes the results of the comparison of the five CNN models that were considered in our study. In general, the performance of the CNNs was quite good with an overall classification accuracy $\approx 85\%$ or higher in all cases. Of the CNNs that were considered, the InceptionV3 [5] was observed to perform the worst, yielding a classification accuracy of 84.69%. Resnet152 [1] was observed to yield the best classification accuracy, outperforming VGG16 [4] and Resnet50 [1] by almost 2%. These results underscore the necessity of formulating deeper CNN architectures, especially when working in this domain.

The confusion matrix for each CNN architecture is presented in Figure 3.4. Most classes are classified with greater than 80% accuracy and several classes exceed 95% accuracy. In all CNN models, there are errors when distinguishing between the classes sand and rubble. These classes share several features in common, and the correct class is in some cases ambiguous. Fortunately, the distinction between these two classes is not of great merit for our ultimate task of determining



(a) VGG16 (b) InceptionResNetV2

(c) InceptionV3 (d) Resnet50

(e) Resnet152 (f) SVM and Texton Dictionary

Figure 2.2: Confusion matrices for various patch-based CNN architectures. We abbreviate *Acropora palmata* as *A. palm*, *Gorgonia ventalina* as *Gorg*, *Orbicella spp.* as *Orb*, *Porites astreoides* as *P. ast*, and *Siderastrea siderea* as *S. sid*.

Table 2.2: Results of the FCNN models. SGD refers to the stochastic gradient descent algorithm.

Architecture	Pixelwise Accuracy	Optimizer	Momentum
FCN8s	50.45	SGD	0.9
Dilation8	62.84	SGD	0.9
DilationMod	64.90	SGD	0.9
DeepLab v2	**67.70**	SGD	0.9

production rates in the reef. All the classes are classified correctly at least a majority of the time among our top performing CNN models.

The SVM-based approach yields an overall accuracy of 84.8% on our dataset, lower than that of our best performing patch-based CNN models. The SVM-based approach also tends to significantly underperform on minority classes such as sea rods, *Siderastrea siderea*, sand, and *Orbicella*.

2.4 Fully Convolutional Neural Network (FCNN) Models

We have shown that patch-based CNNs can estimate the distribution of the various taxa within the coral reefs with greater accuracy than traditional SVM-based approaches. We now focus on fully convolutional neural network (FCNN) models, which represent modifications of the traditional CNNs to provide full semantic segmentation of the input image at the pixel level.

2.4.1 Data Collection

FCNN models for semantic segmentation generally require dense pixelwise ground truth segmentation maps for training purposes. The process of creating ground truth segmentation images for training is often very labor intensive. This is especially true in the case of image data from underwater environments, where corals often contain fine details and image regions are sometimes ambiguous due to poor water clarity. To work around these problems, we created a customized tool to expedite the process of generating ground truth training data. The custom annotation tool

segments a provided image and the user can then annotate the segmented regions with their class labels. Our tool offers two methods of image segmentation: one based on simple linear iterative clustering (SLIC) superpixels [17] and the other based on efficient computation of graph cuts [18]. The program also has a tunable parameter that allows the user to either increase or decrease the level of segmentation, resulting in an oversegmented or undersegmented image. Typically, a user can oversegment the image, annotate its regions, and quickly generate a segmentation map for training purposes. As a user annotates a region, the annotations are propagated to similar regions in its spatial proximity. For instance, if the user annotates a region as sand the tool will automatically propagate the label to other similar regions in its spatial proximity. The tool uses simple RGB histograms and Gabor filter features to measure region similarity and propagates the labels using a k-means clustering algorithm. Finally, the tool offers a manual mode for the user to enter the annotations manually or to correct annotation errors. This tool allowed us to quickly generate 413 dense classification maps for use with our FCNN models [7].

2.4.2 DILATIONMOD

We proposed and tested a modification to the Dilation8 [8] architecture by removing a pooling layer from the Dilation8 architecture. This means that the image is only downsampled to 1/4 of its original size within the network (the downsampling to 1/4 is on account of the remaining two max pooling layers) as opposed to 1/8 in Yu and Koltun's Dilation8 model [8]. The removal of a pooling layer allows the FCNN to preserve the finer details in the input image. This approach requires more memory, but can be accommodated within the memory on an 8GB Nvidia GTX 1080 GPU card when running experiments on our dataset. Furthermore, we introduce dilated convolutions one block earlier in the network (i.e., each convolution layer in the block is dilated by two). Introducing dilated convolution earlier in the network increases the receptive field, counteracting the increase in resolution arising from the removal of a pooling layer. We do not make use of the context module or skip connections. Instead, we upsample the FCNN results using bilinear interpolation. Since we

do not use skip connections or conditional random fields (CRFs) this architecture is very easy to implement.

2.4.3 PREPROCESSING

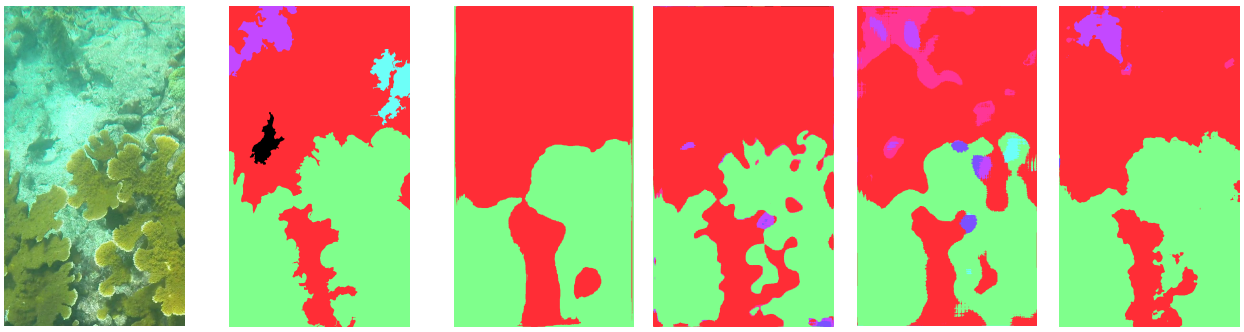
The collected data was preprocessed for use in the FCNN models. Since the images in our dataset are quite large, each image had to be split into four quarters to be used on an Nvidia GTX 1080 GPU with a batch size of one. Since the ground truth segmentation images generated by our tool were in full color, they had to be converted so that each color channel value corresponded to the class label number at that pixel in the image. Since our dataset has 10 classes, the preprocessing outputs images with values 0-9 in their respective color channels. To normalize our data, we subtract the mean RGB value of the training set from each image before passing it to the FCNN.

2.4.4 TRAINING THE FCNN MODELS

We compare the performance of FCN8s [7], Dilation8[8], DeepLab v2 [2], and our modified version of the Dilation8 frontend (i.e., DilationMod) on the task of semantic segmentation of underwater coral reef images. The FCNN weights are initialized using the Imagenet pretrained weights. To retain the benefit of the pretraining, our FCNN models freeze the pretrained weights and train on any additional layers initially with a learning rate of 1×10^{-3}. We use a batch size of one and stochastic gradient descent with a Nesterov momentum term as our optimization technique. We then train the entire model using a learning rate of 1×10^{-4} and weight decay of 1×10^{-6}. Each FCNN model trains for 7,000 iterations to ensure convergence, and the FCNN model with the highest validation accuracy is selected.

2.4.5 PERFORMANCE OF THE FCNN MODELS

Table 2.2 summarizes the results of our comparison of the aforementioned four models. We report the pixelwise accuracy to compare the four methods. Corals contain fine details and consequently



(a) Original (b) Ground Truth (c) FCN8s (d) Dilation8 (e) DilationMod (f) DeepLab v2

Figure 2.3: Outputs of multiple FCNN architectures for a given sample image.

the corresponding image regions are often very thin. Because of this, coral reef semantic segmentation is far more sensitive to downsampling than many other semantic segmentation tasks. The least accurate architecture is FCN8s [7], which only has an accuracy of 50.45%. This result is not unexpected given the downsampling that occurs in the network. While the model makes use of transposed convolution to upsample the image, it cannot adequately recover the fine details required for this task. Dilation8 [8] reports far higher accuracy at 62.84%. Our modified Dilation8 network gives a modest boost to accuracy over the previous two methods, with an overall accuracy of 64.9%. Deeplab v2 [2] is the best performing model on our dataset with an accuracy of 67.7%.

We present the semantic segmentation results of the various FCNN architectures for one of our validation images in Figure 2.3. There is a noticeable disparity between the level of detail preserved by FCN8s and the models that make use of dilated convolution. This is also reflected in the activation maps for each class on this image.

2.5 Conclusion

In this paper, we have shown the effectiveness of deep learning approaches for semantic segmentation of coral reef survey images. This research serves to automate the process of determining the

distribution of organisms and substrates on coral reefs. We have detailed and contrasted two main classes of semantic segmentation based on patch-based CNN models and FCNN models.

We first compared standard CNN architectures for patch-based classification from individual point-based ground truth annotations of training images. The patch-based classification methods can be used for the common task of determining the abundance or paucity of organisms on reefs by leveraging existing segmentation techniques and performing patch-wise classification of each resulting segment. Our best performing CNN model for this task was the ResNet152 [1] architecture, which yielded an accuracy of 90.03%. The previous work of Beijbom et al. [3] using SVMs and texton dictionaries yielded an accuracy of 84.8% on our dataset for this task.

It is important to note that the granularity of classification is much coarser with a patch-based CNN model since it provides a single class label for an entire patch within an image, whereas the FCNN models provide a classification for each individual pixel within an image. The patch-based CNN approaches yield a higher classification accuracy overall. They are, however, limited by the corresponding segmentation algorithm when attempting to localize specific taxa within the coral reef image. Long et al. [7] addressed this tradeoff when proposing the FCN8s architecture, stating that semantic segmentation poses an inherent dilemma between semantics and location in that global information resolves the question of identity, i.e., *what*, whereas local information resolves *where*.

Next, we examined FCNN models, which perform simultaneous segmentation and classification by providing a class prediction at each pixel within an image. We compared four different FCNN models, the best performing of which was the Deeplab v2 architecture, yielding an accuracy of 67.7% on our dense classification dataset. Unlike patch-based CNN approaches, FCNN models do not pose limitations on localization accuracy. Due to the fine granularity of classification, however, the classification accuracy in our tests was below that of the patch-based CNN approaches.

2.6 Future Work and Applications

Since our image data is collected in a serpentine fashion, often from multiple angles so as to capture the entire seafloor, we are able to create semantic maps of entire regions of the coral reef. To create two-dimensional semantic maps of the coral reef regions, each new image can be registered with the result of all previously registered images until all images from a region are processed/registered. The resulting mosaicked image can then be segmented into superpixels. Patches can be extracted from each superpixel and classified using a patch-based CNN architecture. In the case of the FCNN models, the transformation matrices of each image registration can be saved and can then be applied to the corresponding FCNN output for that image. This will result in a mosaicked semantic map for the entire coral reef region.

Currently, we are examining photogrammetric techniques to create a three-dimensional mesh of coral reef regions. We classify mesh faces using the patch-based CNN approaches. The FCNN models presented in this paper use VGG16 [4] as a base architecture that is further enhanced or modified. Future extensions of this work could include applying similar modifications to other network architectures, such as Resnet152 [1]. Finally, since the image data was collected with stereo cameras, future work could look at incorporating disparity information as a channel in the input image. Additionally, deep learning architectures could be developed for leveraging multiple viewpoints to improve classification.

Acknowledgment: This research was funded in part by a Robotics Research Equipment Grant by the Faculty of Robotics and the Office of Vice President for Research, The University of Georgia, Athens, Georgia, to Dr. Bhandarkar and Dr. Hopkinson.

2.7 References

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

[2] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[3] O. Beijbom, P. J. Edmunds, D. I. Kline, B. G. Mitchell, and D. Kriegman, "Automated annotation of coral reef survey images," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Providence, Rhode Island, Jan. 1, 2012.

[4] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[5] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.

[6] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning.," in *AAAI*, vol. 4, 2017, p. 12.

[7] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[8] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *ICLR*, 2016.

[10] J. D. Hedley, C. M. Roelfsema, I. Chollett, A. R. Harborne, S. F. Heron, S. Weeks, W. J. Skirving, A. E. Strong, C. M. Eakin, T. R. Christensen, *et al.*, "Remote sensing of coral reefs for monitoring and management: A review," *Remote Sensing*, vol. 8, no. 2, p. 118, 2016.

[11] L. Burke, K. Reytar, M. Spalding, and A. Perry, *Reefs at risk revisited*. 2011.

[12] K. R. Anthony, "Coral reefs under climate change and ocean acidification: Challenges and opportunities for management and policy," *Annual Review of Environment and Resources*, vol. 41, 2016.

[13] O. Hoegh-Guldberg, P. J. Mumby, A. J. Hooten, R. S. Steneck, P. Greenfield, E. Gomez, C. D. Harvell, P. F. Sale, A. J. Edwards, K. Caldeira, *et al.*, "Coral reefs under rapid climate change and ocean acidification," *science*, vol. 318, no. 5857, pp. 1737–1742, 2007.

[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 three-dimensional reconstruction for archaeology," *Journal of Field Robotics*, vol. 34, no. 4, pp. 625–643, 2017.

[15] R. Ruzicka, M. Colella, J. Porter, J. Morrison, J. Kidney, V. Brinkhuis, K. Lunz, K. Macaulay, L. Bartlett, M. Meyers, *et al.*, "Temporal changes in benthic assemblages on florida keys reefs 11 years after the 1997/1998 el niño," *Marine Ecology Progress Series*, vol. 489, pp. 125–141, 2013.

[16] J. E. Smith, R. Brainard, A. Carter, S. Grillo, C. Edwards, J. Harris, L. Lewis, D. Obura, F. Rohwer, E. Sala, *et al.*, "Re-evaluating the health of coral reef communities: Baselines and evidence for human impacts across the central pacific," *Proc. R. Soc. B*, vol. 283, no. 1822, p. 20 151 985, 2016.

[17] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.

[18] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *Int. J. Comput. Vision*, vol. 59, no. 2, pp. 167–181, Sep. 2004, ISSN: 0920-5691. DOI: 10.1023/B:VISI.0000022288.19776.77.

[19] M. Varma and A. Zisserman, "A statistical approach to texture classification from single images," *International Journal of Computer Vision*, vol. 62, no. 1–2, pp. 61–81, 2005.

[20] T. Treibitz, B. P. Neal, D. I. Kline, O. Beijbom, P. L. Roberts, B. G. Mitchell, and D. Kriegman, "Wide field-of-view fluorescence imaging of coral reefs," *Scientific reports*, vol. 5, p. 7694, 2015.

[21] I. Alonso, A. Cambra, A. Munoz, T. Treibitz, and A. C. Murillo, "Coral-segmentation: Training dense labeling models with sparse ground truth," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2874–2882.

CHAPTER 3

IMPROVING CLASSIFICATION ACCURACY IN DEEP LEARNING SEMANTIC SEGMENTATION MODELS WITH MULTI-VIEW INFORMATION [1]

[1] A. King, S. M. Bhandarkar, and B. M. Hopkinson. To be submitted to 2019 IEEE Winter Conference on Applications of Computer Vision

ABSTRACT

Convolutional neural networks (CNNs) are typically used to make a single classification from a single image or, in the case of fully convolutional neural networks (FCNNs), to generate a semantically segmented output. Often, however, in mapping tasks one may obtain multiple images of a point of interest from different vantage points, as is the case in both stereoscopic image collection and video surveys. In the following, we propose and compare architectures capable of utilizing information from multiple viewpoints to improve classification accuracy of semantic segmentation models. We examine two major classes of architectures. First, we look at extending FCNNs for stereoscopic information. Second, we examine patch-based approaches with multi-view CNNs.

The top-performing fully convolutional approach is our proposed TwinNet architecture, which performs comparably with its baseline architecture, Dilation8 [8], when run only with a left-perspective image, but markedly improves over Dilation8 when run with a stereo pair of images. The top performing patch-based approach is our proposed nViewNet-8 architecture, which outperforms its single-image ResNet152 [1] baseline architecture by 8.72%.

3.1 INTRODUCTION

Modern deep learning approaches to semantic segmentation typically fall into one of two categories. The first major category is fully convolutional neural networks (FCNNs), which segment and classify on a per-pixel basis in one end-to-end trainable network. Second are patch-based approaches that classify existing segments. When semantic segmentation is used in mapping tasks, such as in remote sensing domains or underwater imaging, the images of the underlying objects are often captured from many points of view. In typical approaches to semantic segmentation, only a single viewpoint is utilized to make a classification. The pipeline is typically as follows: images are collected, cleaned, and registered, and then a composite of the images is generated to create a full map of a region. Further analysis is done on the composite map, which has discarded all multi-view information. In this work we propose methods for utilizing this often discarded information with the aim of further improving model accuracy for classification and semantic segmentation.

For FCNN approaches to semantic segmentation we examine utilizing stereoscopic image pairs. We propose and detail a method that generates a disparity map from the left- and right-perspective rectified images. The disparity map is added as a fourth channel, in addition to the three color channels, to the input images to guide the semantic segmentation with three-dimensional disparity information. We then propose the TwinNet architecture, roughly based on siamese networks, which accepts both the left- and right-perspective images as inputs. From these stereo images the network can learn disparity measures or any other spatial features that may be useful in the classification task.

For patch-based approaches to semantic segmentation we examine utilizing a variable number of viewpoints to make a single-entity classification. We create a three-dimensional mesh and perform classification on each mesh face to generate a three-dimensional semantic segmentation. We propose using different voting schemes to improve classification accuracy. Furthermore we propose the nViewNet architecture, which is capable of receiving a variable number of images (with a set maximum number) and learning a combination to ultimately yield a single-entity classification.

We study this problem in the context of coral reef ecology, a field which is often limited by the difficulty inherent in creating accurate maps of diverse marine ecosystems. This work is important, however, because coral reefs across the globe are facing increasing threats, from both natural and anthropogenic stressors. These stressors, which include climate change, ocean acidification, sea level rise, pollutant runoff, and overfishing [12, 13] have combined during the last three decades to cause rapid declines in coral reef ecosystems, resulting in a state of marine environmental crisis [3]. Due to the precarious state of these coral reef ecosystems, advancements in mapping and monitoring technologies are urgently needed to detect and quantify the changes in coral reef ecosystems at appropriate scales of temporal and spatial resolution.

3.2 Background and Related Work

In recent years convolutional neural networks (CNNs) have continued to push the accuracy of image classification models and command large scale image classification tasks [1, 22, 23]. CNNs

have grown deeper over time, which has allowed them to learn more complex patterns, but also presents new difficulties as the gradient updates become smaller and smaller. This phenomenon, known as the *vanishing gradient problem*, is addressed by He et al. [1] in their creation of the ResNet architecture. Our previous work in this problem domain established that the ResNet152 architecture performed particularly well on this classification task [24]. We adopt it as the baseline architecture for many of the proposed models in this work. ResNet makes use of residual convolution blocks that attempt to fit a mapping of the residual as opposed to a direct mapping. This theoretically allows the network to deepen without the gradients becoming vanishingly small.

In our fully convolutional neural network approaches we compare the performance of FCN8s [7] and Dilation8 [8] with and without a disparity channel and further use Dilation8 as a base architecture for TwinNet. In their work on FCN8s, Shelhamer et al. [7] repurpose the VGG16 architecture, intended for classification, for semantic image segmentation. They eliminate the fully connected CNN layers in the VGG16 architecture, replacing them with convolution layers, and make use of transposed convolution to upsample the output. This results in an end-to-end trainable model for semantic image segmentation, eliminating the need for separate segmentation and patch-wise classification phases. The FCN8s architecture requires whole-image ground truth segmentation maps for the purpose of training. The training loss is evaluated by comparing the network output against the ground truth segmentation map.

Yu and Koltun [8] present the Dilation8 architecture for semantic segmentation. They base Dilation8 on the FCN8s architecture [7], but make modifications to further improve the accuracy. Dilation8 removes many of the max pooling layers in the VGG16 base of FCN8s, which means it does not have to rely so heavily on transposed convolution to upsample. Rather than using iteratively larger kernels to maintain a large receptive field, Dilation8 effectively dilates the convolution kernel. Since the kernels still have the same number of parameters, the network maintains a similar amount of computational requirements.

We make use of weight sharing schemes similar to those seen in siamese networks and MVCNN. Siamese networks learn a similarity function between two inputs rather than a simple

classification. To do this, they make use of weight sharing, in which the inputs are both fed through the same network with the same learned weights. They make use of contrastive loss to compare the similarity. We draw on this general idea in our work on TwinNet and nViewNet, which take more than one image as input and shares weights for the initial base architecture.

Su et al. [25] looked at using classification networks for three-dimensional shape recognition in their work on MVCNN. They proposed a network that took inputs from an array of 12 equidistant cameras and pooled the views using an element-wise maximum operation. They showed that when multiple views were pooled in this manner, accuracy increased over single view networks when attempting to classify an image from its shape. We relax the constraints of this network setup to N views from randomly-placed cameras.

3.3 Underwater Stereoscopic Coral Reef Survey of the Florida Keys Image Bank

Our image bank was collected underwater from coral reefs off the Florida Keys by a team of swimmers/divers. An underwater stereo camera rig (GoPro Dual Hero system) was used to collect the underwater video data while swimming over sections of the reef. The rig was carried over the reef in a serpentine pattern in order to capture a complete section of seafloor. Stereo pairs were extracted from the video data at a rate of two frames per second. The resulting 2,391 stereo pairs make up the Underwater Stereoscopic Coral Reef Survey of the Florida Keys image bank (USCSF) and is used for the experiments in this work. This work was conducted under permits from the Florida Keys National Marine Sanctuary (FKNMS-2016-042, FKNMS-2017-035).

Table 3.1: Results of the FCNN stereo and disparity architectures.

Architecture	Pixelwise Accuracy	Input Channels
FCN8s	50.45	3 Color
Dilation8	62.84	3 Color
FCN8s	53.82	3 Color + Disparity
Dilation8	64.02	3 Color + Disparity
TwinNet-LeftOnly	61.93	
TwinNet	**66.44**	

3.4 Extending Fully Convolutional Neural Networks for Use with Stereoscopic Information

3.4.1 Data Collection

To begin the data collection process for FCNN models, it is necessary to create dense pixelwise ground truth segmentation maps for use in training the models. Since our data is stereoscopic, collecting both a left-perspective and right-perspective image, we create ground truths only on the left-perspective images. The creation of these ground truth segmentations can be a time-consuming process. This is particularly true in regard to underwater image data, which can be obscured by poor water clarity. In order to streamline the process of creating ground truth segmentation images, we created a customized tool.

Our annotation tool provides two image segmentation methods: simple linear iterative clustering (SLIC) superpixels and graph cuts. The tool segments images, allows users to annotate regions with class labels, has a tunable parameter to over- or undersegment, and offers a mode for manual annotation. A user can quickly generate segmentation maps upon segmenting and annotating the regions of an image. Our segmentation tool uses RGB histograms and Gabor filter features to measure region similarity and propagates the labels using k-means clustering.

Using this tool, we were able to quickly generate 413 dense classification maps for use with our FCNN models [7]. Our ground truth semantic segmentations classify each pixel into one of the following 10 classes: (1) *Acropora palmata*, (2) *Orbicella spp.*, (3) *Siderastrea siderea*, (4) *Porites astreoides*, (5) *Gorgonia ventalina*, (6) sea plumes, (7) sea rods, (8) algae, (9) rubble, and (10) sand. Furthermore, we employ an *ignore* class for regions that do not fall into one of these categories (such as fish) or for regions that are unclassifiable by an expert. The *ignore* class does not contribute to the loss calculations and is therefore never a classification made by our networks. Additionally, those regions are not used in calculating the accuracy on the validation set.

The first four classes, i.e., *A. palmata*, *Orbicella spp.*, *Siderastrea siderea*, and *P. astreoides*, represent the different species of coral commonly found on reefs in the Florida Keys. The remaining single-species class, i.e., *Gorgonia ventalina*, represents the common sea fan. The remainder of the classes are multi-species classes or general classes.

3.4.2 Preprocessing

Upon collection, the data was preprocessed for use in the FCNN models. Due to the large size of each image (2.7k) in the dataset, it was necessary to split each into four quarters to be used on an Nvidia GTX 1080 GPU with a batch size of one. Because the dataset contains 10 classes, the preprocessing outputs images with values 0-9 in their respective color channels. Finally, we subtract the mean color value of the training set from each image in order to normalize the data before passing it to the FCNN.

3.4.3 FCNN with Disparity Channel

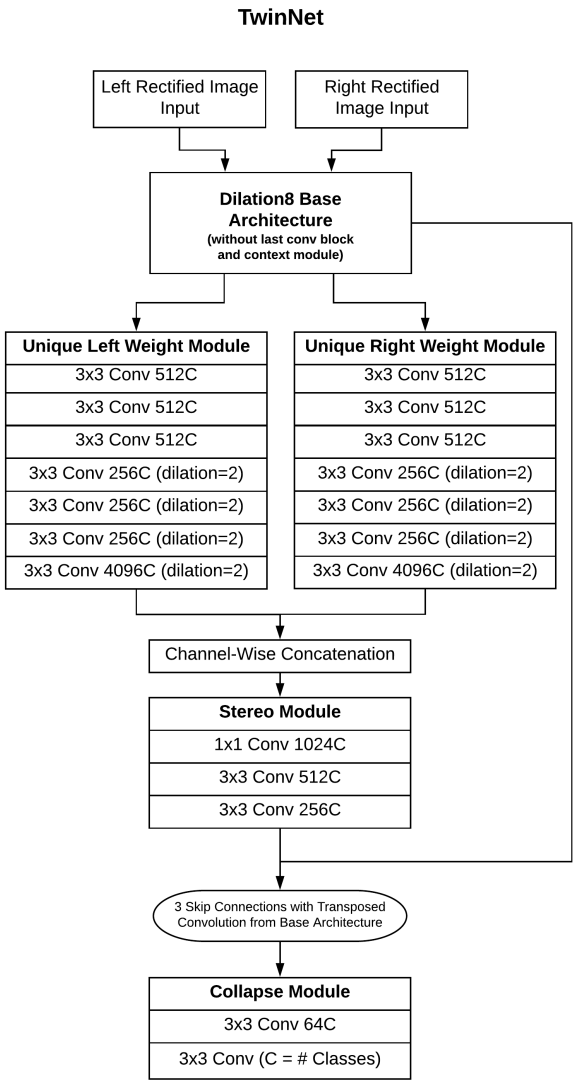
We first examine the use of stereoscopic disparity as a means to leverage multi-view information in fully convolutional neural networks. The images are first rectified with the parameters of our calibrated cameras. We create a disparity map using the semi-Global block matching disparity estimation algorithm proposed by Hirschmuller [26]. We use a uniqueness threshold and block size of 15, a contrast threshold of 0.5 and a disparity range of 64. After a disparity map is created, we inpaint

missing regions on the disparity map using the technique proposed by Telea [27]. The resulting disparity map is then concatenated as a fourth channel on the left-perspective image before it is passed into the fully convolutional neural network. We compare the performance of a standard three-channel input to this four-channel RGB plus disparity input on two models FCN8s [7] and Dilation8 [8].

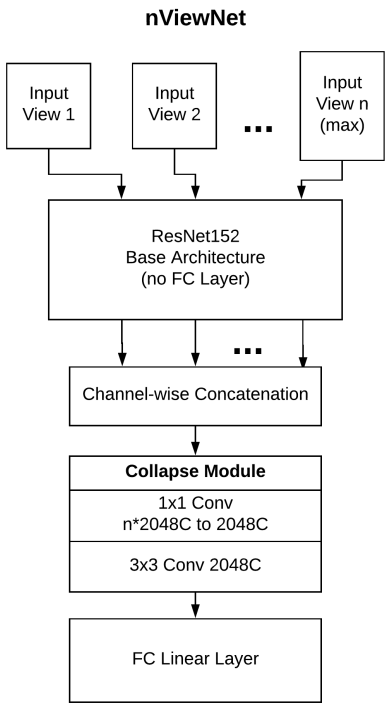
To ensure experimental validity, we separate our dataset into two sets, a training set and a testing set with a 80-20 split, respectively. We train our models with the training set and then report the model performance on the unseen testing set. The overall pixelwise accuracy across all classes is reported. Since Imagenet [22] pretrained weights do not make use of a disparity channel, we train each model from scratch. We train initially with a relatively high learning rate of 1×10^{-3} to quickly learn some initial weights. We use a batch size of one and stochastic gradient descent with a Nesterov momentum term as our optimization technique. We then train the model using a learning rate of 1×10^{-4} and weight decay of 1×10^{-6} to refine the initial learning. Each FCNN model trains for 15,000 iterations to ensure convergence, and the FCNN model with the highest validation accuracy is selected.

3.4.4 TwinNet: Stereo FCNN

Rather than algorithmically deriving features from the relationship between the two stereo images such as is the case with the disparity FCNN, we seek to create an architecture that will learn the best relationship for use in classification. We draw inspiration from the weight sharing schemes in siamese networks [9] and MNCNN [25]. Our base architecture is drawn from the work of Yu and Koltun [8] and their Dilation8 frontend. This base architecture in turn is derived from VGG-16 [4], but utilizes dilated convolutions and less max pooling. The left- and right-perspective images are both fed through the base architecture, and the weights are shared at this point in the network. The left and right outputs are then fed to our stereo module, which allows them to learn weights of their own. The stereo module consists of three convolution layers. Each perspective's stereo module consists of three convolution layers and RELU activations with a two-dilated kernel size of three.



(a) TwinNet Architecture



(b) nViewNet Architecture

Figure 3.1: The TwinNet and nViewNet architectures proposed in this work. Conv stands for convolution layer, C stands for channels, FC stands for fully connected.

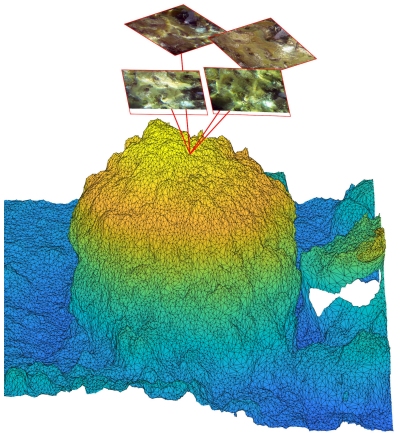


Figure 3.2: A visualization of reprojection from a mesh to multiple images.

The separated outputs of the stereo module are then concatenated on the channel axis and are fed to a collapse module, which uses a convolution layer with a kernel size of one to reduce the number of channels to the total number of classes. At this point, the image is upsampled iteratively through transposed convolution and skip connections until it is returned to its original size (see Figure 3.1 for visualization).

We compare performance of our proposed architecture, TwinNet, to its base architecture, Dilation8. We further compare TwinNet to itself if run with only the left-perspective input. As before, we train our models with the training set and then report the model performance on the unseen testing set. The overall pixelwise accuracy across all classes is reported.

The base architecture weights are initialized using the Imagenet [22] pretrained weights. To retain the benefit of the pretraining, we freeze the base architecture weights and train the additional modules initially with a learning rate of 1×10^{-3}. We use a batch size of one and stochastic gradient descent with a Nesterov momentum term as our optimization technique. We then train the entire model using a learning rate of 1×10^{-4} and weight decay of 1×10^{-6}. Each Stereo FCNN model

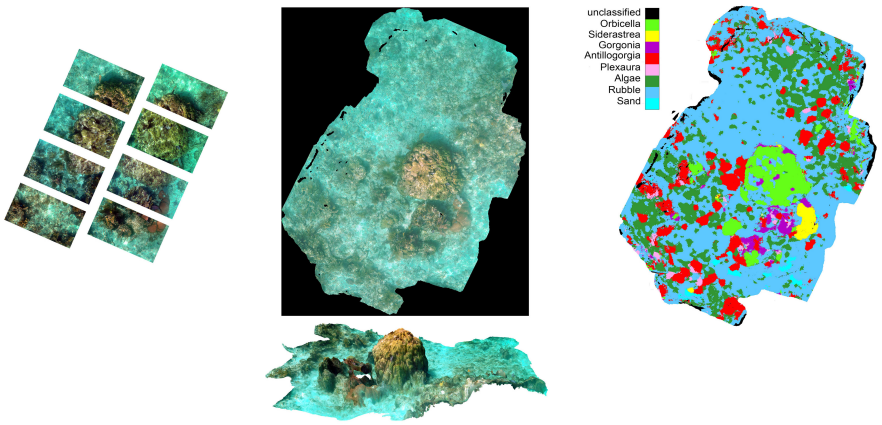


Figure 3.3: Pipeline for 3D reconstruction and annotation of a coral reef ecosystem.

trains for 7,000 iterations to ensure convergence, and the FCNN model with the highest validation accuracy is selected.

3.4.5 Performance of the FCNN Architectures

Table 3.1 summarizes the results of our experiments with extending fully convolutional networks for stereoscopic information. Our disparity FCNNs each give a small boost to accuracy over their corresponding architecture that only utilizes the three color channels. In FCN8s, we see a 3.37% increase in accuracy and in Dilation8 [8] we see 1.18% improvement. This shows that the disparity information may provide at least some benefit to the network in classification. Our TwinNet architecture performs comparably with Dilation8 when run with only the left-perspective image, but has marked improvement over Dilation8 when run with both the left-perspective and right-perspective images. It is therefore reasonable to conclude that the intermediate layers of the model are learning better features from the combination of the two images than are present in only a single image. Furthermore, the accuracy improves over our hand-engineered three color and disparity inputs.

3.5 Extending Patch-Based Approaches to Semantic Segmentation for Multiple Viewpoints

3.5.1 Data Collection

A subset of the collected images from our image bank (USCSF) were annotated by experts to provide ground truth pixel classifications. During the annotation process, an individual pixel in an image is selected in a pseudorandom fashion. The pixel is shown along with its spatial context to an expert who then assigns it to one of the following 10 classes: (1) *Acropora palmata*, (2) *Orbicella spp.*, (3) *Siderastrea siderea*, (4) *Porites astreoides*, (5) *Gorgonia ventalina*, (6) sea plumes, (7) sea rods, (8) algae, (9) rubble, and (10) sand.

We use a photogrammetric processing tool (Agisoft Photoscan) to generate a three-dimensional reconstruction of the underlying coral reef from our image bank (USCSF) and to determine the camera locations from which the images were taken. We assign an ID to each face of the mesh. We match each pseudorandomly-annotated point with its corresponding mesh ID. Other views of the annotated mesh face are obtained by projecting the center of the mesh face into images using a standard projective camera model with extrinsic (camera location and orientation) and intrinsic (focal length, camera center, and radial distortion) parameters obtained through optimization. In short, each mesh ID is assigned a single class and associated with its corresponding location in one or more images. Our final dataset consisted of 6,525 labeled meshes with 138,405 corresponding patches in images.

3.5.2 Voting Networks

We propose an architecture to handle a variable number of views. Since our image bank (USCSF) was extracted from video and was collected in a serpentine fashion, any arbitrary point on the seafloor was likely captured in many images from many points of view. The number of views will vary and so, too, will the camera locations with respect to the point on the seafloor. The first and most obvious approach we compare is a simple voting scheme. We train ResNet152 [1] using a

Table 3.2: Results of the patch-based multi-view architectures.

Architecture	Accuracy	Batch Size
ResNet152	85.54	32
ResNet152 with Simple Voting	90.70	32
ResNet152 with Logit Pooling	91.00	32
nViewNet-4	93.52	16
nViewNet-8	94.26	16

train/test stratification scheme where 80% of the data is used to train the model and 20% is used to test it. Each of the images from the training set and its corresponding class is used to train ResNet152 [1]. The base architecture weights are initialized using the Imagenet [22] pretrained weights. We replace the last layer (the fully connected layer) with a different fully connected layer that has a number of outputs equal to the number of classes in our dataset. To retain the benefit of the pretraining, we freeze the base architecture weights and train the fully connected layer with a learning rate of 1×10^{-3}. We use a batch size of 64 and stochastic gradient descent with a Nesterov momentum as our optimization technique. We then train the entire model using a learning rate of 1×10^{-4} and weight decay of 1×10^{-6}.

Each image corresponding to a mesh face in the validation set is then passed through the trained network. Each image votes on a classification for that face, and the class with the plurality of votes is used. As an alternative, we explore a pooling method that sums the logits, which essentially weights each vote by confidence.

3.5.3 nViewNet

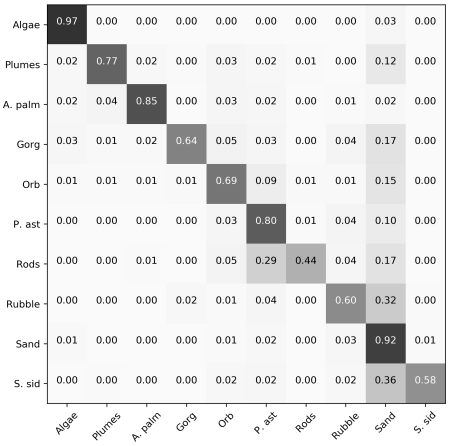
Next we propose the nViewNet architecture to handle a variable number of viewpoint images. nViewNet uses ResNet152 [1] (minus the last fully connected layer) as a base architecture and, in a similar fashion to TwinNet, the base architecture weights are shared across all inputs. To keep

memory constraints constant and for ease of training, we set a cap on the maximum number of viewpoints to be included for classifying each mesh face. If the number of available viewpoints exceeds our maximum the additional image views are ignored and the retained viewpoints should be selected at random. Each viewpoint up to the set maximum is fed through the base, and the outputs are then fed to our collapse module. The collapse module takes two images, each with C channels, as inputs and concatenates them channel-wise. It then reduces the concatenated data, which has $2C$ channels, back to C channels with a two-dimensional convolution and a kernel size of one. Another two-dimensional convolution layer occurs after this, but with a kernel size of three. The collapse module is called recursively to combine pairs of images in a tree-like fashion until only a single output remains (see Figure 3.1). We use a linear transform to reduce the output of the collapse module to a vector of logits with a length equal to the number of classes (see Figure 3.1(b) for the full architecture). In the case where a mesh face is seen in less than the maximum allowable number of viewpoints, we repeat any existing viewpoints until the maximum is reached. We compare using a maximum of four viewpoints (nViewNet-4) and a maximum of eight viewpoints (nViewNet-8) to the previous approaches.

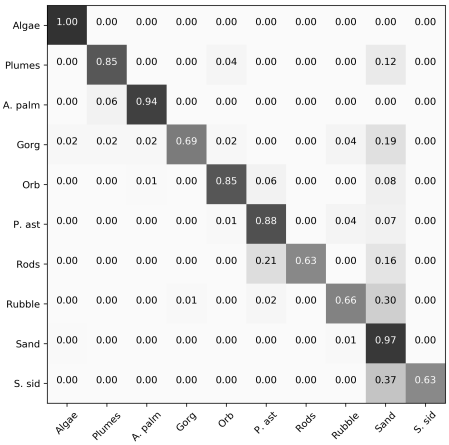
The base architecture weights are initialized using the weights when trained on the viewpoints individually. This initializes the base with feature outputs that are already useful for classifying in this task. We ensure that the training and testing sets contain the same images in the individually trained model and nViewNet to maintain experimental validity. To retain the benefit of the pretraining, we freeze the base architecture weights and train the additional modules initially with a learning rate of 1×10^{-4}. We use stochastic gradient descent with Nesterov momentum as our optimization technique.

3.5.4 Performance of the Patch-Based Approaches

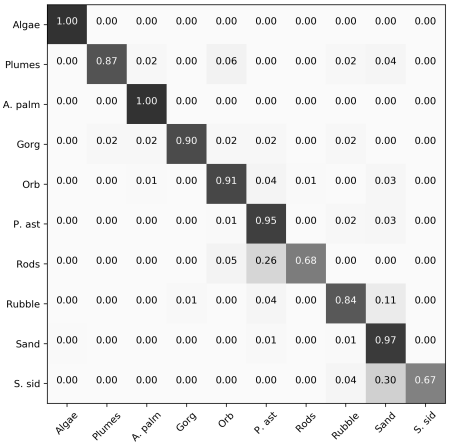
Table 3.2 summarizes the results of our experiments with voting and pooling schemes along with our nViewNet architecture. All methods are observed to greatly outperform their underlying architecture, ResNet152 [1], alone. Logit pooling was essentially tied with simple voting but did techni-



(a) ResNet152



(b) ResNet152 with Logit Pooling



(c) nViewNet-8

Figure 3.4: Confusion matrices for three patch-based multi-view architectures. We abbreviate *Acropora palmata* as *A. palm*, *Gorgonia ventalina* as *Gorg*, *Orbicella spp.* as *Orb*, *Porites astreoides* as *P. ast*, and *Siderastrea siderea* as *S. sid*.

cally improve on it. The best nViewNet architecture outperformed the pooling and voting schemes with 3.26% improvement. This shows that a learned combination provides some benefit to the network in classification. Even though our nViewNet architectures had a maximum cap on the number of viewpoints, both were able to outperform the voting/pooling schemes that use every viewpoint with no cap. The voting/pooling schemes, however, are quite a bit easier to implement and still outperform ResNet152 [1] alone by a large margin.

3.6 Conclusions

We have shown how the often discarded varying points of view common in the data collection for mapping tasks can be used to improve the accuracy of convolutional neural network architectures. We have examined the use of stereoscopic image pairs for FCNN approaches to semantic segmentation. We then proposed a method that uses generated disparity maps from the left- and right-perspective rectified images and adds that disparity map as a fourth channel in addition to the three color channels. This three-dimensional disparity information guides the semantic segmentation. We proposed the TwinNet architecture, roughly based on siamese networks, which accepts both the left- and right-perspective images as inputs.

We have also examined utilizing a variable number of viewpoints for patch-based approaches to semantic segmentation. We created a three-dimensional semantic segmentation by making a three-dimensional mesh and performing classification on each face of the mesh. We proposed the nViewNet architecture, which is capable of receiving a variable number of images (with a set maximum number) and learning a combination to ultimately give a single classification.

Our results indicate that utilizing more than just a single viewpoint to make a classification yields a higher accuracy. Our top-performing FCNN model was our proposed TwinNet architecture, which we ran both with only a left-perspective image and with the full stereo pair. When run with just the left-perspective image, it performed comparably to its base architecture, Dilation8 [8], but when run with both the left-perspective and right-perspective images, it displayed a

marked improvement over Dilation8. This indicates that the additional information from varying points of view is useful in classification and can successfully be utilized to improve performance.

Our top-performing patch-based model was our proposed nViewNet-8 architecture, which utilizes a maximum of eight viewpoints for each mesh face. NViewNet-8 yielded an accuracy of 94.26%, which outperformed the underlying ResNet152 [1] architecture by 8.72%. These performance improvements are not insignificant and are discarded in mapping tasks that first composite and then classify. By utilizing information from multiple points of view these improvements can be realized. Future work could explore methods of utilizing a variable number of images with no maximum cap and no image repetition. Furthermore, we would like to explore an extension of the logit pooling schemes that would discard low confidence predictions and defer to human judgment in such cases.

Acknowledgment: This research was funded in part by a Robotics Research Equipment Grant by the Faculty of Robotics and the Office of Vice President for Research, The University of Georgia, Athens, Georgia, to Dr. Bhandarkar and Dr. Hopkinson.

3.7 REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

[3] O. Beijbom, P. J. Edmunds, D. I. Kline, B. G. Mitchell, and D. Kriegman, "Automated annotation of coral reef survey images," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Providence, Rhode Island, Jan. 1, 2012.

[4] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[7] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[8] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *ICLR*, 2016.

[9] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *ICML Deep Learning Workshop*, vol. 2, 2015.

[12] K. R. Anthony, "Coral reefs under climate change and ocean acidification: Challenges and opportunities for management and policy," *Annual Review of Environment and Resources*, vol. 41, 2016.

[13] O. Hoegh-Guldberg, P. J. Mumby, A. J. Hooten, R. S. Steneck, P. Greenfield, E. Gomez, C. D. Harvell, P. F. Sale, A. J. Edwards, K. Caldeira, *et al.*, "Coral reefs under rapid climate change and ocean acidification," *science*, vol. 318, no. 5857, pp. 1737–1742, 2007.

[22] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[24] A. King, S. M. Bhandarkar, and B. M. Hopkinson, "A comparison of deep learning methods for semantic segmentation of coral reef survey images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 1394–1402.

[25] H. Su, S. Maji, E. Kalogerakis, and E. G. Learned-Miller, "Multi-view convolutional neural networks for 3d shape recognition," in *Proc. ICCV*, 2015.

[26] H. Hirschmuller, “Stereo processing by semiglobal matching and mutual information,” *IEEE Transactions on pattern analysis and machine intelligence*, vol. 30, no. 2, pp. 328–341, 2008.

[27] A. Telea, “An image inpainting technique based on the fast marching method,” *Journal of graphics tools*, vol. 9, no. 1, pp. 23–34, 2004.

CHAPTER 4

CONCLUSION

In this thesis, we have examined the use of deep learning methods for semantic segmentation of images taken in underwater coral reef ecosystems. We have shown that these deep learning methods can be an effective means to automate the object classification and semantic segmentation process, thus determining the abundance of various organisms inhabiting a given coral reef. We have described two major categories of semantic segmentation methods. The first of these, patch-based CNN approaches, work by classifying existing segments. Second, we assessed the utility of FCNN architectures. These architectures are capable of per-pixel segmentation and classification in a single end-to-end trainable network.

Using patch-based methods, we attempted to identify the abundance of organisms on the reefs surveyed. Upon segmentation using existing techniques, we performed a patch-wise classification of each output segment. In our comparison of standard patch-based CNN approaches for classification from ground truth annotations of individual points in training images, our best performing CNN model was the ResNet152 [1] architecture. Resnet152 resulted in 90.03% classification accuracy, while the work of Beijbom et al. [3], using SVMs and texton dictionaries, resulted in an accuracy of 84.8% on the same dataset. The accuracy of Resnet152 is sufficient for use in many of our tasks; still, higher accuracy would allow it to be robustly used with high confidence across more tasks.

In comparing these two classes of methods, we note that the granularity of classification is much coarser with a patch-based CNN model due to the fact that it outputs only one single class label for a whole patch in an image. FCNN architectures, on the other hand, result in per-pixel classification. Although the patch-based methods provide higher overall accuracy, they can often

inherit the limitations of the segmentation algorithm used to localize specific taxa within the coral reef image.

In our discussion of the FCNN architectures, we found that the best of the four compared models was the Deeplab v2 [2] architecture, which yielded 67.7% accuracy on our dense classification dataset. These FCNN methods work by classifying each individual pixel throughout an entire image. Unlike the previously discussed patch-based approaches, FCNN approaches are not limited in terms of localization accuracy. In this specific case, the fine granularity of our classification led to the classification accuracy of FCNN approaches being lower than the accuracy of the selected patch-based methods.

We next examined the use of multi-view image data for improving the accuracy of both FCNN and patch-based CNN methods for semantic segmentation. We show that this data, while typically discarded in traditional approaches that utilize only a single viewpoint for each image, can be used to improve classification accuracy.

For FCNN architectures, we used pairs of left-perspective and right-perspective images to generate a disparity map. This disparity map was then added as a fourth channel to the existing three color channels. We proposed an architecture capable of utilizing stereo pairs as inputs, TwinNet, that is also loosely based on siamese networks. The results of our comparison of TwinNet with other FCNN architectures show that using stereo data can yield a higher classification accuracy than traditional approaches. TwinNet was the best-performing FCNN model in the comparison. We tested the results of TwinNet when trained on only a single left-perspective image as well as with the full stereo pair. We note that when our custom architecture was trained and evaluated with just the left-perspective image, it yielded comparable results to Dilation8 [8], its baseline architecture. When trained and evaluated with both stereo images, however, TwinNet yielded a markedly higher accuracy than Dilation8. These results indicate that the improvement in classification accuracy resulted from the additional viewpoint information utilized by TwinNet.

Next, we examined methods of improving patch-based CNN approaches using image data with a variable number of viewpoints. Using video survey data, we generated a three-dimensional mesh

using a Structure-from-Motion approach and classified each face of the mesh, resulting in a finished three-dimensional semantic segmentation. We proposed the nViewNet architecture, which can receive a varying quantity (with a specific maximum number) of input images and learn a combination to yield a single-entity classification.

Of the patch-based models, our proposed nViewNet-8 architecture, with a maximum of eight viewpoints per mesh face, performed the best, resulting in an accuracy of 94.26%. These results are higher than those of its baseline architecture, ResNet152 [1], which yielded 85.54% accuracy on the mesh face classification task.

In sum, we have explored these various FCNN and patch-based CNN approaches, with and without multi-view data, in the context of semantic segmentation and object classification of coral reef survey images. We have proposed several new architectures to improve on previous methods with the overall goal of mapping and monitoring coral reef ecosystems. These ecosystems provide untold benefits to both the marine organisms that inhabit them as well as to the surrounding coastal communities [11]. We hope that these and future deep learning methods can be used to benefit the field of ecology in general and the health of coral reef ecosystems in particular.

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

[2] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[3] O. Beijbom, P. J. Edmunds, D. I. Kline, B. G. Mitchell, and D. Kriegman, "Automated annotation of coral reef survey images," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Providence, Rhode Island, Jan. 1, 2012.

[4] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[5] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.

[6] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning.," in *AAAI*, vol. 4, 2017, p. 12.

[7] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[8] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *ICLR*, 2016.

[9] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *ICML Deep Learning Workshop*, vol. 2, 2015.

[10] J. D. Hedley, C. M. Roelfsema, I. Chollett, A. R. Harborne, S. F. Heron, S. Weeks, W. J. Skirving, A. E. Strong, C. M. Eakin, T. R. Christensen, *et al.*, "Remote sensing of coral reefs for monitoring and management: A review," *Remote Sensing*, vol. 8, no. 2, p. 118, 2016.

[11] L. Burke, K. Reytar, M. Spalding, and A. Perry, *Reefs at risk revisited*. 2011.

[12] K. R. Anthony, "Coral reefs under climate change and ocean acidification: Challenges and opportunities for management and policy," *Annual Review of Environment and Resources*, vol. 41, 2016.

[13] O. Hoegh-Guldberg, P. J. Mumby, A. J. Hooten, R. S. Steneck, P. Greenfield, E. Gomez, C. D. Harvell, P. F. Sale, A. J. Edwards, K. Caldeira, *et al.*, "Coral reefs under rapid climate change and ocean acidification," *science*, vol. 318, no. 5857, pp. 1737–1742, 2007.

[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 three-dimensional reconstruction for archaeology," *Journal of Field Robotics*, vol. 34, no. 4, pp. 625–643, 2017.

[15] R. Ruzicka, M. Colella, J. Porter, J. Morrison, J. Kidney, V. Brinkhuis, K. Lunz, K. Macaulay, L. Bartlett, M. Meyers, *et al.*, "Temporal changes in benthic assemblages on florida keys reefs 11 years after the 1997/1998 el niño," *Marine Ecology Progress Series*, vol. 489, pp. 125–141, 2013.

[16] J. E. Smith, R. Brainard, A. Carter, S. Grillo, C. Edwards, J. Harris, L. Lewis, D. Obura, F. Rohwer, E. Sala, *et al.*, "Re-evaluating the health of coral reef communities: Baselines and evidence for human impacts across the central pacific," *Proc. R. Soc. B*, vol. 283, no. 1822, p. 20 151 985, 2016.

[17] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.

[18] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *Int. J. Comput. Vision*, vol. 59, no. 2, pp. 167–181, Sep. 2004, ISSN: 0920-5691. DOI: 10.1023/B:VISI.0000022288.19776.77.

[19] M. Varma and A. Zisserman, "A statistical approach to texture classification from single images," *International Journal of Computer Vision*, vol. 62, no. 1–2, pp. 61–81, 2005.

[20] T. Treibitz, B. P. Neal, D. I. Kline, O. Beijbom, P. L. Roberts, B. G. Mitchell, and D. Kriegman, "Wide field-of-view fluorescence imaging of coral reefs," *Scientific reports*, vol. 5, p. 7694, 2015.

[21] I. Alonso, A. Cambra, A. Munoz, T. Treibitz, and A. C. Murillo, "Coral-segmentation: Training dense labeling models with sparse ground truth," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2874–2882.

[22] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[24] A. King, S. M. Bhandarkar, and B. M. Hopkinson, "A comparison of deep learning methods for semantic segmentation of coral reef survey images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 1394–1402.

[25] H. Su, S. Maji, E. Kalogerakis, and E. G. Learned-Miller, "Multi-view convolutional neural networks for 3d shape recognition," in *Proc. ICCV*, 2015.

[26] H. Hirschmuller, “Stereo processing by semiglobal matching and mutual information,” *IEEE Transactions on pattern analysis and machine intelligence*, vol. 30, no. 2, pp. 328–341, 2008.

[27] A. Telea, “An image inpainting technique based on the fast marching method,” *Journal of graphics tools*, vol. 9, no. 1, pp. 23–34, 2004.