IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 2, 2020, accepted August 1, 2020, date of publication August 5, 2020, date of current version August 17, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3014441

Unsupervised Knowledge Transfer for Object Detection in Marine Environmental Monitoring and Exploration

MARTIN ZUROWIETZ AND TIM W. NATTKEMPER
Biodata Mining Group, Faculty of Technology, Bielefeld University, 33615 Bielefeld, Germany
Corresponding author: Martin Zurowietz (martin@cebitec.uni-bielefeld.de)

This work was supported in part by the BMBF Project COSEMIO under Grant FKZ 03F0812C, in part by the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) under Grant 031A537B, Grant 031A533A, Grant 031A538A, 031A533B, Grant 031A535A, Grant 031A537C, Grant 031A534A, and Grant 031A532B, in part by the German Research Foundation (DFG), and in part by the Open Access Publication Fund of Bielefeld University.

ABSTRACT The volume of digital image data collected in the field of marine environmental monitoring and exploration has been growing in rapidly increasing rates in recent years. Computational support is essential for the timely evaluation of the high volume of marine imaging data, but often modern techniques such as deep learning cannot be applied due to the lack of training data. In this article, we present Unsupervised Knowledge Transfer (UnKnoT), a new method to use the limited amount of training data more efficiently. In order to avoid time-consuming annotation, it employs a technique we call ''scale transfer'' and enhanced data augmentation to reuse existing training data for object detection of the same object classes in new image datasets. We introduce four fully annotated marine image datasets acquired in the same geographical area but with different gear and distance to the sea floor. We evaluate the new method on the four datasets and show that it can greatly improve the object detection performance in the relevant cases compared to object detection without knowledge transfer. We conclude with a recommendation for an image acquisition and annotation scheme that ensures a good applicability of modern machine learning methods in the field of marine environmental monitoring and exploration.

INDEX TERMS Object detection, knowledge transfer, deep learning, marine environmental monitoring, image annotation.

I. INTRODUCTION

Digital imaging is nowadays a popular technique in the marine sciences as it is a non-invasive method for monitoring and exploring marine habitats on a large scale (e.g. biodiversity estimation or ecological management). Thanks to recent technological advances in high-resolution digital imaging and digital storage technology, mobile marine observation platforms such as autonomous underwater vehicles (AUV) or ocean floor observation systems (OFOS) are capable to acquire large volumes of imaging data in a short time [1]. The sustainable curation and management of terabyte-scale volumes of marine imaging data is a challenge that has only recently been addressed [2].

The analysis of marine imaging datasets is usually performed manually with dedicated software tools like SQUIDLE+ [3], VARS [4] or BIIGLE 2.0 [5], which are specialized for the the task of manual image annotation. In contrast to other areas of computer science, where image annotation refers to the assignment of semantics to images as a whole (e.g. describing the scene in the image), image annotation in this context refers to the assignment of class labels (e.g. a species name selected from a certain taxonomy) to several points or regions in an image [6]–[8] (see Fig. 1). This type of manual image annotation is a time-consuming and error-prone task [6], [7]. Moreover, usually only domain experts are able to detect the objects of interest (OOI), which can be for example bacterial mats, litter or fauna such as sponges and sea cucumbers, and to select the correct class labels with sufficient accuracy and reproducibility. The growing volume of image data and the limited availability of trained domain experts is a bottleneck problem for the evaluation of marine imaging datasets.

In order to cope with the growing volume of marine imaging data that needs to be analyzed, computer-aided methods have been proposed to automate (or assist in certain steps of)

The associate editor coordinating the review of this manuscript and approving it for publication was Hossein Rahmani.

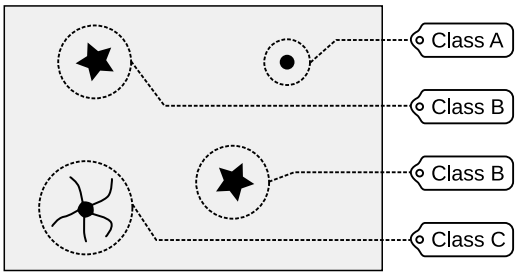


FIGURE 1. Image annotation in the context of marine environmental monitoring and exploration where class labels are assigned to several points or regions in an image. This example shows three classes of marine fauna which were annotated using circles.

image annotation. These include specialized approaches to laser point detection [9], coral reef annotation [10], fish detection [11] or quantification of megafauna [12]. As in many areas of computer vision research, deep learning has become increasingly popular in marine image processing, e.g. for marine object detection [13] or semi-automated image annotation [14]. Most areas of computer vision research focus on large annotated image datasets showing everyday objects from scenes on land. The impressive performance of state-of-the-art deep learning models is based on such datasets (e.g. ImageNet [15] or COCO [16]). The field of marine image processing, however, lacks such large numbers of annotations in images due to the limited availability of domain experts capable of annotating the images. This leads to a ''vicious circle'' in which marine scientists are not able to produce large annotated image datasets because adequate computer support is not available and state-of-the-art methods for adequate computer support cannot be developed because large annotated image datasets are not available.

One approach for situations where precise image annotations are costly to acquire is weakly supervised object detection [17]–[19]. This technique uses weakly labeled images or videos where only the whole image or video frame is labeled instead of a precise region in the image or video frame to train a machine learning model for object detection. Weak labels for images or videos can be created much faster than precise image annotations for all objects in an image. However, weakly supervised object detection methods are unsuitable for datasets where dozens or even hundreds of objects of different classes may occur in a single image, which is not uncommon in marine imaging datasets.

Another technique for dealing with situations where only limited data is available to train a deep learning model is transfer learning [20]. Transfer learning is usually applied in deep learning by reusing the weights of the neural network acquired by training on one ''source'' dataset as the starting point for training on another ''target'' dataset. It has been shown that transfer learning is also effective when the source dataset from which the reused weights are derived is from a completely different visual domain than the target dataset (e.g. everyday objects vs. marine images [14], [21]). In most cases, however, transfer learning still requires supervised training on the target dataset because the (number of) object classes and the high-level visual properties of the objects are different.

In addition to the general problem of a lack of annotated training data, computer-aided marine image annotation suffers from another problem that makes the application of transfer learning difficult. Common scenarios in marine environmental monitoring and exploration are research cruises, where images of the sea floor are captured during multiple deployments of AUVs or OFOSs. Often the image datasets collected on a single cruise cover a specific geographical area and show similar habitats and OOI [22]–[25]. Between different deployments, the visual properties of the photographed habitats and OOI may change due to different cruising speeds of the observation platforms, different camera gear or different distances of the observation platforms to the sea floor. These differences may result in varying degrees of motion blur, distorted colors caused by the water column between the camera and the sea floor, or different scales of OOI of the same class. From a pattern recognition perspective, the distribution of data described by the visual properties of a given class of OOI changes (or drifts) between datasets. This phenomenon is referred to as ''concept drift'' by the computer vision community [26]. In most of the literature, ''gradual'' concept drift is described as a dynamic signal flow that changes continuously over time (e.g. a slowly wearing piece of factory equipment might cause a gradual change in the quality of output parts). In our context, however, a ''sudden'' concept drift is described where the visual properties of OOI change instantaneously from one dataset to the other. Langenkämper *et al.* [27] have recently shown that sudden concept drift can have a strong impact on the performance of deep learning classifiers applied for megafauna image classification collected at the same site but with changing gear and operation. Schoening *et al.* [28] have shown that concept drift is even challenging for purely manual image annotation by trained experts.

Despite the challenges mentioned above, some applications of computer vision for underwater image analysis have been proposed in the last years. Most of these methods focus only on a single class or very few classes of OOI such as fish or corals [13]. Recently we proposed the Machine learning Assisted Image Annotation (MAIA) method [14], which does not distinguish between classes of OOI and can be used in a broader context. MAIA consists of four successive stages. In Stage I unsupervised novelty detection is used to generate a list of possible OOI, which are manually filtered in Stage II. The filtered OOI are used to train a machine learning model for object detection in Stage III. In the last stage, the final detections are again manually filtered and class labels are manually assigned to the OOI to produce the final image annotations. However, all methods, including MAIA, are meant to be applied independently to each new dataset and do not allow the reuse of existing training data for object detection in a new dataset, as they do not account for concept drift between different datasets.

Considering the lack of annotated training data and the high cost of time-consuming manual image annotation, it is desirable to have a computer vision system for automated or assisted image annotation that does not require extensive retraining for each new dataset. Such a computer vision system must be able to adapt to the changes between datasets as described above, where OOI of the same class may differ in their visual appearance. Assisted by such a system, marine scientists would only have to annotate one dataset, and the time required for object detection, which is the most time-consuming part of manual image annotation [6], would be greatly reduced for the remaining datasets of the same geographical area. In this way, the knowledge consisting of images and annotations that were collected previously can be transferred and is not lost.

Knowledge transfer in the context of marine environmental monitoring and exploration has previously been presented by Skaldebø *et al.* [29], who attempt to transfer the knowledge obtained in a simulated underwater environment to the real environment. First, artificial 3D-rendered images are created, showing scenes similar to the real images. Then CycleGAN [30] is used to make the artificial images look more realistic. Walker *et al.* [31] use physics based color correction and scale normalization on underwater images to reduce the generalization error of a DeepLabV3+ model [32] for image segmentation. Similarly, Yamada *et al.* [33] use color correction and image rescaling to enhance their method for unsupervised feature learning of georeferenced sea floor images. All methods are applied to a single dataset and are not used for knowledge transfer to enable cross-dataset machine learning.

In this article we present Unsupervised Knowledge Transfer (UnKnoT), a new method for object detection in marine environmental monitoring and exploration. The method employs a technique we call ''scale transfer'' and enhanced data augmentation to adapt one image dataset to the visual properties of another image dataset and to reuse existing image annotations for object detection. To the best of our knowledge, UnKnoT is the first method that addresses the reuse of existing image annotations for cross-dataset machine learning in marine environmental monitoring and exploration. To evaluate the method, we introduce four fully annotated marine image datasets collected in the same geographical area but with varying gear and distance to the sea floor. Our experiments show that UnKnoT can greatly improve the object detection performance in the relevant cases compared to object detection without knowledge transfer. In combination with the existing MAIA method, UnKnoT can be used instead of novelty detection in Stage I of MAIA to generate more accurate suggestions for OOI if the images of the datasets show the same classes of OOI. Taking this into account, we conclude with a recommendation for an image acquisition and annotation scheme that ensures a good applicability of modern machine learning methods in the field of marine environmental monitoring and exploration.

In the following section, the UnKnoT method is presented in detail, describing the individual steps for scale transfer, data augmentation and object detection (see Section II). Code has been made available with this publication and can be accessed on GitHub.[1] The experimental setup that was used to evaluate the method is presented in Section III, including the four datasets, referred to as S083, S155, S171 and S233. The datasets have been made available with this publication [34]–[37] and can be visually explored in BIIGLE 2.0[2]. The experimental results are summarized in Section IV and discussed in Section V. The manuscript ends with a conclusion about the relevance of our results and the UnKnoT method for marine image annotation.

II. METHODS

In the UnKnoT approach, knowledge is represented by a source dataset D^{s} and annotations that were manually created by domain experts. The knowledge is transferred by transforming D^{s} and the annotations so a deep learning model can be trained to perform object detection on a target dataset D^{t} which has not been annotated. The entire process consists of three consecutive steps which are described in detail in the following sections (see also Fig. 2). In the first step, scale transfer is applied to the images I^{s} of the annotated source dataset D^{s}. This transforms the visible OOI in I^{s} to a similar scale than the OOI in the images of the target dataset D^{t}. A set of annotation patches $A^{\mathrm{s}\to\mathrm{t}}$ is extracted from the scale-transferred images, where each annotation patch is a cropped image centered on an annotated OOI. In the second step, enhanced data augmentation is applied to increase the size and variety of the set of annotation patches $A^{\mathrm{s}\to\mathrm{t}}$, resulting in the set of augmented annotation patches $A^{\mathrm{s}\rightrightarrows\mathrm{t}}$. In the final step, the set $A^{\mathrm{s}\rightrightarrows\mathrm{t}}$ is used to train a Mask R-CNN model [38] which is subsequently applied for object detection on the target dataset D^{t}.

A. SCALE TRANSFER

On most deployments of an AUV or OFOS, the observation platform moves at a fixed distance to the sea floor. This ensures an almost stable scale and illumination of OOI in the images that are captured during the same deployment. The distance to the sea floor may vary between two deployments, though. An OFOS is usually operated much closer to the sea floor than an AUV and even the same observation platform can be operated at different target distances on different deployments. This can result in highly different scales for the same classes of OOI in different image datasets (see Fig. 4). Fully convolutional neural networks for instance segmentation or object detection such as Mask R-CNN [38] are usually scale-invariant because they are trained on large image datasets in which many scales of objects of the same class occur. In this context, however, the scales of OOI of

[1] https://github.com/BiodataMiningGroup/unknot

[2] https://biigle.de/projects/237, login: `unknot@example.com`, password: `UnKnoTpaper`

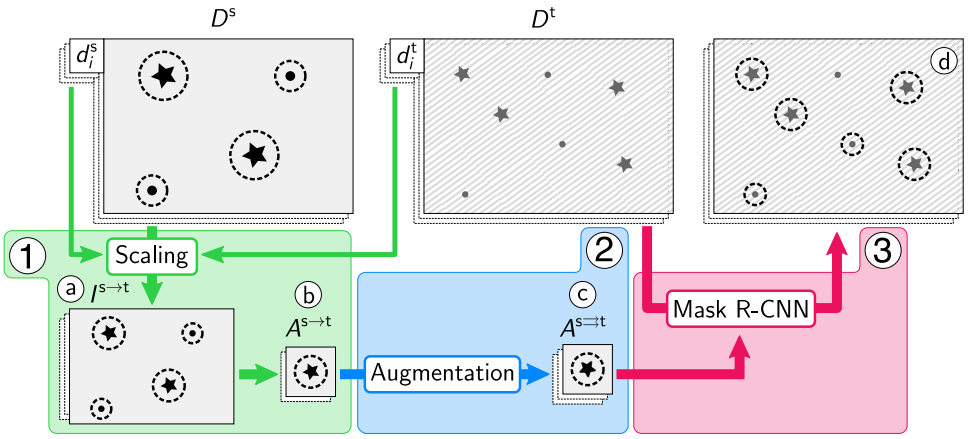


FIGURE 2. The UnKnoT method. (1) Scale transfer: Images from an annotated source dataset D^s are transformed to the set of scale-transferred images $I^{s\to t}$ (a) and the set of annotation patches $A^{s\to t}$ is extracted (b). (2) Data augmentation: The size and variety of the annotation patches $A^{s\to t}$ is increased through data augmentation, resulting in the set of augmented annotation patches $A^{s\rightrightarrows t}$ (c). (3) Object detection: A Mask R-CNN model is trained on $A^{s\rightrightarrows t}$ and applied to the images of D^t to produce the final object detections (d).

the same class and dataset have a very low variance owing to the fixed distance of the observation platform to the sea floor. In addition, the datasets usually have a much lower total number of annotations than in other scenarios.

To mitigate the scale shift between different datasets, scale transfer transforms the images I^s of an annotated source dataset D^s to make the OOI appear at a similar scale than the OOI in the images I^t of the target dataset D^t. The source dataset $D^s = \{(I_i^s, d_i^s)\}$ and the target dataset $D^t = \{(I_i^t, d_i^t)\}$ consist of tuples of an image I_i and the distance d_i of the observation platform to the sea floor when the image was captured. The average distance to the sea floor of the target dataset is denoted as $\bar{d}^t$:

$$\bar{d}^t = \frac{1}{|I^t|} \sum_{i=1}^{|I^t|} d_i^t \tag{1}$$

Each image $I_i^s \in I^s$ has a width of w_i and a height of h_i pixels. To apply scale transfer to an image I_i^s, the scale transfer factor $d_i^{s\to t}$ is calculated first as defined in (2). Next, each image I_i^s is scaled to the width $w_i^{s\to t}$ and height $h_i^{s\to t}$ as can be seen in (3) and (4), respectively, resulting in the set of scale-transferred images $I^{s\to t}$ (see Fig. 2a). A three-lobe Lanczos kernel is applied for downscaling (i.e. $d_i^{s\to t} < 1$) and a cubic filter is applied for upscaling (i.e. $d_i^{s\to t} > 1$) which are the recommended methods of the VIPS image processing library [39].

$$d_i^{s\to t} = \frac{d_i^s}{\bar{d}^t} \tag{2}$$

$$w_i^{s\to t} = w_i \cdot d_i^{s\to t} \tag{3}$$

$$h_i^{s\to t} = h_i \cdot d_i^{s\to t} \tag{4}$$

From each image in $I^{s\to t}$ the annotated OOI are extracted as 512×512 pixel crops which form the set of annotation patches $A^{s\to t}$ (see Fig. 2b). The annotation patches are passed to the next step for data augmentation.

B. DATA AUGMENTATION

Data augmentation is often used to increase the size and variety of data that is available to train a machine learning model. This can often improve the performance of the trained model [40], [41]. In the context of computer vision tasks such as object detection or classification, common augmentation methods include operations such as horizontal or vertical flipping, rotation or blurring of images. Viable augmentation operations highly depend on the visual domain of the image datasets (e.g. vertical flipping makes sense for the image of a football but not for a face).

In case of images of the sea floor captured with an AUV or OFOS, augmentation operations such as flipping, rotation or blurring can be applied. The OOI in the images are mostly living organisms with a symmetric shape, which makes the flipping operations viable. In addition, the OOI in the images are photographed from the top, so they can occur at any rotation angle. Different camera properties, motion of the observation platform or optical distortion by the water column can introduce varying degrees of blur. An object detection model that was trained partially with blurred images through data augmentation can be more robust for these cases.

In case of UnKnoT, the machine learning model is trained with images of one dataset and applied to images of another dataset. The images can be captured with different observation platforms and different cameras, and are usually available as JPEG files. Different camera and storage settings can produce JPEG files with a varying degree of compression, which can introduce characteristic compression artifacts in the images. We propose to use artificial JPEG compression as augmentation operation to make an object detection model more robust for the application on different datasets.

In UnKnoT, data augmentation is applied to the annotation patches $A^{s \to t}$ at each step during training of the Mask R-CNN model (see the following section). For each step, a random selection of zero to all of the following augmentation operations is applied: horizontal flipping, vertical flipping, rotation by 90, 180 or 270 degrees, Gaussian blur with a random standard deviation $\sigma \in [1.0, 2.0]$ and artificial JPEG compression with a random compression factor $c \in [25, 50]$. The set of augmented annotation patches is denoted as $A^{s \rightrightarrows t}$ (see Fig. 2c).

C. OBJECT DETECTION

Object detection is performed in a similar way than in Stage III of the MAIA method [14] which has been shown to be effective in this particular context, with a few differences that are described in the following. In Stage III of MAIA, a Mask R-CNN model [42] is trained on an augmented set of training samples using pre-trained weights of the COCO dataset [16]. The trained model is applied to an image collection for the segmentation of ''interesting'' pixel regions in the images, which are subsequently converted into circle annotations. In UnKnoT, the Mask R-CNN model is trained using the set $A^{s \rightrightarrows t}$ of augmented annotation patches, as well as the pre-trained weights of the COCO dataset. The data augmentation used in Stage III of MAIA is replaced by the enhanced data augmentation described in the previous section. Different to the training configuration of MAIA and [42], a value of 0.85 is used for the `RPN_NMS_THRESHOLD`, which increases the number of region proposals during training. In this context, a higher number of region proposals during training is beneficial for the detection of very small objects in the presence of very large and salient objects in the same image. In addition, a stepped learning rate decay is used to improve convergence of the object detection performance of experiment replicates. For the stepped learning rate decay, the `heads` layers are trained for 10 epochs each with a learning rate of 10^{-3}, $5 \cdot 10^{-4}$ and 10^{-4}, and all layers for another 10 epochs each with a learning rate of 10^{-4}, $5 \cdot 10^{-5}$ and 10^{-5}, resulting in a total of 60 training epochs compared to the 30 epochs of the training configuration of MAIA. One epoch consists of 400 steps and in each step, a batch of five images is processed. Training took about five hours per dataset on a single NVIDIA Tesla V100. Inference is performed on the images I^t of the target dataset in the same way than in Stage III of MAIA [14] (see Fig. 3). The final result is a set of circle annotations, enclosing potential OOI in I^t (see Fig. 2d).

III. EXPERIMENTAL SETUP

Four fully annotated image datasets were created to evaluate the UnKnoT method. The datasets were captured in the same geographical area, showing the same classes of OOI, but with different observation platforms and distances to the sea floor. In addition, a new metric to measure the effectiveness of UnKnoT was created, which accounts for the desired properties of an object detection method for marine image annotation. The method was tested in comprehensive experiments on different combinations of datasets to evaluate the effectiveness of scale transfer and enhanced data augmentation for unsupervised knowledge transfer.

A. DATASETS

The four image datasets used to evaluate UnKnoT are referred to as S083, S155, S171 and S233. Each dataset consists of 550 randomly selected images from the image collections [22] (S083), [23] (S155), [24] (S171) and [25] (S233). The image collections were acquired during the 2015 cruises SO242/1 and SO242/2 of research vessel SONNE at the Peru Basin Disturbance and Colonization (DISCOL) area [43]. The images of the different datasets were captured using different observation platforms (OFOS and AUV) as well as different average distances to the sea floor (see Table 1).

TABLE 1. Properties of the four datasets that were used to evaluate UnKnoT with the observation platform, average distance and standard deviation of the camera to the sea floor, and the number of images and annotations in the train and test splits.

Dataset	Avg. dist. ($\bar{d}^t$)	$\lvert I_{\text{train}} \rvert$	$\lvert A_{\text{train}} \rvert$	$\lvert I_{\text{test}} \rvert$	$\lvert A_{\text{test}} \rvert$
S083 (AUV)	7.62 ± 0.89 m	490	1,808	60	203
S155 (OFOS)	1.70 ± 0.33 m	514	2,107	36	236
S171 (OFOS)	1.61 ± 0.25 m	485	2,061	65	234
S233 (OFOS)	3.33 ± 0.39 m	494	3,719	56	416

The image annotations are based on a subset of ten morphological classes of the fauna identification guide presented in [28] (see Fig. 4). The images were annotated in BIIGLE 2.0 [5] using MAIA [14] with an additional review using the Lawnmower tool to annotate OOI that were missed by MAIA. In total, the datasets contain 10,784 manual annotations on 2,200 images. Compared to datasets of other research areas such as the detection of everyday objects, the datasets presented here may seem rather small. However, the acquisition of annotations in marine images is much more costly, as it requires more training and background knowledge in marine biology. This makes it infeasible to generate datasets as large as e.g. COCO [16] to evaluate machine learning methods in this research area.

The datasets S083, S155, S171 and S233 have been made available with this publication [34]–[37]. Example images with annotations can be found in the supplementary material.

B. EVALUATION METRIC

A common metric to evaluate the performance of an object detection method is the mean average precision [44]. In this context, object detections are produced based on the segmentation output of Mask R-CNN as described in [14] (see Fig. 3). This allows only the calculation of the ''recall'' (i.e. the percentage of OOI that were detected) and the ''precision'' (i.e. the percentage of correct detections in the final result) but does not allow the ranking of the detections, so the mean average precision is not applicable. Another metric which is the harmonic mean of the recall and precision is

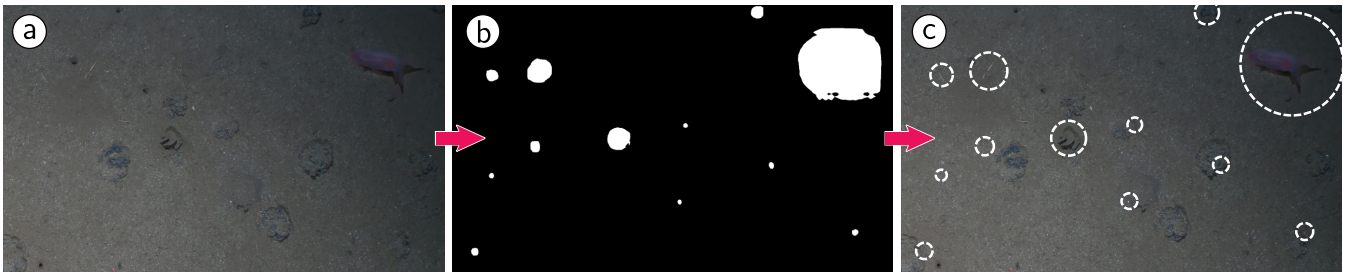


FIGURE 3. Example for inference with Mask R-CNN and the final object detection on a subsection of image `TIMER_2015_09_04at09_13_31IMG_0864.jpg` of the S155 dataset. The image (a) is processed by Mask R-CNN which returns a segmentation mask for "interesting" pixels (b). The regions of interesting pixels are converted to circle annotations for the final detections (c). The full image with manual annotations can be found in the supplementary material.

the F_1-Score [45]. A variant of the F_1-Score is the F_2-Score which puts a higher weight on the recall and which has been used in a similar context to evaluate the object detection performance of the MAIA method [14]:

$$F_2(\text{recall}, \text{precision}) = \frac{5 \cdot \text{precision} \cdot \text{recall}}{(4 \cdot \text{precision}) + \text{recall}} \tag{5}$$

In case of an object detection method for images in marine environmental monitoring and exploration, a minimum of 80% for the recall and 10% for the precision can be considered acceptable [14]. The F_2-Score does not take this into account. For example, it is possible to achieve a higher F_2-Score based on a precision of 20% and a recall of 70% than an F_2-Score based on a precision of 10% and a recall of 80%. In this context, the latter result would be more desirable and should yield a higher score in the evaluation. As a consequence, we do not apply the F_2-Score as the primary metric to evaluate UnKnoT. Instead, we propose the "Logistic Score" (L-Score) as a new metric which is better suited to evaluate marine object detection regarding a minimum recall of 80% and a minimum precision of 10%. The L-Score is the harmonic mean of the two logistic functions L_r to assess the recall and L_p to assess the precision (see (6), (7) and (8)). L_r is centered on the value of 80% recall with a growth rate that yields $L_r(1) \approx 1$ (see Fig. 5a) and L_p is centered on the value of 10% precision with a growth rate that yields $L_p(0) \approx 0$ (see Fig. 5b). The L-Score produces high scores for a recall close to or greater than 80% and a precision close to or greater than 10%, and low scores otherwise (see Fig. 5c).

$$L_r(\text{recall}) = \frac{1}{1 + e^{-0.25 \cdot (100 \cdot \text{recall} - 80)}} \tag{6}$$

$$L_p(\text{precision}) = \frac{1}{1 + e^{-0.5 \cdot (100 \cdot \text{precision} - 10)}} \tag{7}$$

$$L(\text{recall}, \text{precision}) = \frac{2 \cdot L_r(\text{recall}) \cdot L_p(\text{precision})}{L_r(\text{recall}) + L_p(\text{precision})} \tag{8}$$

C. EXPERIMENTS

To evaluate the UnKnoT method, each of the four datasets was separated into train and test splits. The test splits consist of images I_{test} that contain $\approx$ 10% of the annotations of the dataset (A_{test}). The train splits consist of the remaining images (I_{train}) and annotations (A_{train}) of the respective dataset (see Table 1). For evaluation, UnKnoT was applied to a given train split as source dataset D^{s} and a given test split as target dataset D^{t}.

All combinations of using two datasets as D^{s} and D^{t} were evaluated in experiments using the following methods for comparison: UnKnoT ($E^{D^{\text{s}} \to D^{\text{t}}}_{\text{sc,tr,au}}$), UnKnoT without enhanced training compared to MAIA ($E^{D^{\text{s}} \to D^{\text{t}}}_{\text{sc,au}}$), UnKnoT without enhanced image augmentation ($E^{D^{\text{s}} \to D^{\text{t}}}_{\text{sc,tr}}$), UnKnoT only with scale transfer ($E^{D^{\text{s}} \to D^{\text{t}}}_{\text{sc}}$) and the baseline configuration of the MAIA object detection stage without any knowledge transfer ($E^{D^{\text{s}} \to D^{\text{t}}}$). The subscripts "sc", "tr" and "au" refer to scale transfer, enhanced training and enhanced image augmentation, respectively. For all experiments except $E^{D^{\text{s}} \to D^{\text{t}}}$ the combinations $D^{\text{s}} = D^{\text{t}}$ were not evaluated as this would mean knowledge transfer within the same dataset. Each experiment was repeated three times and the average L-Score was calculated as final performance. We denote the average resulting L-Score of the experiments $E^{D^{\text{s}} \to D^{\text{t}}}_{\text{sc,tr,au}}$ and $E^{D^{\text{s}} \to D^{\text{t}}}$ as $L^{D^{\text{s}} \to D^{\text{t}}}_{\text{sc,tr,au}}$ and $L^{D^{\text{s}} \to D^{\text{t}}}$, respectively.

IV. RESULTS

The effect of scale transfer that is applied with UnKnoT can be seen in Fig. 6. In case of source dataset S083, scale transfer magnifies the OOI with a factor of $d_i^{\text{s} \to \text{t}} > 1$ (see Fig. 6 first row). In contrast to that, the size of the OOI of the source datasets S155 and S171 is reduced with a factor of $d_i^{\text{s} \to \text{t}} < 1$ during scale transfer, with the exception of $A^{\text{S155} \to \text{S171}}$ where the size is marginally increased (see Fig. 6 second and third row). In case of S233 as source dataset, the size of the OOI is both increased with S155 and S171 as target datasets and decreased with S083 as target dataset (see Fig. 6 fourth row).

Table 2 shows the average resulting L-Scores of the experiments $E^{D^{\text{s}} \to D^{\text{t}}}$ without knowledge transfer and $E^{D^{\text{s}} \to D^{\text{t}}}_{\text{sc,tr,au}}$ with knowledge transfer. The experiments $E^{D^{\text{s}} \to D^{\text{t}}}$ without knowledge transfer show the highest scores for the cases $D^{\text{s}} = D^{\text{t}}$, where the images of source and target come from the same dataset. The experiments $E^{\text{S155} \to \text{S171}}$ and $E^{\text{S171} \to \text{S155}}$ show almost identical scores to the experiments with $D^{\text{s}} = D^{\text{t}}$ of

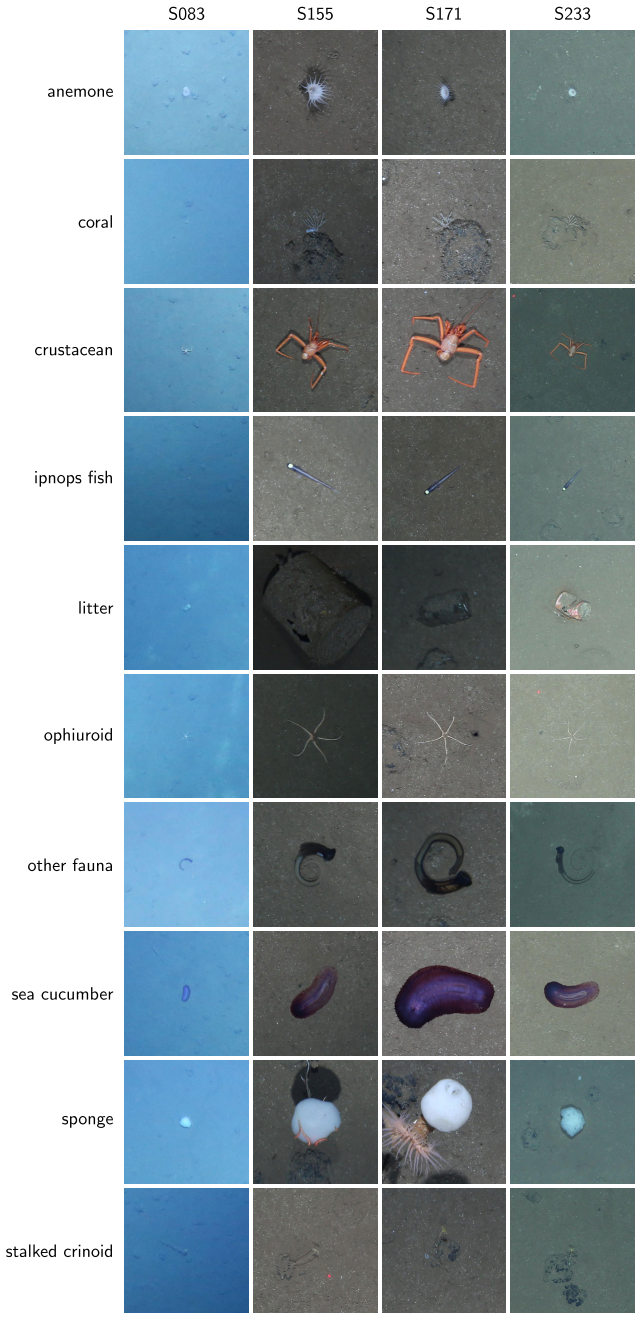


FIGURE 4. Examples for the ten classes of OOI (rows) of each of the four image datasets (columns) that were used to evaluate UnKnoT. Scales of OOI can vary drastically between different datasets.

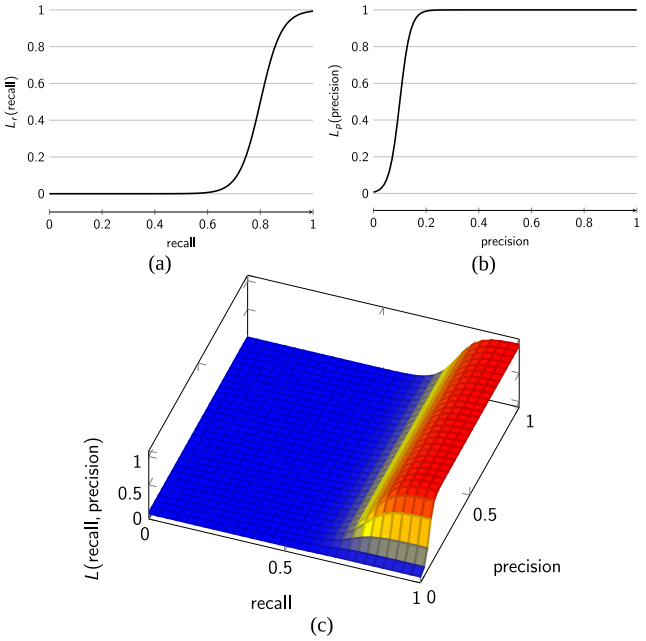


FIGURE 5. The harmonic mean of the two logistic functions L_r (a) and L_p (b) forms the L-Score (c).

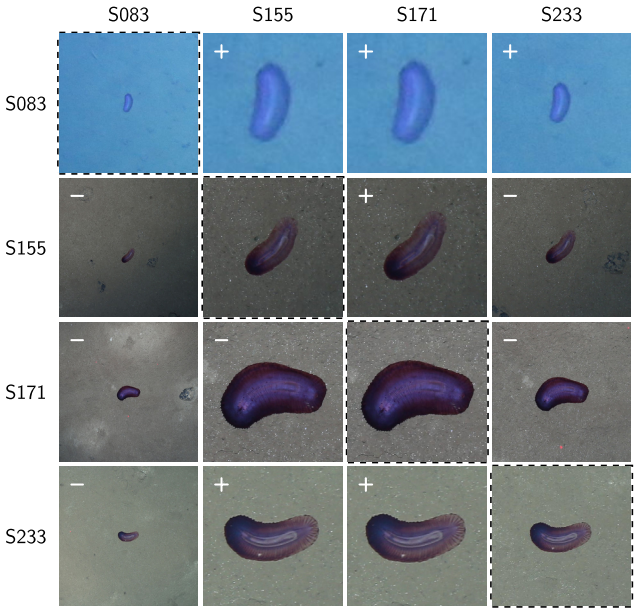


FIGURE 6. Annotation patches of the "sea cucumber" class without scale transfer (dashed outline on the main diagonal) compared to scale-transferred annotation patches of $A^{s\to t}$. The rows denote the source dataset and the columns denote the target dataset (e.g. the patch in the first row and second column is from $A^{S083\to S155}$). Annotation patches produced with a scale transfer factor of $d^{s\to t} > 1$ are marked with a + and patches produced with a scale transfer factor of $d^{s\to t} < 1$ are marked with a −.

these datasets. The experiments $E^{D^s\to S083}$ with $D^s \neq S083$ as well as $E^{S083\to S155}$ and $E^{S083\to S171}$ show a score close to 0.

Eight of the twelve experiments $E^{D^s\to D^t}_{sc,tr,au}$ with knowledge transfer show higher scores than the experiments $E^{D^s\to D^t}$ with the same combination of datasets. The L-Scores are increased by an average of 0.32. However, further inspection of the output of Mask R-CNN reveals invalid segmentation results for the experiments $E^{S083\to S155}_{sc,tr,au}$ and $E^{S083\to S171}_{sc,tr,au}$. In these experiments, the segmentations produced by Mask R-CNN show only crude region proposal boxes instead of the refined regions of a valid segmentation (see Fig. 7). Similarly, the segmentation results for the experiment $E^{S083\to S233}_{sc,tr,au}$ are not as refined as desired. The score of $E^{S233\to S155}_{sc,tr,au}$ is decreased when compared to object detection

TABLE 2. Average resulting *L*-Score of the experiments without knowledge transfer ($L^{D^s \to D^t}$), with knowledge transfer ($L^{D^s \to D^t}_{sc,tr,au}$) and the average increase of the *L*-Score through knowledge transfer. Experiments based on a scale transfer factor of $d_i^{s \to t} < 0.9$ are highlighted.

$D^s \to D^t$	$L^{D^s \to D^t}$	$L^{D^s \to D^t}_{sc,tr,au}$	Increase
S083 → S083	0.95 ± 0.01		
S155 → S083	0.00 ± 0.00	0.85 ± 0.02	0.85
S171 → S083	0.00 ± 0.00	0.79 ± 0.11	0.79
S233 → S083	0.15 ± 0.11	0.93 ± 0.01	0.78
S083 → S155	0.11 ± 0.03	0.00 ± 0.00	-0.11
S155 → S155	0.95 ± 0.01		
S171 → S155	0.92 ± 0.02	0.83 ± 0.01	-0.09
S233 → S155	0.46 ± 0.18	0.32 ± 0.23	-0.14
S083 → S171	0.10 ± 0.02	0.28 ± 0.40	0.18
S155 → S171	0.97 ± 0.00	0.95 ± 0.01	-0.02
S171 → S171	0.96 ± 0.01		
S233 → S171	0.21 ± 0.03	0.91 ± 0.01	0.70
S083 → S233	0.32 ± 0.02	0.71 ± 0.05	0.39
S155 → S233	0.68 ± 0.05	0.92 ± 0.02	0.24
S171 → S233	0.71 ± 0.06	0.96 ± 0.01	0.25
S233 → S233	0.94 ± 0.04		

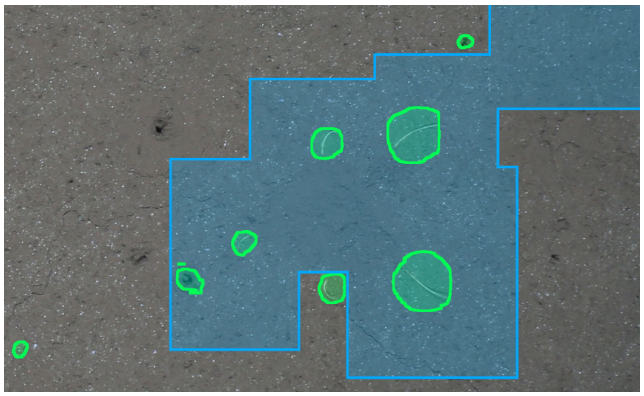


FIGURE 7. Part of an invalid segmentation from the experiment $E^{S083 \to S155}_{sc,tr,au}$ (outlined blue) and a valid segmentation from the experiment $E^{S083 \to S155}$ (outlined green) for comparison. The invalid segmentation shows only crude region proposal boxes instead of the desired refined regions of a valid segmentation. At the center of the image, the five tips of the arms of a burrowed ophiuroid are visible.

without knowledge transfer whereas the score of $E^{S233 \to S171}_{sc,tr,au}$ shows one of the highest valid increases. When the detection is limited to the subset of OOI classes that have an average intra-class area standard deviation of less than 1.5 times their average annotation area (“Coral”, “Crustacean”, “Ipnops fish” and “Ophiuroid”, see Table 3), the *L*-Scores of both experiments converge to 0.58 ± 0.10 (S233 → S155) and 0.86 ± 0.03 (S233 → S171) but are still not equal. All these experiments are exclusively the cases where scale transfer was applied with a factor of $d_i^{s \to t} > 1$. Among the remaining experiments only $E^{S171 \to S155}_{sc,tr,au}$ shows a slightly decreased *L*-Score compared to object detection without knowledge transfer. In this experiment, a scale transfer factor of $0.9 < d_i^{s \to t} < 1$ was applied. The average increase of *L*-Scores of the remaining experiments, where a scale transfer factor of $d_i^{s \to t} < 0.9$ was applied, is 0.58.

The detailed results of all experiments including *L*-Score, recall and precision are presented in Tables 4 and 5.

TABLE 3. Standard deviation of the area of the circle annotations per class and dataset, and the average over all datasets, given as multiples of the average annotation area of the respective class. Rows with an average standard deviation of less than 1.5 are highlighted.

Class	S083	S155	S171	S233	Avg.
Anemone	1.02	0.98	2.10	2.77	1.72
Coral	0.61	1.20	1.74	0.95	1.13
Crustacean	0.45	1.42	1.43	1.06	1.09
Ipnops fish	0.37	0.25	0.29	0.34	0.31
Litter	3.52	0.68	0.83	1.46	1.62
Ophiuroid	0.41	1.08	1.09	0.97	0.89
Other fauna	1.88	3.79	2.61	2.63	2.73
Sea cucumber	0.93	1.07	3.88	1.13	1.75
Sponge	1.12	2.29	1.96	2.54	1.98
Stalked crinoid	0.90	1.99	2.15	1.05	1.52

TABLE 4. Average resulting *L*-Score, recall and precision of the experiments $E^{D^s \to D^t}$ and $E^{D^s \to D^t}_{sc,tr,au}$.

	$D^s \to D^t$	*L*-Score	recall	precision
$E^{D^s \to D^t}$	S083 → S083	0.95 ± 0.01	0.91	0.16
	S155 → S083	0.00 ± 0.00	0.40	0.42
	S171 → S083	0.00 ± 0.00	0.33	0.60
	S233 → S083	0.15 ± 0.11	0.68	0.36
	S083 → S155	0.11 ± 0.03	0.73	0.05
	S155 → S155	0.95 ± 0.01	0.90	0.19
	S171 → S155	0.92 ± 0.02	0.88	0.16
	S233 → S155	0.46 ± 0.18	0.90	0.08
	S083 → S171	0.10 ± 0.02	0.82	0.04
	S155 → S171	0.97 ± 0.00	0.94	0.18
	S171 → S171	0.96 ± 0.01	0.91	0.21
	S233 → S171	0.21 ± 0.03	0.97	0.06
	S083 → S233	0.32 ± 0.02	0.91	0.07
	S155 → S233	0.68 ± 0.05	0.80	0.20
	S171 → S233	0.71 ± 0.06	0.81	0.27
	S233 → S233	0.94 ± 0.04	0.95	0.16
$E^{D^s \to D^t}_{sc,tr,au}$	S155 → S083	0.85 ± 0.02	0.86	0.14
	S171 → S083	0.79 ± 0.11	0.84	0.16
	S233 → S083	0.93 ± 0.01	0.89	0.16
	S083 → S155	0.00 ± 0.00	0.27	0.35
	S171 → S155	0.83 ± 0.01	0.90	0.12
	S233 → S155	0.32 ± 0.23	0.72	0.17
	S083 → S171	0.28 ± 0.40	0.56	0.18
	S155 → S171	0.95 ± 0.01	0.94	0.15
	S233 → S171	0.91 ± 0.01	0.87	0.17
	S083 → S233	0.71 ± 0.05	0.88	0.11
	S155 → S233	0.92 ± 0.02	0.87	0.23
	S171 → S233	0.96 ± 0.01	0.90	0.23

TABLE 5. Average resulting *L*-Score, recall and precision of all experiments with a scale transfer factor of $d_i^{s \to t} < 0.9$.

Experiments	*L*-Score	recall	precision
$E^{D^s \to D^t}$	0.31 ± 0.04	0.60	0.37
$E^{D^s \to D^t}_{sc}$	0.85 ± 0.06	0.87	0.16
$E^{D^s \to D^t}_{sc,tr}$	0.82 ± 0.04	0.86	0.16
$E^{D^s \to D^t}_{sc,au}$	0.86 ± 0.04	0.86	0.19
$E^{D^s \to D^t}_{sc,tr,au}$	0.89 ± 0.03	0.87	0.18

V. DISCUSSION

The UnKnoT method applies knowledge transfer from a source dataset D^s with existing annotations to a target dataset D^t for object detection. The knowledge transfer consists of scale transfer, which adapts the scales of OOI in the source dataset D^s to the scales of OOI in the target dataset D^t,

and of enhanced data augmentation for typical images of the sea floor.

Fig. 6 shows that the scale transfer effectively transforms the scale of OOI of the source dataset to the scale of OOI of the target dataset. First we will review the results obtained for experiments with a scale transfer factor of $d_i^{s\to t} > 1$ (see patches marked with + in Fig. 6). In this scenario, the images of the source dataset have been transformed by upscaling, as the OOI of the target dataset are shown larger and more detailed. In a real setting, the images of the target dataset would have been captured by an AUV or OFOS closer to the sea floor as in the previous dives. In case of S083 as source dataset, the OOI are transformed to a scale that matches the scale of the OOI in the target dataset. However, the scaling blurs the OOI and they do not appear as in focus as the OOI in the target datasets. The results are similar but not as pronounced in case of S233 as target dataset. In the opposite scenario, where the images of the target dataset would have been captured further away from the sea floor than in the previous dives, the images of the source dataset are transformed by downscaling with a scale transfer factor of $d_i^{s\to t} < 1$ (see patches marked with − in Fig. 6). In case of S083 as the target dataset, the scale of the OOI matches the scale of the OOI of the target dataset and the OOI appear in focus. Considering only the visual appearance of the OOI, UnKnoT works more effectively if the source dataset was captured closer to the sea floor than the target dataset and the scale of the annotated OOI is reduced during knowledge transfer. This observation is confirmed by the experimental results.

The experiments $E^{D^s\to D^t}$ with $D^s = D^t$, where the images of source and target come from the same dataset, show the highest average L-Scores. This is to be expected, as Mask R-CNN is trained with OOI that appear most similar to the OOI that should be detected. These experiments can be seen as baseline with the best possible object detection performance in this context. Notably, the experiments $E^{S171\to S155}$ and $E^{S155\to S171}$ show a score almost equal to $E^{S155\to S155}$ and $E^{S171\to S171}$, respectively. Although these datasets differ in the distribution of annotations in the images (cf. $|I_{test}|$ and $|A_{test}|$ in Table 1), both datasets were captured at a similar distance to the sea floor with an OFOS. When Mask R-CNN is trained on one dataset and applied to the other, no knowledge transfer is required to achieve a very good object detection performance. Other notable results are given by the experiments $E^{D^s\to S083}$ with $D^s \neq$ S083, as well as $E^{S083\to S155}$ and $E^{S083\to S171}$ which show a score close to 0. Such a low L-Score is produced if either the recall is bad (i.e. < 80%), the precision is bad (i.e. < 10%) or both. In case of the three experiments $E^{D^s\to S083}$ with $D^s \neq$ S083, the low average recall of 47% is the cause for the low L-Score (see Table 4). Trained with OOI at a much larger scale, Mask R-CNN is unable to achieve an adequate recall in these cases. For the experiments $E^{S083\to S155}$ and $E^{S083\to S171}$, the low average precision of 5% causes the low L-Score (see Table 4). Again, the high difference in the scale of OOI is the cause for the bad object detection performance.

The experiments $E^{D^s\to D^t}_{sc,tr,au}$ can be separated into the same two scenarios as the annotation patches of Fig. 6 mentioned above, where the scale transfer is only effective in the cases where a scale transfer factor of $d_i^{s\to t} < 1$ is applied.

The experiments where the source dataset D^s has a higher average distance to the sea floor than the target dataset D^t belong to the first scenario. Even though UnKnoT produces an improved object detection performance in some of these cases, the segmentation results of Mask R-CNN are invalid (i.e. not as refined as desired) or the object detection performance is highly affected by the intra-class area standard deviation of the annotations. An invalid segmentation (as can be seen in Fig. 7) can be the result of OOI that were highly distorted by a large scale transfer factor $d_i^{s\to t} \gg 1$ so the trained Mask R-CNN model cannot produce a meaningful segmentation for the target dataset. Although the datasets S155 and S171 are very similar in terms of the average distance of the camera to the sea floor, they show very different L-Scores in the experiments $E^{S233\to S155}_{sc,tr,au}$ and $E^{S233\to S171}_{sc,tr,au}$. A closer look at the intra-class area standard deviations of the annotations reveals that the compositions of annotations of some classes differ between these datasets (see Table 3). A high intra-class area standard deviation can be amplified by scale transfer and can potentially result in unrealistically large OOI in the annotation patches of the source dataset. A limited amount of training samples per class and an equally high intra-class standard deviation in the target dataset can lead to highly different object detection performances, even if the source datasets were captured at a similar average distance to the sea floor. When limited only to classes that show an average intra-class area standard deviation of less than 1.5 times their average annotation area (see Table 3), the L-Scores produced by the experiments converge, but are still not equally high. This confirms the observation that UnKnoT is not well suited for cases where the source dataset was captured at a higher distance to the sea floor than the target dataset.

The experiments where the source dataset D^s has a lower average distance to the sea floor than the target dataset D^t belong to the second scenario. Among these cases only $E^{S171\to S155}_{sc,tr,au}$, in which a scale transfer factor of $0.9 < d_i^{s\to t} < 1$ was applied, shows a slightly decreased L-Score compared to object detection without knowledge transfer. This highlights a drawback of the proposed L-Score, as only small changes in the precision and/or recall can cause high differences in the L-Score if the score is already high. In case of $E^{S171\to S155}_{sc,tr,au}$, the lower L-Score is produced by a slightly lower precision of 12% compared to 16% of $E^{S171\to S155}$ and an actually slightly higher recall of 90% compared to 88% of $E^{S171\to S155}$ (see Table 4). Still, even if UnKnoT does not have a negative impact on the object detection performance in this case, it does not improve the performance either. Hence, we only denote the experiments in which a scale transfer factor of $d_i^{s\to t} < 0.9$ was applied

as "relevant". These are the cases with a sufficiently large difference in the average distance of the camera to the sea floor. On average, UnKnoT improves the object detection performance by an L-Score of 0.58 (189%) compared to object detection without knowledge transfer in these cases. Where the experiments $E^{D^s \rightarrow S083}$ with $D^s \neq S083$ produced a bad average recall of 47%, UnKnoT improves the average recall to 86% (see Table 4). Notably, the improved object detection performance is highest for S233 → S083 compared to S155 → S083 and S171 → S083. Also, the object detection performance is improved to a similarly high level for S155 → S233 and S171 → S233, in case of $E_{sc,tr,au}^{S171 \rightarrow S233}$ even surpassing the baseline average L-Score of $E^{S233 \rightarrow S233}$. These results indicate that UnKnoT produces a better object detection performance with a source dataset that was captured at an average distance to the sea floor that is roughly half the average distance of the target dataset.

Considering only the relevant experiments, scale transfer accounts for most of the improvements in the object detection performance. The additional enhanced training configuration of Mask R-CNN and the data augmentation improve the object detection performance even further (see Table 5).

VI. CONCLUSION

Based on the observations and experimental results we draw the following conclusions: If the annotated source dataset and the target dataset are very similar in terms of average distance to the sea floor and observation platform, no knowledge transfer is required to achieve a good object detection performance with a machine learning model such as Mask R-CNN. If the annotated source dataset was captured at roughly half the distance to the sea floor than the target dataset, UnKnoT can be used to greatly improve the object detection performance in an unsupervised way. As the discrepancy in average distances to the sea floor increases, the increase in object detection performance by UnKnoT decreases, but the final object detection is still much better than if no knowledge transfer is performed.

To ensure a good applicability of machine learning methods such as UnKnoT for marine image annotation, we propose a four-step image acquisition and annotation scheme for future studies of the same geographical area:

1) One dataset with images of the sea floor should be captured close to the ground and the current distance to the sea floor should be recorded for each image. The images should be fully annotated in a manual way using a software such as BIIGLE 2.0 [5]. A target distance to the sea floor of 1.7 m should be preferred as OOI are likely to be easy to identify at this distance. Methods to assist image annotation such as MAIA [14] can be used to speed up the image annotation process.
2) The remaining image datasets should be captured at twice the distance to the sea floor than the dataset from Step 1 and should also record the current distance to the sea floor for each image. Image acquisition can be done on a large scale using observation platforms such as AUVs. Following Step 1, the preferred target distance to the sea floor should be 3.4 m. At this distance, the images cover a larger area than the images of Step 1, potentially containing more OOI (cf. $|A_{train}|$ in Table 1).
3) UnKnoT should be used for object detection with the annotated dataset of Step 1 as source dataset and each of the datasets acquired in Step 2 as target dataset.
4) MAIA [14] should be used for the final image annotation of each of the datasets acquired in Step 2, by using the object detection results of Step 3 as training proposals. The object detection results of Step 3 replace the results of the novelty detection stage of MAIA and ensure a highly specialized Mask R-CNN model for each individual dataset in the instance segmentation stage.

This image acquisition and annotation scheme can be an efficient way to produce large volumes of high-quality image annotations in typical scenarios of the field of marine environmental monitoring and exploration.

In summary, we presented UnKnoT, a new method for unsupervised knowledge transfer that allows the reuse of existing knowledge in the form of image annotations for object detection in new marine image datasets that show similar OOI. In addition, we presented the L-Score, a metric that is better suited to evaluate the object detection performance in this particular context. We evaluated the effectiveness of UnKnoT with four fully annotated image datasets, comprising a total of 10,784 annotations on 2,200 images captured in the same geographical area at different distances to the sea floor. Our experimental results have shown that UnKnoT greatly improves the object detection performance compared to object detection without knowledge transfer in the relevant cases. Based on these results, we conclude by recommending a four-step image acquisition and annotation scheme for future studies, which can be an efficient way to produce large volumes of high-quality image annotations in the field of marine environmental monitoring and exploration.

REFERENCES

[1] K. J. Morris, B. J. Bett, J. M. Durden, V. A. I. Huvenne, R. Milligan, D. O. B. Jones, S. McPhail, K. Robert, D. M. Bailey, and H. A. Ruhl, "A new method for ecological surveying of the abyss using autonomous underwater vehicle photography," *Limnol. Oceanography, Methods*, vol. 12, no. 11, pp. 795–809, Nov. 2014.

[2] T. Schoening, K. Köser, and J. Greinert, "An acquisition, curation and management workflow for sustainable, terabyte-scale marine image analysis," *Sci. Data*, vol. 5, no. 1, Dec. 2018, Art. no. 180181.

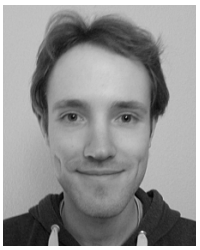
[3] R. Proctor, T. Langlois, A. Friedman, S. Mancini, X. Hoenner, and B. Davey, "Cloud-based national on-line services to annotate and analyse underwater imagery," in *Proc. IMDIS Int. Conf. Mar. Data Inf. Syst.*, vol. 59, 2018, p. 49.

[4] B. Schlining and N. Stout, "MBARI's video annotation and reference system," in *Proc. OCEANS*, Sep. 2006, pp. 1–5.

[5] D. Langenkämper, M. Zurowietz, T. Schoening, and T. W. Nattkemper, "BIIGLE 2.0–browsing and annotating large marine image collections," *Frontiers Mar. Sci.*, vol. 4, p. 83, Mar. 2017.

[6] T. Schoening, J. Osterloff, and T. W. Nattkemper, "Recomia—Recommendations for marine image annotation: Lessons learned and future directions," *Frontiers Mar. Sci.*, vol. 3, p. 59, Apr. 2016.

[7] J. Monk, N. S. Barrett, D. Peel, E. Lawrence, N. A. Hill, V. Lucieer, and K. R. Hayes, "An evaluation of the error and uncertainty in epibenthos cover estimates from AUV images collected with an efficient, spatially-balanced design," *PLoS ONE*, vol. 13, no. 9, Sep. 2018, Art. no. e0203827.
[8] J. M. Durden, B. J. Bett, T. Schoening, K. J. Morris, T. W. Nattkemper, and H. A. Ruhl, "Comparison of image annotation data generated by multiple investigators for benthic ecology," *Mar. Ecol. Prog. Ser.*, vol. 552, pp. 61–70, Jun. 2016.
[9] T. Schoening, T. Kuhn, M. Bergmann, and T. W. Nattkemper, "DELPHI—Fast and adaptive computational laser point detection and visual footprint quantification for arbitrary underwater image collections," *Frontiers Mar. Sci.*, vol. 2, p. 20, Apr. 2015.
[10] O. Beijbom, P. J. Edmunds, D. I. Kline, B. G. Mitchell, and D. Kriegman, "Automated annotation of coral reef survey images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1170–1177.
[11] X. Li, M. Shang, H. Qin, and L. Chen, "Fast accurate fish detection and recognition of underwater images with fast R-CNN," in *Proc. OCEANS-MTS/IEEE Washington*, Oct. 2015, pp. 1–5.
[12] T. Schoening, M. Bergmann, J. Ontrup, J. Taylor, J. Dannheim, J. Gutt, A. Purser, and T. W. Nattkemper, "Semi-automated image analysis for the assessment of megafaunal densities at the arctic deep-sea observatory HAUSGARTEN," *PLoS ONE*, vol. 7, no. 6, Jun. 2012, Art. no. e38179.
[13] M. Moniruzzaman, S. M. S. Islam, M. Bennamoun, and P. Lavery, "Deep learning on underwater marine object detection: A survey," in *Proc. Int. Conf. Adv. Concepts Intell. Vis. Syst.* Antwerp, Belgium: Springer, 2017, pp. 150–160.
[14] M. Zurowietz, D. Langenkämper, B. Hosking, H. A. Ruhl, and T. W. Nattkemper, "MAIA—A machine learning assisted image annotation method for environmental monitoring and exploration," *PLoS ONE*, vol. 13, no. 11, Nov. 2018, Art. no. e0207498.
[15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[16] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. ECCV*. Zurich, Switzerland: Springer, 2014, pp. 740–755.
[17] P. Tang, X. Wang, S. Bai, W. Shen, X. Bai, W. Liu, and A. Yuille, "PCL: Proposal cluster learning for weakly supervised object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 1, pp. 176–191, Jan. 2020.
[18] D. Zhang, J. Han, L. Yang, and D. Xu, "SPFTN: A joint learning framework for localizing and segmenting objects in weakly labeled videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 475–489, Feb. 2020.
[19] D. Zhang, J. Han, G. Guo, and L. Zhao, "Learning object detectors with semi-annotated weak labels," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 12, pp. 3622–3635, Dec. 2019.
[20] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.
[21] E. C. Orenstein and O. Beijbom, "Transfer learning and deep feature extraction for planktonic image data sets," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2017, pp. 1082–1088.
[22] J. Greinert, T. Schoening, K. Köser, and M. Rothenbeck, *Seafloor Images and Raw Context Data Along AUV Track SO242/1_83-1_AUV10 (Abyss_196) During SONNE Cruise SO242/1*. Bremerhaven, Germany: PANGAEA, 2017, doi: 10.1594/PANGAEA.881896.
[23] A. Purser *et al.*, *Seabed Photographs Taken Along OFOS Profile SO242/2_155-1 During SONNE Cruise SO242/2*. Bremerhaven, Germany: PANGAEA, 2018, doi: 10.1594/PANGAEA.890617.
[24] A. Purser *et al.*, *Seabed Photographs Taken Along OFOS Profile SO242/2_171-1 During SONNE Cruise SO242/2*. Bremerhaven, Germany: PANGAEA, 2018, doi: 10.1594/PANGAEA.890620.
[25] A. Purser *et al.*, *Seabed Photographs Taken Along OFOS Profile SO242/2_233-1 During SONNE Cruise SO242/2*. Bremerhaven, Germany: PANGAEA, 2018, doi: 10.1594/PANGAEA.890633.
[26] A. Tsymbal, "The problem of concept drift: Definitions and related work," *Comput. Sci. Dept., Trinity College Dublin*, vol. 106, no. 2, p. 58, 2004.
[27] D. Langenkämper, R. van Kevelaer, A. Purser, and T. W. Nattkemper, "Gear-induced concept drift in marine images and its effect on deep learning classification," *Frontiers Mar. Sci.*, vol. 7, p. 506, Jul. 2020.
[28] T. Schoening, A. Purser, D. Langenkämper, I. Suck, J. Taylor, D. Cuvelier, L. Lins, E. Simon-Lledó, Y. Marcon, D. O. B. Jones, T. Nattkemper, K. Köser, M. Zurowietz, J. Greinert, and J. Gomes-Pereira, "Megafauna community assessment of polymetallic-nodule fields with cameras: Platform and methodology comparison," *Biogeosciences*, vol. 17, no. 12, pp. 3115–3133, Jun. 2020. [Online]. Available: https://www.biogeosciences.net/17/3115/2020/
[29] M. Skaldebo, A. S. Muntadas, and I. Schjolberg, "Transfer learning in underwater operations," in *Proc. OCEANS-Marseille*, Jun. 2019, pp. 1–8.
[30] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.
[31] J. Walker, T. Yamada, A. Prugel-Bennett, and B. Thornton, "The effect of physics-based corrections and data augmentation on transfer learning for segmentation of benthic imagery," in *Proc. IEEE Underwater Technol. (UT)*, Apr. 2019, pp. 1–8.
[32] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. ECCV*, Sep. 2018, pp. 801–818.
[33] T. Yamada, A. P. Bennett, and B. Thornton, "Learning features from georeferenced seafloor imagery with location guided autoencoders," *J. Field Robot.*, pp. 1–16, May 28, 2020, doi: 10.1002/rob.21961.
[34] M. Zurowietz, *S083*, Jan2020, doi: 10.5281/zenodo.3600132.
[35] M. Zurowietz, *S155*, Jan. 2020, doi: 10.5281/zenodo.3603803.
[36] M. Zurowietz, *S171*, Jan. 2020, doi: 10.5281/zenodo.3603809.
[37] M. Zurowietz, *S171*, Jan. 2020, doi: 10.5281/zenodo.3603815.
[38] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2961–2969.
[39] J. Cupitt and K. Martinez, "VIPS: An image processing system for large images," *Proc. SPIE*, vol. 2663, pp. 19–28, Feb. 1996.
[40] S. C. Wong, A. Gatt, V. Stamatescu, and M. D. McDonnell, "Understanding data augmentation for classification: When to warp?" in *Proc. Int. Conf. Digit. Image Comput., Techn. Appl. (DICTA)*, Nov. 2016, pp. 1–6.
[41] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," 2017, *arXiv:1712.04621*. [Online]. Available: http://arxiv.org/abs/1712.04621
[42] W. Abdulla. (2017). *Mask R-CNN for Object Detection and Instance Segmentation on Keras and TensorFlow*. [Online]. Available: https://github.com/matterport/Mask_RCNN
[43] E. J. Foell, H. Thiel, and G. Schriever, "DISCOL: A long-term, large-scale, disturbance-recolonization experiment in the abyssal eastern tropical South Pacific Ocean," in *Proc. Offshore Technol. Conf.*, 1990. [Online]. Available: https://doi.org/10.4043/6328-MS
[44] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.
[45] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manage.*, vol. 45, no. 4, pp. 427–437, Jul. 2009.



MARTIN ZUROWIETZ received the B.Sc. degree in bioinformatics and the M.Sc. degree in informatics in the natural sciences from Bielefeld University, Bielefeld, Germany, in 2013 and 2016, respectively, where he is currently pursuing the Ph.D. degree with the Biodata Mining Group. His research interests include automatic object detection in marine imagery using deep learning methods, assistance systems for manual marine image annotation, and the development of large-scale web-based collaborative image annotation platforms.

TIM W. NATTKEMPER is currently a Professor of biodata mining with the Faculty of Technology, Bielefeld University, Germany. His research interests include the development of methods for the analysis of digital images and video (bioimaging, medical imaging, marine imaging, remote, and sensing). One particular focus of TWNs research is the development of algorithmic approaches to harvest large marine image and sensor data collections for hidden regularities. Two very important aspects are the computational classification/quantification with machine learning and computer vision and the integration of field expert knowledge through modern web-platforms and data-driven visualizations.