Learning Objectness from Sonar Images for Class-Independent Object Detection

Matias Valdenegro-Toro[1]

Abstract— **Detecting novel objects without class information is not trivial, as it is difficult to generalize from a small training set. This is an interesting problem for underwater robotics, as modeling marine objects is inherently more difficult in sonar images, and training data might not be available apriori. Detection proposals algorithms can be used for this purpose but usually requires a large amount of output bounding boxes. In this paper we propose the use of a fully convolutional neural network that regresses an objectness value directly from a Forward-Looking sonar image. By ranking objectness, we can produce high recall (96 %) with only 100 proposals per image. In comparison, EdgeBoxes requires 5000 proposals to achieve a slightly better recall of 97 %, while Selective Search requires 2000 proposals to achieve 95 % recall. We also show that our method outperforms a template matching baseline by a considerable margin, and is able to generalize to completely new objects. We expect that this kind of technique can be used in the field to find lost objects under the sea.**

I. Introduction

Perception is a fundamental problem in Robotics, particularly for underwater environments. Many real-world tasks require a robot to first find an object and then identify it. A more difficult task is to find an object from an abstract description or to find all objects in the scene irrespective if they are from known classes or not.

Finding and detecting objects under water is also a hard problem, mostly due to light absorption by water that degrades the use of optical sensors, and the difficulties associated at interpreting acoustic sensor outputs. There have been large advances at detecting objects in sonar images [1], but they usually make strong assumptions on object shape or require shadow/highlight segmentation [2]. Techniques to detect objects if a medium sized training set are available [3], but many tasks require a robot to find an "novel" object, where training samples are not available a priori. Examples of this are finding submerged airplane wrecks, and detecting marine debris.

This paper deals with the problem of building a class-agnostic object detector for sonar images. In the computer vision literature, this is called detection proposals [4], but in that context they are only used in order to "guide" the object detection process and improve localization results. We believe that detection proposals are also useful in underwater robotics on its own, where the idea is to detect objects, even if their shape or content has not been seen before by the system. In this case, class predictions are not available, but novel objects can still be identified by the detection system as different than background.

This work expands on our previous work [5], where we introduced a basic version of this system. This work introduces the following contributions:

- We propose the use of objectness ranking in order to increase the adaptivity of the system to different environments.
- We introduce a new neural network objectness regressor that requires considerably less parameters and is fully convolutional, resulting in a four times improvement in computation time.
- We perform a more thorough evaluation and compare with the state of the art, showing that our method outperforms other methods and requires less proposals to achieve high recall.

These contributions result in more appropriate technique for underwater robot perception.

II. Related Work

The underwater perception literature contains many techniques to detect objects in sonar images. A very popular option is the use of template matching [1], where a set of templates is cross-correlated with the input image, and this produces maximum correlation where the object is located. A threshold is usually set to avoid false positives.

Another option is using classic computer vision methods, like the boosted cascade of weak classifiers [2], but this only works well in objects that produce large sonar shadows, as Haar features correlate very well with this feature.

Neural networks have also been used [3], where a CNN is trained on image patches and used in a sliding window fashion on a test image. This technique works quite well in terms of accuracy but it produces a large amount of false positives. An end-to-end multi-task approach [6] improves false detections by explicitly modeling the detection process as proposals.

In general, CNN-based techniques are able to model more complex objects than classic computer vision methods. Template matching in particular is not able to model objects more complex than underwater mines, such as marine debris.

The concept of detection proposals is introduced in the computer vision literature [7], where instead of using an expensive sliding window to detect objects, the detection process can be "guided" by a subset of windows that are likely to contain objects. A detection proposals algorithm infers these bounding boxes (also called proposal) from image content. Proposals are also linked to the concept of

[1]Matias Valdenegro-Toro is with the German Research Center for Artificial Intelligence, Robotics Innovation Center. Robert-Hooke-Strasse 1, 28359, Bremen, Germany. matias.valdenegro@dfki.de

"objectness" [8], where the authors define it as "quantifying how likely it is for an image window to contain an object of any class". A set of predefined cues are combined in order to produce objectness, which can be used to discriminate between object and background windows.

EdgeBoxes [9] is a proposals technique that uses a structured edge detector, which extracts high quality edges. Edges are then grouped to produce object proposals that can be scored by predefined techniques. This method is very fast but needs a large amount of proposals to produce high recall. Selective Search [10] takes a different approach, by doing super-pixel segmentation and using a set of strategies to merge super-pixels into detection proposals. It is quite slow but it can achieve very high recall with a medium number of output proposals.

Neural networks have also been used to model detection proposals. The best teachnique is the Region Proposal Network from the Faster R-CNN object detection framework [11]. The RPN module regresses bounding box coordinates and outputs a binary decision corresponding to object vs background. The RPN works quite well on color images and improves the state of the art in the PASCAL VOC 2007/2012 datasets, but we have not been able to train such modules for proposals on sonar images, mostly likely due to the small scale datasets that we have.

While there are established techniques for detection proposals in color images, these are not directly transferrable to sonar images. Bounding box regression techniques cannot be trained unless a large dataset ($\sim$ 1M images) is available for pre-training. The typical dataset of sonar images ranges in the thousands, preventing the use of such techniques. We have developed a simple objectness regressor [5] using neural networks that works well for detection proposals, but it is computationally expensive as features are not shared across neural network evaluations, and simple thresholding of objectness values might not generalize well across environments.

A bigger concern for robot perception is that most detection proposals techniques require a large number of output proposals to achieve high recall. This means that most proposals might not correspond to actual objects in the image, which defeats the purpose of a proposals method over a simple sliding window. In this work we evaluate how high recall can be achieved with a low number of output proposals by ranking objectness instead of using a fixed threshold.

III. Learning Objectness from Sonar Image Patches

Objectness is an abstract concept that quantifies the property that an image window contains an object. A window containing an object should be assigned a high objectness score, while a window with only background should receive a low objectness score.

Our method is based on the idea that objectness can be estimated from an image window/patch. Given ground truth objectness values, an objectness regressor can be trained on such data to learn the relationship between image content and an abstract objectness score concept. This corresponds to a data-driven approach.

Training Data Generation. We compute ground truth objectness as follows. We run a $n \times n$ sliding window with a stride of s pixels in each direction, and for each ground truth bounding box in the image, we assign a positive objectness score o to the sliding window that has the highest Intersection-over-Union score (IoU, Eq 1). We also assign a positive objectness score to any sliding window with IoU ≥ 0.5. This is intended to introduce variety in the range of objectness values.

$$\text{IoU}(A, B) = \frac{\text{area}(A \cap B)}{\text{area}(A \cup B)} \tag{1}$$

The IoU score is commonly used in computer vision to evaluate object detection algorithms [9] [12]. Typically in order of 5 to 10 windows with positive objectness are generated for each ground truth bounding box. To generate negative objectness windows, we randomly sample $N = 10$ windows that have a maximum IoU with the ground truth bounding boxes of ϵ. All negative objectness windows receive a zero objectness score. Positive and negative windows are cropped and stored as a labeled dataset to train an objectness regressor.

The final ground truth objectness score is obtained as:

$$\text{objectness}(\text{iou}) = \begin{cases} 1.0 & \text{if iou} \geq 1.0 - \epsilon \\ \text{iou} & \text{if } 1.0 - \epsilon < \text{iou} < \epsilon \\ 0.0 & \text{if iou} \leq \epsilon \end{cases} \tag{2}$$

The motivation for using Eq 2 is to expand the range of available objectness scores. While the IoU is in the $[0, 1]$ range, obtaining a IoU score close to 1.0 is very unlikely, as it would imply a near-perfect match between the ground truth and the sliding window. We introduce a tolerance where any IoU bigger than $1.0 - \epsilon$ is considered equivalent as the maximum objectness range. The lower threshold is to remove any window that might not have enough intersection with the ground truth. In our experiments we use $\epsilon = 0.2$. Any IoU value between the lower and upper thresholds is kept directly as the objectness score in order to introduce variability into the ground truth objectness scores.

Network Architectures. We use two CNN models that take a 96×96 one-channel sonar image patch as input, and output an objectness score in the $[0, 1]$ range. The first model was previously proposed by us [5] and contains approximately 900K parameters. In order to apply this model to full-size sonar image, we used a sliding window. The second model only contains 20K parameters and is fully convolutional, which allows it to take a full-size sonar image and output objectness for each pixel, sharing computation and avoiding computational performance issues. This model also has the advantage of allowing variable-sized images.

We use the following notation. Conv(n, s) a 2D Convolutional module with n square filters of size s, MP(s) a Max-Pooling module with subsampling size s, FC(n) a fully connected layer with n output neurons.

The first CNN model is based on LeNet [13], with two stacks of convolutional and max-pooling layers, and two fully connected layers. The network architecture is shown in Fig. 1. ReLU is used as activation except at the output layer, where a sigmoid activation is used.

The second model is based on SqueezeNet [14], from which we have derived the Tiny module [15] that allows a model with large expressivity and a low number of parameters. The model shown in Fig. 2 is trained, and for test-time inference, it is modified to construct a fully convolutional version of it. This is done by replacing the last fully connected (FC) layer with a Conv(1, 24×24) layer and reshaping the weights [16] to fit convolutional filters. Given an input image, this model produces an output objectness map that is smaller than the input, due to the use of max-pooling. We up-sample the objectness map back to the original input size with bilinear interpolation. The down-scaling factor is defined by the model architecture (Fig. 2) and the number of max-pooling layers. Minimizing this scaling factor inherent in the model is what motivates the use of a simplified model, with a single fully connected layer.

Training. Both models are trained in the same way, using a mean squared error loss with the ADAM optimizer [17] and a learning rate $\alpha = 0.01$. Training stops when the loss converges, determined by early stopping on a held-out validation set, which usually happens after 15-20 epochs. No pre-training or fine-tuning is performed, and all weights start from random initialization.

IV. Detection Proposals from Objectness Scores

We propose two methods to convert objectness scores into dense detection proposals. First, a 96×96 with stride $s = 4$ sliding window is applied to the input image and objectness scores are computed for each window. Then candidate windows are filtered into detection proposals by a given method:

- **Thresholding**. Any window with objectness bigger than a threshold T_o is output as a proposal. The value of T_o can be tuned in a validation set given a recall target, but in general this parameter determines a trade-off between recall and number of proposals.
- **Ranking**. All candidate windows are sorted by decreasing objectness and the top k ones are output as detection proposals. This method introduces a quality parameter k, which is directly related to recall and can be tuned to maximize recall under a given number of proposals.

The sliding window approach is only used with the CNN model in order to build an objectness map, while the FCN model implicitly does a sliding window as convolution, and outputs the objectness map directly. After deciding proposal windows, we apply non-maximum suppression with a given threshold T_s in order to reduce duplicate detections. We use $T_s = 0.8$ as a good compromise between number of output proposals and recall. Note that our method can potentially produce proposals at multiple scales, but in this work we only report results with a single scale (given by the 96×96 window).

V. Experimental Evaluation

In this section we evaluate our methods and provide comparisons with the state of the art.

Data. We use a marine debris dataset[1] of 2000 full-sized sonar images obtained from an ARIS Explorer 3000 Forward-Looking sonar. They were captured at the Ocean Systems Lab (Heriot-Watt University) water tank and contain marine debris objects such as cans, bottles, tires, etc. After extracting patches using a sliding window with a stride $s = 4$ from 1300 full-sized sonar images, we obtained 51563 training and 22137 validation samples (70%/30% split). The remaining 700 full-size sonar images are used for evaluation and comparison of detection proposal techniques.

Metrics. The typical metric [10] to evaluate detection proposals is the recall:

$$R = \frac{TP}{TP + FN} \tag{3}$$

Where TP is the number of true positives, and FN is the number of false negatives. A proposal is considered a correct if the IoU (Eq 1) between proposal and ground truth bounding boxes is greater than some threshold T_d. The most common value for this threshold is $T_d = 0.5$.

Recall is used because a detection proposal method typically generalizes well and it can generate many bounding boxes that correspond to real objects in the image, but are not labeled as such. Precision is not typically evaluated as unlabeled objects are considered false positives [18] which would skew any evaluation. The area under the ROC curve (AUC) is also not appropriate for the same reasons. Note that this evaluation protocol can be "gamed" due to partially annotated datasets [19].

Another important metric is the number of output bounding boxes (also called proposals), as it is very easy to obtain high recall with a high number of proposals, but too many bounding boxes hurt the applicability of such methods as many boxes do not correspond to real objects. An ideal detection proposals technique would have high recall with a relatively low number of output proposals.

Baselines. There are no detection proposal techniques specifically designed for sonar images, which makes defining a baseline not trivial. We compare against our previous work [5], and we evaluate a number of baselines, namely: EdgeBoxes [9], Selective Search [10], and Cross-Correlation Template Matching [20].

EdgeBoxes extracts high-quality edges and groups them into object proposals at multiple scales. A score threshold is required, and the number of output proposals can also be tuned. We evaluate both parameters by selecting a low threshold 0.0001 at a fixed number of 300 proposals, and using a 0.0 score threshold and varying the number of output proposals.

Selective Search uses a predefined set of strategies that merge super-pixels into detection proposals. We evaluate both

[1]The full dataset is available at `https://github.com/mvaldenegro/marine-debris-fls-datasets/releases/`

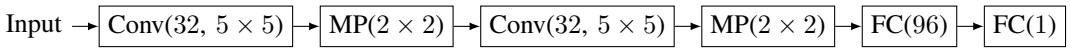


Fig. 1. CNN model based on the LeNet Architecture. All layers use ReLU activation, except the last layer that uses a sigmoid function.

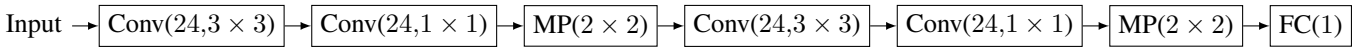


Fig. 2. FCN model based on the Tiny module [15]. All layers use ReLU activation, except the last layer that uses a sigmoid function.

the Quality and Fast configurations, with a variable number of output proposals. For both EdgeBoxes and Selective Search, we used the OpenCV ximproc [2] module implementation.

We also built a detection proposals algorithm using a cross-correlation similarity typically used for template matching. We randomly selected a set of $T = 100$ positive patches from the training set and computed the maximum cross-correlation between an input patch and all templates. This works as a pseudo-objectness measure and we use this score for both thresholding and ranking proposals.

Results. A comparison between CC template matching and CNN/FCN objectness with thresholding is shown in Fig. 3. Our results show that CC TM performs quite poorly, while objectness produced by CNN and FCN perform better, with slowly decreasing recall and number of proposals as T_o is increased. FCN performs slightly worse than CNN, indicated by requiring approximately two times the number of proposals to produce the same recall.

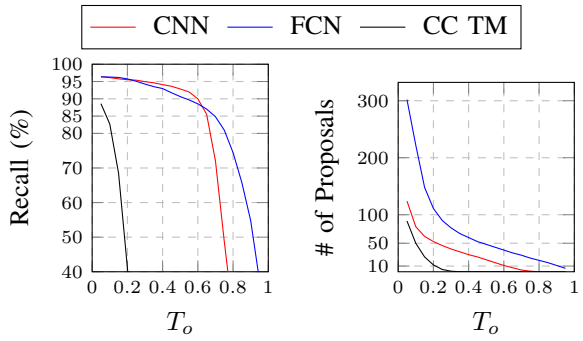


Fig. 3. Objectness thresholding results with CNN, FCN and CC TM objectness. CNN performs slightly better than FCN, while CC TM fails to generalize properly.

A comparison of objectness ranking is shown in Fig. 4. Results show again that CNN produces better objectness, reflected as requiring less proposals, but still FCN objectness can obtain 95 % recall with only 100 proposals per image. CC TM objectness saturates at 88 % recall if more than 110 proposals are output.

Thresholding and Ranking both have a best recall at 95 %, but ranking has the advantage of requiring less output proposals to achieve high recall, 40 for CNN and 80-100 for FCN, while thresholding requires considerable more proposals to achieve the recall target. This indicates that ranking can possibly adapt better to unknown environments, even as the objectness scores change.

[2]Available at https://github.com/opencv/opencv_contrib/tree/master/modules/ximgproc

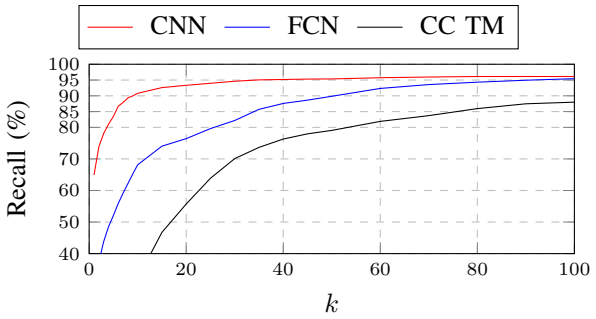


Fig. 4. Objectness ranking results with CNN, FCN, and CC TM objectness. CNN can obtain high recall from 20-40 proposals per image, while FCN requires more to generalize properly. CC TM needs even more proposals to obtain poor recall, which suggests that it is not a good choice for this problem.

Table I provides a global comparison with the state of the art. For each method, we determined the configuration that produces the best recall, the number of output proposals required to achieve such recall, and computation time as evaluated on a AMD Ryzen 7 1700 processor.

EdgeBoxes is by far the fastest method at 0.1 seconds to process one frame, and it produces the best recall, but doing so requires 5000 proposals per image. Selective Search Quality also obtains very good recall but with a large number of proposals. CC TM produces the lowest recall we observed on this experiment.

Our proposed techniques obtain very good recall with a low number of proposals per image. CNN-Ranking produces 96 % recall with only 80 proposals per image, which is 62 times less than EdgeBoxes with only a 1 % absolute loss in recall. Selective Search produces 1 % less recall than the best of our methods, but outputting 25 times more proposals.

In terms of computation time, EdgeBoxes is the fastest. FCN objectness is 4 times faster to compute than CNN objectness, due to the fully convolutional network structure, and it only requires a 1 % reduction in recall. CC Template Matching is also quite slow, at 10 seconds per image.

In Fig. 7 we show a small sample of CNN and FCN detections produced by objectness ranking. Our results show that classical sonar object detection techniques are not really appropriate for detections proposals, as they are slow and cannot achieve high recall. This is likely due to inability to model high object variation.

Number of Proposals vs Recall. Fig. 5 shows a recall comparison between all evaluated methods as we vary the number of output detection proposals. The best compromise between high recall and low number of proposals is CNN with objectness ranking. Cross-Correlation Template Matching

Method	Best Recall	# of Proposals	Time (s)
TM CC Threshold	91.83 %	150	10 ± 0.5
TM CC Ranking	88.59 %	110	10 ± 0.5
EdgeBoxes (Thresh)	57.01 %	300	0.1
EdgeBoxes (# Boxes)	**97.94** %	5000	0.1
Selective Search Fast	84.98 %	1000	1.5 ± 0.1
Selective Search Quality	**95.15** %	2000	5.4 ± 0.3
CNN-Threshold	**96.42** %	**125**	12.4 ± 2.0
FCN-Threshold	**96.33** %	300	3.1 ± 1.0
CNN-Ranking	**96.12** %	**80**	12.4 ± 2.0
FCN-Ranking	**95.43** %	**100**	3.1 ± 1.0

TABLE I. Comparison of detection proposal techniques on Forward-Looking Sonar Images. Our proposed methods obtain the highest recall with the lowest number of proposals. Only EdgeBoxes has a higher recall with a considerably larger number of output proposals.

performs poorly, requiring more proposals than our methods, but not reaching high recall, saturating at 88.59 %. All four methods proposed in this paper reach similar recall values (around 95 %) at 100 proposals per image. In comparison, EdgeBoxes and Selective Search requires one order of magnitude more proposals to produce similar recall.

It is also notable that CNN with objectness ranking can achieve 90 % recall with only 10 proposals per image, while FCN requires around 50 to reach similar recall. No other technique that we have evaluated can reach such high recall less than 50 proposals per image.

Generalization. We now showcase the generalization ability of our proposed methods. For this we use three images that contain unseen objects, namely a Wall, Chain [3], and a rotating platform with a Wrench. We visualize the objectness maps produced by CNN and FCN, which shows the spatial correlation between object position and objectness. Results are shown in Fig. 6. The wall in (a) shows a very good correlation with a high objectness, indicating that both CNN and FCN can produce detections over the wall, even as there is no wall example in the training set. Same effect can be seen in the Chain and Rotating Platform images. We also observe that CNN produces slightly lower objectness values than FCN, but both produce scores that can be easily distinguished from the background.

VI. Conclusions

This work covers the problem of detecting novel objects without class information, which is applicable to the detection of hard to model objects such as marine debris underwater. We have shown a new fully convolutional network to estimate objectness maps from sonar images, and we have proposed objectness ranking to obtain detection proposals from objectness scores.

Our results on a marine debris dataset on Forward-Looking sonar images show that our methods can achieve high recall (95 %) with a low number of output detections (80-300). In comparison EdgeBoxes [9] requires 5000 proposals to obtain 97 % recall, and Selective Search [10] needs around 2000 proposals to obtain 95 % recall. A baseline using classic cross-correlation template matching [1] fails to generalize well and it considerably slower than the novel approaches we propose. These results show that a neural network can learn to predict appropriate objectness values efficiently, while generalizing to completely new objects.

We expect that our results will drive the development of new object detection techniques for sonar images, adding to new capabilities such as finding novel objects, or making underwater robots aware of unknown objects in the environment.

Future work includes evaluating other sensor modalities like side-scan or synthetic aperture sonar, and developing classifiers that can deal with information of unknown object classes.

[3]This image was captured by CIRS, University of Girona.

Acknowledgements

This work has been partially supported by the FP7-PEOPLE-2013-ITN project ROBOCADEMY (Ref 608096) funded by the European Commission, and by the Autonomous Harbour Cleaning project funded by EIT Digital (Ref 18181). The authors would like to thank Leonard McLean for his help in capturing data used in this paper.

References

[1] N. Hurtós, N. Palomeras, S. Nagappa, and J. Salvi, "Automatic detection of underwater chain links using a forward-looking sonar," in *OCEANS-Bergen, 2013 MTS/IEEE*. IEEE, 2013, pp. 1–7.

[2] J. Sawas, Y. Petillot, and Y. Pailhas, "Cascade of boosted classifiers for rapid detection of underwater objects," in *Proceedings of the European Conference on Underwater Acoustics*, 2010.

[3] M. Valdenegro-Toro, "Submerged Marine Debris Detection with Autonomous Underwater Vehicles," in *International Conference on Robotics and Automation for Humanitarian Applications (RAHA)*. IEEE, 2016.

[4] J. Hosang, R. Benenson, and B. Schiele, "How good are detection proposals, really?" *arXiv preprint arXiv:1406.6962*, 2014.

[5] M. Valdenegro-Toro, *Objectness Scoring and Detection Proposals in Forward-Looking Sonar Images with Convolutional Neural Networks*. Springer International Publishing, 2016.

[6] ——, "End-to-End Object Detection and Recognition in Forward-Looking Sonar Images with Convolutional Neural Networks," in *Autonomous Underwater Vehicles (AUV), 2016 IEEE/OES*. IEEE, 2016, pp. 144–150.

[7] I. Endres and D. Hoiem, "Category independent object proposals," in *Computer Vision–ECCV 2010*. Springer, 2010, pp. 575–588.

[8] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 11, pp. 2189–2202, 2012.

[9] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Computer Vision–ECCV 2014*. Springer, 2014, pp. 391–405.

[10] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *International journal of computer vision*, vol. 104, no. 2, pp. 154–171, 2013.

[11] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 91–99.

[12] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448.

[13] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[14] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size," *arXiv preprint arXiv:1602.07360*, 2016.

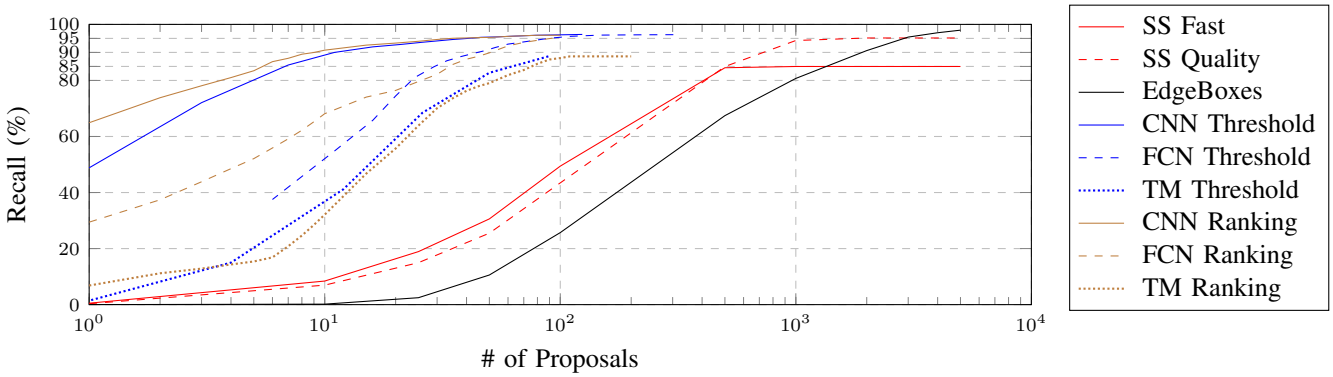
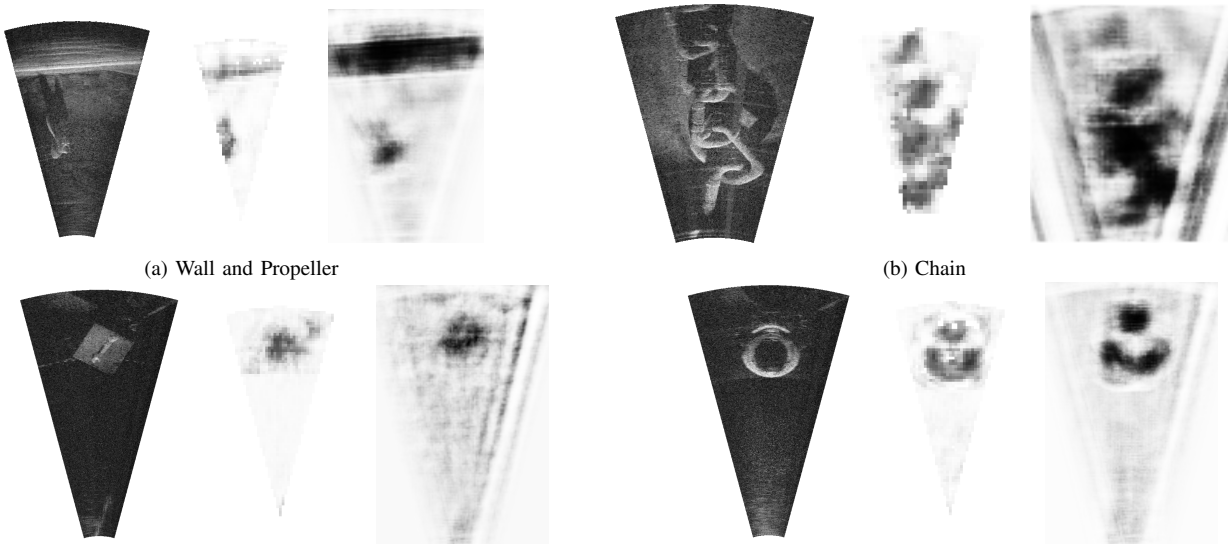


Fig. 5. Effect of the number of proposals on recall for different techniques. State of the art detection proposals methods can achieve high recall but only outputting a considerable number of proposals. Our proposed methods achieve high recall with orders of magnitude less output proposals.



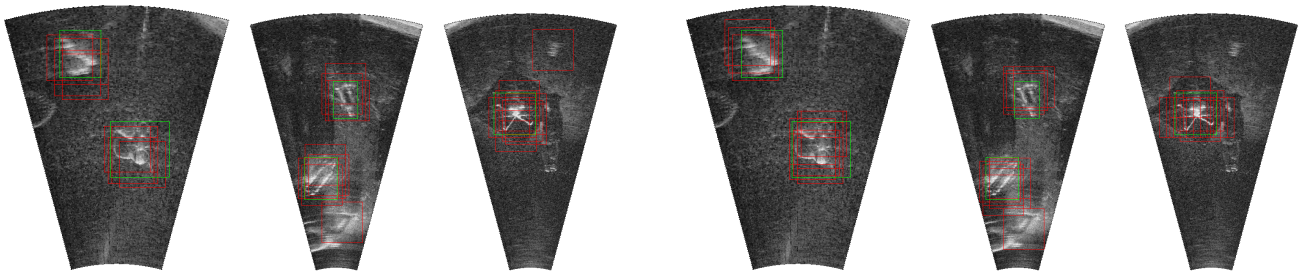
(a) Wall and Propeller

(b) Chain

(c) Rotating Platform with a Wrench

(d) Large Tire

Fig. 6. Visualization of objectness maps produced by CNN and FCN on previously unseen Forward-Looking Sonar Images. In each group: Left is the input image, Center is the CNN objectness map, while Right is the FCN map. Light shades represent low objectness, while Dark ones is high objectness.



(a) CNN

(b) FCN

Fig. 7. Sample detections produced by objectness ranking with CNN and FCN scores. We show the top $K = 10$ scoring detections. Red bounding boxes are detections, while green ones are the ground truth. Note how CNN in some cases detects blob objects that are unlabelled in our dataset.

[15] M. Valdenegro-Toro, "Real-time convolutional networks for sonar image classification in low-power embedded systems," *CoRR*, vol. abs/1709.02153, 2017.

[16] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–3440, 2015.

[17] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[18] J. Hosang, R. Benenson, P. Dollár, and B. Schiele, "What makes for effective detection proposals?" 2015.

[19] N. Chavali, H. Agrawal, A. Mahendru, and D. Batra, "Object-proposal evaluation protocol is' gameable'," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 835–844.

[20] H. Midelfart, J. Groen, and O. Midtgaard, "Template matching methods for object classification in synthetic aperture sonar images," in *Proceedings of the Underwater Acoustic Measurements Conference*, no. S S, 2009.