frontiers in Marine Science

ORIGINAL RESEARCH
published: 14 July 2020
doi: 10.3389/fmars.2020.00506

Gear-Induced Concept Drift in Marine Images and Its Effect on Deep Learning Classification

*Daniel Langenkämper[1], Robin van Kevelaer[1], Autun Purser[2] and Tim W. Nattkemper[1]**

[1] Biodata Mining Group, Faculty of Technology, Bielefeld University, Bielefeld, Germany, [2] Helmholtz Centre for Polar and Marine Research, Alfred Wegener Institute (AWI), Bremerhaven, Germany

In marine research, image data sets from the same area but collected at different times allow seafloor fauna communities to be monitored over time. However, ongoing technological developments have led to the use of different imaging systems and deployment strategies. Thus, instances of the same class exhibit slightly shifted visual features in images taken at slightly different locations or with different gear. These shifts are referred to as concept drift in the domains computational image analysis and machine learning as this phenomenon poses particular challenges for these fields. In this paper, we analyse four different data sets from an area in the Peru Basin and show how changes in imaging parameters affect the classification of 12 megafauna morphotypes with a 34-layer ResNet. Images were captured using the ocean floor observation system, a traditional sled-based system, or an autonomous underwater vehicle, which is used as an imaging platform capable of surveying larger regions. ResNet applied on separate individual data sets, i.e., without concept drift, showed that changing object distance was less important than the amount of training data. The results for the image data acquired with the ocean floor observation system showed higher performance values than data collected with the autonomous underwater vehicle. The results from this concept drift studies indicate that collecting image data from many dives with slightly different gear may result in training data well-suited for learning taxonomic classification tasks and that data volume can compensate for light concept drift.

Keywords: concept drift, deep learning, marine imaging, machine learning, time-series

OPEN ACCESS

Edited by:
Hervé Claustre,
Centre National de la Recherche Scientifique (CNRS), France

Reviewed by:
Simone Marini,
National Research Council (CNR), Italy
Pedro A. Ribeiro,
University of Bergen, Norway

***Correspondence:**
Tim W. Nattkemper
tim.nattkemper@uni-bielefeld.de

Specialty section:
This article was submitted to Ocean Observation, a section of the journal Frontiers in Marine Science

Received: *28 February 2020*
Accepted: *03 June 2020*
Published: *14 July 2020*

Citation:
Langenkämper D, van Kevelaer R, Purser A and Nattkemper TW (2020) Gear-Induced Concept Drift in Marine Images and Its Effect on Deep Learning Classification. Front. Mar. Sci. 7:506. doi: 10.3389/fmars.2020.00506

1. INTRODUCTION

Recent developments in machine learning-based classification and object detection in computer vision has been greatly influenced by deep learning algorithms (LeCun et al., 2015). In "classic" pattern recognition, engineering skills and experiences were necessary to design a pipeline of algorithmic steps to map images to semantic categories using handcrafted feature descriptors and shallow learning architectures. In particular, the initial steps of pre-processing and feature computation required a considerable amount of experience and domain knowledge. At present, the currently available computation power and new concepts for improving the back-propagation learning algorithm allow the training of large multi-layer networks to learn the entire classification process, including signal transformation and feature representation, given the availability of sufficient training data. In parallel, marine research and environmental monitoring have advanced

on the technological level, as new digital imaging hardware with increased storage capacities, higher resolution and improved image contrast, in combination with next-generation research platforms became available. These platforms, such as AUV (autonomous underwater vehicle, Wynn et al., 2014), OFOS (ocean floor observation system, Purser et al., 2018), ROV (remote operating vehicle, Christ and Wernli Sr, 2013), and FUO (fixed underwater observatory, Godø et al., 2014) enable researchers to collect large numbers of digital images from the field, orders of magnitude higher than 20 years ago. For a more in-depth look at image-based monitoring solutions have a look at (Bicknell et al., 2016), Mallet and Pelletier (2014), and Aguzzi et al. (2019). These image collections may provide valuable information on the taxonomic composition of habitat communities. Furthermore, changes in these communities across the spatial and/or temporal domains can be recorded. First attempts to link these two developments have been successful and showed the potential of deep learning in e.g., morphotype detection (Zurowietz et al., 2018), morphotype classification (Smith and Dunbabin, 2007; Gobi, 2010; Beijbom et al., 2012; Bewley et al., 2012; Kavasidis and Palazzo, 2012; Schoening et al., 2012; Langenkämper et al., 2018, 2019; Mahmood et al., 2019; Piechaud et al., 2019) or polyp behavior monitoring (Osterloff et al., 2019). However, all these studies have reported results obtained for data sets collected with the same gear, i.e., with one distinct camera system and the platform for the full analyzed data set. But in large scale studies, for instance, those ranging over a series of cruises and/or years, gear often changes with each deployment or is operated in different ways. As a consequence, the same fauna morphotype may well be recorded in images with transformed features in the contrasting research project data sets. The colors may be shifted and the textural features or some morphotype characteristics may be more or less visible. In addition, some morphotypes might be of lesser abundance in some data sets. These discrepancies in the appearance of particular morphotypes in the different data sets are referred to as "concept drift." Concept drifts can have a significant negative influence on the performance of machine learning classifiers that are trained on one data set, and then re-applied to new "unseen" data (e.g., collected with a different gear), where the performance of the classifier decreases for this new data. As changes in gear and operation for many studies cannot be avoided, the question as to what extent marine imaging can benefit from computer vision research depends on the ability of computer vision systems to compensate for the effects of such concept drifts.

Concept drift problems have been discussed for 20 years inside the machine learning community (see for example the influential early discussion in Schlimmer and Granger, 1986 and in Widmer and Kubat, 1996). A commonly accepted definition of the term "concept drift" from a survey (Gama et al., 2014) is: "In dynamically changing and non-stationary environments, the data distribution can change over time yielding the phenomenon of concept drift." More recent reviews can be found in Žliobaitė et al. (2016) or in Barros and Santos (2018). For this current study, the change in "environment" (referring to the term in the citation above) between different deployments is caused by changes in the gear and its mode of operation. In addition, the location and time of image collection also vary, potentially causing further changes to the visual appearance of the objects of interest in the recorded digital images. To compensate for the common concept drift problem, different methods have been proposed, including ensemble methods (Grachten and Chacón, 2017; Sun et al., 2018; Ren et al., 2019).

In this paper, we investigate the effect of concept drifts on machine learning-based morphotype classification. We present four data sets collected from the same deep-sea seafloor area, with changing gears leading to concept drifts with mild or strong effects. We carried out a series of machine learning studies, starting with a baseline study that applied machine learning classification to standard training, test splits of data from the same set. Next, the concept drift effect was studied by using data from one or more sets and applying the classifier to complementary data from another data set. Finally, we test approaches for compensating the concept drift effects. In addition, we repeated these experiments with a subset of the said data set, which is balanced to eliminate the effect of data imbalance on the results.

2. MATERIALS AND METHODS

2.1. Image Data Sets

The image data sets considered within this study were collected on cruises SO242/1 and SO242/2 of the research vessel SONNE in the Peru basin in the year 2015. During the cruises, four data sets of digital photos were collected from the sea bottom. An OFOS was used for three dives (Purser et al., 2018a,b,c) and an AUV for the fourth (Greinert et al., 2017). The original data is available at PANGAEA (https://pangaea.de/). The relevant parameters of the camera gear used to collect the four data sets are listed in **Table 1**.

While OFOS data sets $\mathcal{D}_1^O$, $\mathcal{D}_2^O$, and $\mathcal{D}_3^O$ were recorded with the same equipment, the fourth data set $\mathcal{D}^A$ was recorded with a different camera carried by an AUV. So the strongest concept drift in feature representation would be expected between any of the $\mathcal{D}^O$ data sets and the AUV data set $\mathcal{D}^A$. The OFOS data set $\mathcal{D}_3^O$ differs from the other sets $\mathcal{D}_{1,2}^O$ in that the OFOS was operating at a significantly higher altitude.

All four data sets were stored and shared for morphotype detection and annotation using BIIGLE 2.0 (Langenkämper et al., 2017). Eight marine biologists from eight different institutions, annotated megafauna morphotypes using a pre-defined set of 23 classes (Schoening et al., 2019).

Each class is a morphotype, with the exception of the one non-biological class "litter". To annotate a morphotype, the users drew a circle around the object using the BIIGLE 2.0 annotation tools. After the annotation task was completed by all users, the 12 most abundant classes were chosen for this study. The other classes were identified as having numbers too low to use for machine learning applications. A threshold of a minimum of 15 samples has been chosen to have at least 3 samples in the test using 20% of all data as test set. For each of the selected 12 classes, one example extracted from data set $\mathcal{D}_3^O$ is shown in **Figure 1A**. The same Figure also shows the abundances of the classes in the four data sets. A square image patch, i.e., the bounding box of the circle annotation, was extracted for all

TABLE 1 | The four data sets were recorded with two kinds of gear (OFOS and AUV) and different camera set ups.

ID	Original name	Gear	Camera	Lens
$\mathcal{D}_1^O$	SO242/2_171-1			
$\mathcal{D}_2^O$	SO242/2_155-1	OFOS	Canon EOS 5D Mark III	Canon EF 24 mm f/1.4L II USM
$\mathcal{D}_3^O$	SO242/2_233-1			
$\mathcal{D}^A$	SO242/1_107-1 _AUV14	AUV	Canon EOS 6D	Canon EF 8–15 mm/4,0 L Fisheye USM

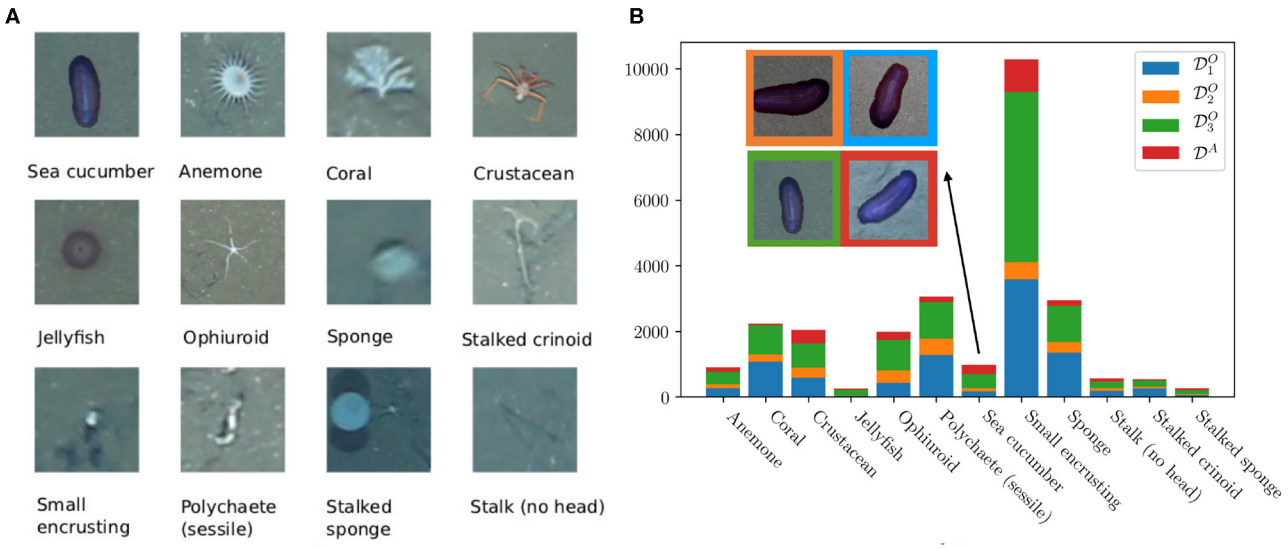

FIGURE 1 | From all 23 detected classes, the 12 most abundant classes of morphotypes were selected for this study. One example of each class from data set $\mathcal{D}_3^O$ is shown in **(A)**. In **(B)** the abundance of each class in each data set is illustrated. In the upper left of the plot, one example of the morphotype "Sea cucumber" from each data set is shown for comparison.

TABLE 2 | The four data sets included between 154 and 311 images.

ID	# images	Image dimensions	Altitude (mean)	Speed (mean)	# patches	# classes
$\mathcal{D}_1^O$	311	$5,760 \times 3,840$	1.6 m ± 0.2	0.5 ± 0.1 knots	9,271	22
$\mathcal{D}_2^O$	206	$5,760 \times 3,840$	1.7 m ± 0.2	0.5 ± 0.1 knots	2,604	22
$\mathcal{D}_3^O$	209	$5,760 \times 3,840$	3.3 m ± 0.2	0.5 ± 0.1 knots	11,533	23
$\mathcal{D}^A$	154	$3,072 \times 2,304/4096 \times 3,072$	4.5–7 m	2.8 knots	2,663	23

The gears were operated with different speed and altitude (both given in estimated mean value). Image patches of the 12 most abundant classes were extracted and used in our study. The total number of extracted patches from these 12 classes are given in the sixth column (see text for details). The last column shows the total number of classes found in each data set (including the ones of lesser abundance).

instances of the 12 classes. These patches were then resized to a uniform size of 256 × 256. In **Table 1**, for each data set the total number of images, the images size, the distance to the ground (average altitude of the OFOS/AUV), the average speed of the OFOS/AUV, the number of annotations of the 12 classes, i.e., extracted patches, and the total number of differently annotated classes is given.

In addition we generated another data set containing only the four classes Crustacean, Sponge, Ophiuroid, Polychaete (sessile), subsampled to the least common abundance in each of $\mathcal{D}_1^O, \mathcal{D}_2^O, \mathcal{D}_3^O, \mathcal{D}^A$, to eliminate the influence of data imbalance on the problem.

2.2. Concept Drift Visualization With t-SNE Projections

To visualize the concept drift for each of the 12 classes, the patches of all annotated objects were projected to the two-dimensional space using dimension reduction. Different methods can be used for dimension reduction and projection, such as Principal Component Analysis (PCA) (Jolliffe, 2011), Self-Organizing Maps (SOM) (Kohonen, 2000), Local Linear Embedding (LLE) (Roweis and Saul, 2000), or Sammon Mapping (Sammon, 1969). Here we applied the t-distributed stochastic neighbor embedding (t-SNE) introduced by van der Maaten and Hinton (2008). The basic idea of t-SNE is that

it minimizes the Kullback–Leibler divergence between two probability distributions. One distribution describes the high-dimensional data point distribution (here the 256^2-dimensional image patches). The second distribution is defined in the low-dimensional space (here the 2D space used for scatter plot visualization). This method has shown good projection results preserving the data topology and avoiding dense cluttering for the majority of points more successfully than other methods, such as the PCA. These features render this approach a good choice for visualizing high-dimensional data.

2.3. ResNet Deep Learning Architecture and Training

Deep residual networks, also referred to as ResNet, have been proposed recently (Kaiming et al., 2015) to overcome limitations of other deep learning frameworks which suffer from the "vanishing gradient problem" (Glorot and Bengio, 2010). This phenomenon leads to a limitation of the number of layers achievable within the network and thus to a limitation of the complexity of the mapping function to be learnt by the network. The excellent results of ResNet architectures in computer vision tasks renders them an appropriate choice for testing deep learning architectures in marine imaging.

In each experiment (see study design) a ResNet was trained with a training data set $\mathcal{D}^t$ consisting of input-output pairs (X, y), where X is an image patch sample and y is the corresponding class label. Thereafter the trained ResNet classifier was used to classify a disjoint test set $\mathcal{D}^v$ for performance assessment. Different selections of samples for $\mathcal{D}^t$ and $\mathcal{D}^v$ were considered with the data collections listed above.

In all runs, the input dimension was 256^2 (i.e., the number of image patch pixels). The architecture of the ResNet, built with TensorFlow (Abadi et al., 2016) was not changed. The network was trained for 250 epochs. As the number of training samples was limited (see **Table 2**) data augmentation was applied to increase the amount of training data. To this end, each patch was flipped left-right and up-down with a chance of 50% for each of both axes. Second, random brightness adjustments of maximum 20% after image-wise standardization (zero mean and unit variance) was applied. Third, random cropping to a size of 224×224 and zero padding back to 256×256 was also applied.

For the experiments with the balanced data sets, we used patches of size 224^2. The ResNet34 of torchvision by PyTorch (Paszke et al., 2019) was used. Data augmentation was omitted to get a bias free result. However, the network was initialized with ImageNet weights.

2.3.1. Performance Assessment

To measure the accuracy of a classifier, we report the F_1-score as well as the macro F_1-score. For this purpose, we define recall R and precision P on the test sets as follows

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

with TP being the number of true positive classifications, FP the number of false positive classifications and FN the number of false negative classifications (Fawcett, 2006). Recall and precision were computed for each morphotype class ω_j and are referred to as R_{ω_j} and P_{ω_j} respectively. Since the data sets show imbalanced class distributions, i.e., the 12 different classes are represented with significantly different amounts of samples (see **Figure 1**), two different accuracy measures are used to assess the accuracy of the trained classifiers in the different studies.

The F_1 measure represents the average accuracy for the entire set of morphotype classes and the accuracy values of all classes are weighted by the class abundances:

$$F_1 = \frac{1}{\sum_j N_{\omega_j}} \sum_j N_{\omega_j} \frac{2 R_{\omega_j} P_{\omega_j}}{R_{\omega_j} + P_{\omega_j}}. \tag{2}$$

where N_{ω_j} is the number of element in class ω_j.

The idea is that each sample is of the same importance independent of the class it belongs to. Thus, the performance for more abundant classes may dominate the overall F_1 value. Of course, such a strong impact of the high abundant classes can be criticized as the accuracy of low abundant classes is ignored which may result in bad results for such classes which motivates us to consider the second accuracy value as well (see below). However, in practice, a computational class assignment is often followed by some manual posterior visual inspection for low abundant and difficult classes so the F_1 measure above is well motivated.

The macro F_1-score $\hat{F}_1$ is motivated by the assumption that all of the classes are of the same interest to the marine biologists and of the same relevance so the accuracy assessment shall not be biased by a small number of most abundant classes. We therefore compute an average accuracy from the class-wise accuracy values, so that every class has the same impact on the overall accuracy assessment independent of its abundance:

$$\hat{F}_1 = \frac{1}{|\omega|} \sum_j \frac{2 R_{\omega_j} P_{\omega_j}}{R_{\omega_j} + P_{\omega_j}}, \tag{3}$$

with $|\omega|$ as the number of different classes (Zheng et al., 2020). The performance measures used were computed using scikit-learn (Pedregosa et al., 2011).

2.4. Study Design

As outlined in the introduction, a number of experiments were conducted to analyze the quality of the data in regard to machine learning-based classification. A visual depiction of the experimental design is shown in **Figure 2**. ResNet34, was used throughout all experiments for comparability. Each one of the Experiments A–E described below was carried out three times and the average accuracy was determined. The accuracy of the classifier was assessed using standard metrics explained above.

2.4.1. Experiment A: Intra-set Study (No Concept Drift)

For this experiment, each data set was investigated separately. So each individual data set was split into a ratio of 80% training data and 20% test data using stratified sampling, i.e., data were sampled in a way such that each class is represented with the same percentage. Splitting one data set into training and test categories had to be carried out with special care. Sometimes an object in an image was annotated by two different experts so it appeared twice in the data set, potentially with a slightly different position

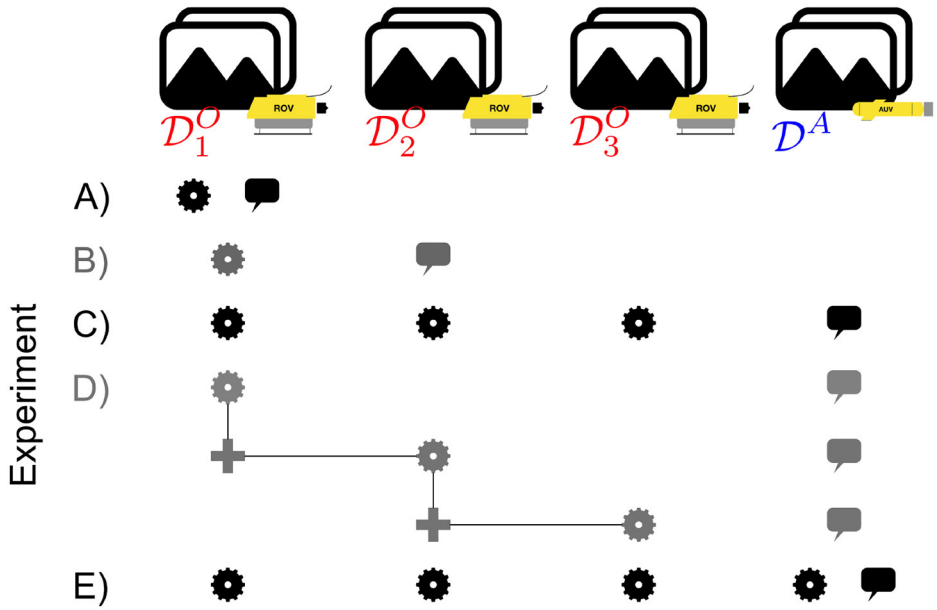

FIGURE 2 | Overview of study design. The cogs mark that this data set is part of the training data and the speech balloons mark the test set. The pluses symbolize the combination of single classifiers to one ensemble. There are multiple experiments for each setup, however, only one instance is shown as an example, i.e., for Experiment B $\mathcal{D}_1^O$ is used as training data and $\mathcal{D}_2^O$ as test data, however, the cases of e.g., $\mathcal{D}_1^O$ (training) and $\mathcal{D}_3^O$ (test) are omitted.

and circle radius. In such cases, both annotations were placed into the training or test set to guarantee that these two sets were truly disjoint. This experiment simulates the case that a part of a big data set is annotated and a classifier is trained to annotate the remaining part. So in these experiments we do not observe any serious concept drift despite small variations in the gear speed or altitude along this transect. Note that most of the studies about machine learning applications to marine image data are carried out this way.

2.4.2. Experiment B: 1 vs. 1 Inter-set Study

For this experiment, the network was trained with all data from one data set $\mathcal{D}^t \in \{\mathcal{D}_1^O, \mathcal{D}_2^O, \mathcal{D}_3^O, \mathcal{D}^A\}$. The trained network was then used to classify all data from another data set $\mathcal{D}^v \in \{\mathcal{D}_1^O, \mathcal{D}_2^O, \mathcal{D}_3^O, \mathcal{D}^A\} \wedge \mathcal{D}^v \neq \mathcal{D}^t$. All in all, four different classifiers were obtained from the four different training data sets. Each classifier was tested on the other three data sets, resulting in 12 sub-experiments. This experiment simulates the case that one previous data set is already annotated and a similar data set should now be automatically annotated with the help of a neural network.

2.4.3. Experiment C: Leave-One-Set-Out Inter-set Study

In this experiment, a classifier was trained with all data from three data sets. Afterwards, the classifier was used to classify the fourth data set. This experiment simulates the case that a series of image collections are already annotated and that a similar data set should now be annotated. All available data with different shifted concepts is used to train the classifier, in order to prepare the network for another concept drift.

2.4.4. Experiment D: Ensemble Classification Heuristic

We implemented a straight forward ensemble classification heuristic that was driven by the results obtained in Experiment B. To classify the data from one set $\mathcal{D}^v$ an ensemble classifier $F(\mathbf{X})$ ($\mathbf{X}$ as one image patch) was constructed from the three classifiers $f_i(\mathbf{X})_{,i=1,2,3}$ that were trained with the three other data sets $\mathcal{D}^t \cap \mathcal{D}^v = \emptyset$, taken from Experiment B (see above). Each patch X from $\mathcal{D}^v$ was classified to one of the 12 classes ω_j according to the following rule:

$$F(\mathbf{X}) = \begin{cases} \omega_j & \text{if } |\{f_i | f_i(\mathbf{X}) = \omega_j\}| \geq 2, \\ f_i(\mathbf{X}) \text{ with } i = \underset{i}{\operatorname{argmax}} |N_i| & \text{else.} \end{cases} \quad (4)$$

To summarize Equation (4), if at least two ensemble members agree, their classification output is chosen. If all three classifiers disagree, the classifier is chosen that was trained by the largest amount of data. Since ensemble methods are reported to work for concept drift problems in other domains, we installed this experiment to either validate or falsify this hypothesis for the domain of marine image informatics.

2.4.5. Experiment E: Leave-One-Set-Out With Adaption

This experiment was carried out in a similar fashion to Experiment C. The training data consisted of the conjunction of three data sets. In addition, we added 20% of all annotations of the fourth data set to the training data to compensate for the concept drift. The trained classifier was applied to the remaining 80% of the fourth data set. When splitting the data of the fourth

data set, we used stratified sampling. This experiment simulated the situation where a large quantity of data is available from past dives and very little data of the new dive has already been annotated. We aimed to investigate here how the classifier can compensate the concept drift using the updated data.

2.4.6. Balanced Data Set Experiments

In addition to the study design presented above, we conducted Experiments A, B, C, and E with a balanced data (cmp. end of section Imager data sets). This way we can analyze the impact of the concept drift without that of data imbalance.

3. RESULTS

The original four different image data sets showed a good overlap in the morphotype composition, especially regarding the 12 most abundant classes chosen for this study. This is important to analyze the impact of concept drift, as otherwise using a classifier trained on one data set to classify a different one would not be possible.

The direct visual browsing and inspection of the images suggests that there might be a concept drift caused by different equipment. In **Figure 1B** it can be seen that the example category "Sea cucumber" from AUV data set $\mathcal{D}^A$ features the lowest contrast and the lowest resolution of details. The number of annotations varied considerably between the data sets. This can be explained by the different locations of the recorded tracks in the habitat, as well as the different altitudes of the AUV/OFOS, i.e., more objects can be found, but they are harder to spot. In the following, we show the results of the t-SNE projections to get a visual representation of the concept drift. This allows us to validate whether a concept drift exists in the data space. Thereafter, we will present the results of the experiments as described above.

3.1. Concept Drift Visualization With t-SNE Projections

In **Figure 3**, we show the 12 scatter plots that were obtained for the 12 different classes using all data sets $\mathcal{D}_1^O$, $\mathcal{D}_2^O$, $\mathcal{D}_3^O$, and $\mathcal{D}^A$. All data sets together were subject to a t-SNE projection into a 2D space. The color of each plotted icon encodes the data set source, following the same scheme as in **Table 1**. Looking at the plots, we observe the trend that the $\mathcal{D}^A$ patches overlap with the $\mathcal{D}_3^O$ most. Let us call this pair $\mathcal{C}_\infty$. In addition, the $\mathcal{D}_1^O$ data seems to overlap with the $\mathcal{D}_2^O$ data, which we call $\mathcal{C}_\in$. This indicates that the members of $\mathcal{C}_\infty$ and $\mathcal{C}_\in$ share some visual qualities within their set. And that the concept drift between $\mathcal{C}_\infty$ and $\mathcal{C}_\in$ is stronger than the drifts between their members.

3.1.1. Experiment A: Intra-set Study

The results of this experiment are shown in **Table 3**. Best results were obtained for the OFOS data set with the higher altitude ($\mathcal{D}_3^O$). The F_1 values for $\mathcal{D}^A$ and $\mathcal{D}_1^O$ are on a similar level. For the results of Experiment A, in most cases observed here the classification performance increases with the amount of data available, i.e., the total number of annotations available (see **Table 2**). For all given data sets, the $\hat{F}_1$ accuracy values were lower when compared with the F_1 accuracy values. The accuracy values for the 12 single classes are provided in **Supplementary Material**. When looking at the results for the balanced data set F_1^{bal} (**Table 2** last row) we see similar results with D_3^O being the best performing data set, except that D_2^O is now performing better than D_1^O and D_A^O.

3.1.2. Experiment B: 1 vs. 1 Inter-set Study

The results of the 12 accuracy measurements are shown in **Table 4**. As expected, the 1 vs. 1 inter-set study produced inferior results compared to those in Experiment A, due to the concept drift. Best results were obtained when $\mathcal{D}_1^O$ or $\mathcal{D}_3^O$ were used for training (see numbers given in boldface). The lowest accuracy values were achieved when the AUV data set $\mathcal{D}^A$ was used for training. Interestingly, a similar altitude helped to compensate a little for the concept drift between AUV and OFOS data, as can be seen when training on $\mathcal{D}_3^O$ and testing on $\mathcal{D}^A$. This combination yielded superior results to others, where $\mathcal{D}^A$ was used as a test set. In addition, this is supported by the t-SNE projection plots, where $\mathcal{D}^A$ and $\mathcal{D}_3^O$ overlap most. Again, the $\hat{F}_1$ results were inferior to the F_1 accuracy values. The results for the balanced data set are shown in **Table S4**. The general observation that the AUV data set $\mathcal{D}^A$ produced the worst results and that the a similar altitude helped to compensate for concept drift still holds true, however the results for $\mathcal{D}_1^O$ and $\mathcal{D}_3^O$ as test sets are ranked differently.

3.1.3. Experiment C: Leave-One-Set-Out Inter-set Study

The four results from this experiment are shown in **Table 5**. In addition to the two accuracy measures, we also list the number of annotations in the union of the three data sets used for training. Comparing the results with those obtained in Experiment A we make the following observations. For three out of four data sets, the performance drops for the F_1 accuracy. However, the results obtained for the $\mathcal{D}_2^O$ data set increase by nine percentage points compared to the intra-set study A. Comparing the $\hat{F}_1$ values in the intra-set Experiment A and the leave-one-set-out inter set Experiment C we notice an increase in the leave-one-set-out inter-set study for data sets $\mathcal{D}_1^O, \mathcal{D}_2^O$ and $\mathcal{D}^A$ with $\hat{F}_1$ accuracy only dropping for $\mathcal{D}_3^O$. Altogether we see that in Experiment C the $\hat{F}_1$ values are lower than the F_1 values, but the difference here is much lower than observed in Experiments A and B. Another observation is that the performance values in Experiment C are always significantly higher than those obtained from Experiment B. The results using the balanced data support the findings stated above.

3.1.4. Experiment D: Ensemble Classification Heuristic

Looking at the results in **Table 6** we can only see small improvements to the results we obtained when we trained the network with just the largest data set in the 1 vs. 1 inter-set study B (see the third row in **Table 4**). For instance, the performance values for the test sets $\mathcal{D}^A$ and $\mathcal{D}_2^O$ do improve a little. However, all the resultant performance values are inferior when compared with the results obtained in Experiment C, when three data sets were joined for one training data set.

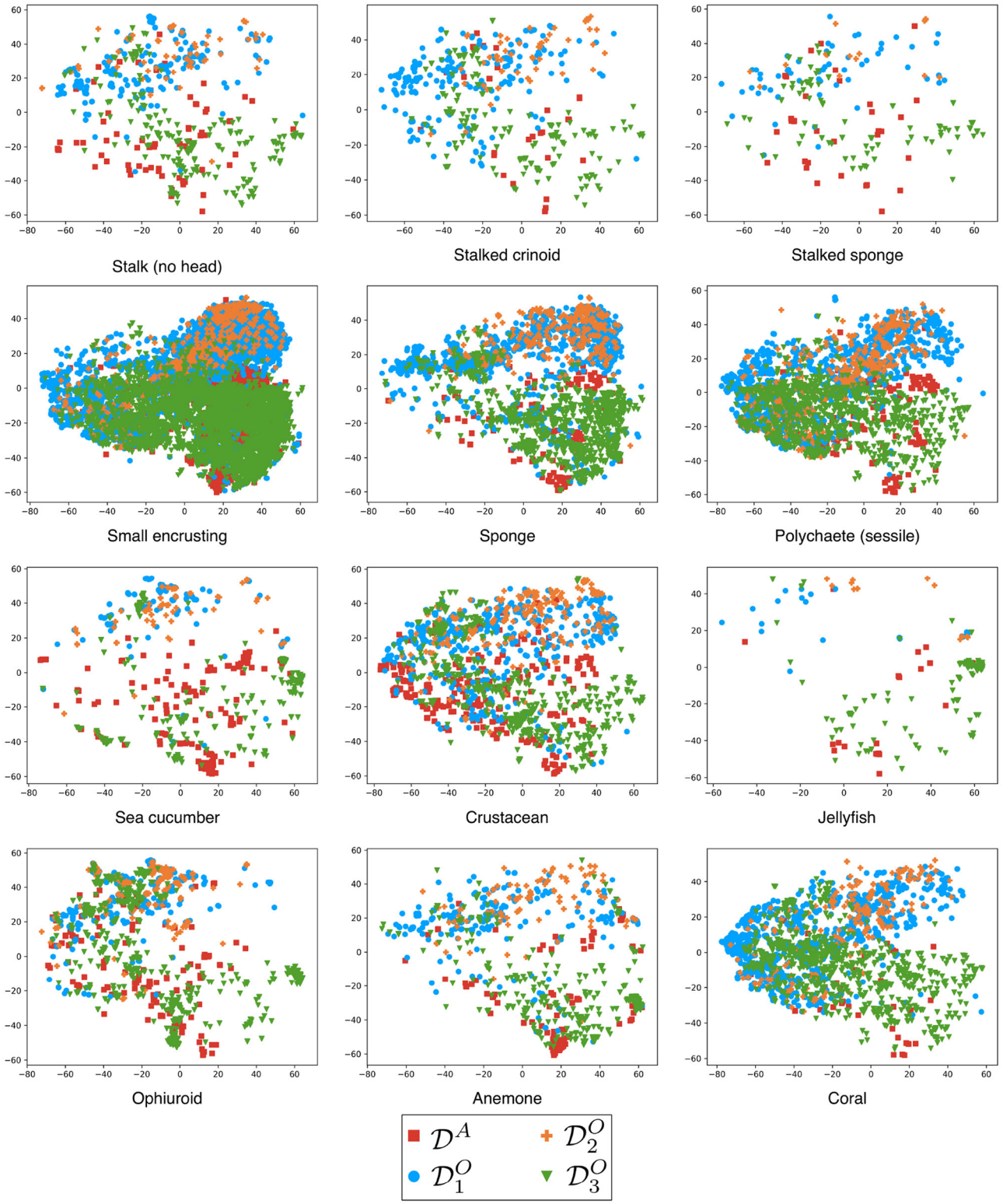

FIGURE 3 | Each scatter plot shows the result of a projection of the image patches of one class for all data sets. The icon color encodes the data set following the same scheme as in **Table 1**.

TABLE 3 | The results of the intra-set study (Experiment A): $\hat{F}_1$ and F_1 scores obtained from training a ResNet-34 on each data set separately and testing it on data from the same data set.

Data set	$\hat{F}_1$	F_1	F_1^{bal}
$\mathcal{D}_1^O$	0.5757	0.7285	0.8333
$\mathcal{D}_2^O$	0.4439	0.6405	0.8753
$\mathcal{D}_3^O$	0.6806	**0.7771**	**0.9202**
$\mathcal{D}^A$	0.4778	0.7015	0.8676

The results for the balanced data set are shown in the last row. Please note that only the F_1-score is shown, here F_1 and $\hat{F}_1$ are the same for the case of a balanced data set. The best result is printed in boldface.

TABLE 4 | The results of the 1 vs. 1 inter-set study (Experiment B): Each row shows the $\hat{F}_1$ and F_1 scores, obtained for the four data sets, of one ResNet-34, trained with one complete data set.

Training set	Test set							
	$\hat{F}_1$				F_1			
	$\mathcal{D}_1^O$	$\mathcal{D}_2^O$	$\mathcal{D}_3^O$	$\mathcal{D}^A$	$\mathcal{D}_1^O$	$\mathcal{D}_2^O$	$\mathcal{D}_3^O$	$\mathcal{D}^A$
$\mathcal{D}_1^O$	–	**0.6253**	**0.5590**	0.3688	–	**0.6902**	**0.7020**	0.5443
$\mathcal{D}_2^O$	0.4723	–	0.4284	0.3270	0.6071	–	0.5530	0.4826
$\mathcal{D}_3^O$	**0.5635**	0.5402	–	**0.4236**	**0.6731**	0.6306	–	**0.6080**
$\mathcal{D}^A$	0.2488	0.3284	0.3616	–	0.4110	0.4060	0.5293	–

In each row the bold face numbers indicate the highest accuracy values achieved for one data set applied as test data.

TABLE 5 | The results of the leave-one-set-out inter-set study (Experiment C): $\hat{F}_1$ and F_1 scores obtained from training a ResNet-34 on three of the four data sets and testing it on the data set that was excluded from the training set.

Test data set	$\hat{F}_1$	F_1	#training patches	F_1^{bal}	#training patches
$\mathcal{D}_1^O$	0.6376	0.7093	16800	0.8166	4820
$\mathcal{D}_2^O$	0.6960	**0.7369**	23467	**0.8342**	5344
$\mathcal{D}_3^O$	0.6209	0.7245	14538	0.8220	3604
$\mathcal{D}^A$	0.4942	0.6705	23408	0.6804	5888

The last two columns refer to the same experiment on the balanced data set. The best result is printed in boldface.

3.1.5. Experiment E: Leave-One-Set-Out Inter-set Study With Adaption

The results in **Table 7** show a small improvement compared to the results of the leave-one-set-out inter-set study in **Table 5**. However, we even see some small decrease in performance for data set $\mathcal{D}_2^O$ which is the smallest data set.

In contrast to the imbalanced data the results for the balanced data set (**Table 7** last row) we see bigger improvements of up to 13% points (D^A) and no decreases in performance.

A summary of the best results for each experiment for each test set is shown in **Table S3**.

TABLE 6 | The results of the ensemble classification heuristic (Experiment D): Average $\hat{F}_1$ and F_1 scores obtained from testing an ensemble of three networks trained in the 1 vs. 1 study applied to the complementary fourth data set.

Test data set	$\hat{F}_1$	F_1^{bal}
$\mathcal{D}_1^O$	0.5516	0.6739
$\mathcal{D}_2^O$	0.5593	0.6447
$\mathcal{D}_3^O$	0.5540	**0.6985**
$\mathcal{D}^A$	0.4454	0.6185

The best result is printed in boldface.

TABLE 7 | The results of the leave-one-set-out inter-set study with a small concept update in Experiment E: $\hat{F}_1$ and F_1 scores obtained from training a ResNet-34 on three of the four data sets and approximately 20% of the fourth, and testing it on the remaining samples of the fourth data set.

Test data set	$\hat{F}_1$	F_1	F_1^{bal}
$\mathcal{D}_1^O$	0.6587	0.7271	0.8341
$\mathcal{D}_2^O$	0.6797	0.7259	**0.8959**
$\mathcal{D}_3^O$	0.6577	**0.7470**	0.8852
$\mathcal{D}^A$	0.5133	0.6885	0.8131

The last row shows the results using the balanced data set. The best result is printed in boldface.

4. DISCUSSION

The results of our experiments show the importance of carrying out such computer vision experiments using data from different dives. The results show limitations for the generalization power of a chosen up-to-date deep learning classification approach training and testing on data from different dives. We also observe a number of interesting trends. Looking at the results of Experiment A, we see that images recorded with a higher altitude seem to gain higher performance values if enough training data is available. Although data sets $\mathcal{D}_1^O$ and $\mathcal{D}_3^O$ have almost the same number of training patches, the performance for the higher altitude data is significantly higher. The same holds for the two small data sets $\mathcal{D}_2^O$ and $\mathcal{D}^A$. The reason for this improved performance may be the reduced motion blur in the higher altitude images. The AUV data is also classified with significantly lower performance than the OFOS data. The AUV is usually operated with a higher speed than the OFOS, so the motion blur is more severe for the AUV acquired data.

When it comes to the concept drift problem, it seems some compensation is possible. The results of Experiment B can be seen as a benchmark for the problems introduced by concept drift. For the given data sets, training a classifier on one of them and classifying one of the remaining data sets did not yield results $>$ 90% not even for balanced data sets from small subsets of classes. Thus, we must assume, that posterior quality assessments and error corrections must be applied by human observers to increase the accuracy. The second main factor may be the training set size. The best results were obtained when the ResNet was trained with one of the two largest data sets $\mathcal{D}_1^O$ or $\mathcal{D}_3^O$. If the three data sets are combined (Experiment C) the results for

the complementary left-out test data set increased significantly (see **Table 5**) when compared to Experiment B. Interestingly, the strongest increase was reported for the small data set $\mathcal{D}_2^O$, even outperforming the intra-set (i.e., concept drift-free) Experiment A. The same is true for three of the data sets regarding the $\hat{F}_1$ measure. So in three cases, the performance values for some less abundant classes benefit from introducing more examples from other data sets.

In our experiment, an ensemble of three separately trained classifiers did not produce improved results. The heuristic approach performed much worse than the leave-one-set-out approach. The leave-one-set-out with adaption approach (Experiment E) produced mixed results compared to the "simple" approach evaluated in Experiment C.

The results from the balanced data set experiments generally support this findings. Nevertheless, the increases by the leave-one-set-out with adaption approach (Experiment E) produce significantly better results. This might be due to more homogeneous data and eliminated data imbalance.

All the presented average classification performance falls short when compared to results from computer vision applications from other domains or public benchmark data (like COCO; Lin et al., 2014). However, we were interested in analyzing concept drift problems induced by image acquisition with different gear of the same area for a mixture of classes. This is the main reason why the results were inferior to other related classification results as the number of training data was rather low, at least for a subset of the classes tested.

In **Tables S1**, **S2**, we show tables of class-wise F_1 and $\hat{F}_1$ values showing acceptable numbers for the classes *Sponge, Small encrusting, Sea cucumber*, and *Ophiuroid* for example. From these numbers, one could conclude that a minimum number of around 500–1,000 examples per class should be collected in the future to gain satisfactory performance values.

5. CONCLUSION

Combining all our observations, we make the following conclusions. Firstly, the concept drift between the four dives is considerable (cmp. **Figure 3**) but can be explained by the differences in survey gear and operation parameters. In this study, the concept is mainly determined by the object distance to the camera and the speed of the imaging platform. Secondly (and not surprisingly) a high number of examples per class for the training set is beneficial for attaining more satisfactory results in the future. Thirdly, combining data from several dives has the clear potential to reduce the impact of concept drift, at least for some low abundance classes. Fourthly, in the context of this particular study, a higher altitude of around 4 m for data collection was found to be preferable than at lower altitudes where motion blur has a greater impact on image quality. Finally, the most simple way of combining data from several dives is by training one network with all the data which achieved the most convincing results. This statement should not be considered as being absolute in the context of other potentially interesting ensemble strategies.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

All authors agree to be accountable for the content of the work. AP selected the data sets, designed the image annotation study, and the morphotype catalog. DL, RK, and TN designed, organized, and carried out the machine learning experiments. AP, DL, RK, and TN discussed and interpreted the results.

FUNDING

This project was financially supported by the BMBF project Miningimpact-DIAS (grant no. 03F0707C) and BMBF project Miningimpact2-COSEMIO (grant no. 03F0812C).

ACKNOWLEDGMENTS

We especially thank the annotators of the images (in alphabetical order): Daphne Cuvelier, Lidia Lins, Yann Marcon, Autun Purser, Erik Simon-Lledo, Inken Suck, James Taylor, and Martin Zurowietz. We acknowledge the financial support of the German Research Foundation (DFG) and the Open Access Publication Fund of Bielefeld University for the article processing charge.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmars.2020.00506/full#supplementary-material

REFERENCES

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., et al. (2016). "Tensorflow: A system for large-scale machine learning," in *12th USENIX Symposium on Operating Systems Design and Implementation (*OSDI *16)* (Savannah, GA), 265–283.

Aguzzi, J., Chatzievangelou, D., Marini, S., Fanelli, E., Danovaro, R., Flögel, S., et al. (2019). New high-tech flexible networks for the monitoring of deep-sea ecosystems. *Environ. Sci. Technol.* 53, 6616–6631. doi: 10.1021/acs.est.9b00409

Barros, R. S. M., and Santos, S. G. T. C. (2018). A large-scale comparison of concept drift detectors. *Inform. Sci.* 451, 348–370. doi: 10.1016/j.ins.2018.04.014

Beijbom, O., Edmunds, P., Kline, D., Mitchell, B., and Kriegman, D. (2012). "Automated annotation of coral reef survey images," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Providence, RI), 1170–1177. doi: 10.1109/CVPR.2012.6247798

Bewley, M., Douillard, B., Nourani-Vatani, N., Friedman, A., Pizarro, O., and Williams, S. (2012). "Automated species detection: an experimental approach

to kelp detection from sea- oor auv images," in *Proceedings of Australasian Conference on Robotics and Automation* (Wellington).

Bicknell, A. W., Godley, B. J., Sheehan, E. V., Votier, S. C., and Witt, M. J. (2016). Camera technology for monitoring marine biodiversity and human impact. *Front. Ecol. Environ*. 14, 424–432. doi: 10.1002/fee.1322

Christ, R. D., and Wernli Sr, R. L. (2013). *The ROV Manual: A User Guide for Remotely Operated Vehicles*. Butterworth-Heinemann.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recogn. Lett*. 27, 861–874. doi: 10.1016/j.patrec.2005.10.010

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., and Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Comput. Surveys* 46:44. doi: 10.1145/2523813

Glorot, X., and Bengio, Y. (2010). "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* (Setti Ballas, CA), 249–256.

Gobi, A. F. (2010). "Towards generalized benthic species recognition and quantification using computer vision," in *OCEANS 2010 IEEE-Sydney*. Seattle, WA: IEEE. 1–6. doi: 10.1109/OCEANSSYD.2010.5603995

Godø, O. R., Johnsen, S., and Torkelsen, T. (2014). The love ocean observatory is in operation. *Mar. Technol. Soc. J.* 48, 24–30. doi: 10.4031/MTSJ.48.2.2

Grachten, M., and Chacón, C. E. C. (2017). Strategies for conceptual change in convolutional neural networks. *arXiv [preprint]* arXiv:1711.01634.

Greinert, J., Schoening, T., Köser, K., and Rothenbeck, M. (2017). Seafloor images and raw context data along AUV track SO242/1_107-1_AUV14 (Abyss_200) during SONNE cruise SO242/1. PANGAEA. In supplement to: Schoening, Timm; Köser, Kevin; Greinert, Jens, An acquisition, curation and management workflow for sustainable, terabyte-scale marine image analysis. *Sci. Data* 5:180181. doi: 10.1038/sdata.2018.181

Jolliffe, I. (2011). *Principal Component Analysis*. Berlin; Heidelberg: Springer Berlin Heidelberg. 1094–1096. doi: 10.1007/978-3-642-04898-2_455

Kaiming, H., Xiangyu, Z., Shaoqing, R., and Jian, S. (2015). Deep residual learning for image recognition. *arXiv [preprint]* arXiv:1512.03385.

Kavasidis, I., and Palazzo, S. (2012). "Quantitative performance analysis of object detection algorithms on underwater video footage," in *Proceedings of the 1st ACM International Workshop on Multimedia Analysis for Ecological Data* (New York, NY: ACM), 57–60. doi: 10.1145/2390832.2390847

Kohonen, T. (2000). *Self-Organizing Maps*. Vol. 30 of Series in Information Sciences. Springer. doi: 10.1007/978-3-642-56927-2

Langenkämper, D., Simon-Lled, E., Hosking, B., Jones, D. O. B., and Nattkemper, T. W. (2019). On the impact of citizen science-derived data quality on deep learning based classification in marine images. *PLoS ONE* 14:e218086. doi: 10.1371/journal.pone.0218086

Langenkämper, D., van Kevelaer, R., and Nattkemper, T. (2018). "Strategies for tackling the class imbalance problem in marine image classification," in *Proc. of CVAUI, ICPR Workshop* (Beijing). doi: 10.1007/978-3-030-05792-3_3

Langenkämper, D., Zurowietz, M., Schoening, T., and Nattkemper, T. W. (2017). Biigle 2.0 - browsing and annotating large marine image collections. *Front. Mar. Sci*. 4:83. doi: 10.3389/fmars.2017.00083

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., et al. (2014). "Microsoft coco: Common objects in context," in *European Conference on Computer Vision*. Zurich: Springer. 740–755. doi: 10.1007/978-3-319-10602-1_48

Mahmood, A., Bennamoun, M., An, S., Sohel, F. A., Boussaid, F., Hovey, R., et al. (2019). Deep image representations for coral image classification. *IEEE J. Ocean. Eng.* 44, 121–131. doi: 10.1109/JOE.2017.2786878

Mallet, D., and Pelletier, D. (2014). Underwater video techniques for observing coastal marine biodiversity: a review of sixty years of publications (1952-2012). *Fish. Res.* 154, 44–62. doi: 10.1016/j.fishres.2014.01.019

Osterloff, J., Nilssen, I., Jarnegren, J., van Engeland, T., Buhl-Mortensen, P., and Nattkemper, T. W. (2019). Computer vision enables short- and long-term analysis of lophelia pertusa polyp behaviour and colour from an underwater observatory. *Sci. Rep*. 9:6578. doi: 10.1038/s41598-019-41275-1

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems* (Vancouver, BC), 8024–8035.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in python. *J. Mach. Learn. Res*. 12, 2825–2830.

Piechaud, N., Hunt, C., Culverhouse, P., Foster, N., and Howell, K. (2019). Automated identification of benthic epifauna with computer vision. *Mar. Ecol. Prog. Ser*. 615, 15–30. doi: 10.3354/meps12925

Purser, A., Marcon, Y., and Boetius, A. (2018a). "Seabed photographs taken along OFOS profile SO242/2_155-1 during SONNE cruise SO242/2," in *Seafloor Images From the Peru Basin Disturbance and Colonization (DISCOL) area Collected During SO242/2*, eds A. Purser, et al. (Bremerhaven: Alfred Wegener Institute; Helmholtz Center for Polar and Marine Research). doi: 10.1594/PANGAEA.890634

Purser, A., Marcon, Y., and Boetius, A. (2018b). "Seabed photographs taken along OFOS profile SO242/2_171-1 during SONNE cruise SO242/2," in *Seafloor Images From the Peru Basin Disturbance and Colonization (DISCOL) Area Collected During SO242/2*, eds A. Purser, et al. (Bremerhaven: Alfred Wegener Institute; Helmholtz Center for Polar and Marine Research).

Purser, A., Marcon, Y., and Boetius, A. (2018c). "Seabed photographs taken along OFOS profile SO242/2_233-1 during SONNE cruise SO242/2," in *Seafloor Images From the Peru Basin Disturbance and Colonization (DISCOL) Area Collected During SO242/2*, eds A. Purser (Bremerhaven: Alfred Wegener Institute; Helmholtz Center for Polar and Marine Research).

Purser, A., Marcon, Y., Dreutter, S., Hoge, U., Sablotny, B., Hehemann, L., et al. (2018). Ocean floor observation and bathymetry system (OFOBS): a new towed camera/sonar system for deep-sea habitat surveys. *IEEE J. Ocean. Eng.* 44, 87–99. doi: 10.1109/JOE.2018.2794095

Ren, S., Zhu, W., Liao, B., Li, Z., Wang, P., Li, K., et al. (2019). Selection-based resampling ensemble algorithm for nonstationary imbalanced stream data learning. *Knowl. Based Syst*. 163, 705–722. doi: 10.1016/j.knosys.2018.09.032

Roweis, S., and Saul, L. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science* 290, 2323–2326. doi: 10.1126/science.290.5500.2323

Sammon, J. W. Jr. (1969). A non-linear mapping for data structure analysis. *IEEE Trans. Comput*. C-18, 401–409. doi: 10.1109/T-C.1969.222678

Schlimmer, J. C., and Granger, R. H. (1986). Incremental learning from noisy data. *Mach. Learn*. 1, 317–354. doi: 10.1007/BF00116895

Schoening, T., Bergmann, M., Ontrup, J., Taylor, J., Dannheim, J., Gutt, J., et al. (2012). Semi-automated image analysis for the assessment of megafaunal densities at the Arctic deep-sea observatory HAUSGARTEN. *PLoS ONE* 7:e38179. doi: 10.1371/journal.pone.0038179

Schoening, T., Purser, A., Langenkämper, D., Suck, I., Taylor, J., Cuvelier, D., et al. (2019). Megafauna community assessment of polymetallic nodule fields with cameras: platform and methodology comparison. *Biogeosci. Discuss*. 1–28. doi: 10.5194/bg-2019-363

Smith, D., and Dunbabin, M. (2007). "Automated counting of the northern pacific sea star in the derwent using shape recognition," in *Digital Image Computing Techniques and Applications, 9th Biennial Conf. of the Australia. Pattern Recognition Soc* (Glenelg, SA), 500–507. doi: 10.1109/DICTA.2007.4426838

Sun, Y., Tang, K., Zhu, Z., and Yao, X. (2018). Concept drift adaptation by exploiting historical knowledge. *IEEE Trans. Neural Netw. Learn. Syst*. 29, 1–11. doi: 10.1109/TNNLS.2017.2775225

van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res*. 9, 2579–2605.

Widmer, G., and Kubat, M. (1996). Learning in the presence of concept drift and hidden contexts. *Mach. Learn*. 23, 69–101. doi: 10.1007/BF00116900

Wynn, R. B., Huvenne, V. A., Le Bas, T. P., Murton, B. J., Connelly, D. P., Bett, B. J., et al. (2014). Autonomous underwater vehicles (AUVs): Their past, present and future contributions to the advancement of marine geoscience. *Mar. Geol*. 352, 451–468. doi: 10.1016/j.margeo.2014.03.012

Zheng, J., Chu, H., Struppa, D., Zhang, J., Yacoub, M., El-Askary, H., et al. (2020). Optimal multi-stage arrhythmia classification approach. *Sci. Rep.* 10, 1–17. doi: 10.1038/s41598-020-59821-7

Žliobaitė, I., Pechenizkiy, M., and Gama, J. (2016). "An overview of concept drift applications," in *Big Data Analysis: New Algorithms for a New Society*, eds N. Japkowicz, and J. Stefanowski (Springer), 91–114. doi: 10.1007/978-3-319-26989-4_4

Zurowietz, M., Langenkmper, D., Hosking, B., Ruhl, H. A., and Nattkemper, T. W. (2018). Maia–a machine learning assisted image annotation method for environmental monitoring and exploration. *PLoS ONE* 13:e207498. doi: 10.1371/journal.pone.0207498

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.