A Comparison of Deep Learning Methods for Semantic Segmentation of Coral Reef Survey Images

Andrew King[1] Suchendra M. Bhandarkar[1,2] Brian M. Hopkinson[3]
[1]Institute for Artificial Intelligence [2]Department of Computer Science [3]Department of Marine Sciences
The University of Georgia, Athens, Georgia 30602, USA
andrewking@uga.edu suchi@cs.uga.edu bmhopkin@uga.edu

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.

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 [9]. 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 [5]. 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 [3, 10]. 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 [4]. 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 [11]. 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 1 depicts the 3D reconstruction of a coral reef accompanied by the semantic classification of its constituent taxa.

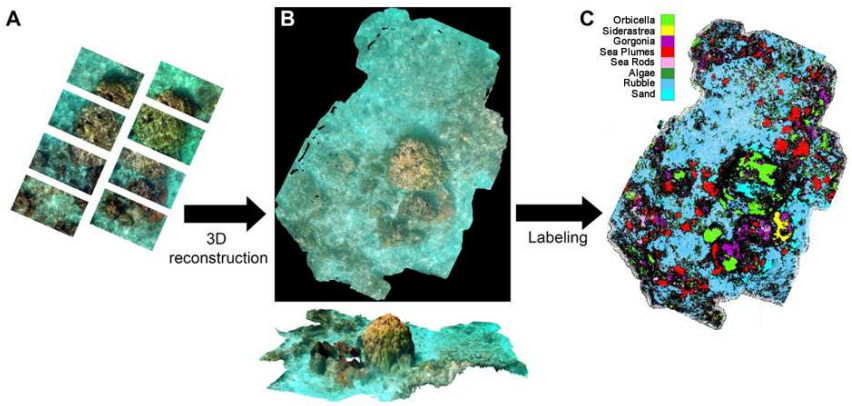


Figure 1: 3D reconstruction and annotation of a coral reef ecosystem.

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 distribution of sessile organisms, including hard corals, soft corals, and algae, and open space for settlement [13, 15]. 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 [14], InceptionV3 [17], InceptionResNetV2 [16], Resnet50 and Resnet152 [8]. 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) [1] or graph cuts [7]. 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 [12], Dilation8 [20], DeepLab v2 [6], and Dilation-Mod, 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. Background

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 [14]. This architecture was proposed in 2014 by Simonyan and Zisserman [14] 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. [17]. 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. [8] 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 [8]. Finally, we also consider the Inception-ResNetV2 architecture proposed by Szegedy et al. [16], which combines the Inception architecture with the ResNet residual block architecture.

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. [12]. 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. [12] use the existing VGG16 classification architecture [14] 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. [12] 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 [20] present a new FCNN architecture termed Dilation8. They base Dilation8 on the FCN8s architecture [12] 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 (1).

$$(F *_l k)(p) = \sum_{s+lt=p} F(s)k(t) \tag{1}$$

Yu and Koltun [20] 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 [20] 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 [20] 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. [12].

The final FCNN model that we consider in this paper is Deeplab v2, proposed by Chen et al. [6]. Chen et al. refine previously proposed FCNN models by employing the ResNet [8] 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 [20] on our dataset of coral reef survey images.

2.3. Related Work

Beijbom et al. [4] 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 [4]. 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 [4]. Beijbom et al. [4] employed a maximum response filter bank in conjunction with a multi-scale 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. [18] 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. [2] 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. [4] 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.

3. Evaluation of Patch-Based CNN Approaches

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 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.

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 [14], InceptionResNetV2 [16], InceptionV3 [17], Resnet50 and Resnet152 [8], to that of the SVM-based and texton dictionary-based approach proposed by Beijbom et al. [4]. 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

Table 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

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. [4] 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.

3.3. Performance of the CNN Architectures

Table 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 [17] was observed to perform the worst, yielding a classification accuracy of 84.69%. Resnet152 [8] was observed to yield the best classification accuracy, outperforming VGG16 [14] and Resnet50 [8] 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 2. 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 production rates in the reef. All the classes are classified correctly at least a majority of the time among our top performing CNN models.

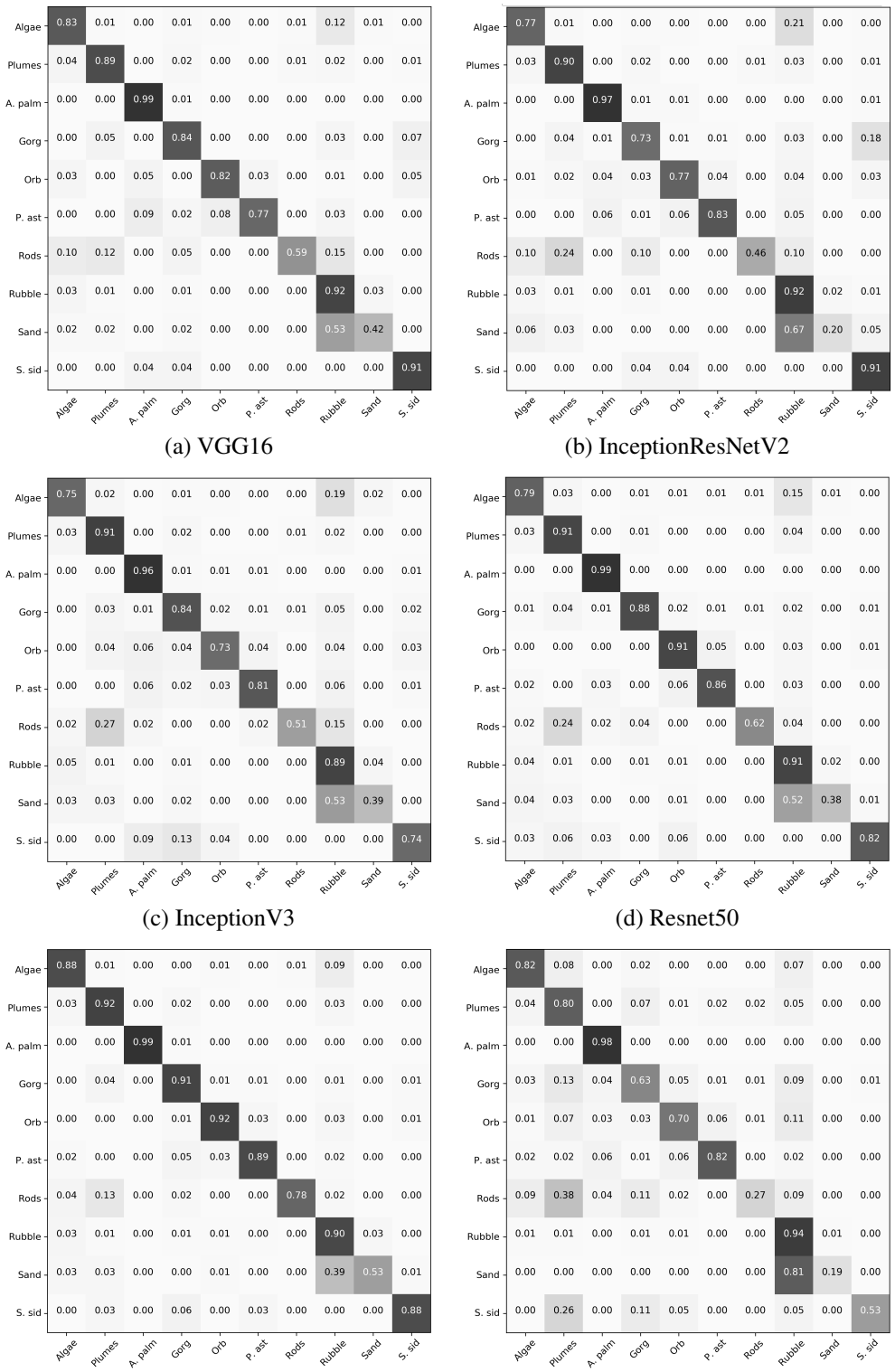
The SVM-based approach yields an overall accuracy of 84.8% on our dataset, lower than that of our best performing patch-based CNN models. The SVM-based approach also tends to significantly underperform on minority classes such as sea rods, *Siderastrea siderea*, sand, and *Orbicella*.

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.

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 [1] and the other based on efficient computation of graph cuts [7]. 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.



(a) VGG16 (b) InceptionResNetV2

(c) InceptionV3 (d) Resnet50

(e) Resnet152 (f) SVM and Texton Dictionary

Figure 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: 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

This tool allowed us to quickly generate 413 dense classification maps for use with our FCNN models [12].

4.2. DilationMod

We proposed and tested a modification to the Dilation8 [20] 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 [20]. 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.

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.

4.4. Training the FCNN Models

We compare the performance of FCN8s [12], Dilation8[20], DeepLab v2 [6], 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.

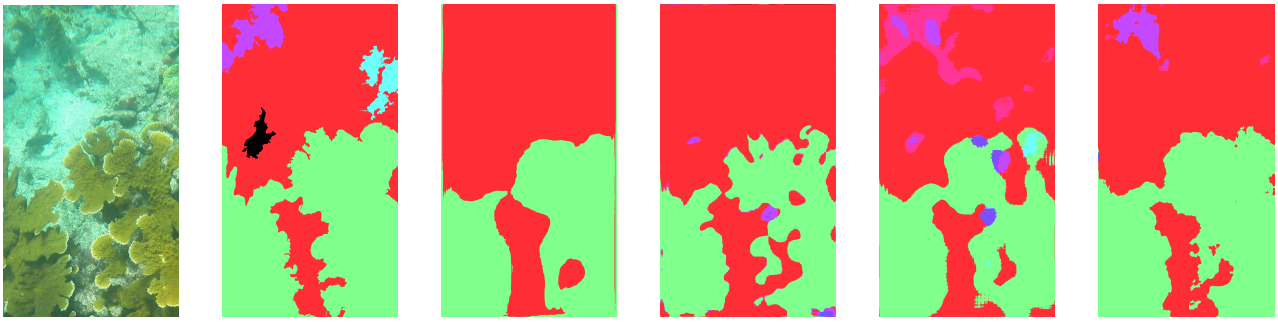
4.5. Performance of the FCNN Models

Table 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 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 [12], 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 [20] 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 [6] 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 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.

5. Conclusions

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 sub-



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

Figure 3: Outputs of multiple FCNN architectures for a given sample image.

strates 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 [8] architecture, which yielded an accuracy of 90.03%. The previous work of Beijbom et al. [4] 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. [12] 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.

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 [14] 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 [8]. 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.

References

[1] 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*, 34(11):2274–2282, 2012.

[2] 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*, pages 2874–2882, 2017.

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

[4] 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, 2012.

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

[6] 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*, 40(4):834–848, 2018.

[7] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *Int. J. Comput. Vision*, 59(2):167–181, Sept. 2004.

[8] 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)*, pages 770–778, June 2016.

[9] 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*, 8(2):118, 2016.

[10] 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*, 318(5857):1737–1742, 2007.

[11] 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*, 34(4):625–643, 2017.

[12] 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*, pages 3431–3440, 2015.

[13] 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*, 489:125–141, 2013.

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

[15] 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*, 283(1822):20151985, 2016.

[16] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, volume 4, page 12, 2017.

[17] 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*, pages 2818–2826, 2016.

[18] 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*, 5:7694, 2015.

[19] M. Varma and A. Zisserman. A statistical approach to texture classification from single images. *International Journal of Computer Vision*, 62(1–2):61–81, 2005.

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