DOI: 10.23784/HN119-02

Deep learning-based detection of marine images and the effect of data-driven influences

An article by MONA LÜTJENS and HARALD STERNBERG

Throughout recent years convolutional neural networks have been applied for various image detection tasks. Training data thereby plays an important role for the performance of those models. Not only the amount of images is crucial but also the number of annotations, classes as well as image dimensions. In view of changing underwater environments, the study of benthic communities is increasingly important especially in the Southern Ocean as they provide a key link for ecosystem shifts. This study concentrates on the automatic detection and classification of benthic species using deep learning. It could be shown that glass sponges, brittle stars and soft corals could successfully be detected even on few input data and highly biased class distributions in varying underwater scenes. Further analyses considering data-driven influences show significant performance declines regarding the training on single objects and classes per image and the evaluation on large image dimensions.

deep learning | automatic detection | underwater imagery | benthos
Deep Learning | automatische Detektion | Unterwasserbilder | Benthos

In den letzten Jahren wurden gefaltete neuronale Netze für verschiedene Aufgaben der Bilderkennung eingesetzt. Die Trainingsdaten spielen dabei eine wichtige Rolle für die Leistungsfähigkeit dieser Modelle. Dabei ist nicht nur die Menge der Bilder entscheidend, sondern auch die Anzahl der Annotationen, Klassen sowie die Bilddimensionen. Angesichts sich verändernder Unterwasserumgebungen wird die Untersuchung benthischer Lebensgemeinschaften vor allem im Südlichen Ozean immer wichtiger, da sie hier vor allem sensibel auf Veränderungen reagieren. Diese Arbeit konzentriert sich auf die automatische Erkennung und Klassifizierung von benthischen Arten mittels Deep Learning. Es konnte gezeigt werden, dass Glasschwämme, Schlangensterne und Weichkorallen selbst bei wenigen Eingabedaten und stark unterrepräsentierten Klassen in unterschiedlichsten Unterwasserlandschaften erfolgreich erkannt werden. Weitere Analysen zu datengetriebenen Einflüssen zeigen deutliche Leistungseinbußen bei einzelnen Objekten und Klassen pro Bild während des Trainings und großen Bilddimensionen während der Evaluation.

Authors
Mona Lütjens is Research Associate at HafenCity University in Hamburg. Harald Sternberg is Professor for Hydrography at HafenCity University in Hamburg.

mona.luetjens@hcu-hamburg.de

1 Introduction

Global ocean temperature rise and ocean acidification are ubiquitous and threaten especially benthic communities in the Southern Ocean where many species survive only in a narrow thermal range (Griffiths et al. 2017). To detect current ecosystem shifts, studies regarding the abundance of megabenthic species can provide information as they are very sensitive to environmental change (Piepenburg et al. 2017). Sponges should be especially investigated as they create and shape habitats for other species like brittle stars and a decrease in sponges might directly lead to a decrease in many other species as well (Mitchell et al. 2020).

One of the main methods to study megabenthic species is through optical imagery. It is a fast and non-destructive sampling method and optical systems are typically mounted on towed or remotely operated vehicles. In light of its advantages, an increasing amount of underwater imagery has emerged raising the need for automatic analytical methods. Recent research in full automatic detection and classification of marine images deploy deep learning algorithms as they show superior results for unconstrained underwater environments, non-iconic images and variant image deformations (Gonzalez-Cid et al. 2017). The latter is one of the main challenges as objects in marine images are greatly changing due to different lightning conditions, rotation of the camera system, lens distortion and noise (Pavoni et al. 2021). To account for this, multilayer convolutional neural network (CNN) models are introduced. Learned features can be recognised regardless of their position or imaging condition and without previous image preprocessing or human supervision. In computer vision tasks, two main methods for recognising multiple objects have emerged: object detection and in-

stance segmentation. The output of an object detector is a set of bounding boxes around detected objects whereas instance segmentation computes pixel-accurate masks around detected objects and is thus able to grasps the shape of objects. Generating training data for instance segmentation is very laborious and masks are typically generated in a second step after the bounding box detection. Since this study simply focuses on the detection of marine species without the necessity to capture shapes of features, instance segmentation was not implemented. Several previous works deal with the classification and detection of fish (Salman et al. 2016; Christensen et al. 2018) or benthic communities (Boulais et al. 2020) using state-of-the art models such as LeNET, SSD via MobileNet and RetinaNet via ResNet50, respectively.

For CNNs the amount of training data is considered to be the main driver for accurate network inference. Also, better results are achieved with deeper layered networks because features can be learned at more diverse levels of abstractions. As more layers of neurons are added to the network, different feature details ranging from low-level features such as lines or dots to high-level features such as common objects or shapes are trained to be recognised. Networks with multiple layers are thus better at generalising because they learn more discriminative features (Pauly et al. 2017). However, deeper layered networks typically consists of several million of parameters, increasing the demand of more training data. Therefore, training data sets are commonly augmented by changing the rotation, sharpness, perspective and brightness (Huang et al. 2019) to produce more input data in a cost and time effective way. In view of successful training, it is further important to consider data related design choices such as number of annotations and classes per image during training as well as the image input size. While considering image sizes ranging from 96 to 224 pixels, it could be shown that the accuracy linearly increases (Mishkin et al. 2017).

This paper investigates the effect of data driven influences on the model accuracy in an attempt to create a road map for optimal input training data with regards to number of annotations and classes per image, class imbalance and image sizes exceeding those in previous mentioned studies. For the detection of benthic morphotypes the state-of-the-art network CenterMask (Lee and Park 2019) via ResNeXt-101 (Xie et al. 2017) was utilised which is trained on the three classes: glass sponges, soft corals and brittle stars.

2 Data

2.1 Underwater imagery data set

A seabed survey to investigate the epibenthos was carried out during the PS118 cruise of RV *Polarstern* in the western Weddell Sea in 2019 (Purser et al. 2021). Seafloor images were obtained using the towed Ocean Floor Observation and Bathymetric System (Purser et al. 2019). For this study images from seven different sampling stations at distinct depths and with diverse seafloor types were used to incorporate various environmental alterations in the network training process. The original 3840 × 5760 sized images were tiled rather than down sampled to 1440 × 960 to keep the input resolution but decreasing the need for computational resources during training. Image annotation for the three object classes was conducted on 1000 images using the web-based annotation tool COCO Annotator (Brooks 2019). The selected image set was split so that 700 images belong to the training set, 100 images to the validation set and 200 images to the test set. After labelling it was evident that a high class imbalance persists because of the 3550 annotations from the training set, 87 % of the labels belong to the class brittle stars, 8 % to the class glass sponges and 5 % to the class soft corals.

Fig. 1: Synthetically derived image compositions by placing cut out foregrounds onto cropped backgrounds

2.2 Data augmentation

Data augmentation was conducted using the image generator COCO Synth (Kelly 2019) which composes new images by placing cut out objects as foreground over plain seafloor images. The foregrounds are randomly altered in brightness, rotation, scale and amount. For training, a total of 12,000 synthetic images were created from 30 foregrounds per class and 30 background images (Fig. 1). It is noted, that the selected foregrounds and backgrounds originate from images that are not part of the original training set mentioned in section 2.1. Also, to alleviate class imbalance 4000 images of the 12,000 images are solely composed of glass sponges and soft corals changing the ratio

to 33 %, 33 % and 34 % for glass sponges, soft corals and brittle stars.

3 Method

3.1 Deep learning architecture

The neural network which was utilised for the detection of benthic species in section 4.1. and 4.2 is the object detector CenterMask (Lee and Park 2019) in combination with the backbone ResNeXt-101 (CM-X-101) (Xie et al. 2017). Backbone refers to the part of the network which is used to extract basic features and creates the feature map representation of the input data. They are typically initialised by ImageNet pre-trained weights. The detection head uses the feature map to perform the task of object detection and classification. It computes bounding boxes on identified objects for each image and calculates the classification confidences. Both architectures used in this study received excellent results in recent benchmarks such as COCO (Lin et al. 2014). For the experiments conducted in section 4.3 and 4.4 the more lightweight backbone VoVNetV2-99 (CM-V-99) (Lee und Park 2019) was used instead of ResNeXt-101 as it comprises fewer network parameters such as weights of connections which reduces the computing time.

3.2 Training details

Training was executed on five NVIDIA Tesla V100 GPUs of a 64-bit Linux machine equipped with an Intel Xeon Gold 6254 CPU @ 3.10 GHz. The base learning rate was set to 0.002. To reduce the effect of early overfitting on highly differentiated data sets, the learning rate was reduced for the first 5080 iterations by one third (Fig. 2). After 25,400 and again after 38,100 iterations the base learning rate was reduced by a factor of ten. The maximum number of iterations one image batch was passed forward and backward through the neural network was 50,800 which corresponds to 20 epochs defined as the number where the entire data set is passed through the network.

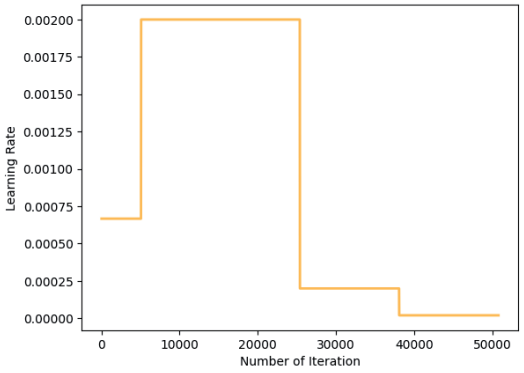

Fig. 2: Learning rate at number of iteration

3.3 Performance metrics

The performance was assessed based on the evaluation metrics adopted for COCO which are based on the average precision and average recall scores (Lin et al. 2014). Both, precision as well as recall are evenly important metrics for the classification of benthic communities. While precision is the ratio of correctly predicted specimen out of all predicted specimen, recall indicates whether all correct specimen could be detected and how many were missed. Consequently, the precision P defines the proportion of false positives FP and the recall R reflects the proportion of false negatives FN. With TP being the number of true positives they can be mathematically computed as follows:

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

Precision and recall scores are then computed into average scores (AP and AR) over all classes and at varying intersection over union (IoU) thresholds which are used to measure the overlap between ground truth and predicted bounding boxes. The defined IoU are 0.5 and the average of ten IoU levels starting from 0.5 to 0.95 with a step size of 0.05 (the latter is further denoted as: .50:.95). AP and AR are also calculated for varying object sizes (small: $< 72^2$ pixels, medium: $> 72^2$ and $< 214^2$ pixels, large: $> 214^2$ pixels) and for different maximum number of detections per image (1, 10, 100). AR_1 computes the mean average recall across all classes and IoU thresholds for images where at most one detection was made while AR_{10} and AR_{100} compute the mean average recall for images where at most ten or at most 100 detections were made, respectively. Additional adopted 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):

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

4 Experiments and results

Main experiments in section 4.1 and 4.2 were executed across varying data sets summarised in Table 1.

Data set	Training set composition
Baseline	Original data set (700 images)
Synth-B	Synthetic images with equal class distribution & Baseline composition (12,700 images)
Synth-GS	Synthetic images composition including extra glass sponges and soft corals & Baseline (12,700 images)

Table 1: Data set compositions for main experiments

The corresponding test runs were performed on the 200 original image test set. Further ablation studies performed in section 4.3 and 4.4 were con-

Model/Data	$AP_{.50:.95}$	$AP_{.50}$	AP_{small}	AP_{medium}	AP_{large}	AR_{1}	AR_{10}	AR_{100}	AR_{small}	AR_{medium}	AR_{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
CM-X-101/Synth-B	48.8	71.0	27.4	39.1	62.8	24.7	58.8	**64.2**	**27.9**	**57.3**	77.1
CM-X-101/Synth-GS	**51.8**	**76.7**	**27.5**	**40.2**	**66.1**	**25.7**	**59.0**	63.9	**27.9**	55.7	**77.9**

Table 2: Summary of detection results with bounding boxes (in percent)

ducted on 20 epochs using a smaller synthetically derived training set of 2000 images with varying image sizes, number of annotations and classes considering the respective experiment. Corresponding testing was performed with respective 300 synthetically derived images.

4.1 Detection results

To investigate the detection results of the trained network, the average precision and average recall exhibit best scores around 76.7 % AP for an $IoU_{.50}$ and 63.9 % AR_{100} on the Synth-GS data set (Table 2). Further, deploying synthetically derived images to support the training increases the performance of $AP_{.50:.95}$ by 17 % and AR_{100} by 16 % emphasising the importance of more input data.

With regards to recall scores at varying numbers of detections, it can be noted that more detections per image will lead to better recall evaluations. Additionally, smaller object sizes receive lower precision as well as recall scores throughout all testing strategies (Table 2). Those low performances might be caused by down sampling operations inside pooling layers that are applied on each feature map in the model. Down sampling output feature maps makes them more robust to changes in the translation of a feature in the image but fewer features might get extracted as resolution decreases with repeated convolutional and pooling layers. Also, there is a relatively large ratio between pixel size and object size for smaller objects which quickly increases the possibility to predict bounding boxes with positional deviations from ground truth boxes. Those positional deviations might already be too large to pass the threshold defined for the IoU.

Considering qualitative results, Fig. 3 shows detection results for various stations with differ-

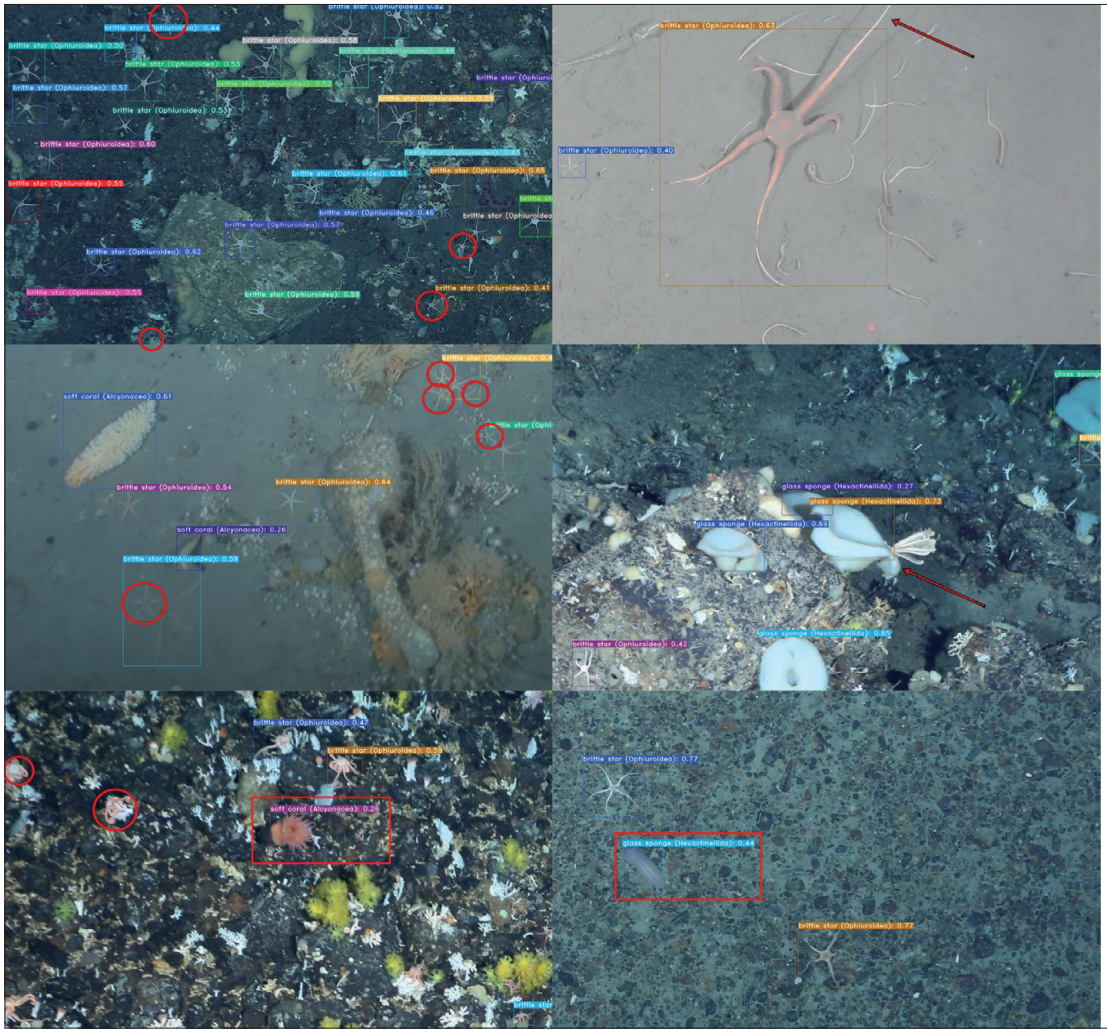

Fig. 3: Detection results and confidences of the model CenterMask – ResNeXt-101. Red circles show missing brittle stars, red arrows indicate inaccurate bounding boxes predictions and red rectangles reveal wrong species detections. Images belong to the test set and originate from different diving locations of the PS118 cruise

Model/Data	Glass sponges		Soft corals		Brittle stars	
	F_1	$AP_{.50:.95}$	F_1	$AP_{.50:.95}$	F_1	$AP_{.50:.95}$
CM-X-101/Baseline	67.7	41.4	70.3	36.8	**80.2**	46.9
CM-X-101/Synth-B	67.8	45.3	69.8	48.8	79.2	52.4
CM-X-101/Synth-GS	**71.4**	**51.4**	**76.8**	**51.5**	79.9	**52.6**

Table 3: Summary of performance results per class (in percent)

ent seafloor types, camera distances, variant illuminations and sharpness. It can be seen, that the trained model is able to correctly detect almost all specimen belonging to the three classes. Even blurred images pose no problem in detection just very small specimen or such that are lying closely to one another might be wrongly detected as one.

4.2 Influence of class imbalance

The influence of class imbalance where class distributions are biased is a known problem in deep learning applications (Guo et al. 2008). There are many approaches to combat class imbalance such as oversampling, undersampling or setting class weights to emphasise minority classes. In this study the underrepresented classes glass sponges and soft corals were oversampled because this method has proven to be very effective (see Guo et al. 2008; Buda et al. 2018). After adopting the data augmentation strategy with additional distributions of underrepresented classes, the F_1 and AP scores are increased by 5 % and 13 % for glass sponges and by 10 % and 6 % for soft corals, respectively (Table 3). Consequently, AP and AR scores across all classes are boosted with a percentage increase of 24 % $AP_{.50:.95}$ and 16 % AR_{100} compared to the Baseline data set

4.3 Influence of number of annotations and classes

Ablation studies with respect to number of annotations show that single annotations per image have a precision reduction of 27 % when evaluating on images with five annotations and 71 % when evaluating on images showing up to 20 annotations (Fig. 4). Training performed on five and

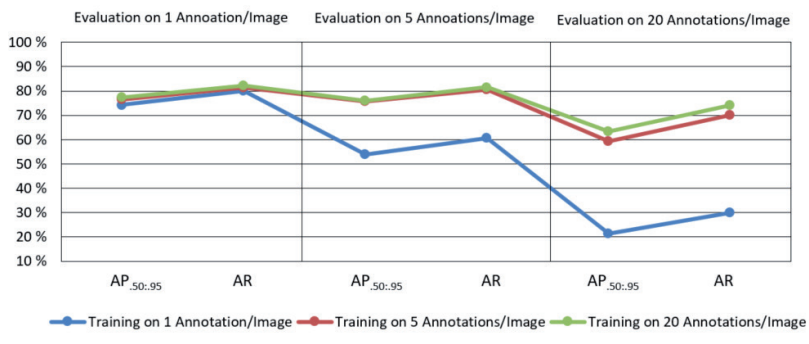

Fig. 4: Number of annotations per image during training and the evaluation regarding average precision and recall across varying numbers of annotations per image

20 annotations show a reduction in precision of 22 % (59.5 % $AP_{.50:.95}$) and 18 % (63.3 % $AP_{.50:.95}$), when testing on images with 20 annotations, respectively. Hence, best AP results are received when training is performed on images with up to 20 annotations, and worse results are scored when images contain only single objects during training. Also, it can be argued that images with lots of specimen not necessarily need to be implemented for training as the gap between 59.5 % $AP_{.50:.95}$ and 63.3 % $AP_{.50:.95}$ is rather low. Overall, precision and recall rates are slightly lower for multiple annotations in comparison to single annotations per image. A reason could be that synthetically derived data sets tend to compose overlapping foregrounds the more foregrounds are being used which poses incorrect detection results as also stated in section 4.1. Considering the number of classes, it is evident that multiple classes per image yield an increase in AP and AR by 280 % and 158 %, respectively (Table 4). Therefore, images with single classes on images should be avoided.

	$AP_{.50:.95}$	AR
Single classes	19.9	31.2
Multiple classes	**75.6**	**80.6**

Table 4: Performance for number of classes per image during training (in percent)

4.4 Influence of image pixel dimension

For the investigation regarding different image pixel dimensions, the original image was tiled into sizes ranging between 1440 × 1280 and 960 × 768 pixels as training for larger image sizes result in GPU memory issues and smaller sizes tend display only single or cut objects for original data. The evaluation was performed also on the original image size of 5760 × 3840 pixels to investigate whether tiling has to be performed also for model inference. In general, the evaluation on image sizes larger than 1440 × 1280 yield a sharp drop in precision (Fig. 5) which might occur because region of interests could be assigned to unsuitable feature levels. Also, as image sizes increase, the more GPU memory and inference time is being used. Meanwhile, same image sizes adopted for both training and evaluation show not necessarily a performance boost which demonstrate that images deployed for the evaluation may vary in size and aspect ratio from the input training set. Highest precision results are achieved by the 1440 × 960 image size trained with adequate (6 GB) GPU memory usage. Further, it can be certain that original images may contain multiple objects and classes.

5 Conclusion and outlook

In conclusion, this study shows that deep convolutional neural networks are a suitable choice to automatically detect and classify benthic species in varying underwater environments. Further, large

amount of training data can be synthetically derived to feed deep networks with sufficient information without risking overfitting. The implemented data augmentation strategy is thus not only useful to extend the input data set but also to alleviate class imbalances boosting the performance considerably. When preparing input data, images not necessarily need to exhibit lots of specimen decreasing time spend for annotation. However, images with single specimen and single classes should be avoided as performance may drop significantly. Therefore, larger image sizes such as 1440 × 960 pixels may be used where chances are high that image tiles contain multiple objects. On the contrary, greater image sizes consume more GPU memory and if image sizes exceed a critical threshold, the precision will drop as region of interests may be assigned to wrong feature levels. Next to this challenge, future studies will incorporate more benthic classes and concentrate on their count. //

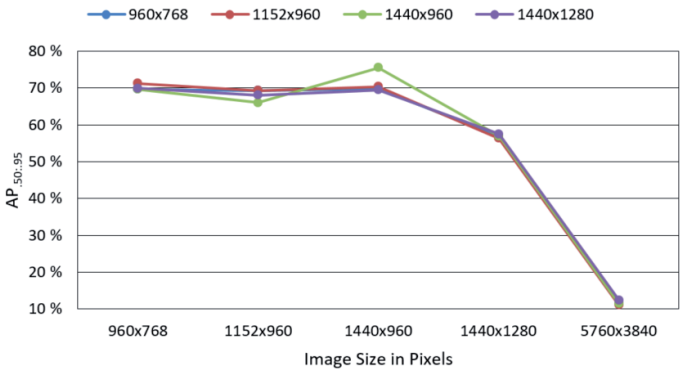

Fig. 5: Average precision regarding the evaluation on various image sizes. Coloured graphs represent the image sizes used for training

References

Boulais, Océane; Ben Woodward et al. (2020): FathomNet: An underwater image training database for ocean exploration and discovery. http://arxiv.org/pdf/2007.00114v3

Brooks, Justin (2019): COCO Annotator: GitHub. https://github.com/jsbroks/coco-annotator/ (accessed 7 January 2020)

Buda, Mateusz; Atsuto Maki; Maciej A. Mazurowski (2018): A systematic study of the class imbalance problem in convolutional neural networks. Neural Networks, DOI: 10.1016/j.neunet.2018.07.011

Christensen, Jesper H.; Lars V. Mogensen et al. (2018): Detection, Localization and Classification of Fish and Fish Species in Poor Conditions using Convolutional Neural Networks. 2018 IEEE/OES Autonomous Underwater Vehicle, DOI: 10.1109/AUV.2018.8729798

Gonzalez-Cid, Yolanda; Antoni Burguera et al. (2017): Machine learning and deep learning strategies to identify Posidonia meadows in underwater images. OCEANS 2017, DOI: 10.1109/OCEANSE.2017.8084991

Griffiths, Huw J.; Andrew J. S. Meijers; Thomas J. Bracegirdle (2017): More losers than winners in a century of future Southern Ocean seafloor warming. Nature Climate Change, DOI: 10.1038/nclimate3377

Guo, Xinjian; Yilong Yin et al. (2008): On the Class Imbalance Problem. 2008 Fourth International Conference on Natural Computation, DOI: 10.1109/ICNC.2008.871

Huang, Hai; Hao Zhou et al. (2019): Faster R-CNN for marine organisms detection and recognition using data augmentation. Neurocomputing, DOI: 10.1016/j.neucom.2019.01.084

Kelly, Adam (2019): COCO Synth: GitHub. https://github.com/akTwelve/cocosynth (accessed 22 February 2020)

Lee, Youngwan; Jongyoul Park (2019): CenterMask: Real-Time Anchor-Free Instance Segmentation. http://arxiv.org/pdf/1911.06667v1

Lin, Tsung-Yi; Michael Maire et al. (2014): Microsoft COCO: Common Objects in Context. http://arxiv.org/pdf/1405.0312v3

Manning, Christopher D.; Prabhakar Raghavan; Hinrich Schütze (2009): Introduction to information retrieval. Cambridge University Press, ISBN: 0521865719

Mishkin, Dmytro; Nikolay Sergievskiy; Jiri Matas (2017): Systematic evaluation of CNN advances on the ImageNet. Computer Vision and Image Understanding, DOI: 10.1016/j.cviu.2017.05.007

Mitchell, Emily G.; Rowan J. Whittle; Huw J. Griffiths (2020): Benthic ecosystem cascade effects in Antarctica using Bayesian network inference. Communications biology, DOI: 10.1038/s42003-020-01310-8

Pauly, Leo; Harriet Peel et al. (2017): Deeper Networks for Pavement Crack Detection. Proceedings of the 34th International Symposium on Automation and Robotics in Construction (ISARC), DOI: 10.22260/isarc2017/0066

Pavoni, Gaia; Massimiliano Corsini et al. (2021): Challenges in the deep learning-based semantic segmentation of benthic communities from Ortho-images. Applied Geomatics, DOI: 10.1007/s12518-020-00331-6

Piepenburg, Dieter; Alexander Buschmann et al. (2017): Seabed images from Southern Ocean shelf regions off the northern Antarctic Peninsula and in the southeastern Weddell Sea. Earth System Science Data, DOI: 10.5194/essd-9-461-2017

Purser, Autun; Simon Dreutter et al. (2021): Seabed video and still images from the northern Weddell Sea and the western flanks of the Powell Basin. Earth System Science Data, DOI: 10.5194/essd-13-609-2021

Purser, Autun; Yann Marcon et al. (2019): Ocean Floor Observation and Bathymetry System (OFOBS): A New Towed Camera/Sonar System for Deep-Sea Habitat Surveys. IEEE Journal of Oceanic Engineering, DOI: 10.1109/JOE.2018.2794095

Salman, Ahmad; Ahsan Jalal et al. (2016): Fish species classification in unconstrained underwater environments based on deep learning. Limnology and Oceanography Methods, DOI: 10.1002/lom3.10113

Xie, Saining; Ross Girshick et al. (2017): Aggregated Residual Transformations for Deep Neural Networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, DOI: 10.1109/CVPR.2017.634

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.