Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnline CONFERENCE PAPER ARCHIVE

IFAC PapersOnLine 54-16 (2021) 76–82

Deep Learning based Detection, Segmentation and Counting of Benthic Megafauna in Unconstrained Underwater Environments

Mona Lütjens* Harald Sternberg*

**HafenCity University Hamburg, Department of Hydrography and Geodesy, Henning-Voscherau-Platz 1, 20457 Hamburg, Germany (e-mail: {mona.luetjens,harald.sternberg}@hcu-hamburg.de)*

Abstract: Assessing and monitoring benthic communities is increasingly important in view of global alteration of marine environments. Deep learning has proven to effectively detect marine specimen in underwater imagery but still face problems with small input datasets, unconstrained environments and class imbalance. This study evaluates a data augmentation strategy to alleviate these limitations. Through synthetically derived image compositions, the entire input dataset was greatly extended from 700 to 12700 images. Additionally, specimen numbers of brittle stars, soft corals and glass sponges are equalized resulting in a mean average precision increase of 24 %. The overall mean average precision for box detections yields 76.7 and for instance segmentation 67.7 at an intersection over union threshold of 0.5. This study shows that deep architectures such as the deployed CenterMask via ResNeXt-101 model can successfully be trained with few original images from varying underwater scenes.

Keywords: object detection, deep learning, data augmentation, marine imagery, benthic megafauna

1. INTRODUCTION

Global alteration of marine environments due to overfishing, pollution, physical habitat destruction and climate change have led to an increasing decline of animal diversity and abundance throughout many marine ecosystems (Jackson, 2008). Especially benthic megafauna in the Southern Ocean are at risk of environmental change and are of significant ecological value as they alter small-scale topography of seabed habitats affecting the entire benthic community (Gili et al., 2006). Assessing the biodiversity and characterisation of benthic communities is increasingly important for identifying vulnerable marine ecosystems and developing conservation strategies.

Marine habitats have been studied based on mainly three practices: physical sampling methods using sledges or trawling, acoustical techniques and optical systems. While physical methods are able to collect specimen on a lower taxonomical scale they destruct the environment and their sampling rate is rather low. Optical systems are more cost effective, robust and precise than acoustic systems which led to a large growing library of digital underwater images throughout recent years paving the way for new research in automatic analytical methods.

Object detection, classification and segmentation of imagery has become a substantial task in the field of computer vision. Prior traditional feature descriptors are often used to extract colour, shape or texture information and are good at detecting specific single objects such as scallops (Dawkins et al., 2013) or lobsters using classifiers such as support vector machine (Tan et al., 2018). However, they are not robust to varying underwater scenes that are exposed to marine snow, water turbidity, lens distortion, sparse, unstable illumination and colour shift due to the survey platform's variation in speed, angle or altitude. Moreover, there is a high variability and invariability between features belonging to the same class and different classes, respectively (Pavoni et al., 2021).

Deep learning (DL) using convolutional neural networks have proven to outperform traditional based object detection (Gonzalez-Cid et al., 2017) as they are more invariant to the deformation of images. Additionally, images can directly be used as input without the necessity of pre-processing. Better results are often achieved using deeper layers thus increasing the number of parameters to several millions. To achieve high performances with those models using images of unconstrained underwater scenes or across varying platforms, a large training dataset size is the most crucial part (Langenkämper et al., 2020), often, however, very time-consuming and costly to establish.

This paper investigates the effect of input image data augmentation and composition strategies in an attempt to overcome limitations of large data set size and class imbalance. Taking account of the results, the state-of-the art anchor-free object detection and instance segmentation model CenterMask (Lee and Park, 2019) via ResNeXt-101 (Xie et al., 2017) will be trained based on a small, highly diverse 700 image dataset. In this work, the detection, segmentation and counting of *glass sponges* (hexactinellids), *soft corals* (primnoids and chrysogorgiids) and *brittle stars* (ophiuroids) will be assessed providing first steps towards future abundance and size estimations of these specimen.

10.1016/j.ifacol.2021.10.076

2. RELATED RESEARCH

Several previous works address object detection and classification of marine scenes using cutting-edge DL architectures such as RetinaNet with ResNet-50 (Boulais et al., 2020) or FDCNet (Lu et al., 2018) showing good classification results on single-labelled or iconic images. Automatic segmentation for benthic fauna has been studied for corals using DeepLabv3+ (Pavoni et al., 2021) and scale worms using U-Net and VGG-16 CNN (Shashidhara et al., 2020). Securing enough training data for DL is crucial as described above hence data augmentation techniques are widely applied. Frequently used techniques include the change of light intensity, sharpness, noise and blurring (Salman et al., 2016) or change of perspective (Huang et al., 2019). Also, rotation and cropping of underwater images have been used (Langenkämper et al., 2020). These methods have proven to be useful, however, only limited image manipulations can be performed. Most notably, the number of original annotated images from most stated works exceed the amount of available training data for this research. Therefore, another technique will be used which changes the entire image composition and adds additional alteration to synthetically generated images. Also, no research regarding instance segmentation and counting of the selected morphotypes is known to the authors.

3. MATERIALS AND METHODS

3.1 Underwater Image Dataset

The image dataset used for this study was collected during the expedition PS118 of the research vessel RV Polarstern in 2019 (Purser et al., 2021). Images were sampled using the towed Ocean Floor Observation and Bathymetry System (Purser et al., 2019) with a flying altitude of approximately 1.5 – 2.5 m above the seafloor. Seven different sampling stations from the western Weddell Sea continental shelf to the northern Powell Basin were selected. Each station area features different substrate types ranging from soft and fine mud sediment to pebbles and complex rocky topography. The original 3840 x 5760 sized images were tiled to 1440 x 960 to maintain resolution but reduce the need for computing power. 1000 images were selected and annotated using the web-based image segmentation tool COCO Annotator (Brooks, 2019). Of the 1000 images, 700 were used as training set, 100 as validation set and 200 as test set. In total, 3550 annotations of the training set were made, of which 87 % belong to the class *brittle stars*, 8 % to the class *glass sponges* and 5 % to the class *soft corals*. For the test and validation set 85 % and 84 % of the annotations belong to the class *brittle stars*, 10 % and 12 % to the class *glass sponges* and 5 % and 4 % to the class *soft corals*, respectively. It is apparent that there is a high class imbalance.

3.2 Data Augmentation

To increase the number of images for training, the image generator COCO Synth (Kelly, 2019) was utilised which composes cut out foreground images of objects over random image backgrounds. Foregrounds are randomly altered in scale, amount, rotation and brightness for each composition (Figure 1). For this study, 30 foregrounds per class and 30 backgrounds from original images were used for training. Images for the compositions were selected from varying stations and differ from previous ones in 3.1. Overall, 12,000 synthetic images were deployed for training of which 2000 images were generated of *glass sponges* and *soft corals* each, to reduce the effect of class imbalance. Now, 33 % of the total annotations belong to the class *glass sponges*, 33 % to the class *soft corals* and 34 % to the class *brittle stars*. In order to emphasize the selected augmentation method, several frequently used image manipulation techniques were additionally performed and compared to the selected method. The following image attributes were altered: brightness, colour tone, contrast, saturation, sharpness and blur (Figure 1). In total, ten different alterations per image were conducted.

Fig. 1. Example images of synthetically derived image compositions (1st row) and traditional data augmentation methods (2nd row, from left to right): original image, blur, low brightness, high brightness, blue colour, green colour, high contrast, low contrast, high saturation, low saturation, sharpness

3.3 Deep Learning Architecture and Training

The deep learning architecture chosen for this study is the anchor-free one stage instance segmentation and object detector CenterMask (Lee and Park, 2019) in combination with the backbone network ResNeXt-101 (Xie et al., 2017).

While object detection is the central task for abundance and assemblage studies, predicting masks will be important for future size estimations and biomass predictions of benthic species. An instance segmentation model was therefore chosen in order to adapt to diverse research questions for future analyses. Also, the model should operate on high inference speed while maintaining strong performances to directly evaluate datasets on board of research vessels for strategic data collection. Since both selected architectures meet the mentioned requirements and further produce excellent results in recent benchmark challenges such as COCO (Lin et al., 2014), they are an appropriate choice for the respective computer vision task of marine imaging.

The backbone ResNeXt-101 used for feature extraction is an advancement of the deep residual network ResNet (He et al., 2016) that has been recently proposed. Based on ResNet, ResNeXt-101 follows the strategy of repeating layers but stacks them parallel rather than sequentially. Thus, resulting in accuracy improvements while reducing the network complexity and number of parameters. For this study the 101 layered network was used.

As a one stage detector, CenterMask does not have a proposal step and prioritizes inference speed. Additionally, as an anchor-free detector, it does not use predefined bounding boxes to identify objects and is thus insensitive to different datasets and hyper-parameters (e.g. input size, scales, etc.). Hence, anchor-free detectors alleviate limitations of objects that have large shape variations and are rather small (Tian et al., 2019) which is ideal for the selected classes used in this research. CenterMask adopts FCOS (Tian et al., 2019) as detection head that directly computes a 4D vector and a class label at each proposed location of different levels of feature maps. Then, the spatial attention-guided mask (SAG-Mask) computes the segmentation masks on each predicted box region using the spatial attention module (SAM) that helps the mask to focus on significant pixels (Lee and Park, 2019).

Training was executed on a 64-bit Linux machine equipped with an Intel® Xeon® Gold 6254 CPU @ 3.10 GHz and 5 NVIDIA® Tesla® V100 GPU. The base learning rate was set to 0.002 and reduced by a factor of 10 after 25400 and again after 38100 iterations. To reduce early overfitting on highly differentiated datasets, the learning rate was also reduced for the first 5080 iterations. Additionally, a weight decay was implemented. The maximum number of iterations was 50800 which corresponds to 20 epochs. All backbone models are initialized by ImageNet pre-trained weights.

3.4 Evaluation Protocol

To assess the performance of the model, the evaluation metrics average precision, average recall, F_1 measure and accuracy are utilized. While the precision P reflects the proportion of false positives FP, the recall R defines the proportion of false negatives FN and can be mathematically expressed as follows:

$$P = \frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{FP})} \text{ and } R = \frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{FN})}, \quad (1)$$

with TP being the number of true positive predictions. In order to classify whether a prediction is a TP or FP, the intersection over union (IoU) threshold is used as it measures the overlap between the ground truth and the predicted bounding box or segmentation mask, respectively. Typical values for IoU thresholds are 0.5 or 0.75. The very common approach to summarize precision and recall into one value is the average precision (AP). The AP for a single class is the averaged precision across all recall levels. Complementarily, the average recall score (AR) averages recall values over all IoU $\in$ [0.5, 1.0] for each class. The mean average precision (*mAP*) and mean average recall (*mAR*) across all classes *C* are defined as (Raphael et al., 2020):

$$mAP = \frac{1}{C}\sum_{i=1}^{C} AP_i \text{ and } mAR = \frac{1}{C}\sum_{i=1}^{C} AR_i \quad (2)$$

As the evaluation metrics used for this research is based on COCO (Lin et al., 2014), it should be noted that *mAP* and *mAR* scores are further denoted as AP and AR for simplicity reasons. They are computed over single (0.5) IoU or the average of then IoU levels starting from 0.5 to 0.95 in steps of 0.05 (the latter is further denoted as AP @.50:.95). AP and AR are also calculated for different object scales (small: $< 72^2$ pixels, medium: $> 72^2$ & $< 214^2$ pixels, large: $> 214^2$ pixels) and for different maximum number of detections per image (1, 10, 100). Object scales deviate from COCO and are adjusted to fit scales in the proposed images.

Additional adopted performance metrics are the *accuracy* to assess the total number of predictions that are correct and the *F_1 measure* which evenly weighs between precision and recall (Manning et al., 2009):

$$accuracy = \frac{\mathrm{TP}+\mathrm{TN}}{(\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN})} and\ F_1 = \frac{2PR}{P+R} \quad (3)$$

4. EXPERIMENTAL RESULTS

This section presents the performance evaluation of different investigated methods to detect, segment and count benthic megafauna across varying models and datasets as listed in Table 1. All test runs are performed on the 200 original image test set with no data augmentation.

Table 1. Abbreviations for training strategies

CM-X-101	CenterMask and ResNeXt-101
CM-V-99	CenterMask and VoVNetV2-99
CM-L-M	CenterMask-Lite and MobileNetV2
M-X-101	Mask R-CNN and ResNeXt-101
R-X-101	RetinaNet and ResNeXt-101

Baseline	Original dataset
Synth	Synthetic image composition with equal class distribution & Baseline
Synth-Blncd	Synthetic image composition including extra glass sponges and soft corals & Baseline (balanced training data set)
Trad. Augm.	Traditional augmentation using brightness (high, low), colour tone (green, blue), contrast (high, low), saturation (high, low), sharpness, blur & Baseline
Fusion	Synth-Blncd with 6,000 images and Trad. Augm. with 7,000 images combined & Baseline

4.1 Evaluation of detection, segmentation and counting

For future benthic assemblage and abundance analyses, the precision score as well as recall are both equally important as specimen should neither be misclassified nor missed. As can be seen in Table 2, the network CM-X-101 trained on Synth-Blncd shows the highest bounding box results of 76.7 % AP @.50 and 59 % AR @10 compared to all other methods. Detections are accurate on various backgrounds, illumination, camera distances and distortions (Figure 2) confirming that automatic detection and classification methods are possible even on small highly diverse input datasets. Also, high AP and AR scores on segmentation masks are achieved, yielding 67.7 % @.50 and 49.1 % @10, respectively (Table 2). Performance for bounding boxes are slightly higher than instance segmentation masks because very coarse object boundaries are drawn on each object including also many irrelevant pixels. Instance segmentation assigns only object relevant pixels to a label and is therefore computationally more advanced.

It can be further noted, that the recall rate of single objects per image is much lower than the recall of multiple objects per image. This accounts for both segmentation masks and bounding boxes. Additionally, smaller objects $<72^2$ reach poorer results than large objects which might be caused by the downsampling in the ResNeXt backbone resulting in fewer features being extracted. Another factor is the relatively large ratio between pixel size and object size for small objects which might often lead to positioning errors when computing IoU.

The performance of the model on different classes can be further investigated with a confusion matrix (Table 3) which is computed using a lower IoU threshold to favour a high recall especially for small objects. At first, it can be seen that no specimen are wrongly classified between classes. Brittle stars have the highest precision of 99 % and *soft corals* have the highest recall of 93 %. On the other hand, *glass sponges* and *brittle stars* are often not detected (high FN) and *glass sponges* and *soft corals* are often misclassified with background clutter (high FP). Using the values for TP, FP and FN, the accuracy amounts to 87 %, 73 % and 58 % for *brittle stars*, *soft corals* and *glass sponges*, respectively, resulting in a mean accuracy of 73 %.

In total, 76 out of 82 *glass sponges* were counted, 49 out of 41 *soft corals* and 597 out of 673 *brittle stars* leading to a percentage variation of -7 %, 20 % and -11 % for each respective class.

4.2 Evaluation of data augmentation strategies

The detection performance of marine organisms improves in both cases using either the data augmentation strategy of synthetic image compositions or the traditional image manipulation techniques. In fact, with the Synth-Blncd dataset, the bounding box AP @.50:.95 result was the most improved with an increase of 24 % over the Baseline dataset.

Generally, AP and AR on bounding boxes are slightly higher on synthetically generated image compositions than traditional augmentation methods (Table 2). The disadvantage of the latter is that not as many images can be created without the risk of overfitting as object compositions are not changing. Synthetic generated images have proven to be a successful alternative and can be generated in a short amount of time as well. However, with regards to instance segmentation performance, synthetic image datasets show inferior results on AP scores especially for small and medium sized objects (Table 2). The reason for this might be the incorrect downsizing of foregrounds as they could appear pixelated in the process of image creation.

Considering the problem of class imbalance, synthetic derived images have the potential to easily balance out numbers of specimen between classes. Table 4 shows that the detection performance of *glass sponges* and *soft corals* are increased by 5-16 % over the Synth dataset. The performance for *brittle stars*, however, is not improved.

Fig. 2. Example images of detection and segmentation results of test set with CM-X-101/Synth-Blncd at varying stations.

4.3 Comparison with other state-of-the-art algorithms

The selected CenterMask via ResNeXt-101 architecture was further compared to other state-of-the art detectors and backbones such as RetinaNet (Lin et al., 2017), Mask R-CNN (He et al., 2017), MobileNetV2 (Sandler et al., 2018) or VoVNetV2-99 (Lee and Park, 2019). Results are demonstrated in Table 2 for bounding boxes and segmentation masks. With regards to competing detectors, both CenterMask and RetinaNet are one stage detectors whereas Mask R-CNN is a two stage detector that utilizes a region-of-interest proposal step which typically prioritizes detection accuracy over inference speed. Moreover, RetinaNet and Mask R-CNN use anchor boxes for feature detection. From the results in Table 2 it is evident that RetinaNet performs nearly as well as CenterMask whereas Mask R-CNN shows a much lower performance considering bounding boxes and segmentation masks. It can be noted that one stage anchor-free detectors can perform just as well on underwater imagery as other detector types.

Furthermore, evaluations on three differently deep backbones were conducted: ResNeXt-101 with 114.3 million parameters, VoVNet-99 with 96 million parameters and MobileNetV2 with 28.7 million parameters (Lee and Park, 2019). Noticeably, MobileNetV2 with fewer layers has considerably lower AP and AR results than the other two. Meanwhile, considering bounding boxes, VoVNet-99 performs nearly as well as ResNeXt-101 and with regards to small objects, AP and AR results are even slightly higher. The problem of detecting small objects is known to increase for very deep backbones as they need more input information to cope with the massive amount of parameters (Nguyen et al., 2020). Small objects are described by fewer pixels and might not be diverse enough to feed the network with sufficient information increasing the changes of overfitting.

Table 2. Summary of detection results with bounding boxes (1st row) and segmentation masks (2nd row/cursive)

Model/Data	$AP^{bbox}_{.50:.95}$	$AP^{bbox}_{.50}$	AP^{bbox}_{small}	AP^{bbox}_{medium}	AP^{bbox}_{large}	AR^{bbox}_{1}	AR^{bbox}_{10}	AR^{bbox}_{100}	AR^{bbox}_{small}	AR^{bbox}_{medium}	AR^{bbox}_{large}
CM-X-101/ Baseline	41.7	68.2	25.3	29.3	54.7	21.6	51.6	55.2	25.4	45.1	70.8
	35.2	*63.0*	*5.70*	*21.2*	*50.7*	*19.4*	*44.4*	*45.2*	*8.90*	*32.7*	*57.9*
CM-X-101/ Synth	48.8	71.0	27.4	39.1	62.8	24.7	58.8	**64.2**	27.9	**57.3**	77.1
	37.3	*63.0*	*4.90*	*24.1*	*51.7*	*20.6*	*47.9*	*49.5*	*7.90*	*41.2*	*59.5*
CM-X-101/ Synth-Blncd	**51.8**	**76.7**	27.5	40.2	**66.1**	**25.7**	**59.0**	63.9	27.9	55.7	**77.9**
	40.8	*67.7*	*4.60*	*27.1*	***54.8***	***22.1***	***49.1***	***50.6***	*7.50*	***41.6***	***62.5***
CM-X-101/ Trad. Augm	48.8	75.0	26.9	38.6	58.5	23.0	55.3	58.9	27.2	50.1	72.6
	40.2	***70.4***	***7.00***	*27.6*	*52.2*	*20.4*	*46.8*	*47.9*	***10.6***	*36.8*	*58.6*
CM-X-101/ Fusion	51.7	74.1	27.1	**42.1**	65.1	24.9	57.6	61.6	27.5	52.2	77.0
	41.5	*69.7*	*6.60*	***29.0***	*53.9*	*21.7*	*47.0*	*48.4*	**10.6**	*37.1*	*59.3*
CM-V-99/ Synth	47.9	72.0	**27.9**	37.0	62.8	23.6	56.6	61.9	28.3	52.6	77.1
	36.9	*62.9*	*4.80*	*23.4*	*49.8*	*20.1*	*46.1*	*47.8*	*7.60*	*38.6*	*58.2*
CM-L-M/ Synth	27.3	48.6	19.1	19.0	40.0	18.3	39.1	43.7	20.0	34.4	59.5
	18.4	*32.6*	*0.60*	*6.30*	*34.6*	*15.2*	*31.8*	*33.4*	*2.00*	*20.3*	*50.5*
M-X-101/ Synth	33.3	53.2	13.2	22.7	53.0	20.7	39.2	40.0	13.2	30.6	60.7
	25.7	*41.6*	*1.70*	*11.8*	*37.9*	*15.2*	*31.8*	*31.9*	*3.80*	*20.5*	*43.6*
R-X-101/ Synth	47.8	70.7	**27.9**	37.1	62.2	24.2	56.6	61.9	**28.4**	53.8	76.7

Table 3. Confusion matrix for CM-X-101/Synth-Blncd

		Predicted				
		Glass Sponges	Soft Corals	Brittle Stars	Background	Recall
Ground truth	Glass Sponges	58	0	0	24	70.1
	Soft Corals	0	38	0	3	92.7
	Brittle Stars	0	0	591	82	87.8
	Background	18	11	6		
	Precision	76.3	77.6	99.0		Accuracy: 73 %

Table 4. Summary of performance results per class

	Glass Sponges				Soft Corals				Brittle Stars			
Model/Data	F^{bbox}_{1}	F^{mask}_{1}	$AP^{bbox}_{.50:.95}$	$AP^{mask}_{.50:.95}$	F^{bbox}_{1}	F^{mask}_{1}	$AP^{bbox}_{.50:.95}$	$AP^{mask}_{.50:.95}$	F^{bbox}_{1}	F^{mask}_{1}	$AP^{bbox}_{.50:.95}$	$AP^{mask}_{.50:.95}$
CM-X-101/ Baseline	67.7	67.7	41.4	46.3	70.3	69.5	36.8	45.5	**80.2**	**67.9**	46.9	**13.9**
CM-X-101/ Synth	67.8	66.2	45.3	46.5	69.8	68.8	48.8	52.6	79.2	65.1	52.4	12.8
CM-X-101/ Synth-Blncd	**71.4**	**69.2**	**51.4**	**54.0**	**76.8**	**74.9**	**51.5**	**55.9**	79.9	64.7	**52.6**	12.6

5. CONCLUSION AND FUTURE STEPS

In conclusion, the used data augmentation strategy of synthetically derived image compositions proved to be a good alternative to frequently used augmentation techniques. Problems such as class imbalance can further easily be alleviated and boost the performance of underrepresented classes. Detection, segmentation and counting of benthic megafauna is a task that can be solved with good performance using few variant original input images using anchor-free one stage detectors. In comparison with other models, it is evident that the detection problem of small objects is a challenge yet to be solved. Additionally, future steps involve the introduction of more benthic morphotypes, improved counting to avert duplications for overlapping images and the allocation of position and water depth to each detected specimen for future assemblage studies.

6. ACKNOWLEDGEMENTS

We thank the annotators of the images: Gavin DMello, Diana Rubio and Seyed Lialestani. Further we thank the captain and crew of RV Polarstern as well as the scientific party of the cruise PS118 for their support. Special thanks go to Autun Purser and Huw Griffiths for their support, the data collection on board and the identification of benthic organisms.

REFERENCES

Boulais, O., Woodward, B., Schlining, B., Lundsten, L., Barnard, K., Bell, K. C. and Katija, K. (2020). FathomNet: An underwater image training database for ocean exploration and discovery. arXiv:2007.00114v3, Cornell University, Computer Science, Computer Vision and Pattern Recognition [cs.CV].

Brooks, J. (2019). COCO Annotator. URL https://github.com/jsbroks/coco-annotator/.

Dawkins, M., Stewart, C., Gallager, S. and York, A. (2013). Automatic scallop detection in benthic environments. *2013 IEEE Workshop on Applications of Computer Vision (WACV),* 160-170. doi: 10.1109/WACV.2013.6475014.

Gili, J.-M., Arntz, W. E., Palanques, A., Orejas, C., Clarke, A., Dayton, P. K., Isla, E., Teixidó, N., Rossi, S. and López-González, P. J. (2006). A unique assemblage of epibenthic sessile suspension feeders with archaic features in the high-Antarctic. *Deep Sea Research Part II: Topical Studies in Oceanography*, volume (53), 1029-1052. doi: 10.1016/j.dsr2.2005.10.021.

Gonzalez-Cid, Y., Burguera, A., Bonin-Font, F. and Matamoros, A. (2017). Machine learning and deep learning strategies to identify Posidonia meadows in underwater images. *OCEANS 2017 – Aberdeen,* 1-5. doi: 10.1109/OCEANSE.2017.8084991.

He, K., Zhang, X., Ren, S. and Sun, J. (2016). Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 770-778. doi: 10.1109/CVPR.2016.90.

He, K., Gkioxari, G., Dollár, P. and Girshick, R. (2017). Mask R-CNN. arXiv:1703.06870v3, Cornell University, Computer Science, Computer Vision and Pattern Recognition [cs.CV].

Huang, H., Zhou, H., Yang, X., Zhang, L., Qi, L. and Zang, A.-Y. (2019). Faster R-CNN for marine organisms detection and recognition using data augmentation. *Neurocomputing*, volume (337), 372-384. doi: 10.1016/j.neucom.2019.01.084.

Jackson, J. B. C. (2008). Ecological extinction and evolution in the brave new ocean. *Proceedings of the National Academy of Sciences,* 11458-11465. doi: 10.1073/pnas.0802812105.

Kelly, A. (2019). COCO Synth. URL https://github.com/akTwelve/cocosynth.

Langenkämper, D., van Kevelaer, R., Purser, A. and Nattkemper, T. W. (2020). Gear-Induced Concept Drift in Marine Images and Its Effect on Deep Learning Classification. *Frontiers in Marine Science*, volume (7). doi: 10.3389/fmars.2020.00506.

Lee, Y. and Park, J. (2019). CenterMask : Real-Time Anchor-Free Instance Segmentation. arXiv: 1911.06667v6, Cornell University, Computer Science, Computer Vision and Pattern Recognition [cs.CV].

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L. and Dollár, P. (2014). Microsoft COCO: Common Objects in Context. In: Fleet D., Pajdla T., Schiele B., Tuytelaars T. (ed.), *Computer Vision – ECCV 2014*, 740-755. Springer, Cham. doi:10.1007/978-3-319-10602-1_48.

Lin, T.-Y., Goyal, P., Girshick, R., He, K. and Dollár, P. (2017). Focal Loss for Dense Object Detection. arXiv: 1708.02002v2, Cornell University, Computer Science, Computer Vision and Pattern Recognition [cs.CV].

Lu, H., Li, Y., Uemura, T., Ge, Z., Xu, X., He, L., Serikawa, S. and Kim, H. (2018). FDCNet: filtering deep convolutional network for marine organism classification. *Multimedia Tools and Applications*, volume (77), 21847-21860. doi: 10.1007/s11042-017-4585-1.

Manning, C. D., Raghavan, P. and Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press, Cambridge. ISBN: 0521865719.

Nguyen, N.-D., Do, T., Ngo, T. D. and Le, D.-D. (2020). An Evaluation of Deep Learning Methods for Small Object Detection. *Journal of Electrical and Computer Engineering*, volume (2020), 1-18. doi: 10.1155/2020/3189691.

Pavoni, G., Corsini, M., Pedersen, N., Petrovic, V. and Cignoni, P. (2021). Challenges in the deep learning-based semantic segmentation of benthic communities from Ortho-images. *Applied Geomatics*, volume (13), 131-146. doi: 10.1007/s12518-020-00331-6.

Purser, A., Marcon, Y., Dreutter, S., Hoge, U., Sablotny, B. and Hehemann, L. (2019). Ocean Floor Observation and Bathymetry System (OFOBS): A New Towed Camera/Sonar System for Deep-Sea Habitat Surveys. *IEEE Journal of Oceanic Engineering*, volume (44), 87-99. doi: 10.1109/JOE.2018.2794095.

Purser, A., Dreutter, S., Griffiths, H., Hehemann, L., Jerosch, K., Nordhausen, A., Piepenburg, D., Richter, C., Schröder, H. and Dorschel, B. (2021). Seabed video and still images from the northern Weddell Sea and the western flanks of the Powell Basin. *Earth System Science Data*, volume (13), 609–615. doi: 10.5194/essd-13-609-2021.

Raphael, A., Dubinsky, Z., Iluz, D. and Netanyahu, N. S. (2020). Neural Network Recognition of Marine Benthos and Corals. *Diversity*, volume (12). doi: 10.3390/d12010029.

Salman, A., Jalal, A., Shafait, F., Mian, A., Shortis, M., Seager, J. and Harvey, E. (2016). Fish species classification in unconstrained underwater environments based on deep learning. *Limnology and Oceanography: Methods*, volume (14), 570-585. Doi: 10.1002/lom3.10113.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. and Chen, L.-C. (2018). MobileNetV2: Inverted Residuals and Linear Bottlenecks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 4510-4520. doi: 10.1109/CVPR.2018.00474.

Shashidhara, B. M., Scott, M. and Marburg, A. (2020). Instance Segmentation of Benthic Scale Worms at a Hydrothermal Site. *2020 IEEE Winter Conference on Applications of Computer Vision (WACV),* 1303-1312. doi: 10.1109/WACV45572.2020.9093574.

Tan, C. S., Lau, P. Y., Correia, P. L. and Campos, A. (2018). Automatic analysis of deep-water remotely operated vehicle footage for estimation of Norway lobster abundance. *Frontiers of Information Technology & Electronic Engineering*, volume (19), 1042-1055. doi: 10.1631/FITEE.1700720.

Tian, Z., Shen, C., Chen, H. and He, T. (2019). FCOS: Fully Convolutional One-Stage Object Detection. *IEEE/CVF International Conference on Computer Vision (ICCV),* 9626-9635. doi: 10.1109/ICCV.2019.00972.

Xie, S., Girshick, R., Dollar, P., Tu, Z. and He, K. (2017). Aggregated Residual Transformations for Deep Neural Networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 5987-5995. doi: 10.1109/CVPR.2017.634.