Estimation of Abundance and Distribution of Salt Marsh Plants from Images Using Deep Learning

J. Parashar[1] S. M. Bhandarkar[1,2] J. Simon[3] B. M. Hopkinson[3]
[1]Institute for AI [2]Dept. of Computer Science [3]Dept. of Marine Sciences
University of Georgia
Athens, Georgia 30602, USA
{jayant.parashar, suchi, jacob.simon25, bmhopkin}@uga.edu

S. C. Pennings
Dept. of Biology & Biochemistry
University of Houston
Houston, Texas 77204, USA
scpennin@central.uh.edu

***Abstract*—Recent advances in computer vision and machine learning, most notably deep convolutional neural networks (CNNs), are exploited to identify and localize various plant species in salt marsh images. Three different approaches are explored that provide estimations of abundance and spatial distribution at varying levels of granularity defined by spatial resolution. In the coarsest-grained approach, CNNs are tasked with identifying which of six plant species are present/absent in large patches within the salt marsh images. CNNs with diverse topological properties and attention mechanisms are shown capable of providing accurate estimations with $> 90\%$ precision and recall for the more abundant plant species and reduced performance for less common plant species. Estimation of *percent cover* of each plant species is performed at a finer spatial resolution, where smaller image patches are extracted and the CNNs tasked with identifying the plant species or substrate at the center of the image patch. For the percent cover estimation task, the CNNs are observed to exhibit a performance profile similar to that for the presence/absence estimation task, but with an $\approx$ 5–10% reduction in precision and recall. Finally, fine-grained estimation of the spatial distribution of the various plant species is performed via semantic segmentation. The *DeepLab-V3* semantic segmentation architecture is observed to provide very accurate estimations for abundant plant species, but with significant performance degradation for less abundant plant species; in extreme cases, rare plant classes are seen to be ignored entirely. Overall, a clear trade-off is observed between the CNN estimation quality and the spatial resolution of the underlying estimation thereby offering guidance for ecological applications of CNN-based approaches to automated plant identification and localization in salt marsh images.**

***Index Terms*—Salt marsh monitoring, convolutional neural networks, network topology, attention mechanism, deep learning, ecological monitoring**

I. Introduction

Ecological studies are often limited in spatial and temporal scale by the time needed to characterize the distribution and abundance of the constituent species. For example, most assessments of plant and invertebrate species distributions are limited to analysis of abundance or presence/absence within small areas of the study site (termed as *quadrats*). The quadrats typically comprise $< 1\%$ of the total study area, due to the intensive effort required to manually analyze these ecological systems [29]. Computer vision and machine learning tools offer the possibility of automating and accelerating this work, making it possible to increase the spatial and temporal resolution of ecological surveys. In particular, deep learning approaches are rapidly gaining popularity for these tasks since they provide a unified computer vision and machine learning framework whose performance typically exceeds that of previous approaches that use predetermined hand-crafted features [8].

In this paper, we examine deep learning approaches for identification, enumeration, and spatial localization of plant species in salt marsh ecosystems. Salt marshes are highly productive, inter-tidal marine habitats found along protected coastlines or behind barrier islands spanning temperate to subpolar regions [22]. Salt marsh sediments store high densities of organic carbon making them important *blue carbon* ecosystems, in which carbon is sequestered that would otherwise be released to the atmosphere increasing atmospheric CO_2 concentrations [21]. Few species inhabit the salt marsh ecosystem due to its harsh conditions. However, on account of the ecosystem's high productivity, the resident species capable of surviving salt marsh conditions often exhibit high abundance. Given their low biodiversity, salt marshes have for long served as model ecosystems that are amenable to both experimental and observational work [6].

Salt marsh plant communities on the east coast of the United States are typically dominated by grasses in the genus *Spartina*, especially at lower marsh elevations as only this genus is capable of tolerating frequent flooding with salt water [23]. At higher marsh elevations, several additional species are also found, many of which are succulents or otherwise adapted to handle the harsh salt marsh conditions. The abundance and distribution of these resident species is commonly assessed using several semi-quantitative methods, employed either in real time in the field or on archived images. One approach is to simply indicate within a small quadrat (e.g. 0.25 m $\times$ 0.25 m) which plant species are present and which are absent. An alternative approach is to estimate the *percent cover* within the quadrat, i.e., the percentage of space occupied by each plant species or substrate, by randomly choosing points (25–100) within a quadrat (on the ground or in the image) and identifying the resident plant species (or substrate) at the chosen point. In this paper, we explore and assess various deep learning approaches to automate the aforementioned tasks on salt marsh images. We specifically explore multi-layer convolutional neural network (CNN) architectures such as

Fig. 1. Example images of marsh plant species (from left to right and top to bottom): *Sarcocornia, Spartina, Limonium, Borrichia, Batis* and *Juncus*.

the *ResNet* [13], *PyramidNet* [12], *residual attention network (RAN)* [28], *DenseNet* [15], *ResNext* [34], and *Inception-V3* [26] in the context of multi-label classification. These CNN architectures are also employed for the percent cover computation task, which is formulated as the more simple image classification problem.

Going beyond what is done in typical time-constrained ecological studies, we also assess the ability of CNN architectures to perform semantic segmentation of salt marsh images. Semantic image segmentation potentially provides more accurate estimates of species abundance and the spatial distribution of plants at a much finer spatial resolution. In this paper, the *DeepLab-V3* CNN architecture is employed for the semantic image segmentation task. As is typically encountered in studies of most ecological systems, the salt marsh images are characterized by fine-grained interleaving of classes, ambiguous class boundaries, and wide-variations in lighting and viewing perspective that complicate automated image analysis procedures, particularly those pertaining to semantic image segmentation. The key contributions of the paper can be summarized as follows: (a) we present a comparison of three distinct approaches, based on presence/absence determination, percent cover computation and semantic image segmentation, to estimate the abundance and distribution of various salt marsh plant species using CNNs, (b) we show a clear trade-off between the precision and spatial resolution of the resulting estimations, (c) we present a comparison of various CNN architectures and show multi-path CNNs to outperform other CNNs for analysis of salt marsh images and finally, (d) we compute the variations in the distributions of various salt marsh plant species across an elevation gradient.

II. Background

A. Deep Learning Applications in Ecology

Previous applications of computer vision and machine learning to the analysis of ecosystem imagery has progressed from traditional pipelines that extract predetermined, hand-crafted features, followed by application of traditional classifiers such as support vector machines (SVMs) to employing end-to-end deep neural networks (DNNs) or deep learning (DL) methods. Beijbom et al. [4], [5] present an excellent example of the traditional approach to automated classification of ecosystem images focusing on coral reef surveys. Employing a maximum response filter bank in conjunction with a multiscale patch/texton dictionary to characterize the features in underwater coral reef images, Beijbom et al. [4] use a traditional SVM-based classifier to categorize image patches as belonging to various reef organism classes. They also outline the many challenges unique to the task of automated analysis of ecological images such as extreme variations in the size, color, shape, and texture of each of the taxa, the organic and ambiguous nature of the class boundaries and significant alterations in ambient lighting and image colors [5]. In the ecological remote sensing, classification approaches have focused almost exclusively on pixel-level spectral information ignoring spatial context, although there have been some notable exceptions [14]. However, classifiers based primarily on pixel-level spectral data are less applicable to *local* ecosystem images which are typically acquired with consumer-grade RGB cameras.

In recent times, CNNs (or ConvNets) and related deep neural networks (DNNs) have revolutionized computer vision, especially semantic image segmentation, feature extraction and classification, and object detection and recognition [17], [19]. The superior performance of CNN- and related DNN-based approaches has led to their rapid adoption in ecological research [8], [30]. CNNs have seen notable success in detection of animals, such as hummingbirds [31] and shorebirds [7] and have resulted in dramatically improved remote sensing methods for specific tasks such as automatic identification and inventory of termite mounds [8] and tree species [2], [32]. Brodrick et al. [8] argue that CNNs may become essential tools for ecologists due to their power, generality, and relative ease of use.

In addition to tackling specific tasks, CNNs are also being increasingly used to provide a broad-scale overview of community composition. Williams et al. [33] have employed CNNs to assess the abundance of major taxa and substrates on coral reefs, achieving classification accuracies similar to those attained by human annotators. Traditional approaches to ecosystem image analysis [4], [5] employ complex feature extraction and classification algorithms that entail extensive knowledge of computer vision and machine learning. In contrast, a CNN-based computational pipeline provides an integrated image analysis framework by leveraging existing, pre-trained CNNs where the feature extractors and classifiers have been automatically learned from training data. Their user-friendliness, superior performance, and the fact that they require minimal background in computer vision and machine learning suggest that CNNs will be widely adopted to accelerate ecological research in the near future.

B. Convolutional Neural Networks (CNNs)

CNNs have become a standard tool for several machine vision tasks such as image classification, semantic image segmentation and automated image captioning. The various

CNN architectures differ in their topological properties across multiple dimensions such as, the type of convolution operation, network depth, spatial dimensions of the network layers, and the width and design of multiple network pathways [16].

1) Inception CNN: The *Inception* CNN employs the principles of variability and modularity to deal with increasing model size and computational costs associated with scaling up of CNNs to address real-world computer vision problems [26]. A key feature of the *Inception* CNN is the introduction of inception layers comprising of multiple-size convolutional filter kernels. Earlier versions of the *Inception* CNN use small-size convolutional filters for computational efficiency whereas subsequent versions construct deeper networks by effectively combining blocks of varying filter sizes using split, transform and merge strategies thereby ensuring an efficient multi-path flow of information [26]. The varying filter sizes capture spatial information at multiple scales which is subsequently combined within a single block using 1×1 convolutional filters [20]. Additionally, the *Inception* CNN uses an auxiliary classifier to deal with the problem of degradation of input between successive network layers [25].

2) Residual Learning CNN: Construction of deeper CNNs causes degradation of the input between successive network layers leading to the *vanishing gradient problem* that severely limits backpropagation learning [25]. Residual learning CNNs, i.e., *ResNets* mitigate this problem by introducing *skip* connections between CNN layers [13]. *ResNets* do not alter the topology of the network connections, rather they simply combine the outputs of alternating layers making it possible to stack several network layers without overfitting the data or having to deal with the vanishing gradient problem. *ResNets* have been shown to be very effective for classification problems on a variety of data sets [27].

3) Densely Connected CNN (DenseNet): Densely Connected CNNs, i.e., *DenseNets* [15], unlike *ResNets*, concatenate the output of every layer with the outputs of all previous layers in a given block. *DenseNets* have compelling advantages in that they alleviate the vanishing gradient problem while strengthening feature propagation to address input degradation between successive layers. Although the number of direct connections increases quadratically in the number of layers, *DenseNets* reduce the number of parameters substantially by encouraging feature reuse.

4) Dual Path Network (DPN): The *dual path network* (DPN) unifies the *ResNet* and *DenseNet* architectures to generate a *higher-order recurrent neural network* (HORNN) [10]. The HORNN architecture introduces recurrent connections between non-neighboring units based on an order hyper-parameter [24]. Since *ResNets* [13] allow for efficient feature reuse whereas *DenseNets* [15] are particularly effective at feature discovery, a multi-path network comprising of *DenseNets* and *ResNets* allows one to combine the advantages of both.

5) ResNext and Pyramid Networks: The *ResNext* [34] is a simple, highly modularized network architecture constructed by repeating a building block that aggregates a set of transformations with the same topology. In essence, the *ResNext* combines the best of the Inception [26] and *ResNet* [13] architectures. The *ResNext* creates blocks with multiple pathways using group convolutions [17] which are later combined using 1×1 convolutional filters [20]. The pyramidal network *PyramidNet* [12] is an enchancement of the *ResNet* wherein the dimensions of the feature map are increased gradually following an arithmetic progression (*additive PyramidNet*) or geometric progression (*multiplicative PyramidNet*). The *PyramidNet* is shown to perform better than the *ResNet* on image classification tasks since it circumvents loss of useful information [12].

6) Semantic Image Segmentation: The aforementioned CNNs are used primarily for image classification where an entire image is classified as belonging to a certain category. However, for more fine-grained analysis, one needs to perform semantic image segmentation where each image pixel is classified as belonging to a certain category. In this paper, we employ the *DeepLab-V3* architecture which leverages *atrous convolution* and *atrous pyramid spatial pooling* to generate pixel-wise labels for semantic image segmentation [9]. Atrous convolution is a special convolution operation that acts over a dilated sub-grid of the input [9].

C. Attention Mechanisms

In order to speed up the CNN training procedure, it is imperative that the CNN models focus on important portions of the image while discarding the irrelevant portions. This can be accomplished via incorporation of an *attention mechanism* within the CNN training procedure. In this work, we employed the *residual attention network* (RAN) architecture, which creates stacks of trunk and mask branches that are trained in an end-to-end fashion. The mask branches perform downsampling (i.e., convolution) and upsampling (i.e., deconvolution) to create attention masks whereas the trunk branches performs only convolution [28]. Upsampling effectively recreates the image dimensions in the RAN after convolution has reduced them.

III. Description of Data Sets

Overhead images of a roughly rectangular section of a salt marsh on Sapelo Island, Georgia, USA were collected in June 2014 using a consumer grade DSLR camera (Nikon D7100) with a wide-angle lens (24 mm). The camera was attached $\approx$ 1.5 m above the ground to a wheeled platform that was pulled across the marsh while high-resolution images (4000 $\times$ 6000 pixels) were continuously acquired at a rate of 1 Hz (1 frame/sec). After the camera had traversed $\approx$ 20–40 m the mobile platform was stopped, and the images acquired over that distance were deemed to comprise a *row*. The imaging platform was then moved $\approx$ 1 m perpendicular to the previous row and pulled again across the marsh to image an adjacent row. A section of marsh covering 80 rows was imaged in this manner moving from a higher marsh elevation (with a low row number) to lower marsh elevations (with high row numbers).

Six plant species are commonly found in salt marshes on Sapelo Island and all six were present in the images: *Spartina alterifora, Juncus roemerianus, Batis maritima, Sarcocornia*

spp., Borrichia frutescens, and *Limonium carolineanum*, all of which are subsequently identified by their genus (Fig. 1). The only plant species found at low marsh elevations is *Spartina*, but all species are found at high marsh elevations with *Spartina* gradually disappearing from the assemblage at the highest elevations giving way to a diverse mix of the remaining species. *Spartina* is a medium to tall grass (height: 0.3–2 m) with wide blades that emerge from a thick stem. The blades appear dark green to light green depending on the time of the day and dead blades attached to the stems are not uncommon. *Juncus* is a tall rush (height $\approx$ 1 m) with smooth cylindrical leaves that are gray-green in color. *Sarcocornia* is a low growing succulent plant with branching stems that lack leaves. Most of the stems are green but shade into red. *Batis* is a dense, succulent shrub with alternate, green to yellow leaves growing to $\approx$ 0.3 m tall in the imaged section of the marsh. *Borrichia* is a shrub with characteristic, grey-green oval leaves and bright yellow flowers at the end of each branch throughout most of the summer. We created three separate data sets for training, validation, and testing for each of the tasks, i.e., presence/absence determination, percent cover computation and semantic image segmentation as shown in Fig. 2.

A. Presence/Absence Determination

In the presence/absence determination task, the input images were divided into 15 sections (3 rows $\times$ 5 columns) as shown in Fig. 2. An expert human annotator then performed multi-label classification, i.e., delineation of all the plants in each image section for randomly selected images from the Sapelo Island marsh image data set. The multi-label classification dealt with seven classes, the six aforementioned plant classes and one *background* class. The most dominant class encountered in the presence/absence task was *Spartina* which was present in more than half of the image sections. In contrast, *Batis* and *Sarcocornia* were present in $\approx$ 17% of the image sections whereas *Limonium, Batis* and *Juncus* were observed to be rare classes, each accounting for $\approx$ 4% of the image sections.

B. Percent Cover Computation

Percent cover, the fraction of space occupied by a particular plant species when viewed from overhead, provides a more refined abundance metric than presence/absence. In the percent cover computation task, 25–50 points were randomly selected in an image. An image patch (512 $\times$ 512 pixels) surrounding each point was presented to an expert human annotator who performed single-label classification of the patch, i.e., labeled the patch based on the class present at the selected point (Fig. 2). There were 9 classes under consideration for this task: the aforementioned six plant classes, *Soil*, *Other*, and *Unknown*. The class *Other* was used to indicate an identifiable entity that did not belong to one of the six plant classes or *Soil*, such as *invertebrates* (crabs, snails, etc.) or portions of the imaging platform captured by the camera. The *Unknown* class denotes a situation where the class underlying the selected point was not identifiable, which typically occurred when the image section was out of focus or heavily shadowed. The dominant classes encountered in this task were *Spartina, Batis* and *Sarcocornia* which collectively accounted for $\approx$ 85% cover in approximately equal proportion. *Borrichia* and *Juncus* were deemed rare classes, each with $\approx$ 6% cover and *Limonium* the rarest class with $\approx$ 3% cover.

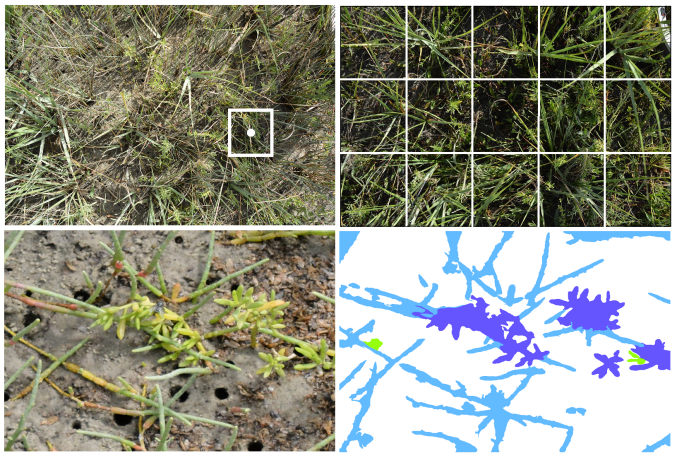

Fig. 2. Image analysis tasks from left to right and top to bottom: percent cover computation, presence/absence determination, semantic image segmentation and the corresponding segmentation masks.

C. Semantic Image Segmentation

The goal of semantic image segmentation is to classify each pixel in an image into one of a set of predetermined categories. To generate training data for the semantic image segmentation task, we used a superpixel labeling tool described in [1]. Each pixel was classified into one of nine classes, which were slightly different than those used for percent cover computation: the aforementioned six plant classes, *dead Spartina* (due to its common occurrence and substantially different appearance compared to *live Spartina*), background (which accounts for the classes *Soil* and *Unknown* in percent cover computation), and *Other* (which accounts for *invertebrates*, and portions of imaging platform captured by the camera).

IV. Experimental Results

A. Performance Evaluation Metrics

Precision, *recall* and *f-1* scores were used as evaluation metrics to compare the various deep learning models both in terms of overall performance and performance on specific classes. For overall performance assessment, both micro- and macro-averaged metrics were computed. Macro-averaged metrics were computed by first computing the precision, recall and *f-1* score metrics for each individual class and then averaging these metrics across all classes. In contrast, micro-averaged metrics were computed by summing the true positives, false positives, false negatives, and true negatives across the entire data set regardless of class and then computing precision, recall, and *f-1* score metrics. In the case of single-label classification (as is done in percent cover computation and semantic image segmentation) micro-averaged precision, recall, and *f-1* score metrics are equal to overall accuracy.

B. Evaluation of Presence/Absence Computation

Approximately 17,000 salt marsh image sections (from roughly 1150 images) were manually labeled for presence/absence of the six plant species. Images in the manually annotated data set were split into training (60%), testing (20%) and validation (20%) data sets. CNNs were trained for multi-label classification on the training data set using the Adam optimizer and binary cross entropy loss function. We initialized the model weights with values pre-trained on *ImageNet* [11] for all models and trained the weights until the loss value stopped declining on the validation data set.

After an initial examination of the performance of a trial model (*ResNet*) on the task, we found the classifier performed reasonably well on most classes, but had trouble identifying *Juncus* and falsely predicted the presence of *Sarcocornia* (and occasionally *Batis*) in low-elevation marsh regions where only *Spartina* was present. *Juncus* was rare and an examination of the classifier showed that *Juncus* was being misclassified as the more common *Spartina* grass. In an attempt to overcome this issue, additional images containing *Juncus* were identified and manually labeled to increase the representation of *Juncus* in the training data set. To address the false positives associated with *Sarcocornia* in the low-elevation marsh regions, additional low-elevation marsh images containing only *Spartina* were manually classified and added to the data sets. The addition of targeted training data helped improve the performance of the trial model (*ResNet*) and we proceeded to evaluate the performance of the remainder of the CNN architectures under consideration.

The seven CNN architectures that we assessed were generally observed to yield similar performance with micro-averaged precision, recall, and *f-1* scores, all exceeding 0.9 (Table I). The macro-averaged metrics were somewhat lower due to poorer performance on the rarer plant species (Table II). The relative performance of individual models was generally consistent across micro- and macro-averaged metrics with *ResNext* yielding the best precision and *f-1* score values and *DPN* the highest recall values. Unlike the other CNN architectures, *DPN* and *ResNext* both use a multi-path strategy which potentially helps to generate complex feature combinations.

The performance of the CNN architectures on individual classes was observed to vary significantly based on the relative abundance of the plant species in the data set (Figs. 3 and 4). *Spartina*, the most abundant species, was classified extremely accurately with precision and recall values exceeding 0.95. The recall value for *Sarcocornia* was also exceedingly high ($>$ 0.95) but the precision was notably lower at $\approx$ 0.9 due to the presence of false positives despite attempts at improvement. Although *Sarcocornia* is commonly present in most image sections, it is often not very abundant with only a few stems present in a given image section, which potentially contributes to the difficulty in achieving high precision values for this class. The quality of predictions for the remaining classes (*Batis, Borrichia, Limonium*, and *Juncus*) was generally observed to follow their occurrence frequency in the manually annotated data sets. While the performance of different CNN architectures at the class level was generally similar, both the *DPN* and *ResNext* exhibited uniquely high individual precision and recall values for multiple plant categories (Figs. 3 and 4). The recall values for DPN for *Limonium* and *Juncus* were at least 5% higher than those of all other CNN models. The precision value for *Limonium* was the highest in the case of the *ResNext*, showing an 8% difference from the next best CNN architecture. The *RAN* performed the worst overall in terms of the *f-1* score. However, the *RAN* was observed to yield a precision value that almost matched that of the *ResNext*. The attention mechanism used in the *RAN* did not seem to hold any advantages for this data set and task.

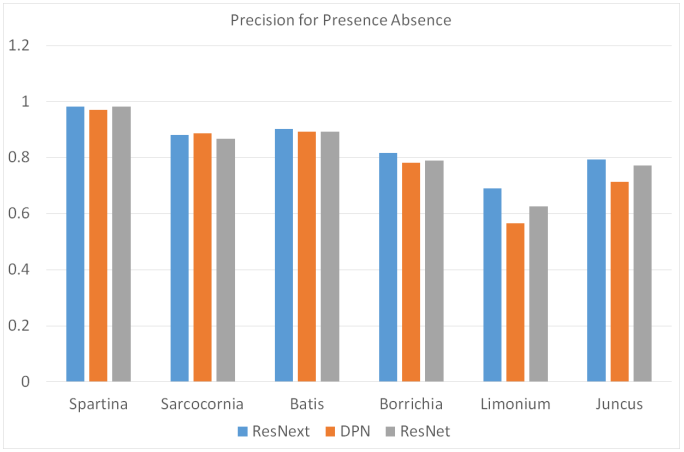

Fig. 3. Precision values for individual classes in presence/absence computation for (a) *ResNext*, (b) *Dual Path Network*, and (c) *ResNet*

TABLE I
PRESENCE/ABSENCE MICRO-AVERAGED RESULTS

CNN type	*Micro precision*	*Micro recall*	*Micro f-1 score*
ResNet101	0.918	0.932	0.925
DenseNet121	0.912	0.930	0.921
DPN92	0.906	**0.938**	0.922
ResNext101	**0.928**	0.931	**0.930**
Inception	0.910	0.923	0.916
RAN	0.924	0.889	0.906
PyramidNet101	0.911	0.923	0.917

TABLE II
PRESENCE/ABSENCE MACRO-AVERAGED RESULTS

CNN type	*Macro precision*	*Macro recall*	*Macro f-1 score*
ResNet101	0.844	0.872	0.858
DenseNet121	0.838	0.868	0.853
DPN92	0.827	**0.884**	0.855
ResNext101	**0.861**	0.867	**0.864**
Inception	0.834	0.845	0.839
RAN	0.848	0.761	0.802
PyramidNet101	0.820	0.840	0.830

C. Evaluation of Percent Cover Computation

To assess the ability of CNNs to automate percent cover computation, approximately 7,500 points were manually labeled in $\approx$ 250 images randomly sampled from the salt marsh image data set. The manually annotated data set was split, at

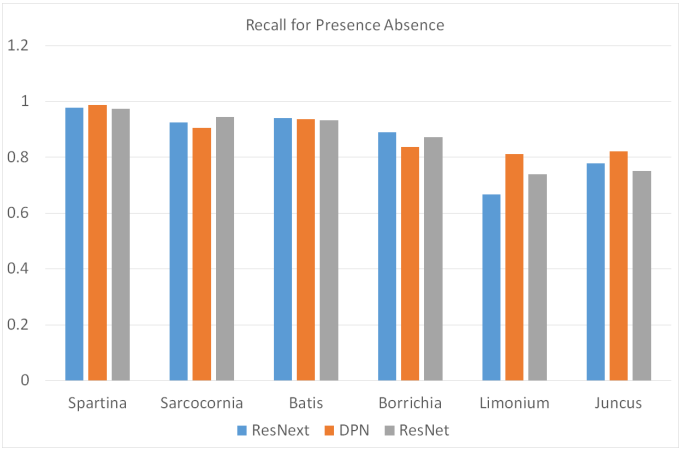

Fig. 4. Recall values for individual classes in presence/absence computation for (a) *ResNext*, (b) *Dual Path Network*, and (c) *ResNet*

the level of an individual image, into training (60%), validation (20%), and test (20%) data sets. The CNN models were trained in a manner similar to that described for the presence/absence computation task except that a multi-class cross entropy loss function was used. Table III reports the results of the same seven CNN architectures used in presence/absence computation, for the percent cover computation task. Since percent cover is a multi-class, single-label classification problem, the micro-averaged precision and recall values are equal and denoted by the term *accuracy*. The values of the macro-averaged precision, recall, and *f-1* scores were observed to be $\approx$ 0.10 lower for percent cover computation compared to presence/absence computation for each of the seven CNN architectures.

A significant challenge for the percent cover computation is the fine-grained interleaving of classes in salt marsh images. Image patches extracted for percent cover computation often contain multiple plant species and the classifier must learn to classify the plant observed in the center of the image patch. Recognizing this issue, we tested an attention-based CNN, i.e., *RAN*, hypothesizing that *RAN* would learn to focus attention on the center of the image patch for the purpose of classification while using the outer regions of the patch for context. However, the performance of *RAN* was generally observed to be inferior to that of the other CNN architectures. The initial patch size used (512 $\times$ 512 pixels) was relatively large so we attempted to reduce the size of the patch to 256 $\times$ 256 pixels to limit the number of classes present in image patches. However, the performance of all the CNN architectures, in terms of recall, dropped precipitously by $\approx$ 0.20 on the smaller patches across all classes. This was likely because the smaller patches did not provide sufficient context that was critical for accurate classification, such as overall leaf shape, instead, forcing the classifier to rely on small-scale texture and color features. As shown in Table III, *ResNext* was observed to achieve the highest overall accuracy and *f-1* score closely followed by *PyramidNet* and *DPN*. The confusion matrix in Fig. 5 resulting from the image patch classification during percent cover computation shows that even the best

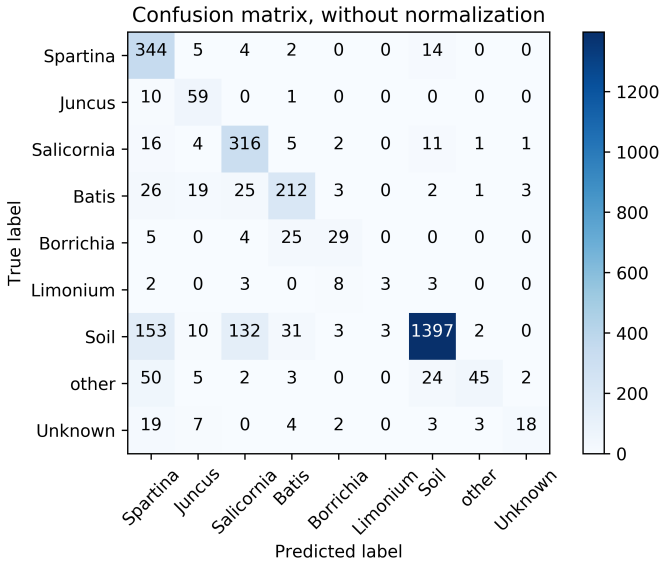

Fig. 5. Confusion matrix for percent cover computation for *ResNext*

performing *ResNext* CNN had particular difficulty recognizing lower abundance classes such as *Juncus, Limnonium*, and, in some cases, *Borrichia.*

TABLE III
RESULTS OF PERCENT COVER COMPUTATION

CNN type	*precision*	*recall*	*f-1 score*	*Accuracy*
ResNet101	0.742	0.700	0.720	0.833
DenseNet121	0.717	0.668	0.692	0.846
DPN92	0.743	0.732	0.737	0.843
ResNext101	**0.767**	0.736	**0.751**	**0.857**
Inception	0.738	0.672	0.703	0.833
RAN	0.713	0.618	0.662	0.851
PyramidNet101	0.751	**0.743**	0.748	0.844

D. Evaluation of Semantic Image Segmentation

Semantic image segmentation is not commonly employed in ecological research due to the labour intensive nature of labelling entire images at the pixel level. However, automated semantic segmentation offers potentially unprecedented spatial resolution in the field of ecology allowing novel insights into spatial relationships amongst organisms as well as computation of a more accurate abundance metric. We manually labeled $\approx$ 200 salt marsh image sections for semantic image segmentation using a super-pixel segmentation and labeling tool developed in [1]. The manually labeled data set was split into training (60%), validation (20%), and testing (20%) sets. The *DeepLab-V3* CNN [9] was trained on the training data set for 100 epochs, while retaining the model that performed best on the validation set.

The training of the *DeepLab-V3* CNN employed a stochastic gradient descent optimizer with a learning schedule for weight decay. The best *mean Intersection-over-Union* (*mIoU*) measure was achieved with the *ResNet* backbone of *DeepLab-V3* which was more than twice of that achieved using the *Inception* backbone. The *mIoU* measure achieved using the *ResNet*

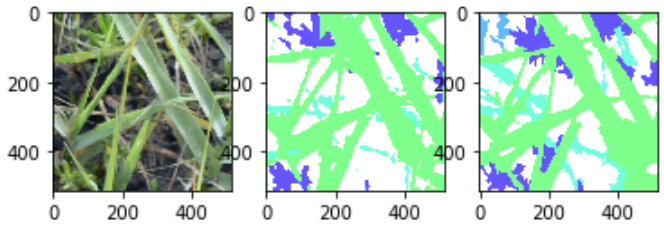

Fig. 6. Semantic segmentation results for *DeepLab-V3*, showing from right to left : original image, output mask and target mask

backbone on the test data set was 0.54 with an overall pixel accuracy of 85%, and macro-averaged precision and macro-averaged recall scores of 0.570 and 0.607 respectively. For abundant classes such as *Spartina, Sarcocornia* and *Batis*, the semantic segmentation was highly accurate as illustrated in Fig. 6 and reflected in the high overall pixel-level accuracy (Table IV). However, the less abundant classes were either poorly predicted (*Borrichia*) or neglected entirely (*Limonium* and *Juncus*), which negatively impacted the macro precision and macro recall metrics. The current performance of the semantic segmentation approach is sufficiently accurate that it could be applied in ecological research to estimate the abundance and spatial distribution of abundant classes, but it clearly needs to be further improved before it can be used to study the rarer taxa. Additional training data on these less abundant classes is likely the best way to improve the performance of semantic image segmentation.

E. Comparison of Approaches

Three distinct tasks, i.e., presence/absence computation, percent cover computation, and semantic image segmentation, were evaluated as alternative approaches to assess the abundance and distribution of plants in salt marsh images. There are multiple potential scientific applications for these distinct approaches and our intent was not necessarily to determine the *best* approach but rather to explore the performance of these approaches and identify the underlying trade-offs. When the best performing CNN architecture was employed, all of the approaches performed reasonably well at classifying the abundant classes, i.e., *Spartina* and *Sarcocornia*, with $> 85\%$ precision and recall. The performance was observed to decline for less abundant classes. The decline was more dramatic in the case of percent cover computation and semantic image segmentation. For example, semantic image segmentation was observed to completely ignore the *Limonium* and *Juncus* classes whereas presence/absence computation attained relatively high values of precision and recall for both classes compared to the other approaches. On many ecological tasks inter-agreement between human annotators is 70-90% indicating that in many cases our automated classifiers are likely to be as accurate as humans [5].

In essence, there was a clear trade-off observed between the performance (in terms of precision and recall) and spatial resolution of estimation with higher-resolution approaches exhibiting lower performance. Table IV illustrates this trade-off showing that *accuracy* (i.e., micro-averaged *f-1* score) and the *f-1 score* (macro-averaged *f-1* score) decrease from the presence/absence computation task to the semantic image segmentation task whereas the *estimation resolution* (i.e., estimations per pixel) increases. One caveat of these comparisons is that different data sets were used for each approach. We put roughly the same effort in terms of human annotation hours into producing each manually annotated data set. Consequently, the comparison in Table IV represents trade-offs for a similar amount of training data. It is expected that the performance of any of the approaches could be increased, to some extent, with additional training data. However, the difficulty of accurately detecting salt marsh plants at high resolution might require machine learning approaches capable of representing concepts such as ambient lighting, object boundaries and object shapes.

TABLE IV
COMPARISON OF RESULTS FROM DIFFERENT APPROACHES

Approach	***Model***	***Accuracy***	***f-1 score***	***Resolution***
Presence/Absence	*ResNext*	0.929	0.853	e-7
Percent Cover	*ResNext*	0.857	0.770	e-3
Segmentation	*DeepLab-V3*	0.849	0.587	1

F. Application example: Plant distribution across an elevation gradient

As a demonstration of the potential use of these methods, we employed the *ResNext101* classifier designed for presence/absence computation for all the images from the Sapelo Island marsh data set. The images were split into 15 sections (3 rows $\times$ 5 columns), as is done in the training procedure. The presence/absence classifier was used to determine which plant species were present in each image section. The total number of image sections (ranging from 0 to 15) in which a plant was present in each image was used as a semi-quantitative metric of plant abundance. This semi-quantitative index was averaged over all images for a given row to produce an estimate of plant abundance as a function of the row number, and plotted for all 80 rows as shown in Fig. 7. The results show a diverse plant community in the high-elevation marsh regions (row numbers < 25) transitioning to *Spartina* dominance in the low-elevation marsh regions (row numbers $>$ 35) which is consistent with the expected distribution. Some spurious *Sarcocornia* predictions were observed in the low-elevation marsh regions (row numbers > 35) and efforts are currently underway to provide additional training data in this section to improve the quality of estimations. Nonetheless, this example illustrates the presence/absence computation method's potential to rapidly assess plant community structure across environmental (i.e., elevation) gradients in a salt marsh. Future work aims to apply these tools to assess changes in plant community structure over time at this Sapelo Island site and to extend the spatial scale of the sampling to additional sites.

V. FUTURE WORK

In future work, we plan to introduce cross-talk between percent cover computation, presence/absence computation and

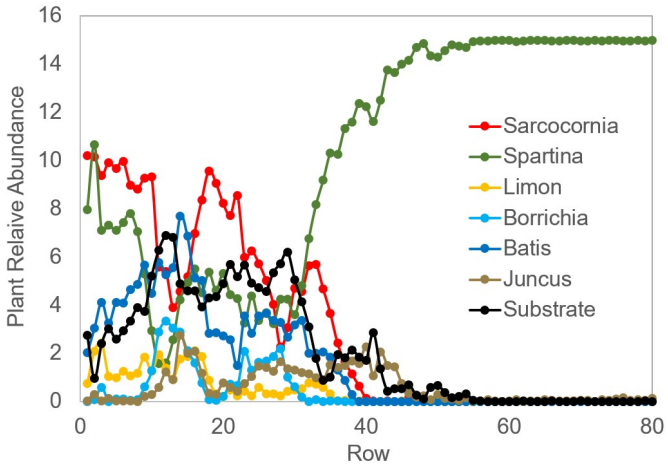

Fig. 7. Row-wise distribution of plants across an elevation gradient

semantic image segmentation. Rather than comparing these approaches based on results on their individual test data sets, we plan to devise a common testing data set to ascertain the relative performance of these approaches on producing a spatial plant class distribution. Since our plant classes are interweaved and require deep understanding of many high-order concepts, we argue that the best CNN models for our approach must possess generality. We also hypothesize that the best way to create neural networks capable of generality is through neuromodulatory approaches [3]. Our proposed data set will provide a test for novel image analysis approaches.

Acknowledgment

This research was supported by grants from the US National Science Foundation (OCE-1832178, DBI-2016741).

References

[1] A.C. King, S. M. Bhandarkar, and B. M. Hopkinson, A comparison of deep learning methods for semantic segmentation of coral reef survey images, *Proc. IEEE CVPR*, 2018.

[2] E. Ayrey, and D.J. Hayes, The use of three-dimensional convolutional neural networks to interpret LiDAR for forest inventory, *Remote Sensing*, 10(4):649, pp. 1-16, 2018.

[3] S. Beaulieu, L. Frati, T. Miconi, J. Lehman, K.O. Stanley, J. Clune, and N. Cheney, Learning to continually learn, *arXiv:2002.09571*, 2020.

[4] O. Beijbom, P.J. Edmunds, D.I. Kline, B.G. Mitchell, and D. Kriegman, Automated annotation of coral reef survey images, *Proc. IEEE CVPR*, pp. 1170-1177, 2012.

[5] O. Beijbom, P.J. Edmunds, C. Roelfsema, J. Smith, D.I. Kline, B.P. Neal et al., Towards automated annotation of benthic survey images: variability of human experts and operational modes of automation, *PLoS One* 10(22), 2015.

[6] M.D. Bertness, and A.M. Ellison, Determinants of Pattern in a New England Salt Marsh Plant Community, *Ecological Monographs*, 57(2) pp. 129-147, June 1987.

[7] C. Bowley, A. Andes, S. Ellis-Felege, and T. Desell, Detecting wildlife in uncontrolled outdoor video using convolutional neural networks, *Proc. IEEE Intl. Conf. E-Science* pp. 251-259, 2016.

[8] P.G. Brodrick, A.B. Davies, and G.P. Asner, Uncovering ecological patterns with convolutional neural networks, *Trends Ecol. Evol.* 34, pp. 734-745, 2019.

[9] L-C Chen, G. Papandreou, F. Schroff, and H. Adam, Rethinking atrous convolution for semantic image segmentation, *arXiv:1706.05587*, 2017.

[10] Y. Chen, J. Li, H. Xiao, X. Jin, S. Yan, and J. Feng, Dual path networks, *Proc. NIPS*, pp. 4470–4478, 2017.

[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, ImageNet: A large-scale hierarchical image database, *Proc. IEEE CVPR*, 2009.

[12] D. Han, J. Kim, and J. Kim, Deep pyramidal residual networks, *Proc. IEEE CVPR*, pp. 6307–6315, 2017.

[13] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, *Multimedia Tools and Applications*, 77(9), pp. 10437–10453, Dec. 2015.

[14] J. Heinzel, and B. Koch, Investigating multiple data sources for tree species classification in temperate forest and use for single tree delineation, *Intl. Jour. Applied Earth Observation and Geoinformation* 18, pp. 101-110, 2012.

[15] G. Huang, Z. Liu, L. Van Der Maaten, and K.Q. Weinberger, Densely connected convolutional networks, *Proc. IEEE CVPR*, pp. 2261–2269, 2017.

[16] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, A survey of the recent architectures of deep convolutional neural networks, *arXiv:1901.06032*, 2019.

[17] A. Krizhevsky, I. Sutskever, and G. Hinton, Imagenet classification with deep convolutional neural networks, *Proc. NIPS*, 2012.

[18] Y. Lecun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, and L.D. Jackel Backpropagation applied to handwritten zip code recognition, *Neural Computation*, 1(4), pp. 541-551, 1989.

[19] Y. LeCun, Y. Bengio, and G. Hinton Deep learning, *Nature*, 521, pp. 436-444, 2015.

[20] M. Lin, Q. Chen, and S. Yan, Network in network, *arXiv:1312.4400*, pp. 1–10, 2013.

[21] E. McLeod, G.L. Chmura, S. Bouillon, R. Salm, M. Björk, C.M. Duarte, C.E. Lovelock, W.H. Schlesinger, and B.R. Silliman, A blueprint for blue carbon: toward an improved understanding of the role of vegetated coastal habitats in sequestering CO_2", *Frontiers in Ecology and Environment*, 2011.

[22] W. Mitsch, and J. Gosselink, *Wetlands*, Wiley, 5th edition, 2015.

[23] S.C. Pennings, M-B Grant, and M.D. Bertness, Plant zonation in low-latitude salt marshes: disentangling the roles of flooding, salinity and competition, *Jour. Ecology*, 2004.

[24] R. Soltani, and H. Jiang, Higher-order recurrent neural networks, *arXiv:1605.00064*, 2016.

[25] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, Going deeper with convolutions, *Proc. IEEE CVPR*, pp. 1–9, 2015.

[26] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, Rethinking the Inception architecture for computer vision, *Proc. IEEE CVPR*, pp. 2818–2826, 2016.

[27] H. Touvron, A. Vedaldi, M. Douze, and H. Jegou, Fixing the train-test resolution discrepancy, *Proc. NIPS*, 2019.

[28] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, Residual attention network for image classification, *Proc. IEEE CVPR*, 2017.

[29] K. Wasson, K. Raposa, M. Almeida, K. Beheshti, J.A. Crooks, et al. Pattern and scale: evaluating generalities in crab distributions and marsh dynamics from small plots to a national scale, *Ecology*, 100(10), 2019.

[30] B.G. Weinstein, A computer vision for animal ecology, *Jour. Animal Ecology*, 87, pp. 533-545, 2018.

[31] B.G. Weinstein, Scene-specific convolutional neural networks for video-based biodiversity detection, *Methods in Ecology and Evolution*, 9, pp. 1435-1441, 2018.

[32] B.G. Weinstein, S. Marconi, S. Bohlman, A. Zare, and E. White, Individual tree-crown detection in RGB imagery using semi-supervised deep learning neural network, *Remote Sensing*,11:13, 2019.

[33] I.D. Williams, C.S. Couch, O. Beijbom, T.A. Oliver, B. Vargas-Angel, B.D. Schumacher, and R.E. Brainard, Leveraging automated image analysis tools to transform our capacity to assess status and trends of coral reefs, *Frontiers in Marine Science*, 6:10.3389/fmars.2019.00222, 2019.

[34] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, Aggregated residual transformations for deep neural networks, *Proc. IEEE CVPR*, pp. 5987–5995, 2017.