TOWARDS DETECTING FLOATING OBJECTS ON A GLOBAL SCALE WITH LEARNED SPATIAL FEATURES USING SENTINEL 2

Jamila Mifdal[1,*], Nicolas Longépé[1], Marc Rußwurm[2]

[1]Φ-lab, European Space Agency, ESRIN, 00044 Frascati, Italy ({jamila.mifdal,nicolas.longepe}@esa.int)
[2]Chair of Remote Sensing Technology, TUM, Arcissstraße 21, 80333 Munich (marc.russwurm@tum.de)

KEY WORDS: Marine Litter Detection; Floating Objects; Sentinel 2; Oceanography; Deep Learning; CNN

ABSTRACT:

Marine litter is a growing problem that has been attracting attention and raising concerns over the last years. Significant quantities of plastic can be found in the oceans due to the unfiltered discharge of waste into rivers, poor waste management, or lost fishing nets. The floating elements drift on the surface of water bodies and can be aggregated by processes, such as river plumes, windrows, oceanic fronts, or currents. In this paper, we focus on detecting big patches of floating objects that can contain plastic as well as other materials with optical Sentinel 2 data. In contrast to previous work that focuses on pixel-wise spectral responses of some bands, we employ a deep learning predictor that learns the spatial characteristics of floating objects. Along with this work, we provide a hand-labeled Sentinel 2 dataset of floating objects on the sea surface and other water bodies such as lakes together with pre-trained deep learning models. Our experiments demonstrate that harnessing the spatial patterns learned with a CNN is advantageous over pixel-wise classifications that use hand-crafted features. We further provide an analysis of the categories of floating objects that we captured while labeling the dataset and analyze the feature importance for the CNN predictions. Finally, we outline the limitations of trained CNN on several systematic failure cases that we would like to address in future work by increasing the diversity in the dataset and tackling the domain shift between regions and satellite acquisitions. The dataset introduced in this work is the first to provide public large-scale data for floating litter detection and we hope it will give more insights into developing techniques for floating litter detection and classification. Source code and data are available at `https://github.com/ESA-PhiLab/floatingobjects`.

1. INTRODUCTION

Marine litter consists of all human-created trash discharged in the ocean, such as cigarettes, bags, beverage bottles. According to the United Nations Environment Program[1], roughly 70% of marine litter such as glass and metal sinks to the ocean floor. A portion of the marine litter, which in many cases contain plastic, floats on the surface and can be detected by its spectral signature if aggregated into patches (Biermann et al., 2020; Topouzelis et al., 2019; Themistocleous et al., 2020). Initiatives across the world such as the UN Sustainable Development Goal 14 and the EU Marine Strategy Framework Directive's descriptor 10 encourage improving the ocean's health. Moreover, with the rapid scientific advances in the machine learning field, multiple initiatives aim at automating marine litter detection in the sea. These goals could be reached with proper monitoring of waste in the ocean based on scientific evidence on the existence of floating objects and their quantification. In many cases, marine litter pollution originates from land-based sources that enter the oceans and marine environments through rivers. Extreme weather events also contribute to transporting human waste into the sea. In fact, during rainy periods floods help carry trash into rivers that end up into the ocean. Floating debris causes a variety of harmful effects on marine life, biodiversity, and human life. In fact, marine organisms can ingest or become entangled in floating debris (Garaba and Dierssen, 2018; Carpenter et al., 1972). Moreover, some materials, such as plastic are very resilient to degradation and they might persist in the marine environment for at least 400 years.

Research on macro-debris detection is recent as managing human waste in the ocean is becoming one of the most pressing environmental challenges nowadays (Eriksen et al., 2014). In general, there is a lack of understanding of floating debris detection in the open sea due to the limited monitoring capabilities (Garaba and Dierssen, 2018). Floating objects drift due to winds and ocean currents. This requires monitoring with data at high temporal frequency. At large-scale, this data is provided by sensors, such as Sentinel 2, with a moderate spatial resolution of 10 meters at which the detection of floating objects is challenging. High-resolution alternatives, such as UAV acquisitions, have been proposed in the literature (Wolf et al., 2020; Papakonstantinou et al., 2021) but scale poorly when monitoring hundreds of kilometers at frequent intervals. When floating objects agglomerate in the middle of the sea, it becomes challenging and even impossible to track them with drones or satellites. Also, between the Great Pacific Garbage Patch with at most 100 kg/km^2 of plastic mass (Lebreton et al., 2018), and the spatial/temporal variability of phenomena found in coastal areas, the detection of marine litter at sea is a great challenge.

2. REMOTE SENSING FOR MARINE LITTER

Satellites and drones can be used to track floating objects on water bodies. In this work, we focus on the use of Sentinel 2 data which contains bands with a spatial resolution of up to 10m. The Sentinel 2 data is provided following two-level of processing: L1C top-of-atmosphere and L2A bottom-of-atmosphere. The L1C data has 13 bands including one band for clouds detection. The L2A data has 12 bands that are atmospherically corrected. We use both data types for better generalization.

* Corresponding Author
[1] https://wedocs.unep.org/rest/bitstreams/17739/retrieve accessed 2021-04-26

3. RELATED WORK

Important work towards gathering spectral responses of marine litter has been conducted in the Plastic Litter Project (Topouzelis et al., 2019) on the coast of Mytilene in Greece and a similar initiative in the harbor of Limassol, Cyprus (Themistocleous et al., 2020). Both projects deployed targets of floating objects in the sea and acquired imagery by unmanned aerial vehicles (UAV) at the same time as the overpass of multispectral Sentinel 2 satellite. Similar studies on coastal regions with aerial imagery (Moy et al., 2018; Garaba and Dierssen, 2018) showed that it was possible to detect and map floating macro debris in the open ocean with optical data (Hu et al., 2015; Aoyama, 2016; Garaba et al., 2018; Topouzelis et al., 2019; Maximenko et al., 2019). Recently, Biermann et al. (2020) introduced a Floating Debris Index (FDI) that measures the discrepancy between an interpolated near-infrared reflectance with the measured response. This discrepancy highlights the presence of plastic debris on Sentinel 2 images. Similarly, Themistocleous et al. (2020) defined a Plastic Index (PI) as the ratio of near-infrared and red which was effective in detecting deployed plastic targets off the shore of Cyprus. Using the ratio of near-infrared and red is conceptually similar to the Normalized Difference Vegetation Index (NDVI) which was also used as a discriminatory feature by Biermann et al. (2020). Nonetheless, visual inspections and the use of statistical data analysis techniques are still used (Lebreton et al., 2018). In terms of methods, Wolf et al. (2020) also used a Convolutional Neural Network (CNN) but focused on high-resolution UAV images for the detection and quantification of plastic litter. While UAV imagery provides imagery of high-quality that is well-suited for a machine learning approach, the availability of UAV imagery is inherently limited due to the acquisition costs. To address this, Papakonstantinou et al. (2021) proposed a citizen-science platform to upload imagery of plastic litter. The marine litter detection field is developing fast as the collection of trash in the ocean is becoming urgent. The detection of floating objects on the sea surface can be expensive when UAV data is acquired by drones that require a personal presence on the field for analyses. This makes the UAV data of floating objects difficult to acquire. In our work, we focus on Sentinel 2 imagery as it is globally available and free of charge which is essential for a remote sensing technology to guide clean-up operations of plastic with dedicated ships, as done by Ruiz et al. (2020). Compared to the UAV-driven approaches for targeted detection of plastic litter, we aim at detecting the general class of floating objects on the sea surface using globally available medium-resolution Sentinel 2 imagery.

In this work, we

- train and evaluate a CNN to learn spatial features for floating object detection,
- compare the neural network models with shallow methods trained on recently proposed classification features, i.e., NDVI + FDI,
- aggregate and publish a large-scale hand-labeled dataset of floating objects which is, to the best of our knowledge, the largest and most diverse dataset on floating objects available.

4. DATASET

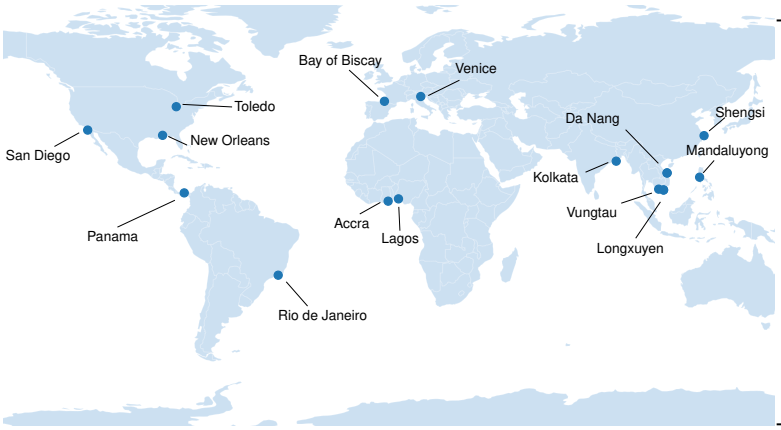
Modern data-driven methods require diverse datasets to obtain robust solutions that work under varying acquisition conditions on a global scale. In this section, we outline the design decisions we took while building a large-scale annotated dataset that can be used for the CNN baseline described in the next section.

4.1 Definition of Floating Objects

Let us first clarify the primary objective of the dataset and define floating objects. In-situ studies (Topouzelis et al., 2019; Themistocleous et al., 2020) have shown that only aggregations of floating objects are detectable with the coarse 10m resolution of Sentinel 2. Hence, methods rely on aggregation processes, such as river plumes, ocean currents, or windrows to accumulate various floating objects, such as plastics, pumice, algae, seaweed, seawater, and timber. These sub-categories of objects can be separated by their spectral responses in some cases. However, these spectra are always mixed and have a permanent background water signal which makes the distinction between water and floating objects difficult. It is common to use spectral features, such as the NDVI or the FDI, are easier to apply since they are expressed in closed forms which is not the case for spatial features. In this work, we shift our focus from spectral characteristics towards the spatial patterns that the aggregation processes leave on the water surface. We resort to Convolutional Neural Networks (CNNs) to learn the spatial features from annotated data and focus on a binary classification problem of floating objects versus non-floating objects. By concentrating on spatial features on this generalized problem, we can capture the characteristics of objects by aggregating a diverse dataset of globally distributed examples. A large-scale data-driven approach can be a step towards constructing a floating-object detector that automates the process of detecting shapes on the water surface. This detector could sift through large quantities of satellite imagery and isolate floating objects in the open water bodies which would facilitate and accelerate the task of analyzing the composition of the detected elements.

4.2 Data-Driven Feature Learning for Floating Objects

Classical model-driven machine learning approaches typically use a two-step process: first problem-specific features are manually defined. Then a problem-agnostic classification is performed in this hand-designed feature space. For instance, (Biermann et al., 2020) discovered the effectiveness of FDI (alongside NDVI) for floating object detection and used a problem-agnostic Naïve Bayes classifier for their gathered dataset. The discovery of problem-specific features, such as the FDI index, is driven by deep oceanographic domain knowledge and usually targets few individual spectral bands. The manual design of spatial features that use the entire spatio-spectral information in the data is often more difficult, if not impossible. Hence, data-driven learning with deep neural networks approaches this problem from a different perspective: instead of using expert knowledge to design specific features, we encode our knowledge in the labeled dataset by visually identifying floating objects to the best of our understanding using our visual system with the domain knowledge we obtain from literature and visualization of images with specific features and color schemes. A deep neural network can then be optimized on the labeled dataset to approximate and automate the effort that we put into hand-labeling the images. By using a 2D-CNN

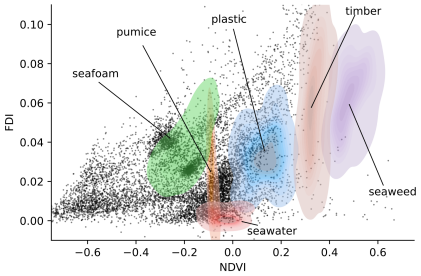


(a) Regions of visually labelled floating objects

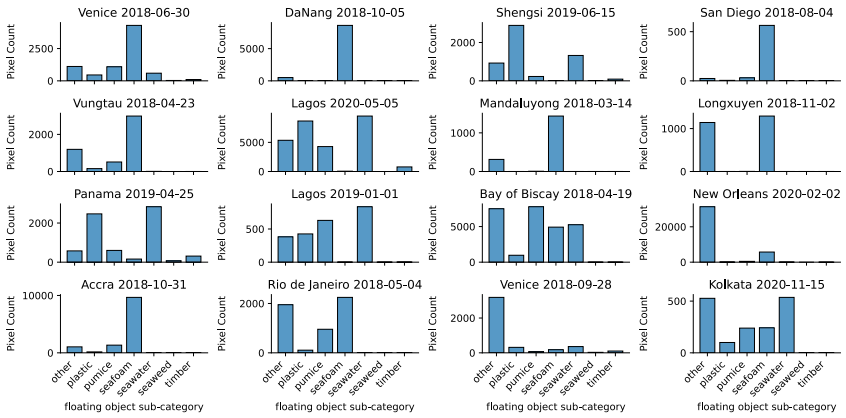
Region	Date	# lines	fold train △, val □, test ○					
			1	2	3	4	5	6
Bay of Biscay	2018-04-19	824	△	△	△	△	□	△
Da Nang	2018-10-05	51	○	△	△	△	△	□
Kolkata	2020-11-15	15	△	□	○	□	△	△
Longxuyen	2018-11-02	33	○	○	△	△	○	△
Panama	2019-04-25	182	□	△	△	△	○	△
Rio de Janeiro	2018-05-04	114	□	△	△	□	△	○
San Diego	2018-08-04	15	△	□	○	○	□	△
Toledo	2019-12-21	35	△	△	△	△	△	△
Accra	2018-10-31	99	□	○	△	△	△	△
Lagos	2019-01-01	30	○	○	□	△	△	□
Lagos	2020-05-05	188	△	△	□	△	△	△
Mandaluyong	2018-03-14	13	△	△	□	△	△	○
Venice	2018-06-30	97	△	△	○	□	□	○
Venice	2018-09-28	67	△	△	△	○	○	△
Vungtau	2018-04-23	79	△	□	△	○	△	□

(b) Summary of the sites used for labelling, their date of acquisition by Sentinel 2, the number of labels extracted from each site and specification of the use of the site: training or testing.

Figure 1. Overview on the Regions of floating objects in this Dataset



(a) Labelled pixels from our dataset (black) overlayed by class distributions of floating objects identified by (Biermann et al., 2020)



(b) Analysis of our labelled pixels categorized by the dataset labelled by (Biermann et al., 2020)

Figure 2. Sub-categorization of our floating-objects dataset using few hundred labelled pixels selected by Biermann et al. (2020).

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-V-3-2021-285-2021 | © Author(s) 2021. CC BY 4.0 License.

for efficient spatial-pattern learning and appropriate data augmentation techniques we can make sure that the deep learning model learns spatial features i.e., patterns in the pixel neighborhoods to identify floating objects.

4.3 Data collection method

For the data collection process, we visualized Sentinel 2 imagery at coastal areas which are likely to contain floating objects in Google Earth Engine (GEE). We referred to newspapers, social media, and articles that reported the existence of floating material on the sea surface. We followed the same approach as Biermann et al. (2020); Ruiz et al. (2020) by identifying several coastal regions, shown in Fig. 1a, where we found objects present at one date, but not in another. For each selected area we manually assessed the likelihood of floating objects by RGB representation along with the FDI and NDVI indices. Similarly, we focused on hints for ocean processes that can aggregate objects, such as windrows, ocean currents, river plumes. We used lines to label the identified objects and stored them with the image data as Sentinel 2 scenes at L1C top-of-atmosphere processing and bottom-of-atmosphere L2A level if these were available in the GEE catalog.

4.4 Label Analysis

Let's now analyze the floating-object labels that we gathered in the dataset to get a deeper understanding of the underlying diversity of the data. Since no labeled data is publicly available, we reconstructed the 195 labelled pixels from Figure 2 at Biermann et al. (2020) that categorize "plastic", "pumice", "seafoam", "seawater", "seaweed", "timber" by their FDI/NDVI characteristics. We plot the kernel-density distributions from these data points in Fig. 2a. These data distributions are well-separable since they were gathered in idealized conditions, i.e., specific atmospheric correction, manual selection of single pixels based on expert knowledge. In black, we show 10000 (out of 157319) floating-object pixels from our dataset that were gathered in the wild on realistic acquisition scenarios, i.e., L1C and L2A data, and under diverse atmospheric conditions in the presence of haze and clouds. We see that none of the idealized data distributions of seafoam, pumice, plastics, and even seawater align well with the gathered data in realistic conditions. This demonstrates the difficulty of transferring knowledge from a small-scale (in terms of the number of pixels), labor, and expertise expensive dataset, which obtained near-perfect accuracy in idealized conditions, to a realistic large-scale application scenario. Nonetheless, we can still use this data to obtain a general intuition on the nature of floating objects in our dataset. Since many floating objects in our dataset are out of distribution, we decided to use a class-wise Gaussian kernel density with a small bandwidth of 0.01 to conservatively add all pixels with a density threshold lower than 5 to the class "other". In Fig. 2, we split the resulting classification by region to obtain a sub-categorization of the diverse nature of floating objects on the different areas represented in the dataset. These results based on the categorization by Biermann et al. (2020) indicate that we captured plastic-like objects in some scenes, such as Panama, Lagos, and Shengsi, while many floating objects that we labeled also appear to be natural seafoam. This analysis, however, is limited by the inherent difficulty of finding accurate labels for a diverse group of floating objects as can be seen in the false detection of "pumice" in the Bay of Biscay which is unrealistic. It also demonstrates the difficulty of applying data from an idealized scenario on a real-world application on large-scale global data while still providing some insight into the inherent nature and diversity of floating objects in the dataset. Still, some of our manually gathered labels show feature characteristics of plastics which motivates our problem. After all, this necessitates the need for a robust large-scale floating-object detector that can be used as an initial step before further categorizations can be made.

5. METHODS

We implemented, trained, and evaluated a U-Net (Ronneberger et al., 2015) CNN for the problem of floating object detection and the data-driven feature learning of spatial characteristics. U-Net-based networks are composed of two sub-networks. The encoder downsamples the original image to a high-level representation of the entire scene. The decoder reconstructs the label on the original resolution from this high-level representation combined with intermediate features from the encoder via skip connections. This "what-and-where" strategy makes U-Net-type networks successful in applications where the scale of objects does not vary much. U-Nets were originally designed for biomedical image segmentation and have been employed widely in remote sensing. For instance, U-Net-based networks with different encoder components dominated the Spacenet 6 (Shermeyer et al., 2020) challenge of building footprint detection. We chose a U-Net model for this problem of floating object detection as we consider it to be a suitable and easy-to-access baseline for future work. We compared it with several shallow-learning methods on a hand-design feature space proposed for this problem, such as the Naïve Bayes classifier used in (Biermann et al., 2020). We also used, for comparison, Random Forest (RF) (Breiman, 2001) and Support Vector Machine (SVM) (Boser et al., 1992) which are supervised machine learning algorithms that can be used for classification and regression tasks.

6. EXPERIMENTS

In this section, we emphasize the technical details of the necessary steps before the experiments. We talk more specifically about the preparation of the dataset once exported from GEE and we highlight the way some technical issues are tackled to improve the training and the predictions.

6.1 Implementation and Training Details

The inspection of regions suspected to contain floating objects and their labeling was curried out on GEE. Further data processing, as well as the model training, were done in PyTorch. A few data-augmentation techniques were applied such as rotation, flipping, and adding spatial and spectral noise. The deep learning model was trained with a batch of 80 images of size 128×128 for 50 epochs with a learning rate of 0.001. To ensure that we obtained a model that generalized on unseen data,

Method			Metrics		
	input	spat.	acc.	f1	κ
SVM-Machine	NDVI+FDI	✗	58.82	0.67	0.17
Random Forest	NDVI+FDI	✗	58.83	0.69	0.17
Naïve Bayes	NDVI+FDI	✗	60.81	0.53	0.21
CNN (U-Net)	12 S2-bands	✓	**84.28**	**0.81**	**0.68**

Table 1. Metrics for assessing available methods to classify this dataset

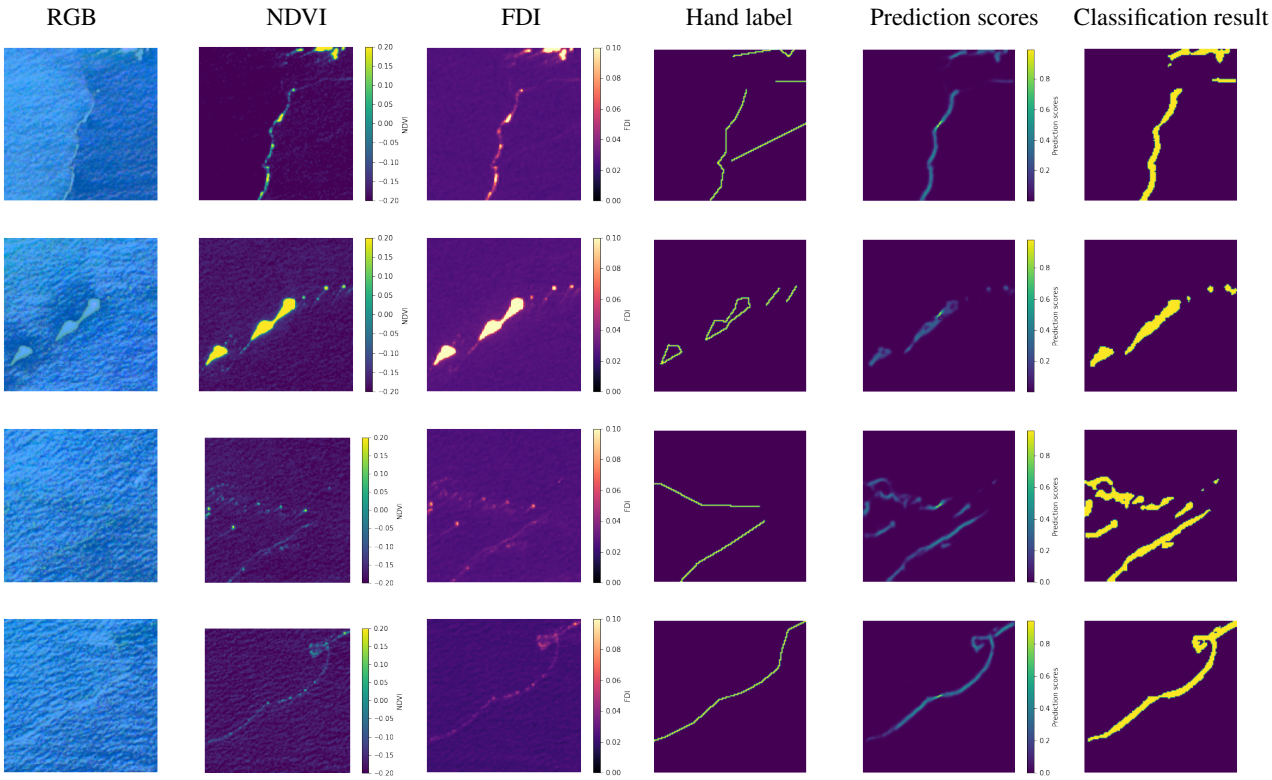


Figure 3. Examples from the dataset. RGB, NDVI, FDI, hand-labels and neural network predictions

we only stored the model if the validation loss decreased. This can be seen as a form of early stopping even though we always iterate through the entire 50 epochs.

Dataset implementation. Let's highlight some implementation details on the dataset: we stored each region as a Sentinel 2 image with associated floating-object labels in lines. During training and validation, we centered on individual line segments and crop the Sentinel 2 scene with a given output size of 128×128 pixels. We used L2A bottom-of-atmosphere data 50% of the time if it was available and rasterized the labels given the positions of the pixels of the cropped Sentinel 2 image. If the labels form closed rings, we assigned a floating-object label to the interior of this polygon. While testing, we split the original Sentinel 2 scene into 480×480 pixel tiles with a 64-pixel overlap that we sequentially predicted with a trained model. We also performed test-time augmentation by predicting the scores multiple times with different flipped and rotated input images. We merged the overlap between adjacent tiles smoothly. Given the georeference, we could combine patches again to retrieve a prediction score for each Sentinel 2 pixel.

Data Augmentation. We artificially increased the diversity of representations in the training dataset by data augmentation and flipping the training images vertically and horizontally 50% of the time. Similarly, we rotated the images in random multiples of 90 degrees and cropped the images from 256×256 to 128×128 pixels on random locations to avoid floating-object labels in the central pixel in all training images. We added random noise spatially and spectrally with the noise level being the standard deviation of the Sentinel 2 image used for training. For the spatial noise, we generated arbitrarily a 2D image with the spatial dimensions of the spectral bands. Then we multiplied this 2D image with the noise level and we added it to each band of the Sentinel 2 image. For the spectral noise, we generated a vector with the length being the number of bands, we multiplied it by the noise level and then we added this vector to all the pixels belonging to the same spatial coordinates. When a bottom-of-atmosphere scene was available, we randomly mixed top-of-atmosphere and bottom-of-atmosphere to further increase the diversity and improve the generalization to unseen regions.

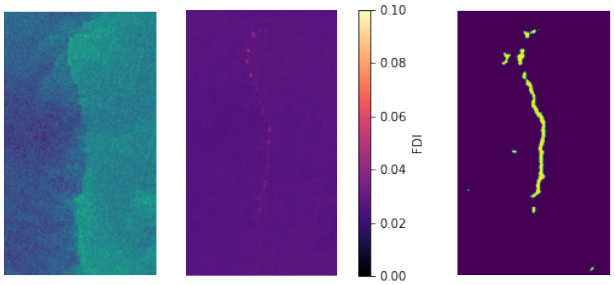
Train/Test Splits. Since we cropped the Sentinel 2 scenes with rasterized labels dynamically over individual line segments, we obtained a significant overlap between images. For this reason, we resorted to a region-wise split where we assigned some scenes/regions randomly to the training/validation/test partitions. This, however, may lead to shifts in data representations which is to some degree expected, as globally distributed scenes vary due to different types of floating objects, e.g., see Section 4.4, acquisition conditions, such as variations in atmospheric conditions. We addressed this issue by six-fold cross-validation but would like to investigate this problem further in future work by either increasing the dataset diversity through enlarging the dataset or using other domain adaptation or transfer learning approaches.

Class Imbalance. In the collected dataset, there are more pixels from the water class than pixels belonging to the floating objects class. To address this class imbalance issue, we use a weighted Binary Cross Entropy loss

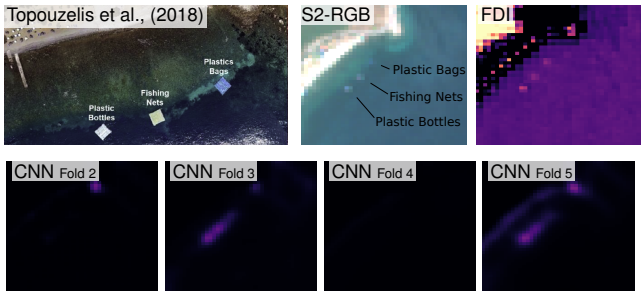
$$H(y, \hat{y}; \alpha) = -\alpha y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})$$

with labels $y \in \{0, 1\}$ and predictions $\hat{y} \in [0, 1]$ where $\alpha > 1$ increases the loss for wrong classifications of the positive class of floating objects. An additional strategy to address this class imbalance is to tune the threshold parameter on the prediction scores $\hat{y}$ to determine a binary floating-object label. Since wa-

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-V-3-2021-285-2021 | © Author(s) 2021. CC BY 4.0 License.



(a) Classification on Scotland (Biermann et al., 2020).



(b) Targets of the Plastic Litter Project 2018 (Mytilene, Greece). (Topouzelis et al., 2019)

Figure 4. U-Net predictions on images from other projects.

ter pixels are significantly more common, we found out that the model underestimated the prediction scores of floating objects. A threshold of 0.5 to assign the floating-object label to the continuous prediction score is too conservative in many cases. To address this, we could determine a better-suited threshold by measuring the classification performance on the validation set.

Hard Negative Mining. The training dataset contains solely images that always contain floating objects in some pixels. During test time, however, we would like to predict entire Sentinel 2 scenes containing other objects, such as land, clear water, ships, etc. Hence, we enriched the dataset dynamically with *hard negative examples* (Hughes et al., 2018; Tang et al., 2017) by randomly choosing patches within the available Sentinel 2 scenes.

6.2 Results

Let us now compare the CNN model to shallow learning models commonly used for this problem and provide qualitative examples. For an objective comparison, we compared the CNN model to the shallow classifiers using three metrics for the evaluation process: accuracy, f1-score, and the kappa coefficient. We also applied the CNN model trained on our dataset on images from two projects by (Biermann et al., 2020) and (Topouzelis et al., 2019). An analysis of the results is provided below.

6.2.1 Quantitative Comparison to Pixel-Wise Classifiers

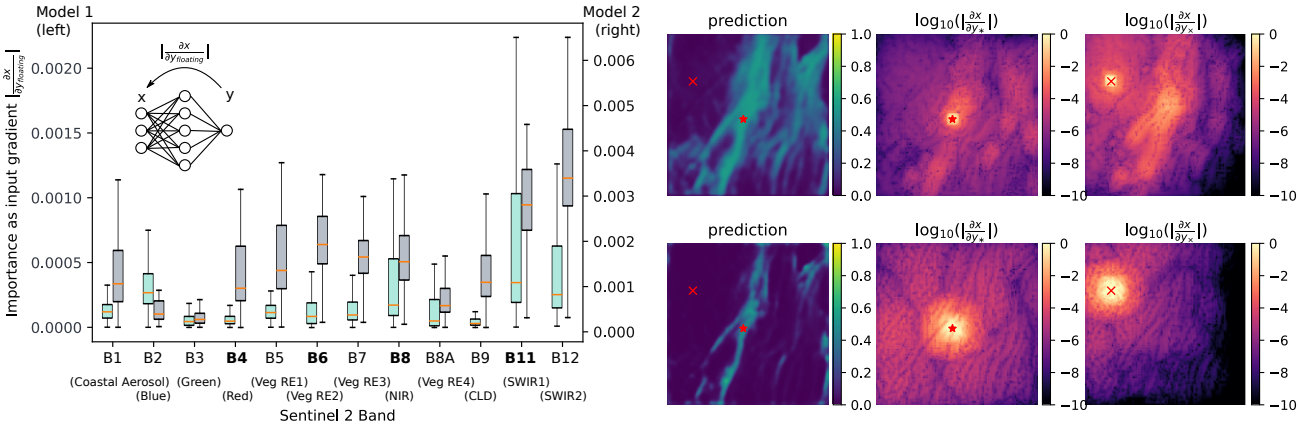
In Table 1, we compare the U-Net model with the pixel-wise machine learning classifiers. The SVM, RF, and Naïve Bayes classifiers were trained on a balanced dataset from the training regions while we used regular predictions from the U-Net model. We compare all models on a balanced dataset by randomly sampling the same number of water and floating object pixels from the respective images of the test regions. Performance on the validation regions was used to determine the respective model hyperparameters, i.e., $\gamma = 10^{-3}, C = 30$ for the SVM, and 1000 estimators for the random forest with a depth of 2. Following Biermann et al. (2020), we optimized and predicted the shallow learning models on the designed FDI and NDVI feature space while the U-Net models used the raw input space of 12 Sentinel 2 bands. From the comparison in the table, we can see that the U-Net model outperforms the shallow-learning models in overall accuracy, the f1-score, and the kappa coefficient. Given that the U-Net model has access to contextual spatial information through the 2D convolutional layers, it seems reasonable that it outperforms the shallow-learning models that can only process each pixel separately without information of the local neighborhood.

6.2.2 Qualitative Comparison

We provide further results of floating objects detection at different regions in Fig. 3. On the latter, we present the RGB images along with their FDI and NDVI representations, the masks based on the geometrical shapes detected by the FDI index, the prediction scores, and finally the classification result. Let us start from the top, left to right: the FDI and NDVI representations of the four RGB images contain different geometrical shapes. The first RGB image is mostly composed of a line with some patches at the top. In the interest of time, we labeled these patterns as continuous lines. Even though the labels are only roughly accurate, the prediction scores and the classification results follow the geometrical shapes accurately. The second row shows three main patches in the FDI and NDVI representations that appear to be influenced by the current. Even though only the exterior line is labeled as floating objects, the model can generalize and predict the entire patch accurately. This shows that the deep learning model could successfully capture the spectral response of the floating patch. On the fourth row, we can see a circular current that appears on the FDI and NDVI representations but was not labeled accurately. Nonetheless, the pattern was captured accurately by the prediction. From the results of Fig. 3 and the analysis above, we see that the model could produce reasonable predictions even though the labels do not represent the actual shapes of the floating objects accurately. We also notice that the model can generalize on the general shape of floating objects without over-fitting on artifacts from the inaccurate labeling process.

Let us now apply the U-Net model to scenes used in related work. We show the result of our predictor on an image from the work in Biermann et al. (2020) where the existence of plastic is suspected. Fig. 4a shows a Sentinel 2 image captured on the 20th of April on the year 2018 in Scotland, its FDI presentation showing the presence of floating objects and the classification result after applying the deep learning algorithm. We could see that the geometrical shape on the classification result is successfully detected and quite consistent with the shape highlighted by the FDI index. We also validated our model by applying it on the Sentinel 2 image from a scene captured during the Plastic Litter Project 2018 (Topouzelis et al., 2019). This project provides some of the few confirmed labels of plastic litter publicly available. Even though the targets of plastic bags, fishing nets, and plastic bottles were 10m by 10m in size and visible in the UAV acquisition, they are only barely visible on the Sentinel 2 scene. We classified this scene with all six models trained on different train/test folds. However, only two CNN models predicted some floating-object scores, while four others show no classifications, as shown in Fig. 4b. The fact that the two models trained on coarse floating-object labels produced prediction scores on these comparatively small target pixels is encouraging, but also highlights the difficulty of predicting plastic litter with a coarse spatial resolution of ten meters.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-V-3-2021-285-2021 | © Author(s) 2021. CC BY 4.0 License.



(a) Sentinel 2 bands utilized by two CNN models. Bands utilized in the Floating Debris and NDVI indices are bold-faced for reference.

(b) Perceptive field of two CNN models (Model 1 top, Model 2 bottom) while predicting the pixels at positions × and ⋆.

Figure 5. Analysis of the input feature importance on two trained U-Net Models. Figure (a) shows the band-importances by the input-gradient signal which reveals that a broar range of Sentinel 2 bands is utilized for the prediction. In (b) and (c), we analyze the local perceptive field of the U-Net models and see that Model 2 uses a larger pixel neighborhood to make a prediction.

6.3 Feature Importance

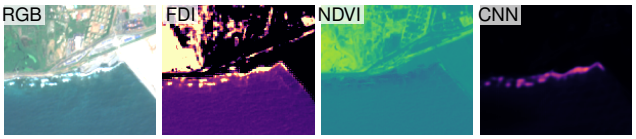
Let us now focus on the U-Net model itself and analyze the learned features. Data-driven learning allows the model to extract features from the raw input data solely based on the labeled data without making hard a-priori assumptions on the expected importance of the spectral bands. We can analyze the most important input features x by exploiting the differentiable characteristics in deep learning models by backpropagating the gradient signal $\frac{\partial x}{\partial y}$ from the predicted labels $\hat{y} = y_{floating}$ to the input tensors (Zhou et al., 2016). This provides an estimate of the importance of input bands by asking: "how should the input x have changed to change the prediction $y_{floating}$?". The learned features and feature weights can vary between models with identical settings since each deep neural network is optimized from random initialization. Hence, we report the feature importances evaluated on two trained neural network models.

Band Importance. In Fig. 5a, we plot the estimated averaged input gradients of two trained U-Net models over averaged floating label pixels on 200 images from the test set. Since we labeled the dataset while referencing NDVI and FDI indices, we would expect the deep learning model to approximate these features. If this would be the case, we would see a high influence on the same bands that were used in the calculation of these features which we highlighted in boldface. While the models utilized these bands to some degree, also other bands were considered. For instance, the blue and coastal aerosol bands (B1, B2) are not used in the calculation of NDVI and FDI but influenced the neural network classification. We speculate that these bands are important to identify pure (blue) water pixels. The neural networks also utilized all near-infrared bands (B5-B8) and the second short-wave infrared (B12) while the hand-designed features use one band from these groups only.

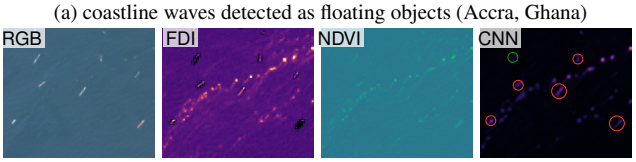
Pixel Importance. In Fig. 5b we further analyze the feature importance of two trained models by calculating the gradients to the input image with respect to single-pixel predictions at two points ⋆ and ×. In contrast to the hand-designed features of NDVI and FDI, CNNs learn the spatial patterns for the classification of floating and non-floating objects. With this analysis, we can visualize the perceptive field of the trained CNNs and evaluate how much spatial context these models utilized for their predictions. While Model 2 used a larger spatial neighborhood, Model 1 drew its features from a smaller perceptive field. This seems to affect the prediction quality where the estimates of Model 2 appear more accurate. Both models utilize large-scale spatial features to a small degree which can be seen in the general background structure visible in the gradient images. This is a sign that these U-Net models also utilize deeper higher-level features in the inner layers and do not solely rely on the initial skip connections.

7. LIMITATIONS

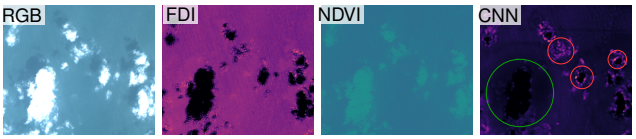
Training a neural network on a globally distributed dataset is a challenging task that requires a large dataset that is diverse enough to generalize on new unseen areas. We train and evaluate on different regions to obtain an estimate of the generalization performance of our model. Still, the number of the scenes in this dataset is limited and regions vary significantly. We see this variability during training where models improve steadily on the train regions but start to overfit early with accuracy stagnating on the unseen validation and test regions. Fig. 6 shows several examples of systematic failure cases we observed in the model predictions. The RGB representation and FDI/NDVI indices provide some context for interpretation and visual comparison to the state-of-the-art (Biermann et al., 2020). In Fig. 6a, we observe that waves and coastlines were predicted incorrectly as floating objects. Also, the FDI index shows large values in these cases. However, the CNN could suppress the signal from the land pixels in contrast to FDI and NDVI that show high responses. Man-made objects like ships are also confused with floating objects, as shown in Fig. 6b. These objects also cause some responses in NDVI and FDI. In-



(a) coastline waves detected as floating objects (Accra, Ghana)



(b) some ships detected as floating objects (Accra, Ghana)



(c) some clouds detected as floating objects (Da Nang, Vietnam)

Figure 6. Failure cases we observed in the predictions

terestingly, the U-Net CNN predictions are not wrong for all ships in this scene, even though all ships appear to have a similar spectral response. One ship (green circle) does not cause a response in the floating-object prediction score. This indicates that the CNN utilizes some spatial features that cause different responses for the ships in this scene. Clouds similarly can cause false activations in the prediction scores, as shown in Fig. 6b. Similar to the previous example, the CNN has correctly identified one large cloud as a not-floating object while missing the fringes of smaller clouds. In contrast, the spectral FDI and NDVI must, by design, produce similar responses on all clouds since the geometric shapes of the objects do not influence these indices. This indicates that the CNN learned spatial features on the geometry of clouds.

8. OUTLOOK CHALLENGES

In light of the limitations mentioned above, we identify several research directions to improve the model's performance. Increasing the diversity of regions by adding additional sites to the dataset will likely help deep learning models to generalize to unseen sites. A more targeted negative example strategy that includes ships and clouds may be helpful to encourage the model to learn these patterns and suppress the prediction scores whenever ships or clouds are present. Additionally, refining the label quality will have a positive impact on model performance. Beyond simply increasing the quality and quantity of the available labels, further techniques to tackle the domain shift between regions via, for instance, targeted data augmentation could be investigated. Also, suitable model initializations could be found for better generalization.

9. CONCLUSION

In this work, we provided a hand-labeled Sentinel 2 dataset for floating objects detection on the sea surface as one step towards identifying and eventually collecting marine litter. We evaluated a baseline U-Net model that learned spatial characteristics of floating objects. The qualitative results showed that the deep-learning-based model was able to predict correctly the geometrical shapes even if the labels were inaccurate or absent. The feature importance analysis on band level showed that more Sentinel 2 bands can be utilized for floating object detection than the ones that are employed by current hand-designed features. The analysis of the perceptive field of the CNNs and the good performance compared to pixel-wise classifiers showed that spatial features are useful for detecting floating objects on the sea surface. However, the high number of false positives, some of which we show in the Limitations section, makes this CNN not suitable for a stand-alone detection of floating objects, yet. We aim to improve the data diversity and label quality towards this issue in future work. Nonetheless, we believe that providing the first large-scale open dataset for this problem along with pre-trained models is a step towards a large-scale and accurate detection of floating objects on a near real-time basis that can utilize the publicly available Sentinel 2 imagery to its full potential.

References

Aoyama, T., 2016. Extraction of marine debris in the sea of japan using high-spatial-resolution satellite images. *Remote Sensing of the Oceans and Inland Waters: Techniques, Applications, and Challenges*, 9878, International Society for Optics and Photonics, 987817.

Biermann, L., Clewley, D., Martinez-Vicente, V., Topouzelis, K., 2020. finding plastic patches in coastal Waters using optical Satellite Data. *Scientific reports*, 10(1), 1–10.

Boser, B. E., Guyon, I. M., Vapnik, V. N., 1992. A training algorithm for optimal margin classifiers. *Proceedings of the fifth annual workshop on Computational learning theory*, 144–152.

Breiman, L., 2001. Random forests. *Machine learning*, 45(1), 5–32.

Carpenter, E. J., Anderson, S. J., Harvey, G. R., Miklas, H. P., Peck, B. B., 1972. Polystyrene spherules in coastal waters. *Science*, 178(4062), 749–750.

Eriksen, M., Lebreton, L. C., Carson, H. S., Thiel, M., Moore, C. J., Borerro, J. C., Galgani, F., Ryan, P. G., Reisser, J., 2014. Plastic pollution in the world's oceans: more than 5 trillion plastic pieces weighing over 250,000 tons afloat at sea. *PloS one*, 9(12), e111913.

Garaba, S. P., Aitken, J., Slat, B., Dierssen, H. M., Lebreton, L., Zielinski, O., Reisser, J., 2018. Sensing ocean plastics with an airborne hyperspectral shortwave infrared imager. *Environmental science & technology*, 52(20), 11699–11707.

Garaba, S. P., Dierssen, H. M., 2018. An airborne remote sensing case study of synthetic hydrocarbon detection using short wave infrared absorption features identified from marine-harvested macro-and microplastics. *Remote Sensing of Environment*, 205, 224–235.

Hu, C., Feng, L., Hardy, R. F., Hochberg, E. J., 2015. Spectral and spatial requirements of remote measurements of pelagic Sargassum macroalgae. *Remote Sensing of Environment*, 167, 229–246.

Hughes, L. H., Schmitt, M., Zhu, X. X., 2018. Mining hard negative samples for SAR-optical image matching using generative adversarial networks. *Remote Sensing*, 10(10), 1552.

Lebreton, L., Slat, B., Ferrari, F., Sainte-Rose, B., Aitken, J., Marthouse, R., Hajbane, S., Cunsolo, S., Schwarz, A., Levivier, A., Noble, K., Debeljak, P., Maral, H., Schöneich-Argent, R., Brambini, R., Reisser, J., 2018. Evidence that the Great Pacific Garbage Patch is rapidly accumulating plastic. *Scientific Reports*, 2018.

Maximenko, N., Corradi, P., Law, K. L., Van Sebille, E., Garaba, S. P., Lampitt, R. S., Galgani, F., Martinez-Vicente, V., Goddijn-Murphy, L., Veiga, J. M. et al., 2019. Towards the integrated marine debris observing system. *Frontiers in marine science*, 6, 447.

Moy, K., Neilson, B., Chung, A., Meadows, A., Castrence, M., Ambagis, S., Davidson, K., 2018. Mapping coastal marine debris using aerial imagery and spatial analysis. *Marine pollution bulletin*, 132, 52–59.

Papakonstantinou, A., Batsaris, M., Spondylidis, S., Topouzelis, K., 2021. A Citizen Science Unmanned Aerial System Data Acquisition Protocol and Deep Learning Techniques for the Automatic Detection and Mapping of Marine Litter Concentrations in the Coastal Zone. *Drones*, 5(1), 6.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical image computing and computer-assisted intervention*, Springer, 234–241.

Ruiz, I., Basurko, O. C., Rubio, A., Delpey, M., Granado, I., Declerck, A., Mader, J., Cózar-Cabañas, A. et al., 2020. Litter windrows in the south-east coast of the Bay of Biscay: an ocean process enabling effective active fishing for litter.

Shermeyer, J., Hogan, D., Brown, J., Van Etten, A., Weir, N., Pacifici, F., Hansch, R., Bastidas, A., Soenen, S., Bacastow, T. et al., 2020. Spacenet 6: Multi-sensor all weather mapping dataset. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 196–197.

Tang, T., Zhou, S., Deng, Z., Zou, H., Lei, L., 2017. Vehicle detection in aerial images based on region convolutional neural networks and hard negative example mining. *Sensors*, 17(2), 336.

Themistocleous, K., Papoutsa, C., Michaelides, S., Hadjimitsis, D., 2020. Investigating Detection of Floating Plastic Litter from Space Using Sentinel-2 Imagery. *Remote Sensing*, 12(16), 2648.

Topouzelis, K., Papakonstantinou, A., Garaba, S. P., 2019. Detection of floating plastics from satellite and unmanned aerial systems (Plastic Litter Project 2018). *International Journal of Applied Earth Observation and Geoinformation*, 79, 175–183.

Wolf, M., van den Berg, K., Garaba, S. P., Gnann, N., Sattler, K., Stahl, F., Zielinski, O., 2020. Machine learning for aquatic plastic litter detection, classification and quantification (APLASTIC-Q). *Environmental Research Letters*, 15(11), 114042.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., 2016. Learning deep features for discriminative localization. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2921–2929.