Earth and Space Science

RESEARCH ARTICLE
10.1029/2019EA000960

Automated River Plastic Monitoring Using Deep Learning and Cameras

Colin van Lieshout[1,2,3], **Kees van Oeveren**[1], **Tim van Emmerik**[1,4], **and Eric Postma**[2,5]

[1]The Ocean Cleanup, Rotterdam, The Netherlands, [2]Jheronimus Academy of Data Science, 's-Hertogenbosch, The Netherlands, [3]Soda science, 's-Hertogenbosch, The Netherlands, [4]Hydrology and Quantitative Water Management Group, Wageningen University, Wageningen, The Netherlands, [5]Cognitive Science and AI, Tilburg University, Tilburg, The Netherlands

Key Points:
- The proposed automated monitoring method locates river plastic on images reliably
- The method generalizes reasonably well to new locations and would benefit from a larger data set
- Automated method counts agree reasonably with manual methods

Correspondence to:
C. van Lieshout,
colin.vanlieshout@theoceancleanup.com

Citation:
van Lieshout, C., van Oeveren, K., van Emmerik, T., & Postma, E. (2020). Automated river plastic monitoring using deep learning and cameras. *Earth and Space Science*, *7*, e2019EA000960. https://doi.org/10.1029/2019EA000960

Received 21 OCT 2019
Accepted 21 JUN 2020
Accepted article online 28 JUL 2020

Abstract Quantifying plastic pollution on surface water is essential to understand and mitigate the impact of plastic pollution to the environment. Current monitoring methods such as visual counting are labor intensive. This limits the feasibility of scaling to long-term monitoring at multiple locations. We present an automated method for monitoring plastic pollution that overcomes this limitation. Floating macroplastics are detected from images of the water surface using deep learning. We perform an experimental evaluation of our method using images from bridge-mounted cameras at five different river locations across Jakarta, Indonesia. The four main results of the experimental evaluation are as follows. First, we realize a method that obtains a reliable estimate of plastic density (68.7% precision). Our monitoring method successfully distinguishes plastics from environmental elements, such as water surface reflection and organic waste. Second, when trained on one location, the method generalizes well to new locations with relatively similar conditions without retraining (≈50% average precision). Third, generalization to new locations with considerably different conditions can be boosted by retraining on only 50 objects of the new location (improving precision from ≈20% to ≈42%). Fourth, our method matches visual counting methods and detects ≈35% more plastics, even more so during periods of plastic transport rates of above 10 items per meter per minute. Taken together, these results demonstrate that our method is a promising way of monitoring plastic pollution. By extending the variety of the data set the monitoring method can be readily applied at a larger scale.

1. Introduction

Marine plastics are a widespread concern because of their persistence and negative impact on the marine ecosystem and human health (Jambeck et al., 2015; Lebreton et al., 2017; Schmidt et al., 2017; van Emmerik & Schwarz, 2020). Plastics account for over 80% of anthropogenic litter observed in rivers (González-Fernández et al., 2018). Larger plastics fragment into microplastics and can be ingested by wildlife (Cole et al., 2013; Thompson et al., 2004). Land-based plastics are assumed to be a main source of marine plastic pollution, as they get transported into the ocean by rivers (Jambeck et al., 2015; Lebreton et al., 2017). Developing mitigation strategies requires better understanding of the spatiotemporal distribution of plastic transport. Various in situ and modeling approaches to river plastic monitoring methods have been proposed. In situ methodologies include human visual counting (van Emmerik et al., 2018; González-Fernández & Hanke, 2017; van Calcar & van Emmerik, 2019), debris sampling using nets (Rech et al., 2014) and debris sample collection from existing infrastructure such as a regional network of floating debris-retention booms (Gasperi et al., 2014). Although such methods provide site-specific data, they are unsuitable for application at different locations for extended periods of time because of their labor-intensive nature and sampling-equipment requirements (van Emmerik & Schwarz, 2020). Modeling approaches provide an alternative by relying on secondary data, for example, data on mismanaged plastic waste, geography, population density, and hydrology, to estimate the input of riverine plastic into the oceans (Lebreton et al., 2017; Schmidt et al., 2017; Tramoy et al., 2019). Such methods provide a first-order estimation of the global and local contributions of river plastic emission but rely heavily on approximations based on a small number of in situ assessments (Lebreton et al., 2017; Schmidt et al., 2017). Given these limitations, an alternative monitoring method is needed to determine the spatiotemporal distribution of plastic transport in a more reliable and feasible way.

Video camera technology may enable an alternative method, as video cameras are commonly used in monitoring systems (Monge-Ganuzas et al., 2017; Ruiz et al., 2020). Products as security cameras, drones, and smartphones are easily accessible globally. Using video cameras to record the water surface makes it possible to observe (near) floating plastics in turbid rivers, which are expected to make up the majority of total riverine plastic transport (van Emmerik et al., 2019). By mounting a video camera perpendicular to the water surface at a height of about 4 to 9 m from the water surface, it is feasible to observe macroplastics (5 cm and larger), which constitutes to ≈90% of the total plastic mass in the oceans (Lebreton et al., 2018).

While it is feasible to manually locate plastic debris in images, automating plastic detection from images is a challenging task. Modern artificial intelligence (AI) technology called deep learning offers the opportunity to make this possible. In fact, deep learning is responsible for recent breakthroughs in signal and image processing (LeCun et al., 2015). Specifically, the deep learning variant called *convolutional neural networks* (CNNs) have achieved successful performances on a wide variety of visual tasks (Bengio et al., 2013; LeCun et al., 2015). As all machine learning methods, CNNs require to be trained on a data set of examples. The performance of CNNs depends crucially on the size and quality of this so-called training set. As for the size of the data, deep learning performance typically increases with the size of the training set up to a certain point where additional data do not lead to much further improvement. The required size depends on the classification task at hand, and hence, the data set size required for a task has to be determined empirically.

In this paper, we propose an automated monitoring method for the in situ detection and quantification of floating macroplastic in rivers. Our monitoring method is based on image data captured by an off-the-shelve digital video camera and processes the images using deep learning technology. The images of floating plastics in rivers are the example images on which the method is trained. Recently, a CNN algorithm for the classification of floating marine plastic debris has been proposed by Kylili et al. (2019). Their CNN classified objects as bottles, buckets, or straws but required cropped images centered around the objects. Our monitoring method is directly applicable to realistic data by detecting arbitrary plastic debris from images of a large river segment as captured by a video camera.

Regarding the quality of the data, the training data should contain a representative sample of examples to the proper classification of novel data. As a case in point, modern deep learning object-recognition algorithms perform very well on the recognition of everyday objects such as toothbrushes and other bathroom objects. However, these algorithms have shown to perform badly on the recognition of such objects in different regions around the world (DeVries et al., 2019). A toothbrush in a low-income country may look quite different from one in a high-income country, where most images for training the algorithm originate from.

For our monitoring task, data quality implies that our training data should be representative of as many occurrences of floating plastics as possible. The only way to determine if the training set is sufficiently representative for the task of plastic classification is to perform experiments on various training sets covering different locations and situations. In this way the degree to which the trained method generalizes to new situations can be assessed.

In our study we experimentally evaluated our monitoring method for quantifying plastic pollution. The method was trained on image data collected at five locations in Jakarta, Indonesia. Three experiments were performed. Experiment I determined the precision of our method and how it depends on both the amount of training data as well as training algorithm settings. Experiment II assessed how well of our method generalizes to new locations, by training it on one location and testing it on other locations. Finally, Experiment III compared the monitoring results of humans and our method to establish the degree to which human estimates match those of our method.

This study is the first to propose and evaluate a proof of principle for a practically applicable plastic-waste monitoring method.

2. Methods

Figure 1 provides an overview of all processing steps discussed in this section. The floating plastic data set was used for training our monitoring method, Experiments I and II. The visual counting data set from human in situ monitoring was used for Experiment III.

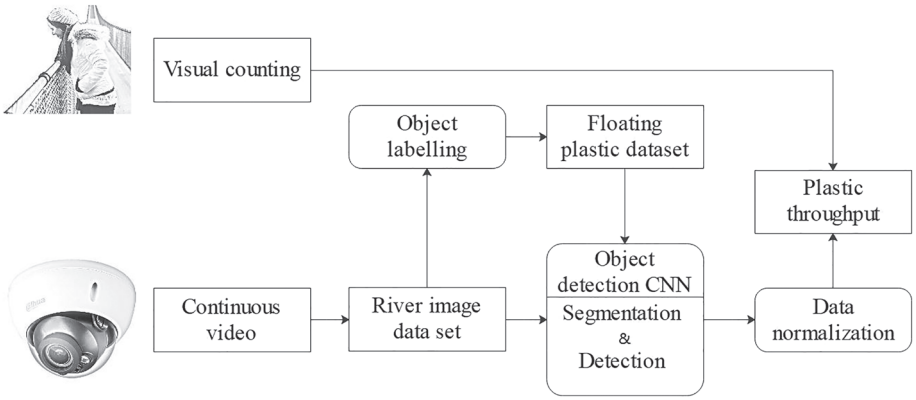

Figure 1. Overview of all data collection and processing steps.

2.1. Data Sets

2.1.1. Data for Training the Monitoring Method

Like all machine learning algorithms, our monitoring method is trained and evaluated on a training set and test set, respectively. The training set ensured that the method is trained to perform the task of detecting floating plastic waste; the test set provided previously unseen examples to validate how well it performs on previously unseen data.

The data sets used in this study were collected by a video camera mounted on bridges at five different waterways in Jakarta, Indonesia, from 30 April to 12 May 2018. The five locations are henceforth referred to as Locations A to E. Figure 2 shows two examples of the monitoring setup consisting of a video camera that is mounted perpendicular to the water surface. At each location, the camera recorded continuous sequences of a 1080p video stream with H.265 compression for 3 to 10 days. Despite the high compression quality, H.265 still exhibits perceivable encoding artifacts (Lin et al., 2019), which may negatively affect our detection results. The same camera was used for all experiments, the Dahua Easy4ip IPC-HDBW1435EP-W. An overview of the camera specifications can be found in Table 1.

Figure 3 shows a map of the five locations. Table 2 summarizes the characteristics of the data collected at each of these locations. First, the amount of plastic per image varied substantially with a difference of plastic objects per image between Location A and Location E of 21 and 0.5, respectively. Second, only one location (E) had waves, whereas the others had relatively still water surfaces. Third, for three of the five locations (B, C, and D) images contained different levels of organic material such as leaves or branches. These add complexity to the environment because of their variability in shape, size and color. Fourth, at some locations (A and B) organic and plastic clutter together in debris patches, making individual objects harder to distinguish. Fifth, camera altitude (i.e., distance from the water) differs, which is relevant as more distant monitoring

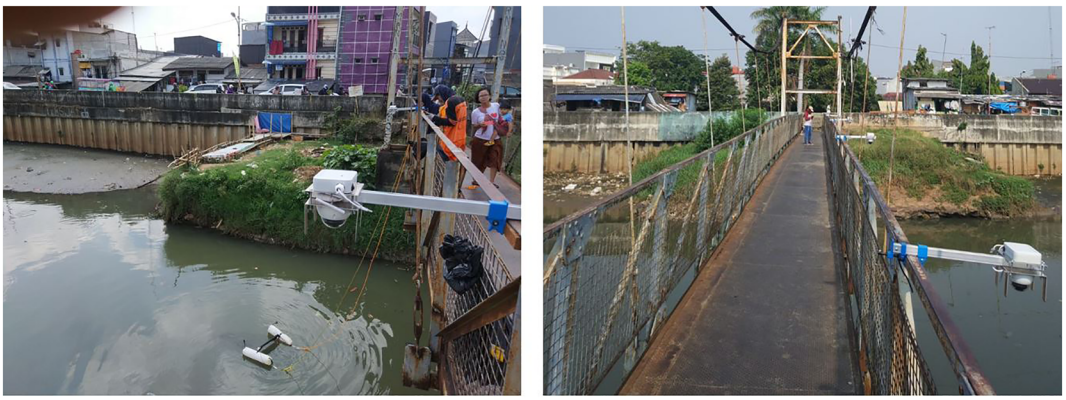

Figure 2. Camera setup Jakarta, Indonesia.

Table 1
Specifications Dahua Easy4ip IPC-HDBW1435EP-W

Sensor size	Resolution	Focal length	View angle	Framerate	Type	Compression
1/3″	4 MP	2.8 mm	106°	10 fps	MJPEG	H.265

implies fewer pixels for similar sized objects. Locations A and D are relatively close to the water, B and D are somewhat more distant, while Location E is by far most distant. Visual inspection was used to categorize qualitative variables into *no*, *some*, and *many*. Example images that illustrate location subset differences can be found in Appendix A.

To train our monitoring method, two data sets were needed: a data set of images of the water surface as captured by the camera and a data set of cropped and labeled images of objects floating on the water. These data sets are referred to as the *river image data set* and the *floating plastic data set*, respectively.

2.1.2. River Image Data Set

From the 26 days of video footage collected on the five sites, 1.272 JPEG images were selected. Individual images taken at the same location were separated by at least 5 min, to avoid visual overlap between the images.

2.1.3. Floating Plastic Data Set

Through Zooniverse, a citizen science web portal (Citizen Science Alliance), volunteers labeled all 1.272 images of the river image data set manually by drawing rectangular boxes around image regions that contain plastic waste. The 14,968 rectangular boxes so obtained constitute the floating-plastic images. All labels created by the Zooniverse volunteers were visually inspected and, if necessary, corrected by one of the authors (C. v. L.) in order to guarantee label quality and consistency.

Figure 4 illustrates how the floating plastic data set was subdivided into subsets for experimentation. In the left column, subsets are labeled according to the five Monitoring Sites A to E. For each location we distinguish between three types of subsets: *Total Set X*, *Train Set X*, and *Test Set X*, where X represents one of the locations. For each site several train and test subsets are specified (e.g., A.1 and A.2). The relative sizes of these subsets are illustrated by the lengths of the horizontal bars in the right columns. The rightmost column specifies the sizes in terms of percentages of the total data set. The horizontal span of the bars indicate which part of the total data set (top bar) they cover. For instance, the bars of *Total Set A* and *Total Set B* do not overlap because they do not share data, whereas *Train Set A.4* is part of *Train Set A.5*, and therefore, their bars do overlap.

All experiments were performed on the raw images. Preprocessing was not applied (e.g., filtering or color correction).

2.1.4. Visual Counting Data

To be able to compare the automated monitoring results to in situ visual counting, we used the data set presented by van Emmerik et al. (2019). For the visual counting measurements, observers stood on bridges looking downward at a river. All floating plastic items within a predefined part of the river cross section were

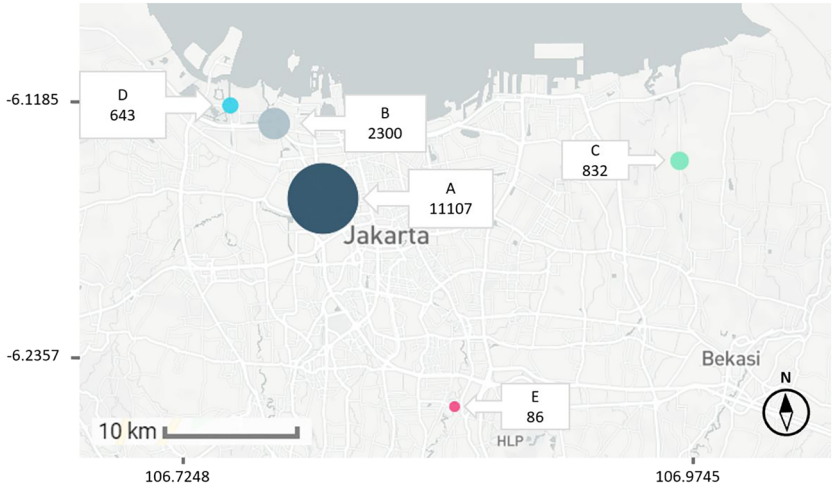

Figure 3. Monitoring locations, with the number of objects available per location.

Table 2
Characteristics of the Data Sets of Each of the Five Locations

Monitoring location	River	Location ID	Number of observation days	Number of images	Number of objects	Mean objects per image	Altitude (m)	Presence of waves	Organic debris	Debris patches
BKB-Grogol	Ciliwung	A	10	528	11,107	21	4.5	no	no	some
BKB-Angke	Ciliwung	B	3	92	2,300	25	5.5	no	some	many
BKT	Various	C	3	280	832	3.4	6.5	no	many	no
Cengkareng	Pesanggraha	D	7	208	643	3.1	4.0	no	some	no
Haryono	Ciliwung	E	3	164	86	0.5	8.0	many	no	no

counted for a specified amount of time (2 min). Each unique piece of plastic that flows underneath the bridge was counted using a clicker counter device. Total counts were then converted to a plastic count per minute per meter river width. The counts took place at the exact same locations and times as the video recording. More details on the data collection and subsequent analysis can be found in van Emmerik et al. (2019).

2.2. Object Detection CNN: Segmentation and Detection

The core of our monitoring method consisted of two stages: a segmentation stage and a detection stage. Both were implemented using the Tensorflow Object Detection API (Huang et al., 2017). In the segmentation stage an object-detection CNN selected promising image regions from a river image. In our case, promising regions corresponded to regions that are likely to contain plastic objects. Generally, only a small proportion of the river image contained plastic, and typically, there were multiple objects in one image. For such tasks, object detection CNNs are highly suitable. These networks are able to locate multiple objects within an image, distinguishing them from the background (Girshick, 2015).

In the detection stage, a second CNN detected the selected image regions that contain plastic waste. Figure 5 illustrates both stages by showing an example input image (left), the result of the segmentation stage (middle), and the result of the detection stage (right).

For the segmentation stage, we used a so-called Faster R-CNN (regional-convolutional neural network) (Ren et al., 2015), which is known for its high accuracy (Girshick, 2015). The Faster R-CNN was trained on the river image data set. The detection stage was realized by an Inception v2 network pretrained on the COCO (Common Objects in Context) data set, which is among the best-performing object-classification CNNs (Ioffe & Szegedy, 2015; Lin et al., 2014; Rosebrock, 2017; Yosinski et al., 2014). The Inception v2 was trained on the output of the segmentation stage.

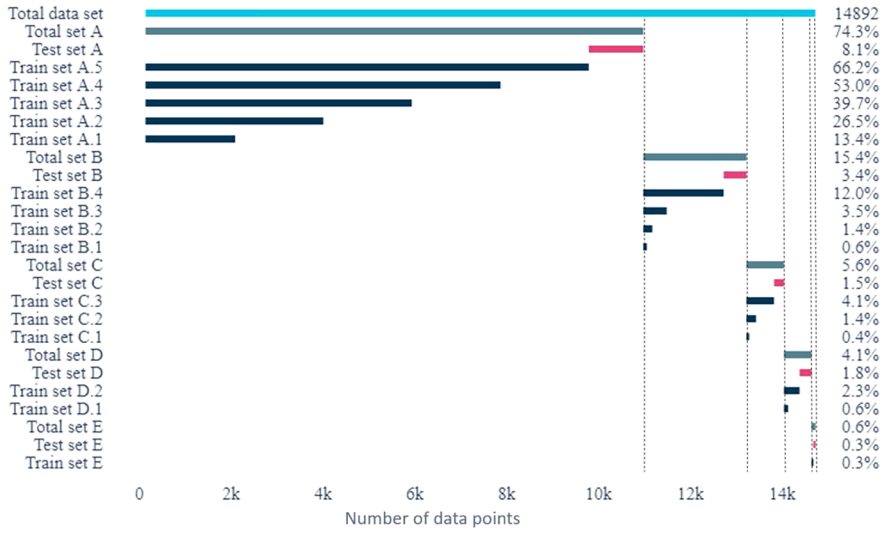

Figure 4. Illustration of the subsets of the floating plastic data set used in the experiments. The left column lists names of subsets as referred to throughout the study, where A to E refer to different locations as found in Table 2. The right column lists percentages of data contained in each subset, compared to the total number of labels shown on top. The lengths of the horizontal bars represent subset size.

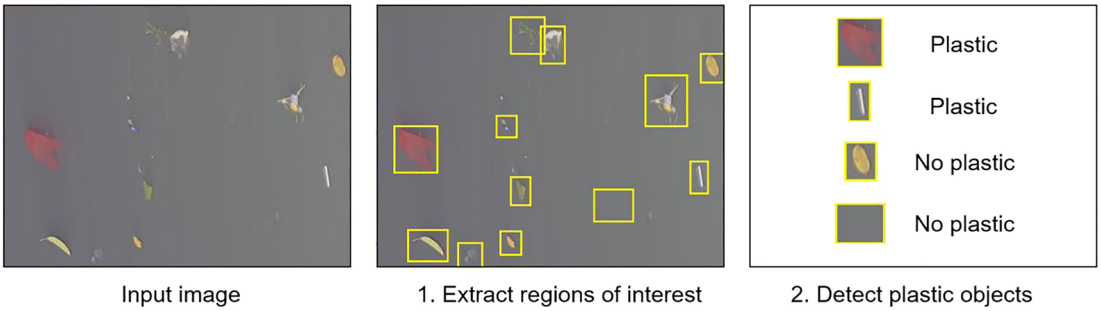

Figure 5. Given a river image as input image (left), the monitoring method consists of two stages: (1) the segmentation stage that predicts which regions of the image are of interest (i.e., likely contain a plastic object) and subsequently (2) the detection stage that determines which of the suggested regions actually contain plastic and which do not.

2.3. Three Experiments

We performed three experiments to determine: the precision of the automated monitoring and how it depends on data set size (Experiment I), how well the image-based monitoring performance generalizes to novel locations (Experiment II), and how the image-based monitoring performance compares to human monitoring performance (Experiment III).

2.3.1. Experiment I: Precision and Effect of Data Set Size

To assess the precision of our method, we trained it on our largest data set (A). In deep learning experiments it is customary to explore different settings of the learning algorithm to optimize the performance. We explored the following settings: (i) data augmentation, (ii) fixed versus adaptive learning rate, and (iii) training procedure.

Data augmentation is a method to virtually extend the size of the training set by manipulating images, for example, adding horizontally and vertically flipped copies of the original images. Data augmentation is known to boost performance (Bengio et al., 2013; Perez & Wang, 2017). The learning rate specifies the rate at which the parameters (weights) of the CNN are updated. Depending on the task, a fixed or adaptive (slowly decreasing) learning rate can be beneficial to performance. The fixed learning rate is set to 0.0002. The training procedure specifies how the parameters of the CNN are updated. We experimented with both the momentum and ADAM procedures (LeCun et al., 2015; Ruder, 2016).

CNNs are data hungry and typically require thousands of examples per class (LeCun et al., 2015). We experimented with the largest location Subset A and its five constituent subsets of decreasing size (i.e., A.5, A.4, A.3, A.2, and A.1, see Figure 4) to estimate how data set size affects monitoring precision.

2.3.2. Experiment II: Generalization to New Locations

To establish the generalization of our method, that is, the extent to which it can be applied to different locations and environmental conditions, we measured the precision of our method when being trained on one Location X and tested on another novel Location Y. In addition, we determined for the CNN trained on Location X, to what extent a small amount of incremental training on data from the novel Location Y improves the performance on Location Y. In this way we were able to estimate the performance of our method on new locations and situations.

2.3.3. Experiment III: Automated Versus Human Performance

In this experiment, human counted plastic object data were compared to our monitoring method. A visual counting sessions determined rate of plastic objects passing by, rather than the amount of plastics in a given field of view which the object detection CNN yields.

To obtain a rate estimate for our monitoring method, 82 one-minute video clips were processed image wise for Location A. The images in these clips were not part of Training Set A and hence are images that were previously unseen to our method. For each plastic object prediction, in addition to bounding box coordinates, our method provided a confidence score ranging from 0 to 1. The confidence score represents how confident the CNN was that the object is plastic, closer to 1 being more confident. We included all predictions with a confidence score of at least 0.5.

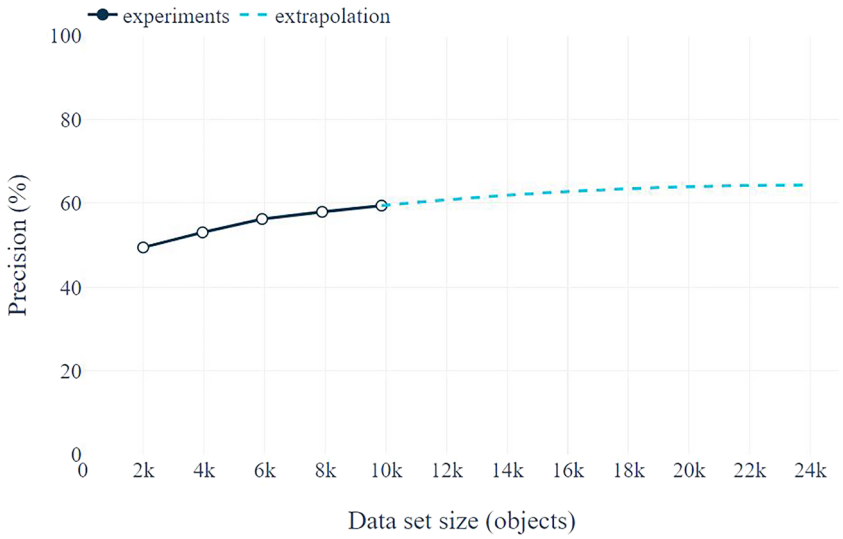

Figure 6. Effect of data set size on detection performance and extrapolation to large data sets.

Data normalization. In order to enable comparative evaluation of automated versus human counting, we applied a processing step to automatically count plastics in videos when objects crossed a horizontal line at the center of the video, to mimic visual counting behavior. Furthermore, all counts were normalized for observation area between camera and observers field of view.

2.4. Evaluation

For Experiments I and II, performance of the monitoring method was evaluated using average precision (Salton & McGill, 1986; Rosebrock, 2017), henceforth referred to as *precision*. The precision metric is the percentage of examples that are correctly detected as being plastic. The value of the precision ranges from 0% (*worst*) to 100% (*best*). We estimate that pure guessing would yield a precision of about 10%. Any value above that indicates that our method performs better than chance level. We do not expect perfect performance (precision = 100%), due to the complexity of the task. The visual appearance of water surface in the outdoors is subject to high variability due to environmental factors, such as scattering of sunlight, brightness variations resulting from overcast conditions, and distortions due to wind and sediment transport. Additionally, plastic debris composition is highly variable (van Emmerik et al., 2018) and not always easily distinguishable from other types of waste. For instance, a study involving the automatic detection of underwater waste by means of an autonomous underwater vehicle reported a precision ranging between 65% and 85% in detecting plastics (Fulton et al., 2019). Hence, we consider a precision within this range as a successful performance of our method.

For Experiment III, we compared the counting results of our monitoring method with those of the human counters. A data normalization step was implemented to express river plastic flow rate as the number of plastic particles per minute per meter river width.

It is important to note that for Experiment III the evaluation cannot rely on a ground truth, because both the human and automated counts are estimates. Therefore, we examine the similarity of the counts obtained by both methods. To this end we compared the counts in three ways. First, determined the coefficient of determination, R^2, that is, the proportion of variance in one count that is predictable from the other. Second, we computed the difference between the mean counts of each method. Finally, we compared the differences in spread of both counts in terms of standard deviations.

3. Results

3.1. Results of Experiment I: Precision and Effect of Data Set Size

The precision achieved by our method when trained on the largest training set (A.5), without optimizing its settings, equals 59.4%. This is below the range we consider as successful (65–85%). Still, we consider this a promising performance, especially because it can be readily further enhanced by optimizing the settings

Table 3
Overview of the Results Obtained by Varying the Settings

Data augmentation	Learning rate	Optimization function	Precision %
none	fixed	momentum	59.4
horizontal flip	fixed	momentum	60.6
h + vertical flip	fixed	momentum	63.0
h + v flipping	fixed	adam	65.7
h + v flipping	adaptive	adam	**68.7**

(see below). The effect of data set size becomes clear by comparing performances on the subsets of increasing size, that is, A.1, A.2., A.3., A.4, and A.5. This yields a precision of 49.4% on Train Set A.1, gradually increasing up to 59.4% on Subset A.5, as illustrated by the curve in Figure 6. As the curve reveals, enlarging the data set contributes to the precision while the contribution becomes smaller with growing data set size. Employing a conservative extrapolation of this trend, we expect a precision of 64.3% at a data set size of about 24k images. This is illustrated by the dashed extension of the curve.

The performance of our monitoring method is readily enhanced by optimizing the settings. Our exploration of the three different settings yielded the following results. With data augmentation, by adding horizontally and vertically flipped images, the precision is increased to 63%. Using the Adam optimizer rather than the default optimizer with momentum results in an increase from 63% to 65.7%. Applying learning rate decay rather than a fixed learning rate improves the precision up to 68.7%. Altogether, these optimization methods raise the precision with 9.3%. The optimization results are summarized in Table 3. The best result is printed in boldface (68.7%) and represents a more than 9% improvement over the original version of our method. With optimization, our monitoring method performs successfully.

It is highly likely that the improvement due to optimization applies to all data set sizes. Therefore, applying optimization to the data set size results, the dashed extrapolation in Figure 6 would shift upward by about 9%. Training our optimized method with 24k images is expected to yield a precision of around 73%.

3.2. Results of Experiment II: Generalization to Novel Locations

Figure 7 visualizes how the performances obtained at Rivers B–E compare to the performance achieved at River A by our optimized method in Experiment I. The latter is represented by the horizontal dashed line at 68.7% ("baseline model River A"). The performances obtained at Rivers B–E are shown as open circles in the graph. Without any additional training ("0" on the horizontal axis), the best precision (in the sense of "nearest to the performance at River A") is obtained for River C (top circle on the left) and the worst for River E (bottom circle on the left).

Figure 7. Visualization of generalization performance of the monitoring method as a function of the number of location-specific training examples. The dashed line represents the baseline, where testing is done on the same location as on which the monitoring method is trained (A). The solid lines represent evaluation on test sets of different locations while trained exclusively on Train Set A.5 (the number of additional training examples is 0) or with a train set of the new location appended to Train Set A.5 (the number of additional training examples is larger than 0). See text for further details.

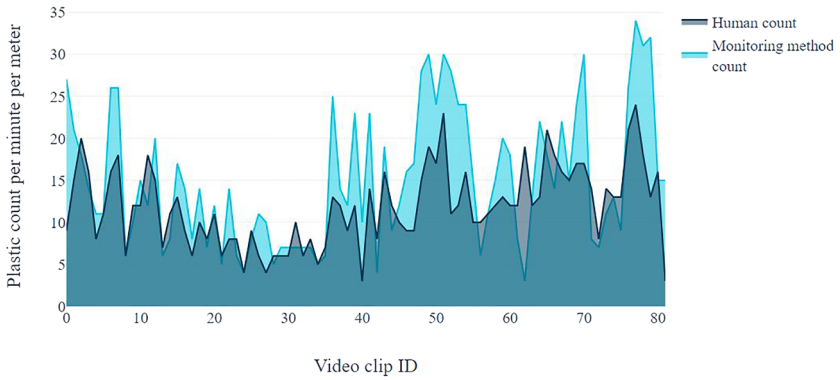

Figure 8. Scatter plot showing the relationship between human counts (vertical axis) and the counts of the monitoring method (horizontal axis). The 82 counts are represented by the black dots. The green line represents the linear 1:1 relationship, and the red line is the best linear fit to the data.

Including additional training examples from the novel location has a positive effect on the performance at that location, as reflected in the mainly positive slopes of the curves when moving from left (0 additional training examples) to right (50, 200, 600, and 1,800 additional training examples).

Three observations are made concerning these results. The first observation is that when predicting on subsets different from the one our method is trained on, performance is reduced to at best a precision of 54% (D) and at worst a precision of 20.5% (E). Train Subset D (moderate plastic density, no waves, low altitude, moderate organic frequency, and no debris patches) is most similar to Train Subset A and therefore suffers the least in performance loss. Subsets B (high plastic density, no waves, moderate altitude, some organics, and many debris patches) and C (moderate plastic density, no waves, moderate altitude, high organic frequency, and no debris patches) are more different from Location A and hence clearly suffer in terms of performance. Train Subset E (low plastic density, many waves, high altitude, no organic material, and no debris patches) differs the most from Train Subset A and hence shows the largest performance loss. These results show that training on images from a single location is insufficient to achieve a good generalization to other locations. The images of Location A are not fully representative for those in the other locations.

The second observation is that the loss in performance can be mitigated by inclusion of location-specific training data. Including only about 50 additional training examples of Train Set E boosts the precision at Location E with 21.3%. For Locations B and C, some improvement is observed with the inclusion of additional location-specific training examples, but both seem to saturate at a maximum value. For Location D, the addition of about 200 examples brings the performance at the same level as for Location A.

The third and final observation is that the precision increases for all locations by including more training examples, which is in agreement with the results of Experiment I.

3.3. Results Experiment III: Human Versus Machine

To measure the agreement of the human counts versus our method's counts, we computed the coefficient of determination, R^2. We found $R^2 = 0.43$, which indicates a reasonable agreement between both counts. Figure 8 presents the scatter plot of the counts of the 82 one-minute video clips obtained by visual counting by humans (vertical axis) and by our method (horizontal axis). A perfect agreement would result in all counts to be positioned on the diagonal (blue line). However, the observed agreement is represented by the best linear fit (red line), which reflects the higher counts of the monitoring method as compared to the human counters.

On average, 34.6% more plastic is detected with our automated method compared to the human counters. The counts of our method have a higher spread than those of human counters, with a standard deviation of 8 compared to 4.7, respectively.

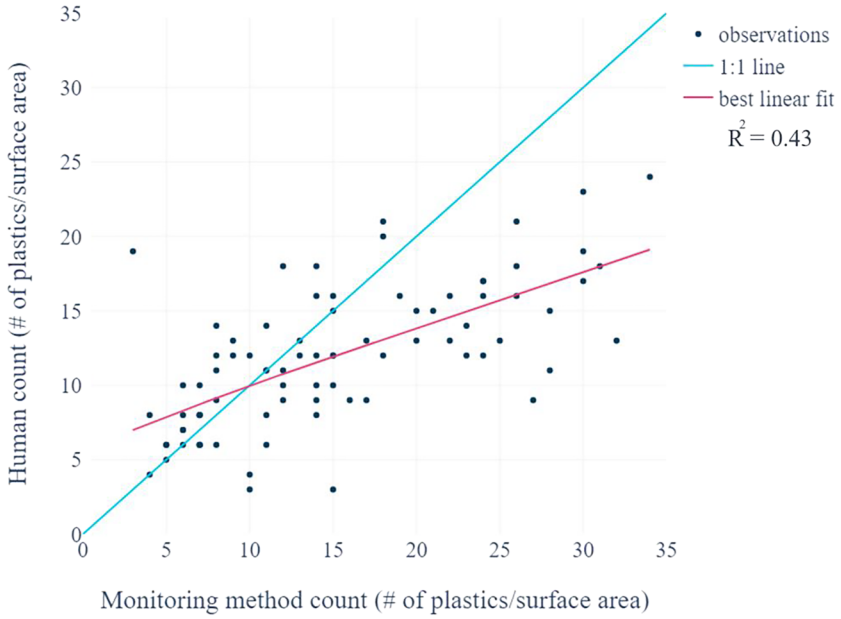

Figure 9. Comparison of the counts by humans and the monitoring method for the 82 samples.

Figure 9 illustrates the high variability within both data sets. We do see that variability in both data sets tends to move together (as reflected in the R^2 value), though variations in the monitoring method's predictions are more pronounced (as reflected in the standard deviations).

Interestingly, the monitoring method predicts relatively higher quantities for video clips with large quantities. In the clips of some monitoring locations, observers had to count plastic objects floating at rates of up to 35 objects a minute per meter river width. Given that the monitoring area is up to 8 m in length, this poses a considerable challenge to human counters. Therefore, the differences observed in Figure 9 might be explained by a limitation on how many objects per minute a human observer can realistically count.

Taken together, these results suggest that the monitoring method's counts are correlated with those of humans. Our results even suggest that the automated monitoring method is better able to count high rates of plastics per time unit. As stated before, there is no ground truth according to which we can determine which counts are more reliable. However, these results suggest that the monitoring method has a reasonable agreement in terms of counting with those of humans.

4. Discussion

Our study assessed the feasibility of the application of deep learning, with the goal of using image sensors to monitor macroplastics on water surfaces. In this section we discuss points of improvements and the implications of our results.

4.1. Points of Improvement

Our monitoring method can be improved along three lines: data set, sensor (video camera), and the segmentation and detection algorithms.

4.1.1. Data Set Improvements

The data set used in this study is unique in the sense that it is the first of its kind. Still, we see two main lines of improvement: increasing the number and variety of locations and improving the reliability of human labeling.

By increasing the number and variety of locations, the monitoring method can be trained on a more representative set of examples, which results in improved performance. We have shown that some fine tuning to the local situation, by performing some additional training on new locations, is feasible. However, the fine

tuning requires additional computational resources. It would be preferred to have a robust monitoring method that does not need fine tuning. Future research focused on increasing the number and variety of the data can establish the feasibility of such a generic monitoring method.

The second line of improvement concerns the reliability of human selection of image regions containing plastic. The abilities of volunteers to distinguish plastic from environment and organic material in the images may vary considerably. As objects can clutter, the distinction between plastic objects and background (including organics) can be hard to make. During the checking and correcting of the selected regions by volunteers, many of these problems became apparent. Future work should be directed at improving the reliability of the selected regions by mutual checks among multiple assessors.

4.1.2. Sensor Improvements

While RGB video cameras are readily available worldwide, there are at least four limitations in terms of their imaging quality. The first limitation is the H.265 video compression employed that induces perceivable encoding artifacts (Lin et al., 2019) that may impede image quality and hence counting precision.

The second limitation is the limited resolution, contrast, and color depth of off-the-shelve video cameras that also negatively affect counting precision. A substantial proportion of the image regions selected by the Zooniverse volunteers were only a few pixels in width and height. Higher resolution, contrast, and color depth may positively affect the processing of these images.

The third limitation concerns the RGB sensor. Monitoring with RGB video cameras is only feasible at daylight hours. To achieve 24-7 monitoring different sensor types can be experimented with Biermann et al. (2020).

The fourth and final limitation is the lack of sensitivity to plastic located below the water surface. It is unknown what percentage of the plastic waste is floating under the water surface. Cross validation with conventional methods can be used to approximate, for example, the relation between visible and invisible proportions of debris in different environmental circumstances (Zaat, 2020).

4.1.3. Segmentation and Detection Improvements

We suggest two main improvements for the segmentation and detection stages of the monitoring method.

The first improvement concerns the constituent CNNs of both stages. Our specific implementation of the monitoring method was based on the combination of the faster R-CNN for segmentation and the Inception V2 CNN for classification. The rapid developments in deep learning and the efficiency and accuracy of deep learning strongly suggests that more recent variants of these CNNs may further improve the precision of the monitoring task.

The second improvement concerns data augmentation. Horizontal and vertical flipping proved effective data augmentation techniques, which is in line with other research stating that data enrichment results in better performances (Perez & Wang, 2017). Other techniques such as image distortion can be explored in the future.

4.2. Implications of Our Results

The main implication of our work is that the automated monitoring of river plastic is feasible. We have shown our method to be successful in the detection of plastic and to have a reasonable agreement with flow-rate estimates of humans. While a considerable performance drop can be expected on new circumstances, the monitoring method is able to retain over two thirds of its predictive accuracy for three out of four new locations, even without any additional training. If we permit some retraining, generalization is improved considerably.

We expect the success of our method to generalize beyond the trained locations. This expectation is based on the widespread experience with machine learning and especially deep learning that larger and more representative data sets lead to better performances (Sun et al., 2017).

While acknowledging the aforementioned limitations, the results reported for our monitoring method suggest that it provides an effective measurement method on plastic throughput on water surfaces at scale. Moreover, the method allows for standardization and centralization more easily compared to decentralized human based sampling methodologies. Humans do not excel in performing a repetitive, high-pace, task. As image sensors are relatively affordable, many devices are eligible to become an input feed for data collection. While we used the same camera for all data collection, other cameras, and drones (Geraeds et al., 2019; Niu

et al., 2019) can all be considered. More so, many of these devices are equipped with additional sensors such as a GPS, enabling inclusion of spatial information and other metadata to eliminate user bias.

5. Conclusion

Deep learning technology applied to images can enable water surface plastic monitoring to an extent where estimates are obtained that are expected to be more reliable and consistent than human monitoring. We are confident that the performance of our monitoring method will generalize to other locations and situations, provided that the training set is enlarged to incorporate a wider variety of situations.

Riverine plastics are a major concern to the environment, aquatic life, and also human health. While it is expected that rivers are a major contributor of plastic emission into the ocean, the spatiotemporal distribution is yet to be fully understood. Existing methods for river plastic monitoring either lack scalability or depend on many assumptions. This study presented a proof of principle showing that an automated method based on deep learning is feasible and can enable obtaining reliable insights on floating macroplastics in rivers around the world.

Appendix A: Camera Footage Sample per Location

We provide some camera imagery of each location in Figures A1–A5. This should give the reader some impression of what data are used in this study and on the location characteristic differences discussed.

Figure A1. Location A, the main and baseline location with high plastic density, no waves, low altitude, little organics, and only some debris patches.

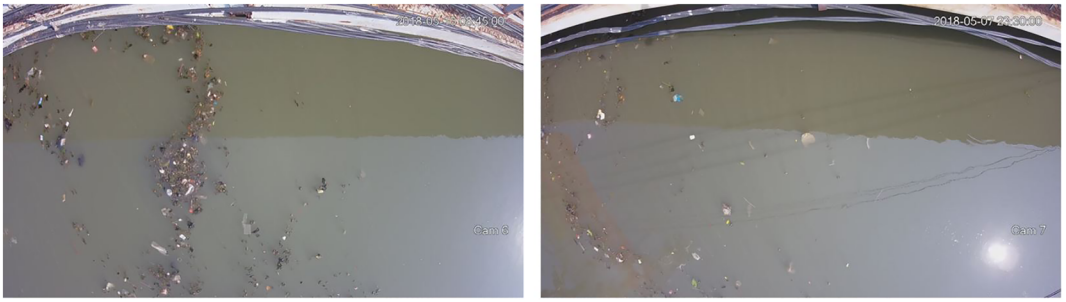

Figure A2. Location B, high plastic density, no waves, moderate altitude, some organics, and many debris patches.

Figure A3. Location C, moderate plastic density, no waves, moderate altitude, high organic frequency, and no debris patches.

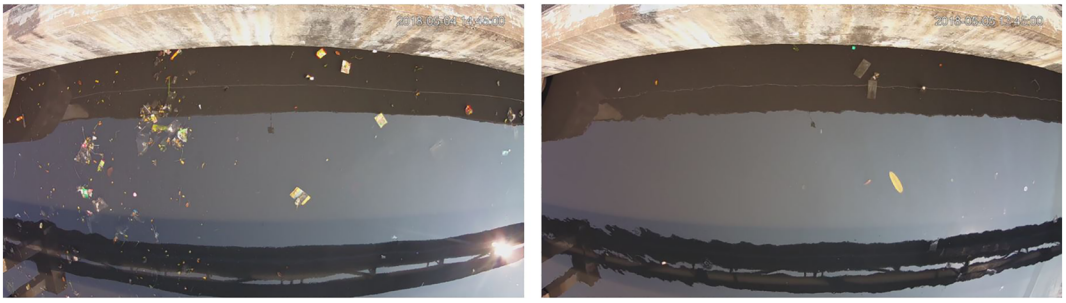

Figure A4. Location D, moderate plastic density, no waves, low altitude, moderate organic frequency, and no debris patches.

Figure A5. Location D, low plastic density, many waves, high altitude, no organic material, and no debris patches.

Data Availability Statement

Data and code for this research are available in Cyliesho (2020) with GNU GPLv3 license (10.5281/zenodo.3817117). Please contact us directly for more information.

Conflict of Interest

C. v. L., K. v.e., and T. v. E. are or used to be employed by The Ocean Cleanup. C. v. L. is cofounder of Soda science.

Acknowledgments

We would like to thank the donors of The Ocean Cleanup who helped funding this study and the Microsoft AI for Earth program for providing the computing resources necessary to execute this study. Additionally, we would like to thank everyone who contributed to the data collection and labeling procedures as without that this study would have been impossible to execute. This publication uses data generated via the http://www.zooniverse.org platform, development of which is funded by generous support, including a Global Impact Award from Google, and by a grant from the Alfred P. Sloan Foundation. The object detection model was optimized and trained using the Tensorflow Object Detection API (Huang et al., 2017). We thank the two anonymous reviewers, whose comments helped to improve the manuscript considerably.

References

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(8), 1798–1828.

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.

Cole, M., Lindeque, P., Fileman, E., Halsband, C., Goodhead, R., Moger, J., & Galloway, T. S. (2013). Microplastic ingestion by zooplankton. *Environmental Science & Technology*, *47*(12), 6646–6655.

Cyliesho (2020, May). colinvanlieshout/riverplasticdetection: Automated river plastic monitoring using deep learning and cameras. https://doi.org/10.5281/zenodo.3817117

DeVries, T., Misra, I., Wang, C., & vander Maaten, L. P. J. (2019). Does object recognition work for everyone?*CVPR workshop on computer vision for global challenges*.

Fulton, M., Hong, J., Islam, M. J., & Sattar, J. (2019). Robotic detection of marine litter using deep visual detection models, *2019 International Conference on Robotics and Automation (ICRA)* (pp. 5752–5758).

Gasperi, J., Dris, R., Bonin, T., Rocher, V., & Tassin, B. (2014). Assessment of floating plastic debris in surface water along the Seine River. *Environmental Pollution*, *195*, 163–166.

Geraeds, M., van Emmerik, T., de Vries, R., & bin Ab Razak, M. S. (2019). Riverine plastic litter monitoring using unmanned aerial vehicles (uavs). *Remote Sensing*, *11*(17), 2045.

Girshick, R. (2015). Fast R-CNN, *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1440–1448).

González-Fernández, D., & Hanke, G. (2017). Toward a harmonized approach for monitoring of riverine floating macro litter inputs to the marine environment. *Frontiers in Marine Science*, *4*, 86.

González-Fernández, D., Hanke, G., Kideys, A., Navarro-Ortega, A., Sanchez-Vidal, A., Brugère, A., et al. (2018). Floating macro litter in European rivers-top items (Ph.D. Thesis).

Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., et al. (2017). Speed/accuracy trade-offs for modern convolutional object detectors, *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 7310–7311).

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167.

Jambeck, J. R., Geyer, R., Wilcox, C., Siegler, T. R., Perryman, M., Andrady, A., et al. (2015). Plastic waste inputs from land into the ocean. *Science*, *347*(6223), 768–771.

Kylili, K., Kyriakides, I., Artusi, A., & Hadjistassou, C. (2019). Identifying floating plastic marine debris using a deep learning approach. *Environmental Science and Pollution Research*, *26*, 17,091–17,099.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436.

Lebreton, L., Slat, B., Ferrari, F., Sainte-Rose, B., Aitken, J., Marthouse, R., et al. (2018). Evidence that the Great Pacific Garbage Patch is rapidly accumulating plastic. *Scientific Reports*, *8*(1), 4666.

Lebreton, L., Van der Zwet, J., Damsteeg, J.-W., Slat, B., Andrady, A., & Reisser, J. (2017). River plastic emissions to the worlds oceans. *Nature Communications*, *8*, 15611.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., et al. (2014). Microsoft COCO: Common objects in COntext, *European Conference on Computer Vision* (pp. 740–755).

Lin, L., Yu, S., Zhao, T., & Wang, Z. (2019). PEA265: Perceptual assessment of video compression artifacts. CoRR (Computing Research Repository).

Monge-Ganuzas, M., Gainza, J., Liria, P., Epelde, I., Uriarte, A., Garnier, R., et al. (2017). Morphodynamic evolution of Laida beach (Oka estuary, Urdaibai Biosphere Reserve, southeastern Bay of Biscay) in response to supratidal beach nourishment actions. *Journal of Sea Research*, *130*, 85–95.

Niu, G., Li, J., Guo, S., Pun, M., Hou, L., & Yang, L. (2019). SuperDock: A deep learning-based automated floating trash monitoring system, *2019 IEEE International Conference on Robotics and Biomimetics (ROBIO)*(pp. 1035–1040).

Perez, L., & Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. arXiv preprint arXiv:1712.04621.

Rech, S., Macaya-Caquilpán, V., Pantoja, J. F., Rivadeneira, M. M., Madariaga, D. J., & Thiel, M.(2014). Rivers as a source of marine litter—A study from the SE Pacific. *Marine Pollution Bulletin*, *82*(1-2), 66–75.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks, *Advances in neural information processing systems* (pp. 91–99).

Rosebrock, A. (2017). *Deep learning for computer vision with Python*. New York: Pyimageseach.

Ruder, S. (2016). An overview of gradient descent optimization algorithms. arXiv preprint arXiv:1609.04747.

Ruiz, I., Barurko, O., Epelde, I., Liria, P., Rubio, A., Mader, J., & Delpey, M. (2020). Monitoring floating riverine pollution by advanced technology. *EGU General Assembly*, *21*, EGU2019–EGU18999.

Salton, G., & McGill, M. J. (1986). *Introduction to modern information retrieval*. New York, NY: McGraw-Hill, Inc.

Schmidt, C., Krauth, T., & Wagner, S. (2017). Export of plastic debris by rivers into the sea. *Environmental Science & Technology*, *51*(21), 12,246–12,253.

Sun, C., Shrivastava, A., Singh, S., & Gupta, A. (2017). Revisiting unreasonable effectiveness of data in deep learning era, *The IEEE International Conference on Computer Vision (ICCV)*.

Thompson, R. C., Olsen, Y., Mitchell, R. P., Davis, A., Rowland, S. J., John, A. W. G., et al. (2004). Lost at sea: Where is all the plastic? *Science*, *304*(5672), 838–838.

Tramoy, R., Gasperi, J., Dris, R., Colasse, L., Fisson, C., Sananes, S., et al. (2019). Assessment of the plastic inputs from the seine basin to the sea using statistical and field approaches. *Frontiers in Marine Science*, *6*, 151.

van Calcar, C. J., & van Emmerik, T. H. M. (2019). Abundance of plastic debris across European and Asian rivers. *Environmental Research Letters*, *14*(12), 124051.

van Emmerik, T., Kieu-Le, T.-C., Loozen, M., van Oeveren, K., Strady, E., Bui, X.-T., et al. (2018). A methodology to characterize riverine macroplastic emission into the ocean. *Frontiers in Marine Science*, *5*, 372.

van Emmerik, T., Loozen, M., van Oeveren, K., Buschman, F., & Prinsen, G. (2019). Riverine plastic emission from Jakarta into the ocean. *Environmental Research Letters*, *14*(8), 84033.

van Emmerik, T., & Schwarz, A. (2020). Plastic debris in rivers. *Wiley Interdisciplinary Reviews: Water*, *7*(1), e1398.

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks?*Advances in neural information processing systems* (pp. 3320–3328).

Zaat, L. (2020). Below the surface: A laboratorial research to the vertical distribution of buoyant plastics in rivers (Master's Thesis), the Netherlands.