SEMANTIC SEGMENTATION OF UNDERWATER SONAR IMAGERY WITH DEEP LEARNING

Maryam Rahnemoonfar & Dugan Dobbs*

School of Engineering and Computing Sciences, Texas A&M University-Copus Christi, TX

ABSTRACT

Majority of deep learning methods are developed for RGB imagery. However, for many applications such as detecting objects underwater other types of sensors such as sonar or radar are required. One of the most precise sensors to map the seagrass disturbance is side scan sonar. Here we developed a new deep learning framework based on dilated convolution, dense module, and inception to perform semantic segmentation for automatic extraction of potholes in underwater sonar imagery. We tested our proposed approach on a collection of underwater sonar images taken from Laguna Madre in Texas. Experimental results in comparison with the ground-truth and state-of-the-art semantic segmentation methods show the efficiency and improved accuracy of our proposed method.

1. INTRODUCTION

In recent years Convolutional Neural Networks (CNN) have been widely used in computer vision research including classification [1, 2], object recognition [3], counting [4, 5] and semantic segmentation [6]. Majority of deep learning methods are developed for RGB imagery. However for many applications such as detecting objects under water [7, 8] or under ice [9, 10] other types of sensors such as sonar or radar are required. This research investigate developing a novel semantic segmentation technique based on dense inception network for identifying pothole in seagrass.

The widespread loss of seagrass beds is largely caused by the rapid expansion of human populations around coastal waterways. Detection of seagrass with optical remote sensing (both satellite imagery and aerial photography) is complicated by the fact that light is attenuated as it passes through the water column and reflects back from the benthos.

Underwater acoustics mapping produces a high definition, two-dimensional sonar image of seagrass ecosystems. The intensity and contours of the image are then determined by the amount of time a sound wave takes to return to the transducer and how the waves were reflected.

Several pattern recognition techniques have been applied on sonar images mainly for detecting concrete objects on sandy sea floor based on hierarchical MRF model [11], active contour [12], Bayesian classifier [13], Gauss-Markov random field model and level-set [14].

Side scan sonar (SSS) has been recently used for detecting the extent of seagrass beds and mapping its disturbance [7, 8] based on mathematical morphology and level set. all of the aforementioned techniques are based on hand-crafted feature engineering.

Here we developed a new deep learning framework based on Dilated convolution, Densenet, and Inception to perform semantic segmentation for automatic extraction of potholes in underwater sonar imagery. We tested our proposed approach on a collection of underwater sonar images taken from Laguna Madre in Texas. Experimental results in comparison with the ground-truth and state-of-the-art semantic segmentation methods show the efficiency and improved accuracy of our proposed method.

2. METHODOLOGY

The overall architecture of our method is depicted in Figure 1. It includes three different modules: 1) Atrous block, 2) Dense module, 3) DeconvXY block. The specific parameters of the network is shown on Table1.

Dilated (Atrous) Convolutions: Atrous convolutions help to build scale invariance and provide the network with a larger viewing window. We used this module to extract multi-scale information about our original image. This assists the early level dense modules with contextual information. Our Atrous Block is comprised of a series of convolutions with different dilations that share kernel weights. Gradients are calculated only for the first convolution. Each output is concatenated together, and then a final convolution is performed in order to help disrupt the gradient shifts from our unconventional method and compress the information to prevent parameter explosion in our coming Dense Blocks.

Dense module: Dense convolutional networks (DenseNet) is build with feed-forward connections between each layer to every other layers [15]. Such network is designed to alleviate the vanishing-gradient problem while maximizing the information flow between network layers. It requires fewer parameters by reducing redundant feature map learning. Moreover, each layer has direct access to the gradients from the loss function and the original input signal which will provide an

* Corresponding author.

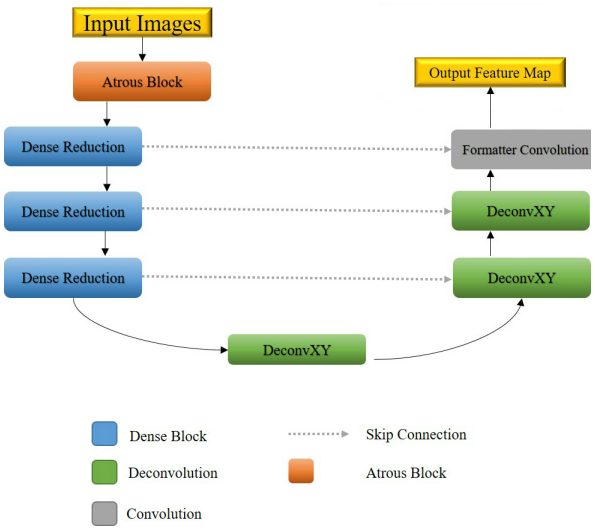

Fig. 1. Our Network Architecture

Network State	Shape of State
Images	(440,200,1)
Output of Atrous Block	(440,200,12)
Output of DenseBlock1 (Pre-Pooling)	(440,200,92)
Output of DenseBlock1 (Post-Pooling)	(220,100,92)
Output of DenseBlock2 (Pre-Pooling)	(220,100,172)
Output of DenseBlock2 (Post-Pooling)	(110, 50,172)
Output of Denseblock3 (Pre-Pooling)	(110, 50,300)
Output of DenseBlock3 (Post-Pooling)	(22, 10,300)
Output of DeconvXY3	(110, 50, 36)
Output of DeconvXY2	(220,100,150)
Output of DeconvXY1	(440,200,160)

Table 1. Network State Sizes

implicit formulation between layers. This is also shown to perform some regularization effect which may be helpful for over-fitting problems.

DeconvXY: Our proposed Inception-Deconvolution, referred to as DeconvXY due to the way channels perform their deconvolutions can be seen on Figure 2. This module uses three different channels of transposed convolutions which upsample the network at different rates. The output feature map of each transposed convolution is the inverse of how much upsampling occurs. As an example, a 1×1 network with 4 feature maps, $(1 \times 1 \times 4)$, being upsampled by a rate of 2 would become $(2 \times 1 \times 2)$ and $(1 \times 2 \times 2)$ after the first transposes. The final shape would be $2 \times 2 \times 1$ at the end of each channel. These outputs are concatenated together and then have dropout performed with a keep percentage of .85, then fed to the next module in the network. This type of data representation ensures that these modules are limited in scope to translating information into different spatial dimensions, rather than creating new inferences.

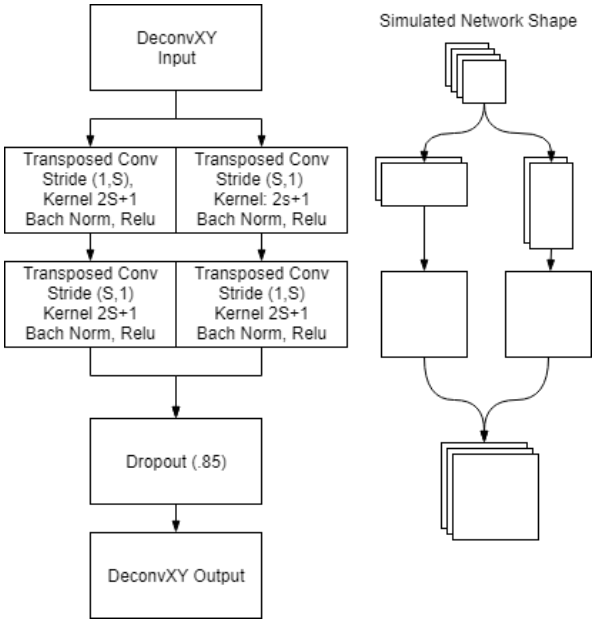

Fig. 2. Inception-Deconvolution Module (DenconvXY)

3. EXPERIMENTAL RESULTS

Data was collected from the seagrass beds of the Lower Laguna Madre in southern Texas in 2016 from an average depth of 75 cm. A specialized side scan sonar unit was constructed consisting of a towfish with two Lowrance Structure Scan HD LSS-2 transducers, a Dual Beam 200 kHz down- imaging transducer connected to a Humminbird 998C HD SI control unit. A total of six transects approximately 5000 x 60000 pixels and overlapping by 50% were processed individually to image an area of approximately 88,000 m2. The images used in this experiment were large transects. Each transect was 600 meters long, and spaced 20 meters apart with a horizontal swath of approximately 50 m. Total area covered between all six transects was 88,000 square meters (150mX600m).

To apply our proposed method on these images, they were re-sized to a uniform width of 4400 then chopped into varying amounts of 4400×2000 images, resulting in 167 total images, with 117 for training and 50 for testing. These images are augmented by flipping along the horizontal and vertical axes, increasing the number to 468 and 200.

For training, we evaluate our loss using cross entropy and optimize using the Adam optimizer, a learning rate of 1e-3, and a batch size of 8. The normal training, validation, testing format is replaced with training, testing for each epoch until the error begins to increase. The optimal amount of training on this dataset for our architecture was 14 epochs. For evaluating metrics we use accuracy, intersection over union (IOU), precision, recall, and F1 score for each class then calculate the mean per class metrics.

Detailed quantitative results can be viewed on Table 3,

Metric	Formula
Acc_n	$\frac{TP_n+TN_n}{TP_n+TN_n+FP_n+FN_n}$
IOU_n	$\frac{TP_n}{TP_n+FP_n+FN_n}$
$Prec_n$	$\frac{TP_n}{TP_n+FP_n}$
Rec_n	$\frac{TP_n}{TP_n+FN_n}$
$F1_n$	$2\frac{Prec_n*Rec_n}{Prec_n+Rec_n}$

Table 2. Evaluation Metrics

Class	Seagrass	Pothole	Mean
Accuracy	95.84	95.78	95.81
Precision	82.81	96.89	89.85
Recall	70.61	98.42	84.51
IOU	61.58	95.40	78.49
Specificity	98.47	74.36	86.42
F1	76.22	97.68	86.93

Table 3. Our Per Class Results

and formulas used to obtain these results can be seen on Table 2. Mean per class metrics between our proposed method and FCN [16] can be viewed on Table 4. Training was completed in under two hours utilizing a single machine leveraging a single GTX-1080-TI.

A side by side comparison with FCN [16], ground truths, and the proposed network can be found on Figure 3.

A large benefit of our proposed architecture is the smoothness and lack of artifacting in the output feature map due to skip connections. In fact, after the Atrous Block, the gradient flows almost unimpeded to every other point in the network. The output from the Atrous Block is maintained through each Dense Reduction with max pooling, through skip connections, and through each convolution in each Dense Reduction.

Our proposed network outperforms FCN on every metric, however, the point where the proposed network excels are the number of parameters. The proposed network uses approximately 9.2 million, while FCN [16] utilizes 141 million parameters. While FCN is better at extracting features like boat scars, the increased overhead and downgrade of metrics is not effective enough.

Network	Ours	FCN
Accuracy	**95.81**	95.38
Precision	**89.85**	88.90
Recall	**84.51**	82.22
IOU	**78.49**	76.16
Specificity	**86.42**	84.49
F1	**86.93**	85.16

Table 4. Mean Per Class Metrics

4. CONCLUSION

In this project we developed a novel architecture for semantic segmentation of pothhole in sonar imagery. Our method outperformed the state-of-the-art techniques in terms of both accuracy and efficiency. While our proposed network outperforms previous networks in this segmentation task, feature extraction of argumentative features could be improved. Other avenues of improvement are further optimization of the network to decrease parameters, increase portability, and speed. The intended goal of this system is to be able to deploy it on a mobile platform in order to generate real time sensor analysis for environmental and fishery preservation.

5. REFERENCES

[1] C. Sheppard and M. Rahnemoonfar, "Real-time scene understanding for uav imagery based on deep convolutional neural networks," in *Geoscience and Remote Sensing Symposium (IGARSS), 2017 IEEE International*. IEEE, 2017, pp. 2243–2246.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.

[3] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, *Simultaneous Detection and Segmentation*. Cham: Springer International Publishing, 2014, pp. 297–312.

[4] M. Rahnemoonfar and C. Sheppard, "Deep count: fruit counting based on deep simulated learning," *Sensors*, vol. 17, no. 4, p. 905, 2017.

[5] ——, "Real-time yield estimation based on deep learning," in *Autonomous Air and Ground Sensing Systems for Agricultural Optimization and Phenotyping II*, vol. 10218. International Society for Optics and Photonics, 2017, p. 1021809.

[6] M. Mostajabi, P. Yadollahpour, and G. Shakhnarovich, "Feedforward semantic segmentation with zoom-out features," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 3376–3385.

[7] M. Rahnemoonfar, A. F. Rahman, R. J. Kline, and A. Greene, "Automatic seagrass disturbance pattern identification on sonar images," *IEEE Journal of Oceanic Engineering*, 2018.

[8] M. Rahnemoonfar, M. Yari, A. Rahman, and R. Kline, "The first automatic method for mapping the pothole in seagrass," in *Proceedings of the IEEE Conference on*

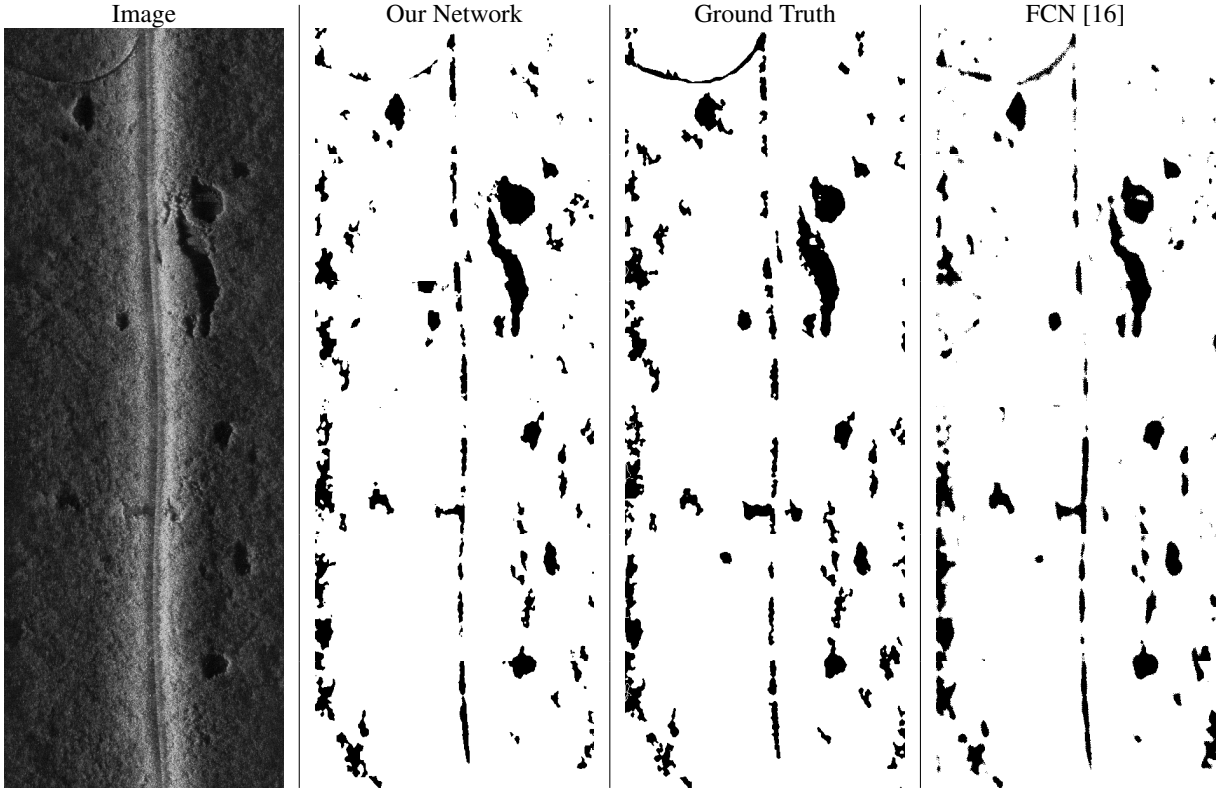

Fig. 3. Comparison between our result (second column), Ground Truth, and FCN [16]

Computer Vision and Pattern Recognition Workshops, 2017, pp. 73–80.

[9] M. Rahnemoonfar, G. C. Fox, M. Yari, and J. Paden, "Automatic ice surface and bottom boundaries estimation in radar imagery based on level-set approach," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 9, pp. 5115–5122, 2017.

[10] M. Rahnemoonfar, A. A. Habashi, J. Paden, and G. C. Fox, "Automatic ice thickness estimation in radar imagery based on charged particles concept," in *Geoscience and Remote Sensing Symposium (IGARSS), 2017 IEEE International*. IEEE, 2017, pp. 3743–3746.

[11] M. Mignotte, C. Collet, P. Perez, and P. Bouthemy, "Sonar image segmentation using an unsupervised hierarchical mrf model," *IEEE transactions on image processing*, vol. 9, no. 7, pp. 1216–1231, 2000.

[12] M. Lianantonakis and Y. R. Petillot, "Sidescan sonar segmentation using active contours and level set methods," in *Oceans 2005-Europe*, vol. 1. IEEE, 2005, pp. 719–724.

[13] W.-M. Tian, "Automatic target detection and analyses in side-scan sonar imagery," in *Intelligent Systems, 2009. GCIS'09. WRI Global Congress on*, vol. 4. IEEE, 2009, pp. 397–403.

[14] X.-F. Ye, Z.-H. Zhang, P. X. Liu, and H.-L. Guan, "Sonar image segmentation based on gmrf and level-set models," *Ocean Engineering*, vol. 37, no. 10, pp. 891–901, 2010.

[15] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," 2016. [Online]. Available: http://arxiv.org/abs/1608.06993

[16] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 07-12-June-2015, pp. 3431–3440, 2015.