AGU ADVANCING EARTH AND SPACE SCIENCES

Water Resources Research

RESEARCH ARTICLE
10.1029/2025WR040605

Exploring the Transferability of Image-Based Algorithms for River Plastic Detection: The Value of Small Mixed Data Sets

Khim Cathleen Saddi[1,2,3], Tim H. M. van Emmerik[4], Domenico Miglino[1], Matteo Poggi[5], Francesco Isgrò[6], Paolo F. Tasseron[4,7], Luigi Daniele[8], and Salvatore Manfreda[1]

[1]Dipartimento di Ingegneria Civile, Edile e Ambientale (DICEA), Università degli Studi di Napoli Federico II, Napoli, Italy, [2]Scuola Universitaria Superiore IUSS Pavia, Pavia, Italy, [3]Department of Civil Engineering and Architecture, Ateneo de Naga University, Naga, Philippines, [4]Hydrology and Environmental Hydraulics Group, Wageningen University, Wageningen, The Netherlands, [5]Dipartimento di Informatica - Scienza e Ingegneria, Università di Bologna, Bologna, Italy, [6]Dipartimento di Ingegneria Elettrica e delle Tecnologie dell'Informazione (DIETI), Università degli Studi di Napoli Federico II, Napoli, Italy, [7]Amsterdam Institute for Advanced Metropolitan Solutions, Amsterdam, The Netherlands, [8]Consorzio di Bonifica Integrale Comprensorio Sarno, Nocera, Italy

Abstract Using large image data sets has been the conventional strategy to improve object detection, but it increases annotation effort and training cost and does not guarantee robust transfer to new sites. Here we quantify the value of a *small, diverse* training set for floating macroplastic detection by jointly evaluating performance, computational cost, annotation effort, and cross-site transferability. We compile four river-camera data sets from Indonesia, The Netherlands, and Vietnam (training/internal validation) and Italy (external validation, single site and single day data from a long-term camera monitoring system), harmonized into 13 litter classes and a five-level tiering scheme (progressive class aggregation). We train YOLOv7 and YOLOv8 models and compare site-specific data sets with a merged "Mixed" data set (999 images) spanning heterogeneous environmental conditions. Results show that data sets with more diverse backgrounds (Type II; e.g., The Netherlands) achieve higher performance per annotation than homogeneous data sets (Type I; e.g., Indonesia, Vietnam), whereas naïvely merging data sets can degrade internal validation unless accompanied by feature-aware filtering. Class aggregation substantially increases overall detection skill, with gains consistent across data sets when moving from fine (Tier 4) to coarse (Tier 0) label spaces. Finally, internal validation does not reliably predict external-site performance, underscoring the need for transferability-aware data set design and evaluation. Overall, our findings emphasize that *data diversity and curation*, rather than data set size alone, are key levers to scale river plastic detection toward broader deployment.

Plain Language Summary Large image data sets are often used to improve the performance of object detection algorithms, but collecting and labeling them requires substantial manual and computational resources. Here we show the value of a small, diverse "Mixed" data set (999 images), helping balance annotation workload and processing time while improving algorithm capability to work on new sites. Such camera-based plastic detection can enable wider monitoring of floating litter in rivers. We trained and internally validated models using four data sets, and assessed transferability using a fifth, independent data set for external validation. Using YOLOv7 and YOLOv8 (You Only Look Once), we considered 13 litter classes and five tiered label groupings to investigate how performance changes as data sets are combined or simplified. Our results demonstrate the high potential of using small mixed data sets in a larger scale river plastic detection, which can be implemented in bridge-based monitoring systems. We also show that strong internal performance does not necessarily translate to robust performance at other sites, and merging data sets can reduce accuracy unless a data-filtering step is applied. Overall, we look forward to increased interest in data set quality development so we can bridge scaling up gaps, and optimize our efforts in global river plastic detection monitoring.

Key Points:

- Using a small mixed data set provides added value by optimizing the balance between computational cost, labor cost, and transferability
- Merging data sets saturates the internal validation performance but proves to be beneficial when detecting more plastic classes are desired
- The traditional performance metrics (precision and recall) don't reflect model transferability

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

K. C. Saddi,
khimcathleen.saddi@unina.it

Citation:

Saddi, K. C., van Emmerik, T. H. M., Miglino, D., Poggi, M., Isgrò, F., Tasseron, P. F., et al. (2026). Exploring the transferability of image-based algorithms for river plastic detection: The value of small mixed data sets. *Water Resources Research*, *62*, e2025WR040605. https://doi.org/10.1029/2025WR040605

Received 7 APR 2025
Accepted 26 FEB 2026

Author Contributions:

Conceptualization: Khim Cathleen Saddi, Tim H. M. van Emmerik, Salvatore Manfreda
Data curation: Khim Cathleen Saddi, Domenico Miglino
Formal analysis: Khim Cathleen Saddi
Funding acquisition: Salvatore Manfreda
Investigation: Khim Cathleen Saddi, Domenico Miglino, Paolo F. Tasseron
Methodology: Khim Cathleen Saddi
Project administration: Salvatore Manfreda
Resources: Luigi Daniele
Software: Khim Cathleen Saddi
Supervision: Tim H. M. van Emmerik, Salvatore Manfreda
Validation: Khim Cathleen Saddi

1. Introduction

Considering that recent projections of plastic leakage in the environment show a 50% increase from 2020 levels—30 metric tons by 2040 (OECD, 2024)—balancing the time and resources required to carry out harmonized global plastic monitoring to provide timely aid to policy is critical (Lusher & Primpke, 2023). Despite the alarm raised in the 1970s about the growing plastic pollution (Ryan, 2015), and with the discovery of the Great Pacific Garbage

Visualization: Khim Cathleen Saddi
Writing – original draft: Khim Cathleen Saddi
Writing – review & editing: Tim H. M. van Emmerik, Domenico Miglino, Matteo Poggi, Francesco Isgrò, Paolo F. Tasseron, Luigi Daniele, Salvatore Manfreda

Patch (GPGP) in 1997 (Moore et al., 2001), a comprehensive global data set on the spatial and temporal distribution of plastic pollution remains unavailable. Such data is essential for advancing our understanding of global plastic transport dynamics. Current research on plastic transport and fates is still in its infancy. It was generally assumed that rivers only serve as transport hubs for most of the plastics found in the marine environment (Lebreton & Andrady, 2019). However, recent studies have found that rivers play a more complex role, acting as sinks and reservoirs instead (van Emmerik et al., 2022). Over the years, various plastic monitoring strategies have been developed to study plastic transport and fate, including seasonality, both using traditional and advanced technologies (Savoca et al., 2022; van Emmerik et al., 2019; Veettil et al., 2022). The most recent direction leverages fixed camera systems coupled with machine learning (ML) for optical detection of riverine plastics, increasingly supported by citizen science (Manfreda et al., 2024). This approach is promising for global plastic monitoring because it has successfully been implemented in other monitoring applications such as urban traffic (Pérez et al., 2019), crowd detection (Sánchez et al., 2020), and crop growth (Kuwata & Shibasaki, 2015). These applications have worked well primarily due to their ease in implementation, low labor requirements, and high temporal resolution at a relatively low-cost. However, it is crucial to note that the transferability of models and replicability of methods across different study sites is essential to support ending global plastic pollution (UNEP, 2024).

In computer vision, Transferability or Transfer learning mainly refers to a model's ability to apply pre-trained information to a new task (Djolonga et al., 2021; Thrun & Pratt, 1998). This is beneficial when new knowledge data is limited or expensive to acquire (Fouquet et al., 2023). In this study, the concept of model transferability is defined as the capability of a model to extract new knowledge (data from site B) using both pre-training and training knowledge (data from site A). Very few plastic detection studies discussed and explored the concept of model transferability (Armitage et al., 2022; Jia et al., 2024; Maharjan et al., 2022; Pérez-García et al., 2024). Most of the works are often focused on either algorithm development (Cortesi et al., 2021; Wolf et al., 2020), or new site implementation (van Lieshout et al., 2020). The most popular AI algorithms applied to plastic detection in recent years include Random Forest (RF; e.g., Gonçalves et al., 2020), Convolutional Neural Network (CNN; e.g., Fallati et al., 2019) and You Only Look Once (YOLO; e.g., Lin, 2021). Depending on the model used, the computational cost (time required for training a model) and inference time (time required to detect an object) varies. Most ML models require high computational power for training, often involving multiple GPUs for this phase and still requiring a single one for the inference task to be carried out at deployment (Al-Jarrah et al., 2015), which serves as a barrier for large-scale implementation of object detection in real- and near-real time. The YOLO family of object detectors gained a huge interest in the computer vision community and beyond, due to its fast inference time (Jia et al., 2023). These object detectors are commonly pre-trained and validated using benchmark data sets (i.e., PASCAL VOC, Imagenet, MS COCO, or Open Images) with sizes ranging over four to six orders of magnitude (22k–14 million images) yielding a mean average precision (mAP) of 53.7%–83.5% (Salari et al., 2022; Zou et al., 2023). Except for PASCAL VOC, these benchmark data sets have plastic-related classes identified as either Plastic bag or bottle among other classes (Deng et al., 2009; Everingham et al., 2010; Kuznetsova et al., 2020; Lin et al., 2014). However, prior to plastic detection implementation, these detectors still have to be trained (fine tuning) with a dedicated plastic data set requiring manual annotation which is both laborious, prone to user error, and time consuming. Most algorithms require large datasets—typically in the range of two orders of magnitude (2.5k–5k images)—to achieve a mean average precision (mAP) of approximately 31.6%–81% (Fulton et al., 2019; Huang et al., 2021). This requirement places significant pressure on the resource allocation and hinders large scale implementation of river plastic detection. In addition, although increased data set sizes can improve performance to some extent, results eventually tend to plateau (Zhu et al., 2016).

Looking into publicly available plastic image data sets could be beneficial to minimize resource allocation requirements. However, most of these data sets are commonly acquired through field campaigns restricted to one or few sites (e.g., van Lieshout et al., 2020). This practice often generates data sets which contain site-specific features (e.g., highly turbid water, vegetation transport). In fact, the plastic background (water) is heavily influenced by water-related environmental disruptions (i.e., the effects of solar glint, sky-mirroring, suspended pollutants) affecting plastic detection (Marye et al., 2025). In the same manner, limited plastic classes (only one or a few classes were used in most studies to date) would restrict the amount of information that can be extracted for long-term monitoring. Conversely, incorporating a broader range of plastic classes would require additional labor costs due to the increased annotation requirement workload. These features play a crucial role in data set

development since generally the lack of feature diversity leads to homogenous data sets. In addition, models trained with homogeneous data sets are generally more difficult to transfer to other homogenous data sets due to their high degree of dissimilarity (Chui et al., 2023). This led to increased interest toward heterogeneous transfer learning (HeTL) which usually requires the development of heterogeneous data sets which have wider data feature coverage. These data sets may be created by merging data sets (Day & Khoshgoftaar, 2017; Khan et al., 2024) or collected through a global citizen science approach (e.g., CrowdWater, 2026). However, developing heterogeneous data sets without applying any feature selection process is crucial. All features are considered relevant during the model training process which may result in lower overall performance, given that less performing features could affect the higher performing features (Kachouri et al., 2010).

To shed light over these challenges, this work explores the transferability of image-based algorithms using five data set sources (Indonesia, Vietnam, The Netherlands, and Italy). In Section 2, the overall methodology was presented including data set packaging, model training, and validation (internal and external). These data sets were categorized as either data sets having homogeneous (Type I) or diverse (Type II) backgrounds since it is too early to identify whether a data set is heterogeneous at this point. Also included in this section are the 13 litter classes used, which are mostly taken from dominant plastics found in the environment and other litter observed in the data sets. Since these classes have varying annotation count, which was expected to affect the class detection performance, it was decided to explore clustering certain classes into five tiers (Tiers 0–4). Presented in Section 3, the class- and tier-based model performance using traditional performance metrics (Precision, and Recall), performance trends when downsizing a mixed data set, and computational and labor costs. It was also presented how different models perform when transferred to external data sets. Lastly, we provided our conclusion and critical future directions in Section 4.

This work highlights the added value and transferability of small data sets in river plastic detection. Specifically, we focused on the data set quality, and understanding the level of information that can be extracted using machine learning (ML), rather than on the algorithm itself—whose development is already advancing rapidly (Zou et al., 2023). Developing small mixed data sets would allow for advancement of river plastic detection and facilitate large-scale implementation at a lower cost, regardless of future algorithmic advancements.

2. Materials and Methods

2.1. Method Overview

Understanding how a model can be effectively transferred from the training site to other sites, regardless of their environmental characteristics, is essential. This ensures a model's capacity to be implemented on a global scale. Highlighting the importance of transferability, this work used both Type I (Indonesia, Vietnam) and Type II (The Netherlands, and Italy) data set sources, two YOLO base models (version 7; Wang et al., 2023 and version 8; Jocher et al., 2023), five tiers, three traditional performance metrics (precision, recall, and mAP), two validation (internal and external), and three proposed scalability metrics (computational cost, labor cost, and transferability; see Figure 1).

Data set preparation started with manual filtering of images with information. The filtered data was uploaded in Roboflow (Dwyer et al., 2026) for online annotation labeling for each data set. Data cleaning was also performed to double check annotations (both from roboflow and existing data set labels) prior to data set packaging in YOLO format. All models were trained and validated with their respective data set packages. An external validation (for transferability evaluation) was performed using an external data set package (Italy). Using the proposed scalability metrics, the viability in scaling both Type I and Type II data sets were explored. More detailed discussions of the methodology were presented in the following sections.

2.2. Plastic Classes and Tiers

Plastic types found in different environments vary depending on urbanization, hydrological conditions (de Carvalho et al., 2021), and other environmental factors. This makes plastic detection in rivers more complex since there are various plastic categories that are currently being used in plastic monitoring worldwide (i.e., polymer type, size, sectoral use), of which recently the most used is the OSPAR categories (Wenneker & Oosterbaan, 2010). Dominant plastics by polymer type found in different aquatic ecosystems include Polyethylene (PE), Polypropylene (PP), and Polystyrene (PS; Schwarz et al., 2019).

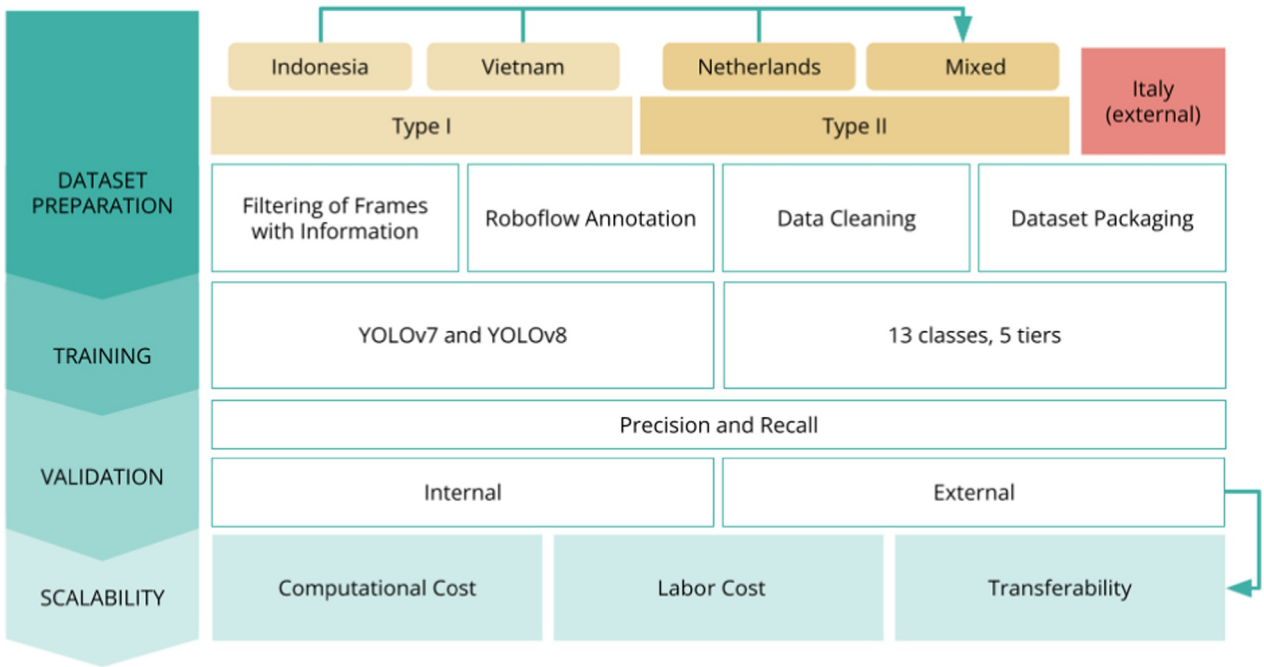


Figure 1. Workflow for river plastic detection models.

Considering future implementation with citizen science, and the need for brevity in identifying plastic objects, this work used 13 plastic classes, and then later clustered them into tiers to understand the extent of information which can be extracted utilizing object detection (see Figure 2). A hierarchical plastic class aggregation was enforced using a tier-based system to investigate performance gains and losses when plastic classes are merged, especially with the class imbalance problem. This is a classic ML problem where a model becomes biased to a few classes which have relatively more annotations than the others (Japkowicz & Stephen, 2002). By merging the classes, our approach attempts to partially address the class imbalance, while also allowing the model to learn diverse features in aggregated classes. This would yield critical insights for future implementation requirements. Due to the scarcity of river plastic data, the plastic classes used in this work were based on dominant plastics found as ocean litter (Morales-Caselles et al., 2021) which were randomly listed, and observations within the data sets. Tier 4 presents the most complete classes (13) used in training the models. With one level lower, some plastic classes were merged into class 0-Unclassified Plastic Objects, until all classes were reclassified as class 0 (Tier 0). These classes can be categorized as level 2–3 in the EU Joint List of litters hierarchical system (Galgani et al., 2013).

2.3. Training Data Sets

Organizing the amount of annotation needed for each class to yield the required detection will allow optimizing the efforts in river plastic detection. In reality, this is a challenging task since it is inevitable that river plastic data sets have class biases depending on the acquisition site. Unless these class requirements are identified, then the creation of a heterogeneous data set satisfying these minimums would be optimized.

There were four data set sources that were utilized in this work (Indonesia, Vietnam, The Netherlands, and Italy). Summarized in Table 1 their properties, (i.e., image count, annotations present, etc.) that shows the degree of similarity and dissimilarity of these data sets. Plastic data sets from Indonesia (van Lieshout et al., 2020) and Vietnam (L. Schreyers et al., 2022) are both publicly available. They both have a 90° camera angle (from the vertical axis), while The Netherlands has a mix of 45° and 90°. Most publicly available plastic data sets on, and near water are always challenged by the use of 90° view angle (Gnann et al., 2022). This serves as a limiting factor in river plastic detection since plastic objects tend to change geometry (e.g., rotate, submerge, or geometrically transform) as they move along the river during transport. In such cases, the model will only be learning the top view data features, restricting learning. On the other hand, using data with a 45° camera angle allows the model to

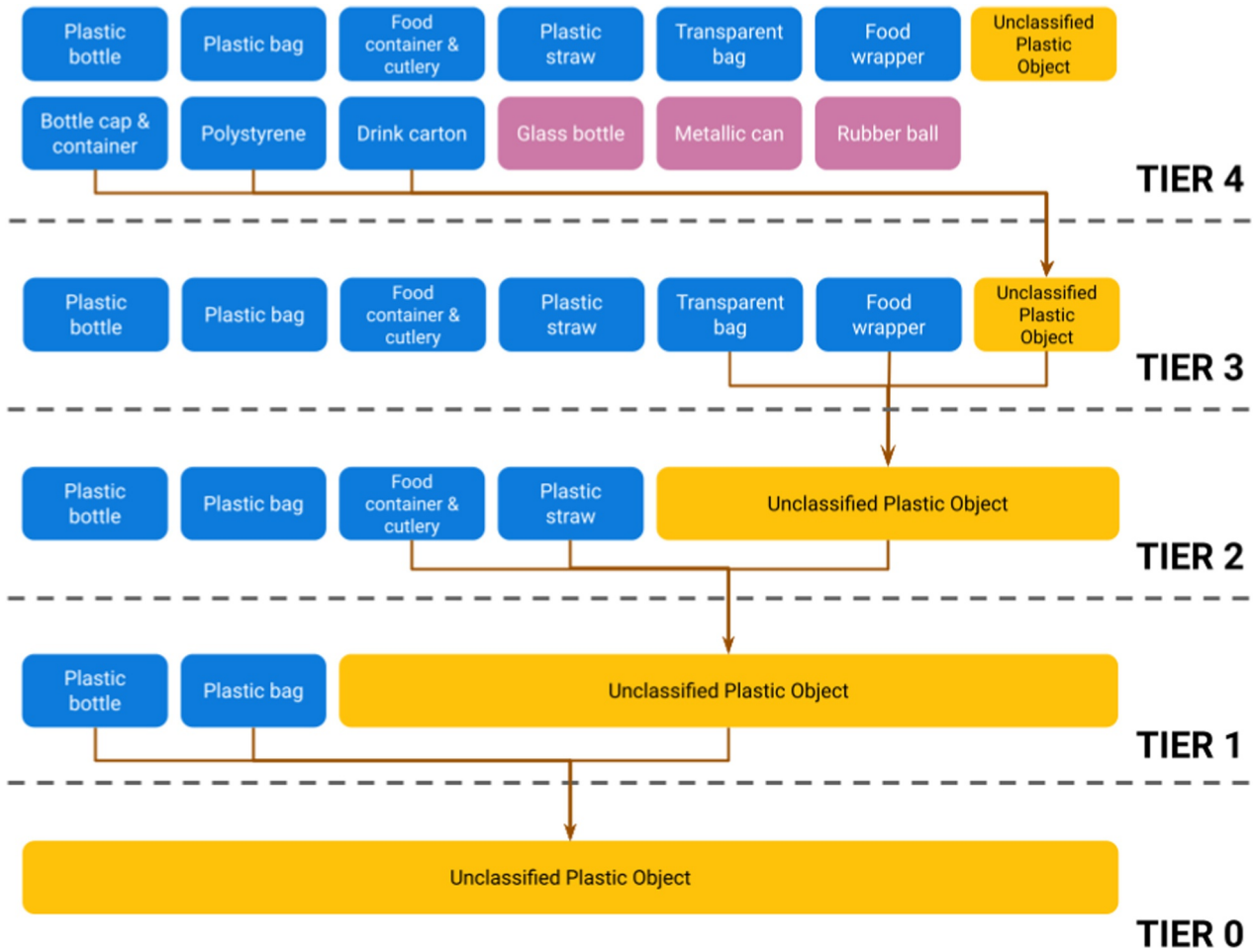


Figure 2. Overview of Plastic Class Tiers. Classes from Plastic bottle through Polystyrene (Tier 4) were based on the most dominant plastic types found in the global environment (Morales-Caselles et al., 2021), then later other classes were added as site observations of other litter progressed (Drink Carton through Rubber ball). Tier-based models were generated based on these clusters. Note that non-plastic classes (Glass bottle, Metallic can, and Rubber ball) were not included in Tiers 0–3.

Table 1
River Plastic Data Sets Description

Plastic data set/Description	Indonesia (Jakarta)	Vietnam (Ho Chi Minh)	The Netherlands (Amsterdam)	Italy (Sarno)
Image count	525	206	268	40
Annotations	10,921	2401	1890	292
Environmental site feature	Sky mirroring	Vegetation transport	Algae abundance, sky mirroring	Low plastic count
Camera angle	90°	90°	45° and 90°	45°
Camera to water surface distance	4.5–8 m	10 m	~5 m	~2–3 m
Input resolution (in pixels)	1920 × 1080	2795 × 2095	4000 × 3000	1920 × 1080
Sensor	Dahua Easy4ip IPC-HDBW1435EP-W	DJI Phantom 3 UAV, FC6310 camera	GoPro Hero 5	Brinno TC200
Remark		Training and Validation		External Validation Only

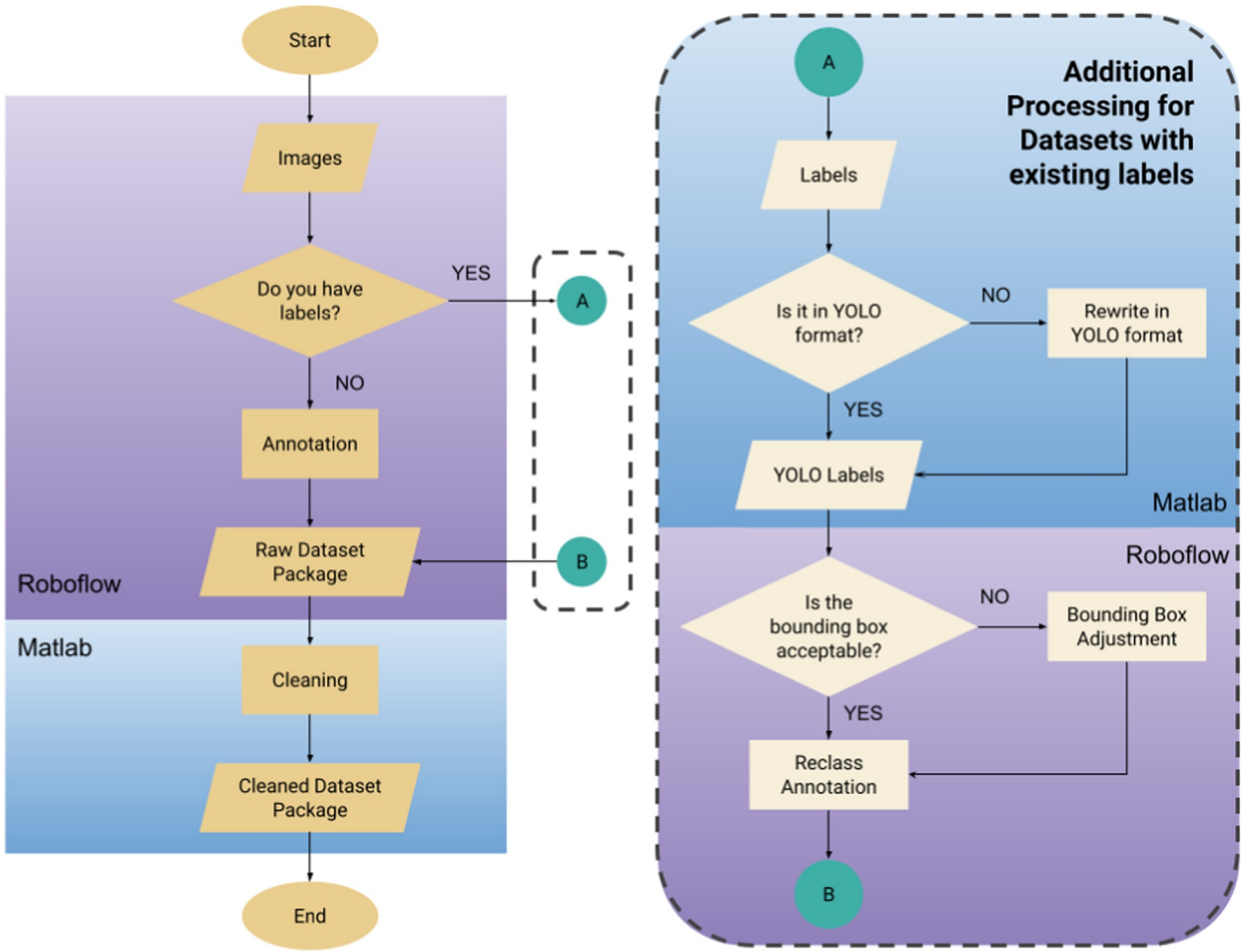


Figure 3. Data set preparation showing the additional processing for data sets with existing labels.

learn features in isometric views (3D) which are more robust in terms of geometrical diversity, and are more relevant in future integration of citizen science in the pipeline. The prevalence of capturing objects (in this case plastic litter items) in 3D as observed in citizen-captured images in rivers (e.g., CrowdWater, 2026), follow a normal tendency to tilt the camera toward the object especially in young adults (Thomson & Uddin, 2023), but not to the extent of reaching 90°, except when the object is directly under the sensor. In this work, we investigated the learning capability of these data sets, especially when data sets with 90° limited training views were applied to the Italy data set (external validation) which has the 45° camera angle.

Indonesia and Vietnam data sets were tagged as Type I, while The Netherlands data set was identified as Type II since they have more diversity in the background information. Another Type II data set (Mixed) was created by merging Indonesia, The Netherlands, and Vietnam data sets. These four data sets were used as training data sets. All annotations (15,212) were manually created for all data sets in Roboflow including the adjustment of the existing annotations (one class only) for Indonesia, which required cleaning and reclassification (Figure 3). We decided to use a variety of camera systems from UAV, fixed and portable cameras with different camera angles to test the capability of our models across geometrical variations, which has been one of the challenges in the field of computer vision (Pham & Smeulders, 2005). All data sets follow the 70-20-10 data split accounting for training, validation, and test. Figure 4 presents the sample training data images per data set. See Text S1 and Table S1–S3 in Supporting Information S1 for more details on the data set packaging.

10.1029/2025WR040605

Figure 4. Sample images of the Training Data sets taken from (a) Indonesia, (b) Vietnam, and (c) The Netherlands (see Figure S1 in Supporting Information S1 for the sample training images per class).

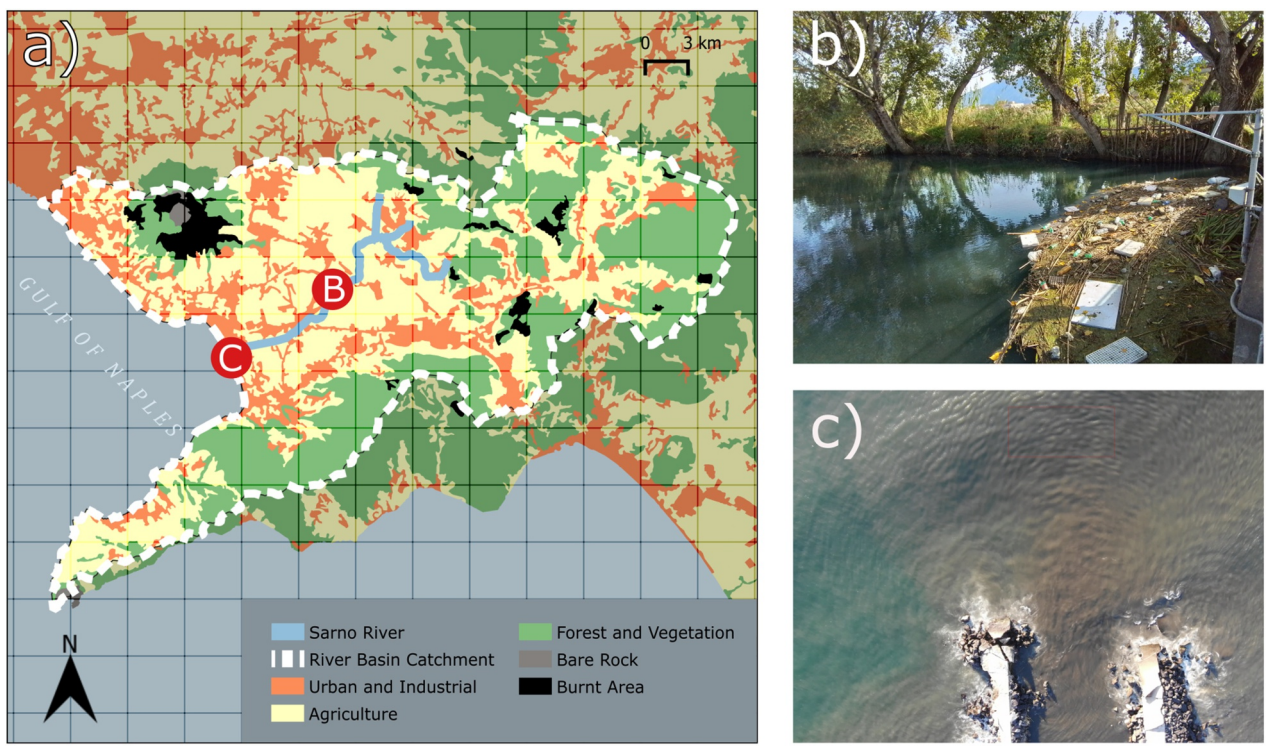


Figure 5. Overview of the Sarno River Basin and its current pollution problem showing (a) land use and cover of the river basin based on Corine Land Cover 2018 (EU, 2024), (b) long-term monitoring station for Riverwatch Project, and (c) Sarno river mouth. The land cover surrounding the river shows heavy agricultural and industrial use with a high concentration of urban areas toward the river mouth. Observations of plastic and other pollutant transport accumulating in bridges require constant effort to unclog upstream, and high variability of water color can be observed in the river mouth downstream.

2.4. The Study Case Used to Test Model Transferability: The Sarno River

Situated in the Campania Region adjacent with Mt. Vesuvius, South of Italy, Sarno River is part of a river basin district where 71%–80% of water bodies failed to achieve good ecological status in Europe for 2021 (EEA, 2024). In Figure 5, we clearly illustrated the extensive industrialization in the region (i.e., tanneries, agricultural, chemical) which highly influenced its pollution level (Thiombane et al., 2018). In fact, despite the significant efforts for restoration (e.g., Braioni et al., 2009), diffuse pollution sources and plastic pollution are still present brought by the intense anthropogenic activities surrounding the catchment.

An RGB camera capturing images at a 15-second time step was installed under the Riverwatch project at Point B. The Riverwatch project (https://sites.google.com/view/riverwatch) utilizes citizen science to build a dense network of river monitoring stations using a mix of fixed cameras and smartphones for pollution monitoring along the Sarno river with a special focus on plastic. To test transferability, a single day camera data was taken from this long-term camera setup to comprise the external validation—Italy data set.

2.5. Manual Annotation

Manual annotations were carried out for the data sets The Netherlands, Vietnam, and Italy using the Roboflow platform (Dwyer et al., 2026). The time spent annotating one frame depends on the number of features present, the number of defined classes, and the technical capacity of the operator carrying out the annotation. As existing annotations were already available for the Indonesia data set, additional steps were taken to harmonize these data with the other data sets, including label formatting, bounding box calibration, and reclassifying the annotations (Figure 2). Environmental disturbances (i.e., solar glint) were not annotated in this study despite the risk of generating false positives, especially when these artifacts visually mimic other classes. The different labor cost across tiers was not measured since all classes (Tier 4) were already defined during the annotation process, and the reclassification of the classes to generate other tiers (Tier 0-Tier 3) was generated in a python notebook (see Availability Statement section).

2.6. Training, Validation, and Performance Metrics

Jupyter notebook was used to implement the YOLO-based training and validation with NVDIA RTX A4500 GPU. This facilitated a more efficient model training than a free cloud-based GPU due to the iteration nature of the computing methodology (target to produce 140 models). A cloud-based service with GPU computing credits was used during the early phase of the model training. However, due to its limited GPU computing credits, shifting to offline GPU computing proved to be more viable. All models reported in Section 3 were retrained in the offline GPU. Training epochs were set to 300, however, it is important to note that the training in YOLOv8 stops when there is no significant increase after the last 100 epochs by default (patience = 100), resulting in runtime savings, unlike in YOLOv7.

There were two validations that were conducted to assess the performance of the models produced: (a) internal validation—model is validated with the validation data of the same data set used for training, and (b) external validation—model is validated with an external data set (the Sarno river in Italy). These two types of validations will provide an overview of the potential transferability of the trained model on similar environments as the trained data, as well as on completely different environments.

The metrics adopted are the Precision (P) and Recall (R). Usually, most object detection work pays heavy attention to the mean average precision (mAP) when assessing the performance of their models (Padilla et al., 2020), despite that it was previously argued that mAP may not be very good indicators of object detector performance (Redmon & Farhadi, 2018). In the context of plastic detection, high mAP means that the model may be good at detecting the extent of the plastic object within a frame (good detection box fit), independent of whether the plastic item was classified correctly or not. Hence, it was decided to put more emphasis on basic metrics, specifically P (Equation 1) and R (Equation 2) instead of the mAP (Equation 3), yet the latter was also reported for easy comparison with other studies. P and R are based on three detection types: true positive (TP, correct detection included in ground truth), false positive (FP, incorrect detection based on ground truth), and false negative (FN, undetected ground truth; Padilla et al., 2020), while mAP is simply the average AP (area under the P-R curve) over all classes (Oksuz et al., 2018).

Table 2
Point Scale Description of Plastic Data Set Scalability Metrics

	Metric/Scale		
Interpretation	Computational cost (in minutes)	Labor cost (in hours)	Transferability (average of P and R)
Highly scalable	<10	<40	>0.70
Moderately scalable	10–20	40–60	0.60–0.70
Scalable	20–30	60–80	0.50–0.60
Limited scalability	>30	>80	<0.50

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$\text{mAP} = \frac{1}{N}\sum_{i=1}^{N} \text{AP}_i \tag{3}$$

2.7. Plastic Data Set Scalability Metrics: Computational Cost, Labor Cost and Transferability

Currently, there are no available scalability metrics which could similarly capture our goal to assess the viability of data sets (both Type I and Type II) to be applied on global plastic monitoring efforts. In this work, there were three metrics used to measure plastic data set scalability: (a) Computational cost, (b) Labor cost, and (c) Transferability. Table 2 provides an overview of the point scales used to describe these scalability metrics. These point scales were drawn from the experience in this work, while also considering recent algorithm development.

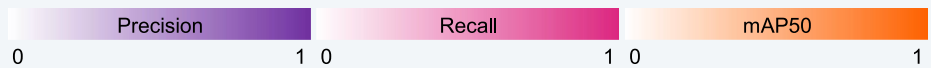
Computational cost (or training runtime) is vital in computer vision. This time element depends heavily on hardware resources (i.e., machines with high computing power) which are often scarce in the global context, especially considering the direction toward Open-World Detection (i.e., out of domain detection, zero-shot detection; Zou et al., 2023). This direction idealizes that models are able to recognize objects outside of their training data (similar to the external validation in this work) and even beyond their class assignments. However, until now, the latter goal would still require iterative addition of new data (incremental training or fine-tuning) to improve models (Feng et al., 2022; Zhang et al., 2022). This training iteration approach would require low computational cost, especially when this is implemented automatically on site (e.g., automatic model training after collecting a certain threshold data). Here, it was defined that a data set which has a computational cost less than 10 min is highly scalable while those with computational costs exceeding 30 min was considered with limited scalability.

Labor cost (in hours) is one of the biggest barriers in computer vision which is heavily associated with the manual annotation work (Adnan et al., 2023). There are recent advances in ML that try to address this challenge by exploring semi-supervised learning (e.g., Li et al., 2022; Tang et al., 2021) where artificial labels are generated to train the models which would try to predict augmented versions of the unlabeled data (Sohn et al., 2020). These approaches were already demonstrated with data sets with one or few litter classes (<10) but not with data sets having >10 complex plastic classes similar in this work. Hence, manual annotation is still necessary, especially when validations are used for incremental learning (model detected objects are added to the training data set). Here, it was defined that a data set which requires a labor cost less than 40 hr (1 week—1 operator, on an 8-hr work week) is highly scalable while those with labor costs exceeding 40 hr (2 weeks—1 operator) was considered with limited scalability. In this work, the manual labor in data collection is not included in the labor cost.

Lastly, data set transferability allows data sets to generate models (trained on one data set) which can also be applied on external data sets, advancing efforts to global plastic detection monitoring. In this work, transferability is measured using the external validation performance (P and R).

Table 3
Performances in Terms of Precision (P), Recall (R), and Mean Average Precision at 50% (mAP50) Across Different Tiers (All Classes)

	Indonesia			The Netherlands			Vietnam			Mixed		
Tier	P	R	mAP50	P	R	mAP50	P	R	mAP50	P	R	mAP50
Tier 0	0.55	0.42	0.43	0.66	0.45	0.49	0.57	0.38	0.39	0.49	0.44	0.40
Tier 1	0.57	0.23	0.21	0.50	0.38	0.37	0.28	0.23	0.22	0.38	0.29	0.25
Tier 2	0.40	0.19	0.15	0.56	0.28	0.26	0.54	0.18	0.19	0.28	0.24	0.18
Tier 3	0.36	0.22	0.17	0.52	0.26	0.27	0.27	0.18	0.14	0.28	0.26	0.20
Tier 4	0.33	0.20	0.13	0.52	0.31	0.34	0.35	0.17	0.16	0.32	0.21	0.18



3. Results and Discussion

3.1. Tier-Based Analysis

It is common for data sets containing multiple classes to have class imbalance problems, meaning the annotation count for some classes is higher than for others. This imbalance is particularly critical in object detection, as establishing interdependencies between classes remains a significant challenge (Oksuz et al., 2020). In general, the models prioritize minimizing errors of prevailing classes (usually with high annotation) which results in higher errors in less prevailing classes (Saini & Susan, 2023). For most object detectors, including YOLO, the overall model performance is generated by averaging all class performances. In this way, low performing classes (usually with relatively less annotation count) would negatively influence the overall model performance (Krawczyk, 2016). In Table 3, we presented the Tier-based performance trends for all training data sets showing the effects of class aggregation on the overall model performance. The model performance decays as the tiering progresses (from Tier 0–4) which varies from ∼0.04–0.5 across all data sets and tiers considering P, R, and mAP50 values.

The partial class aggregation impacts model performance. A decreasing performance trend as the tiering increases was expected since there are more classes and less annotation per class in Tier 4 than in Tier 0. But at the same time, higher tiers also have more defined features that are specific to certain classes rather than clustered in lower tiers. So, the relative increase in performance in Tier 2 (rather than the expected decrease from Tier 1) for both The Netherlands, and Vietnam was unexpected which may suggest that clustering classes may lead to an increase in plastic detection performance to a certain extent, especially when there is a big disparity in annotation count across classes. Observed differences among different tiers provided an overview of quantitative performance tradeoffs developing approaches with relatively high performance with poor information (Tier 0), and relatively low performance with rich information (Tier 4). This is not evident in most river plastic detection studies to date since they only used simple classes like general litter or plastic (Jia et al., 2023; van Lieshout et al., 2020). These observations were drawn upon a single tiering design. A deviation in the plastic tiering system (e.g., plastic class aggregation) might influence a change in these trends, which are deemed critical to be investigated in future work.

3.2. Class-Based Analysis

The Precision-Recall (PR) relationship of different classes (all data sets) and tiers (Mixed data set only) is described in Figure 6. A perfect classifier is characterized by a curve that traverses the upper right corner (with 100% for both precision and recall) of a PR graph (Miao & Zhu, 2022). It was observed that PR curves of dominant classes (curves that are much closer to the perfect PR curve) vary depending on the data set. Presented in Table 4 the class-based performance of the Tier 4 models, where a data annotation and performance variability can be observed per class and per data set source. These trends were clustered into three groups based on the number of annotation and Precision value: LAHP—low annotation, high precision ($A \leq 1{,}000$, $P > 0.5$); LALP—low annotation, low performance ($A \leq 1{,}000$, $P \leq 0.5$); HALP—high annotation, low precision ($A > 1{,}000$, $P \leq 0.5$). LAHP classes from at least two data sets include Bottle cap and container, Drink carton, Glass bottle,

10.1029/2025WR040605

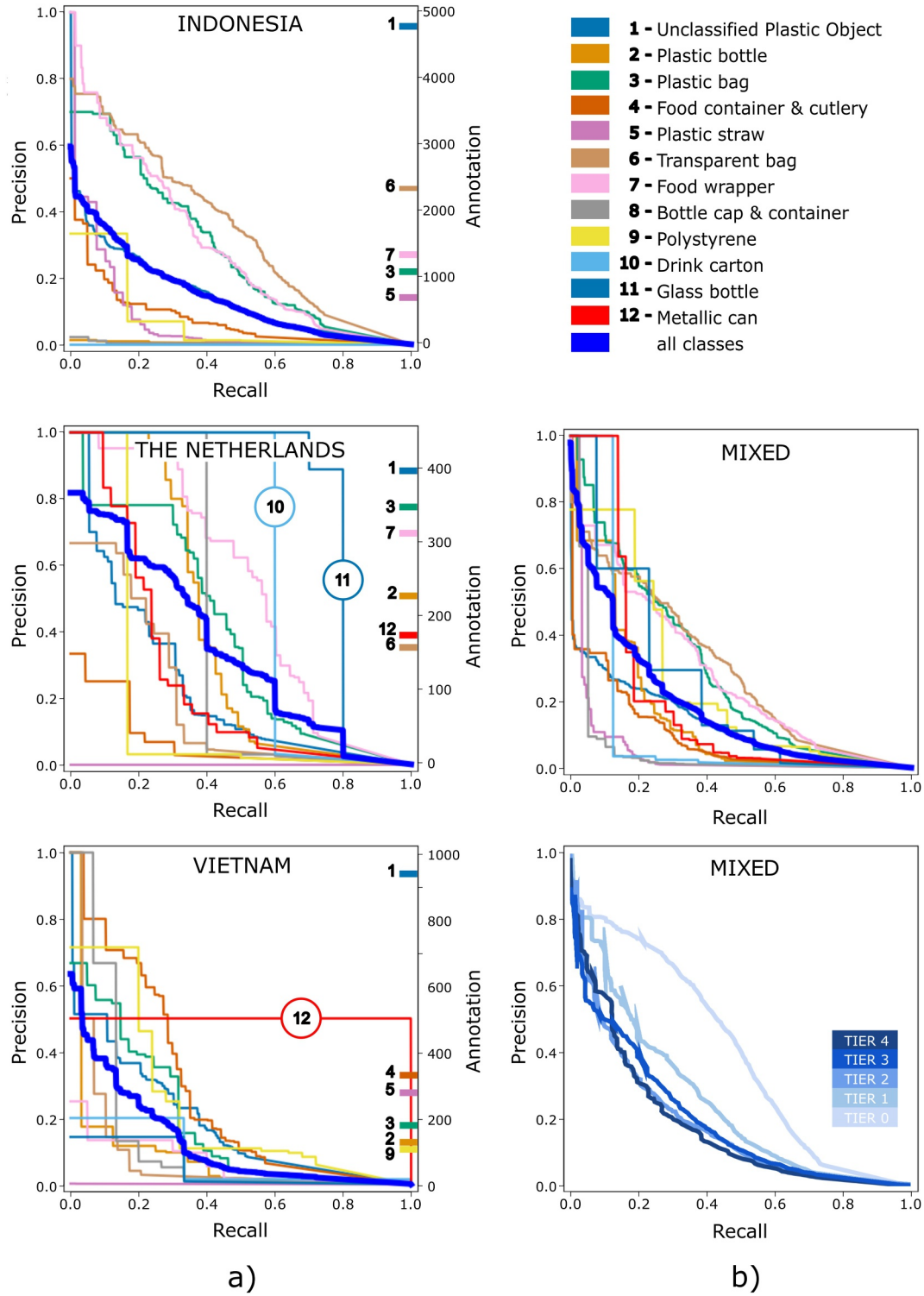


Figure 6. Precision-Recall relationship of Data set Sources with (a) standalone data sets (class-based), and (b) Mixed data set (class-based and tier-based). Classes with annotation incidence >5% of the total annotation count were reflected in the right *y*-axis. Generally, in exception of UPO, classes with high annotation (Indonesia: Transparent bag, Food wrapper, and Plastic bag; The Netherlands: Food wrapper and Plastic bag; Vietnam: Food container and cutlery) showed better precision-recall relationship than other classes. High precision but low recall trends were mostly observed in classes from The Netherlands (Bottle cap and container, Polystyrene, Drink carton, and Glass bottle). The class Metallic can for Vietnam showed a horizontal precision-recall curve which can be attributed to its very low annotation incidence (train-valid-test split: 2-1-1).

Table 4
Characteristics of the Different Performance Trends of Plastic Classes Distinguished Relating to Precision

	Indonesia		The Netherlands		Vietnam		Mixed	
Class	P	R	P	R	P	R	P	R
Unclassified plastic object	HALP	HALR	LALP	LALR	LALP	LALR	HALP	HALR
Plastic bottle	LALP	LALR	LAHP	LALR	LALP	LALR	LALP	LALR
Plastic bag	HALP	HALR	LAHP	LALR	LALP	LALR	HALP	HALR
Food container and cutlery	LALP	LALR	LALP	LALR	LALP	LALR	LALP	LALR
Plastic straw	LALP	LALR	LALP	LALR	LALP	LALR	HALP	HALR
Transparent bag	HALP	HALR	LAHP	LALR	LALP	LALR	HALP	HALR
Food wrapper	HALP	HAHR	LAHP	LAHR	LALP	LALR	HALP	HALR
Bottle cap and container	LAHP	LALR	LAHP	LALR	LALP	LALR	LALP	LALR
Polystyrene	LALP	LALR	LALP	LALR	LALP	LALR	LALP	LALR
Drink carton	LAHP	LALR	LAHP	LAHR	LALP	LALR	LALP	LALR
Glass bottle	–	–	LAHP	LAHR	LAHP	LALR	LALP	LALR
Metallic can	–	–	LAHP	LALR	LAHP	LALR	LALP	LALR

Note. LAHP- low annotation, high precision (A ≤ 1,000, P > 0.5); LALP—low annotation, low precision (A ≤ 1,000, P ≤ 0.5); HALP—high annotation, low precision (A > 1,000, P ≤ 0.5); and recall: LAHR-low annotation, high recall (A ≤ 1,000, R > 0.5); LALR—low annotation, low recall (A ≤ 1,000, R ≤ 0.5); HAHR—high annotation, high recall (A > 1,000, R ≥ 0.5), HALR—high annotation, low recall (A >1,000, R ≤ 0.5; see Figure S3 in Supporting Information S1 for the scatterplot of class-based performance trends).

and Metallic can. These classes are all geometrically regular, as they are transported along the river. The rest of the LAHP was from The Netherlands data set which include Plastic bottle, Plastic bag, Transparent bag, and Food wrapper. However, considering other data sets, some of these classes had different trends (Plastic bottle—LALP for Indonesia and Vietnam; Plastic bag—HALP for Indonesia, LALP for Vietnam; Food wrapper—HALP for Indonesia, LALP for Vietnam).

Data annotation requirements strongly vary per plastic class. Dominant classes observed in Indonesia have relatively high annotation count (A > 1,000—Plastic bag, Food wrapper, Transparent bag) than with dominant classes from Vietnam (A < 500—Food container and cutlery, Polystyrene), and from The Netherlands (A < 500—Glass bottle). These classes from stand alone data sets, both with low (A < 500) and high (A > 1,000) annotations, seemed to also influence the Mixed data set (i.e., Plastic bag, Food wrapper, Transparent bag, Glass bottle and Polystyrene). In addition, we also showed how the tiering system affected the PR relationship which shifted from concave (Tier 0, 1 class) to convex (Tier 4, 12 classes; Figure 6b) for the Mixed data set. Models with convex PR curves, despite the high precision (more true positives detection than false positives), generally have low Recall (less ground truth detection). In this context, concave PR curves are more desirable. Apart from being closer to the perfect PR curve, concave PR curves generally represent high Precision when Recall is high–which may sometimes mean less annotations present (Cook & Ramadas, 2020).

Data annotation requirements strongly vary per site. The varying performance trends of the same classes in different data sets provide an added challenge in generalizing plastic class annotation requirements. In The Netherlands, dominant classes have low annotation count. This LAHP trend may be attributed to its background diversity, which other data sets lack. On the other hand, this means that these classes for data set sources Indonesia and Vietnam may need more annotation, like other LALP classes across all data set sources (Food container and cutlery, Plastic straw, Polystyrene). As expected, the more clustered class Unclassified plastic object was tagged as HALP, which means that it may require more annotation than LALP classes.

Applying the same principle with recall, we observed four different trends and clustered into groups: LAHR—low annotation, high recall (A < 1,000, R > 0.5); LALR—low annotation, low recall (A < 1,000, R ≤ 0.5); HAHR—high annotation, high recall (A > 1,000, R > 0.5); HALR—high annotation, low recall (A > 1,000, R < 0.5). Generally, most LAHP are also LALR (i.e., Plastic bottle, Plastic bag) which was expected due to their low annotation count. At the same time, most HALP are also HALR like Unclassified plastic object, Plastic bag and

Transparent bag, except for Food wrapper (Indonesia) which is HAHR due to its high annotation count. Both The Netherlands and Vietnam had the LALP-LALR trend for Unclassified plastic object (UPO), while LALP-LALR was observed for all data set sources (excluding Mixed) for classes Food container and cutlery, Plastic straw, and Polystyrene.

In simpler terms, despite the relatively low annotation incidence across classes, the Type II data set (The Netherlands) yielded better following the preferred performance trend LAHP/LAHR (low annotation, high precision or recall) in most classes than a Type I data set (Indonesia) which mostly has HALP/HALR (high annotation, low precision or recall). However, mostly LALP/LALR trends were observed in Vietnam which is another Type I data set. It was also clear that the Mixed data set was heavily affected by the low performance of Indonesia (evident in classes UPO, Plastic bag, Transparent bag, and Food wrapper) where fraction annotation is about 67%–90%. However, this was not observed with Plastic straw despite the 68% fraction annotation with the Mixed data set. This suggests that increasing data set size by merging single site data sets, without in-depth analysis of data features and their effect on the Mixed data set performance, may degrade the data set quality instead of the desired performance improvement. Though our trend clustering splits classes which have either less than or more than 1,000 annotations, it should be emphasized that for The Netherlands data set, Food wrapper (A = 317), Drink carton (A = 15), and Glass bottle (A = 38) were all LAHP-LAHR despite the extremely low annotation count. This shows that with the proper level of environmental conditions diversity, some classes can obtain high performance despite low annotation (see Figure S2 in Supporting Information S1 for full class-based performance trends, and Figure S3 in Supporting Information S1 for class-based performance trends highlighting annotation count).

3.3. Performance of Small Data Sets

Training most deep learning foundational models on a custom data set (fine tuning) would require large data set size requirements and intensive computational resources to reach high performance (Alshalali & Josyula, 2018), which can be at par with their performance in benchmark data sets (Salari et al., 2022; Zou et al., 2023). In the case of YOLOv8, despite the relatively small data sets used in this work for training and also considering the class complexity present (n = 13), high precision and recall performance (defined previously as ≥0.5) were still observed in some cases. Presented in Figure 7 the precision and recall performance of Tier 4 models from different data sets as they progressed by epoch for YOLOv8. These curves are characterized by smaller vertical distances between each learning step which yields a smoother curve than YOLOv7 results (see Figure S4 in Supporting Information S1 for model performance results of YOLOv7). Since these curves show a steady model learning, the majority of model performance analysis carried out in the next sections are anchored on YOLOv8.

The learning curve of relatively smaller Type II data sets (e.g., The Netherlands) can potentially perform at par with larger Type I data sets (e.g., Indonesia) considering traditional performance metrics (P, R, and mAP50). This follows a similar concept presented with decision tree ensembles (Gashler et al., 2008). More specifically, the high annotation count present in Indonesia with at least one order of magnitude higher than both The Netherlands and Vietnam, did not provide a similar higher trend in performance. The continuous increase in the Recall and mAP50 performances of YOLOv8 for The Netherlands and Mixed data sets, even after epoch 200, suggests that these data sets (both type II) provide richer information for learning, than in Indonesia and Vietnam which are both Type I. Noticeably, the dropping and plateauing performance trends in Indonesia and Vietnam follows the performance saturation commonly present in site specific models (Zhu et al., 2016).

3.4. The Role of Downsizing a Mixed River Plastic Data Set

In Figure 8, we illustrated how performance is affected by the increasing data set size for sample classes with D1 (20 images) as the base data set. Looking closely into the four trends observed, recall and mAP50 trends are similar with each other than precision and recall trends. It can also be observed that a linear increase in annotation count does not guarantee a similar linear increase in performance, and can be a source of unnecessary or redundant data (e.g., plastic straw). Precision performance peaks can be observed mainly at D6 for plastic bag, D5 for plastic straw, and D7 for transparent bag, with about 23%, 2,200%, and 309% increase in precision from the base D1. While plastic bottle also peaked at D2 for precision, due to its periodic upward and downward, computing for the mean was deemed to be a better option for analysis (D1–D10) with about 466%. Contextualizing the observed

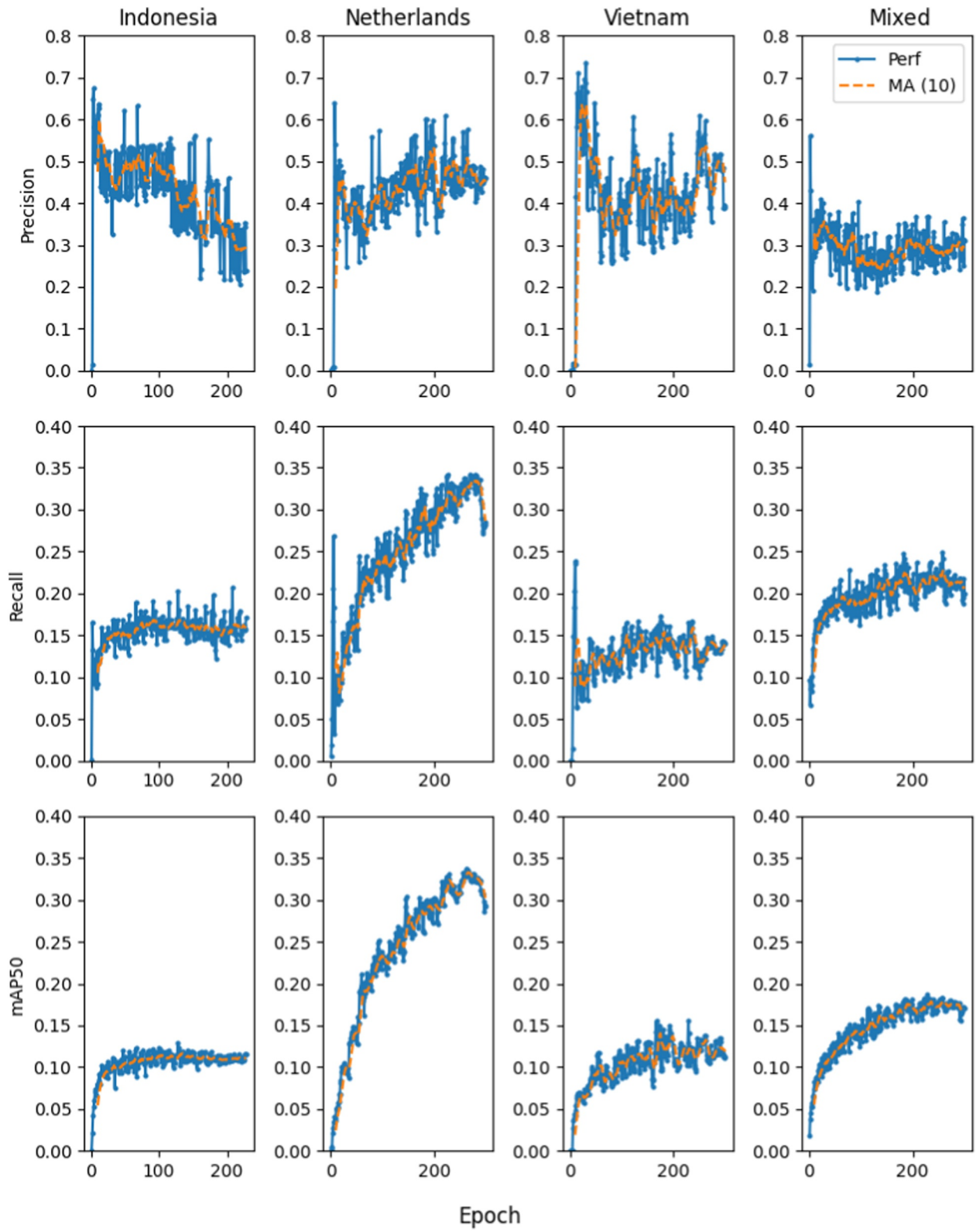


Figure 7. Model Performance using YOLOv8 across different Data set Sources (Tier 4).

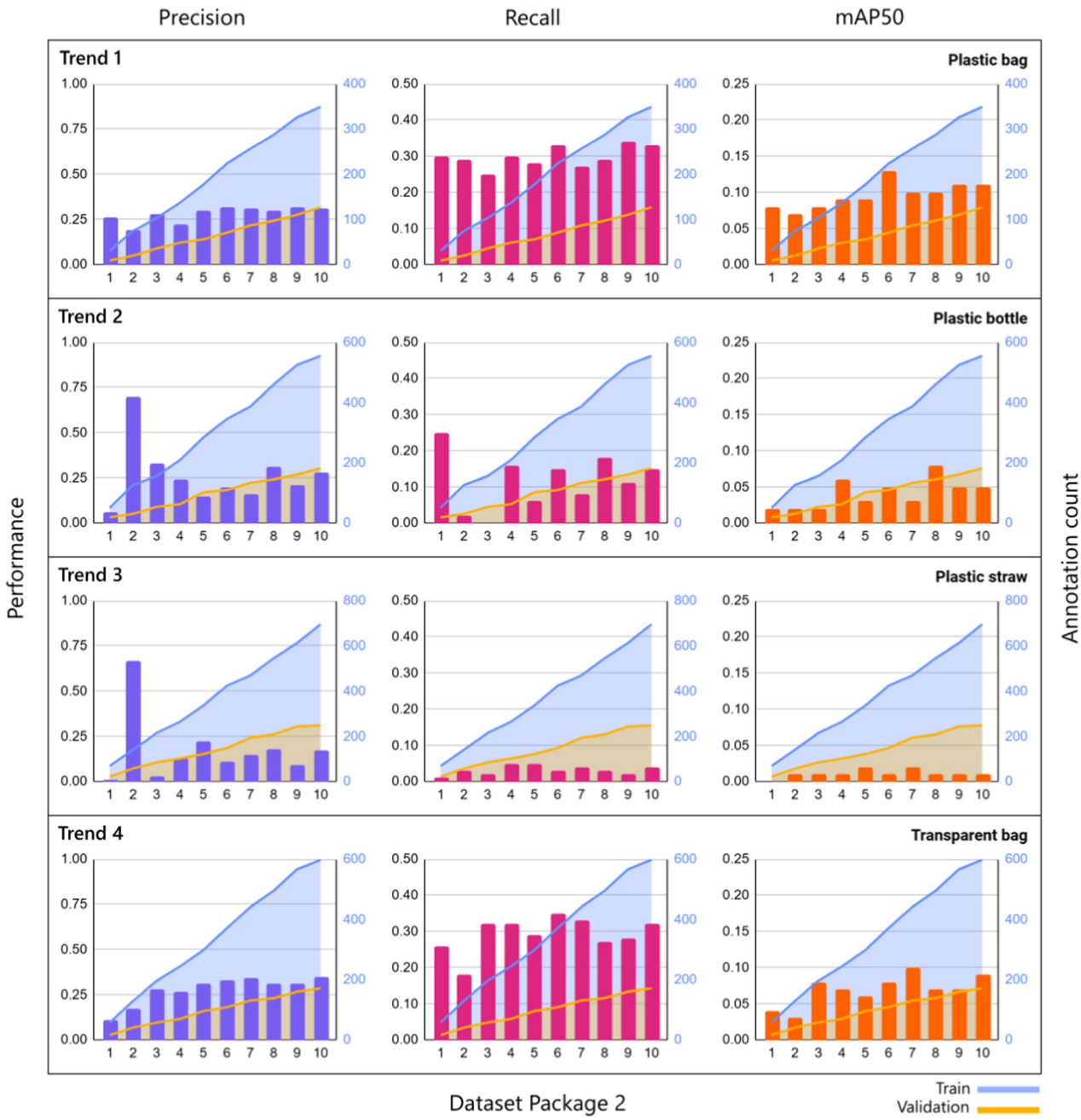


Figure 8. Performance trends with increasing Data set size using Data set Package 2—Mixed Data set (DP2-Mixed). Four performance trends were observed across all classes and were illustrated using selected classes: Trend 1—upward (Plastic bag), Trend 2—periodic upward and downward (Plastic bottle), Trend 3—horizontal (Plastic straw), and Trend 4—upward with fluctuations (Transparent bag). All classes have annotation incidence of >400 in the Train and Valid of DP2-Mixed-10 data set. DP2-Mixed-1 to DP2-Mixed-10 has an incremental data set size (20 images were added in each increment step) randomly generated from the Mixed data set for three iterations. All values were taken from the performance average (P, R, and mAP50) of the iterated data sets.

trends, the use of D7 ($n = 140$) images, will prove to be more beneficial than using D10 ($n = 200$) for the case of transparent bag.

Using a relatively small data set always has performance tradeoffs. Depending on tolerable performance requirements, by exploring the Data set Package 2 (120 data sets in total), we showed that downsizing a mixed data set is possible, and may still be at par with a relatively large one. In addition, it was clearly presented here that analyzing class annotation requirements with incremental annotation count can be beneficial to avoid unnecessary

Figure 9. Model Transferability using YOLOv8 for overall model performance with external data set (IT-Italy) and other training data sets. Note that all single site data sets (IN-Indonesia, N-The Netherlands, and V-Vietnam) are included in the model training of the Mixed (M) data set.

labor and computational effort, especially when desired performance is already achieved (see Figure S5 in Supporting Information S1 for precision percent growth for all classes).

3.5. Model Transferability

The model performance through validation (internal) is a direct approach to measure how well the model captures the knowledge it learned. However, to measure the performance of a model to capture a related knowledge, or a completely different knowledge, from the learned knowledge is yet to be established. This work explored the effectiveness of using traditional ML performance metrics (P, R, and mAP) to understand how well the internal validation performance represents the external validation performance. Figure 9 shows the overall model transferability across all tiers.

Internal validation performance doesn't equate to external validation performance. Inspecting Tier 0, the performance of Indonesia in the internal validation ($P_{IN} = 0.52$, $R_{IN} = 0.45$, $maP50_{IN} = 0.43$) was not expected to yield a higher transferability performance in the external validation ($P_{IN\text{-}IT} = 0.91$, $R_{IN\text{-}IT} = 0.17$, $maP50_{IN\text{-}IT} = 0.54$) for Precision and mAP50, but not for Recall (see Table S3 in Supporting Information S1 for the varying validation image and instance count for each data set). In fact, Indonesia yielded the highest external validation precision performance followed by Vietnam ($P_{V\text{-}IT} = 0.39$), Mixed ($P_{M\text{-}IT} = 0.39$), and The Netherlands ($P_{N\text{-}IT} = 0.38$), respectively. On the other hand, despite the highest performance of The Netherlands in the internal validation ($P_N = 0.66$, $R_N = 0.45$, $maP50_N = 0.49$), its performance in the external validation ($P_{N\text{-}IT} = 0.38$, $R_{N\text{-}IT} = 0.13$, $maP50_{N\text{-}IT} = 0.27$) was the lowest among single site data sets. Its precision performance when transferred to other single site data sets decreased for Indonesia ($P_{N\text{-}IN} = 0.41$) but increased for Vietnam ($P_{N\text{-}V} = 0.77$) while both recall and mAP50 values decreased ($R_{N\text{-}IN} = 0.05$, $R_{N\text{-}V} = 0.07$, $mAP_{N\text{-}IN} = 0.22$, and $mAP_{N\text{-}V} = 0.42$). While for Tier 4, all internal validation performance dropped when transferred to other sites. Nevertheless, it was clearly demonstrated that models utilizing a small mixed data set (i.e., The Netherlands, $n = 352$ annotations for Plastic bag class) can also be transferred to external sites.

The Mixed data set showed superiority in model transferability when classes are not aggregated. As the tiering system progressed (Tier 0–4), the saturation of the external validation performance (Italy) and cross-validation (other training data sets) highlighted the added value of using the Mixed data set which has higher heterogeneity. It should be noted that the Mixed models partially were trained with all data sets except Italy. In fact, looking at Tier 4, the validation precision performance of the Mixed model for Indonesia and The Netherlands is higher ($P_{M\text{-}IN} = 0.18$, $P_{M\text{-}N} = 0.37$, $P_{M\text{-}V} = 0.35$) when compared with all performance results based on single site data sets. This shows the added value of having a wider data set coverage which can be obtained from merging single site data sets to increase precision performance.

Model transferability performance varies depending on both the internal training source and the external site of application. Type I data sets (i.e., Indonesia) usually yield higher transferability performance in areas with similar environmental conditions (i.e., External-Italy) and perform poorly to areas which are very different (i.e., The Netherlands and Vietnam; Chui et al., 2023). Also, given the similar camera angle, and regional proximity, it was expected that Indonesia and Vietnam would perform better when transferred to each other, but it was not the case. The model from Indonesia worked better when transferred to Vietnam than in The Netherlands. On the other hand, both The Netherlands and Vietnam worked better with each other than Indonesia (see Table S4 in Supporting Information S1 for full classes validation performance). This phenomenon may be attributed to the lack of diversity in the water background (i.e., different turbidity levels, presence of vegetation or algae) in Indonesia. Thus, it was more challenging for Indonesia to be applied to Vietnam, where plastics were being transported mostly on top of floating vegetation or The Netherlands, where plastics were surrounded with algae or other shallow water vegetation (see Figure S6 in Supporting Information S1).

These findings suggest that traditional performance metrics (P, R, and mAP) from the training data set may not accurately reflect the transferability of models, hence the need to understand other indicators and metrics (i.e., diversity of background information). This work only uses one site for external validation (Italy), where the model transferability analysis was highly anchored. The model transferability method can be improved with the use of diverse data sets used in future work.

3.6. Developing a Heterogeneous River Plastic Data Set

3.6.1. Plastic Data Set Optimisation

Most work in the AI community focuses on developing the highest-performing algorithm in a test site, which usually results in overfitting and low transferability. Taking Indonesia as a use case, it was observed that there was a decrease in precision performance from the original work which uses a CNN-based algorithm ($P = 0.67$; van Lieshout et al., 2020) to our YOLO-based single class Tier 0 model ($P = 0.52$; without data augmentation and hyperparametrization). However, when we transferred our YOLOv8-Indonesia model (Tier 0) to the external (Italy) and other data sets (The Netherlands and Vietnam), we observed a varying precision performance increase and decrease ($P = 0.30$–0.91). This simple comparison using the threshold variability in the precision performance between changing the algorithm (~22%) and transferring the model to other sites (~42%–175%), shows that for this case, the data set quality weighs more than changing the algorithm.

Further considering the different approaches carried out, the precision performance in the internal validation of The Netherlands data set (Type II) yielded the highest results for both tier-based (Table 3) and class-based (Figure 6). However, when considering the external validation ($P_{N\text{-}IT}$ = 0.38, $R_{N\text{-}IT}$ = 0.13, $mAP50_{N\text{-}IT}$ = 0.27), The Netherlands yields the lowest precision and recall, and second lowest mAP. In addition, when this model (Tier 0 only) was cross-validated (external validation with other training data sets) with Vietnam, while its performance increased for precision ($P_{N\text{-}V}$ = 0.77; Figure 9), a huge drop was observed for recall ($R_{N\text{-}V}$ = 0.07), and a minimal drop for mAP50 ($mAP_{N\text{-}V}$ = 0.42; see Table S4 in Supporting Information S1). On the other hand, despite the saturated precision performance of models from the Mixed data set for both tier-based (P_M = 0.32–0.49) and class-based (P_M = 0.32) approaches when compared to those trained with single site data sets, its transferability was more consistent when tiering progressed (Tier 0–4) utilizing more classes. Factoring in the general observation that transferability of single site data sets drops as the tiering progresses, highlights that the use of Mixed data set provides a huge benefit due to its wider coverage when transferability is considered. YOLOv8 performance is overall better than YOLOv7 when considering runtime, and consistency of performance with epoch across all data sets used. This section only focused on the results of models from YOLOv8 and highlighted different factors affecting class-based, tier-based and transferability performance in river plastic detection, as well as limitations of our methods and recommendations for future work.

In terms of scalability of the single site data sets in the context of the external validation (Italy), considering computational cost, and labor cost, both Type I (Indonesia, and Vietnam) and Type II (The Netherlands) data sets were either moderately scalable or highly scalable. Considering transferability, only Indonesia yielded Scalable (Tier 0 only). Though it is not appropriate to compare the scalability of the Mixed data set with the single site data sets, for context it was found scalable in labor cost and not scalable for both transferability (despite yielding the highest transferability across all data sets for Tier 4) and computational cost (see Table S5 in Supporting Information S1).

3.6.2. Computational Cost

The computational cost of all data set sources, reflecting tiers and base model versions used is presented in Figure 10. The big disparity in the runtime (about one order of magnitude) when comparing YOLOv7 and YOLOv8 was observed. Using YOLOv8, all data set sources required similar runtime, despite having a relatively lower size in terms of both annotation and image count in The Netherlands (Type II) than Indonesia and Vietnam (Type I). The Mixed data set Tier 4 on the other hand, required about 10.6% percent more runtime than the sum of all Tier 4 single site data sets (Indonesia, The Netherlands, and Vietnam). When all tiers were summed, the Mixed data set needed about 18% more runtime than the sum of the single site data sets (considering runtime of all tiers). Further, when considering downsizing of the Mixed data set, about 73% (~6 hr) additional computation and labor cost was found between D6 and D10 (see Table S6 in Supporting Information S1). All computational costs presented are limited to training runtime using a machine with NVDIA RTX A4500 GPU.

3.6.3. Other Processing Cost

Processing existing labels into YOLO format including data cleaning also requires additional labor and computational cost. Summarized in Table 5 all costs carried out in our analysis, excluding the labor cost spent in developing the codes and minimal validation runtime. For the Indonesia data set which already has a single class annotation, other processing costs included (a) conversion of the existing labels into YOLO format, (b) adjustment of the bounding boxes (bbox), and (c) reclassification of the annotations. This cost around 0.26 s per bbox for Indonesia, which is higher than the 0.20 s per bbox from scratch (i.e., The Netherlands and Vietnam).

It was also observed that about 76% of the total computational and labor cost to generate a model using the Mixed data set (data set package 1 only) was attributed alone to Indonesia, with about 71% fraction of annotation. This follows that creating data sets with large numbers of annotations may require more computational and labor cost, but whether all these data contribute to performance increase is still yet to be determined. Also, considering the runtime trends of both single site and Mixed data sets, merging of single site data sets should be balanced with desired runtime. This is very beneficial when real time and near-real time applications (Ahmed et al., 2021) are desired for long-term monitoring implementation.

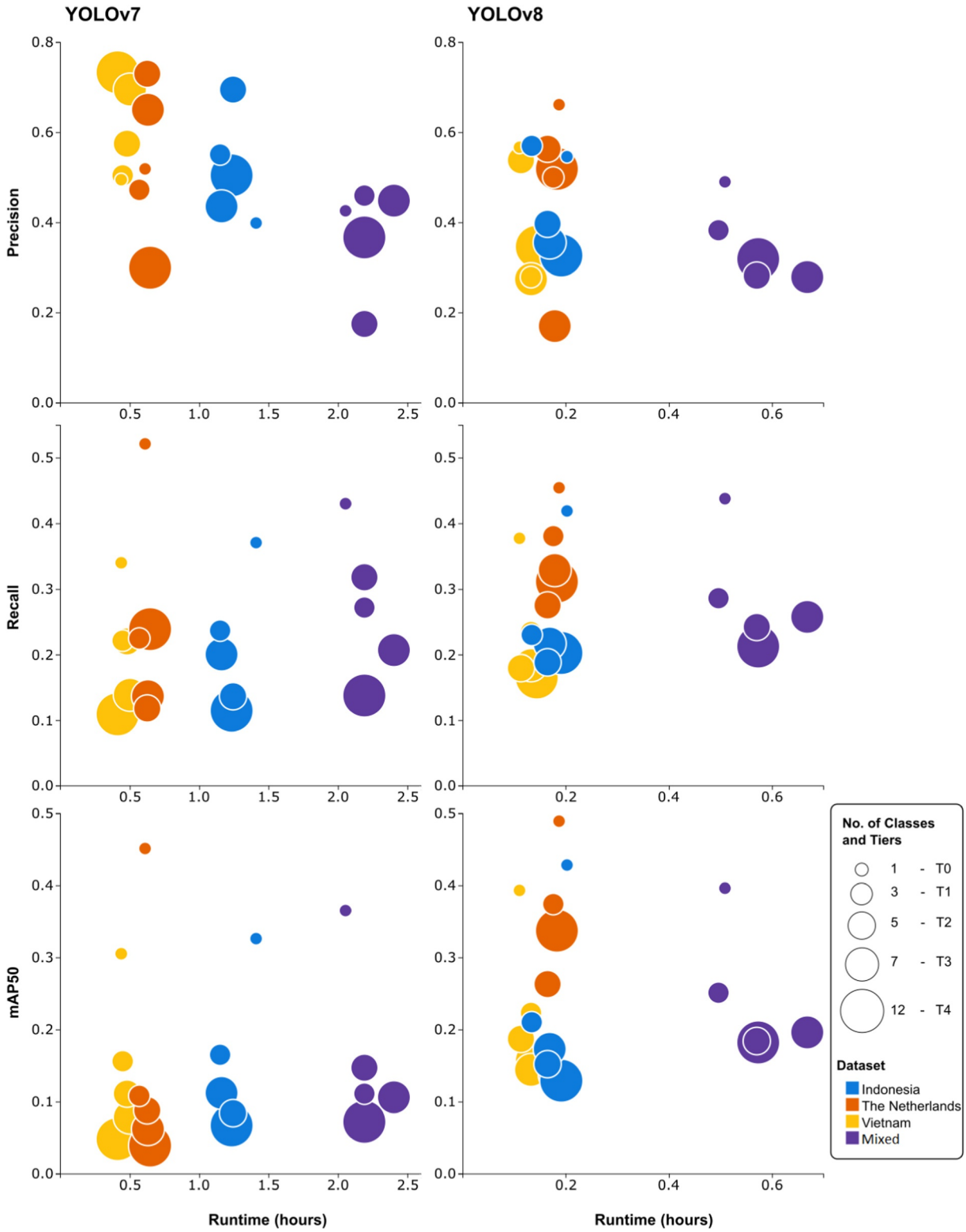


Figure 10. The computational cost in relation to different performance metrics (Precision, Recall, and mAP50), Tiers, and Data sets using YOLOv7 and YOLOv8.

Table 5
Overview of the Computational and Labor Cost (Tier 4—YOLOv8)

	Data set/Cost (in hours)			
Cost description	Indonesia	The Netherlands	Vietnam	Mixed
Annotation Count	10,921	1,890	2,401	15,212
Computational Cost				
Data set Package 1	0.192	0.183	0.144	0.574
Data set Package 2	2.214	2.426	2.243	5.927
Labor Cost				
Annotation		6.464	8.212	included
Cleaning	0.003	0.001	0.001	0.004
Other Processing Cost				
Label into YOLO format	0.003			
Bounding Box Adjustment	22.752			
Reclass Annotation	24.522			
TOTAL	75.864			

3.7. Limitations

Our methodology involved a class- and tier-based analysis of river plastic detection, utilizing both YOLOv7 and YOLOv8 base models. As this work pioneers the exploration of complex plastic classes ($n = 13$) and hierarchical class clustering (five-tier aggregation system), several limitations should be considered when replicating or extending this framework.

First, although Type II data sets—characterized by greater environmental and background diversity—consistently demonstrated superior transferability across sites, the specific drivers of this improvement were not quantitatively isolated. The present analysis focused on empirical performance comparisons rather than feature-level attribution or correlation-based investigation of data set properties (e.g., background heterogeneity metrics, annotation density per class, or intra-class visual variability). We hypothesize that increased environmental variability may act as an implicit regularization mechanism, reducing overfitting to site-specific background features. However, this interpretation remains inferential and should not be considered a formally demonstrated causal relationship.

Second, the intermediate tier configurations (Tiers 2–3) exhibited performance instability, including instances of near-zero precision. These collapses likely reflect non-optimized class aggregation strategies rather than intrinsic limitations of hierarchical class grouping. The current tiering system progressively aggregated semantic classes but was not designed according to visual similarity, material properties (e.g., rigid vs. flexible plastics), transparency, or geometric characteristics. Alternative grouping strategies based on physically or visually grounded criteria may yield more stable intermediate-tier performance. Therefore, the observed instability should be interpreted as a limitation of the aggregation design adopted here, rather than as evidence against tier-based learning in general.

Third, some plastic classes are inherently more heterogeneous than others. For example, the *Food container and cutlery* class includes both plastic utensils and food containers, resulting in broader intra-class variability compared to other categories. This structural imbalance may partially influence class-level performance variability.

Fourth, this study did not implement stepwise annotation-based data set packaging. In practice, annotation counts per image vary substantially, and threshold-based balancing strategies can be complex to implement. Instead, the analysis focused on performance growth trends relative to data set package size (Figure 8) and annotation volume (see Figure S3 in Supporting Information S1). While this provides insight into scalability behavior, more controlled annotation-level balancing could refine future analyses.

Fifth, external validation was limited. The independent Italy data set used for external validation consisted of imagery from a single day, and class diversity on that day was restricted. Although additional cross-validation

experiments using the other data sets (Figure 9) were included to provide contextual performance comparison, the temporal and hydrological representativeness of the external validation remains limited. Consequently, conclusions regarding large-scale transferability should be interpreted as preliminary. Broader validation across multiple seasons, discharge conditions, and geomorphological contexts would be required before operational deployment at regional or global scales.

Sixth, the effect of camera-to-water surface distance on detection performance was not explicitly measured. Training images varied in resolution (~1920–3000 pixels width), while camera-to-water distance variability was approximately 8 m. Prior research indicates that high-resolution imagery (2000–4000 pixels width) can support reliable YOLO-based detection up to ~40 m, with performance decreasing by approximately 25% when image resolution is reduced by half (Hao et al., 2023). Nevertheless, distance-related performance sensitivity was not systematically assessed in this study.

Seventh, annotations were limited exclusively to litter classes. Environmental disturbances such as solar glint were intentionally not labeled. In instances where such artifacts visually resemble floating plastic, the model may generate false positives. However, the magnitude of this effect was not quantitatively measured, and its influence on performance metrics remains unquantified.

Finally, the plastic data set scalability metric introduced in this work is tailored to the iterative background and packaging strategy implemented here, as well as to the projected active-learning pipeline envisioned for future deployments. Its generalizability to other detection tasks, alternative model architectures, or fundamentally different data acquisition strategies requires further validation and potential refinement.

Taken together, these limitations delineate the boundaries within which the results should be interpreted. While the findings demonstrate that strategically mixed small data sets can enhance cross-site transferability, the mechanisms underlying this behavior, the stability of intermediate-tier aggregation, and the robustness under broader environmental variability require further targeted investigation.

3.8. Final Recommendations and Consideration

There are still a lot of gaps to realize a reliable long-term monitoring system for plastic detection in rivers. Therefore, the next steps and needs are highlighted in this section with the aim to identify future directions for the present research.

Using Type II data sets with high diversity in feature information. It is counterproductive to use higher epochs with plateauing training results. For classes that are more problematic (i.e., UPO, Plastic bag, Transparent bag) than others with more defined geometry (i.e., Drink carton), the possibility to fuse the detection data or batch detection in the future is promising. When using data sets with existing annotations, depending on their quality (i.e., bbox fit, classes), annotating from scratch may prove to be more viable in terms of labor cost. Given that performance in the own data set doesn't equate to a similar transferability performance, it is crucial to establish other scalability metrics in future work, on top of the metrics we used in this work (computational cost, labor cost, and transferability). Another possible metric which can be explored is ranking the quality of input data based on features, which may or may not be known to the base data set. Parameterization of input images in future works may prove one way to avoid blindly packaging a data set (i.e., ranking feature information included in images, classification of images based on site conditions and turbidity levels). One option for feature ranking is through lighting augmentation (during training and detection) which was previously explored to increase detection performance (Saddi et al., 2024). Depending on the level of information needed, partial class aggregation may be used to yield a satisfactory increase in performance. Generally, the coarser the class information (<Tier 4), the higher the performance, except when classes have enough annotation and feature that satisfies the minimum detection requirement of a class.

Improving class imbalance by controlling the annotation incidence per class. Acknowledging that class performance of single site data sets contributes to the performance of the Mixed data set, improvements of class performance before or after merging of data sets was necessary to be explored. In this context, the following steps are suggested to better understand class performance trends in the future relating to precision and recall:

- *LAHP/LAHR*—it is essential to establish whether its peak trend was already reached by increasing annotation count. This can be verified with iterative data set packaging with incremental increase of annotation size (discussed in full detail in Section 3.4).
- *LALP/LALR*—more data is needed to understand and reach its peak trend.
- *HAHP/HAHR*—other data set sources, which have more background diversity (Type II; i.e., crowdsourced data) should be explored to reduce the annotation incidence.
- *HALP/HALR*—these classes are problematic since despite the high annotation, their performance remains low. Clustering to other classes may prove to be useful since in this way, these classes will still be included in the detection. Other remote sensing augmentation techniques (i.e., image processing) may be beneficial to increase their precision and recall performance.

Building plastic data sets using generative AI, with structured tiering systems, and using data from rivers with high turbidity levels. Recent advances in the use of generative AI presents a huge potential in building plastic data sets in the future considering the gaps in data collection, and addressing the class imbalance problem. Another way to optimize data set performance is to develop a structured tiering system which will aggregate classes to optimize the overall model performance, and potentially produce a consistent performance decay when tiering progresses from single class (Tier 0) to complete classes (Tier 4). Data fusion between Tier 0 and Tier 4 will also prove beneficial in future auto detect-to-annotation deployments. Further, most of the training and validation data are from highly turbid waters, which may have affected the detection of plastics that are partially submerged or those that reflect the water spectra (i.e., Plastic bag, Transparent bag). High turbidity in the rivers can be expected to provide a solid (almost unicolor) background that highlights the plastic features. This is similar to the effects of shadow presence, provided by buildings and riverbank structures, minimizing the lighting which a plastic object can reflect (see Figure S7 in Supporting Information S1). However, the presence of other environmental disturbances (i.e., sky-mirroring) which tend to blend more the plastic features in the water, makes the detection more complicated (see Figure S8 in Supporting Information S1). In this context, using environmental disturbances as a potential data feature (i.e., filter, preprocessing) in future development of small mixed data sets might be beneficial.

Applying XAI for object detection models. It is also important to understand how computer vision models learn features for discrimination to avoid data set overfitting (Ying, 2019) and improve generic detection capabilities, just like how humans visually learn to recognize objects. The Netherlands data set, which is about half the size of Indonesia, has more variability in the acquisition location, providing more variability in the background information (i.e., algae abundance, sky-mirroring, water color). Vietnam data set also provided an added value due to the high presence of vegetation which is being transported together with the plastics. There are also classes with relatively low annotation count but yields high PR (e.g., Plastic bottle in The Netherlands). XAI presents an opportunity to explain why these smaller data sets performed better than a much larger Indonesia data set, and potentially determine data features that allowed these small data sets to perform better than large ones (i.e., background information).

Implementing Long-term river plastic detection. The method covered in this work can be deployed to measure projected surface area and detection (class and quantity). The potential of relating the geometric relationship of projected surface area to the plastic mass (Kataoka et al., 2024), provides a huge advantage to also derive the plastic mass transport (e.g., in kg/h), a unit that can be easily associated with plastic production, and emissions (e.g., Meijer et al., 2021; L. J. Schreyers et al., 2024). This approach to track plastic transport will be an improved version of previous works using plastic tracers (e.g., Sanness Salmon et al., 2023). In this way, global efforts on plastic monitoring, including river plastic detection approaches, can be harmonized.

4. Conclusions

This study explores the transferability of feature detection algorithms for monitoring river plastic pollution, with a primary focus on data set structure and quality rather than model architecture development. Using different data sets from various rivers in Indonesia, Italy, The Netherlands, and Vietnam with diverse environmental conditions and camera systems, our work presented an in-depth investigation on how river plastic detection from local sites can be transferred to other sites using both class- and tier-based analyses.

Our findings confirm that transferability is strongly influenced by data set composition. Developing relatively small but strategically mixed datasets—combining diverse plastic classes and heterogeneous background

conditions—improves both internal consistency and external validation performance. Simply merging data sets from single sites provided both a low performance in the internal validation, yet a wider detection capability for external validation (Italy—single site, and single day data from a long-term camera monitoring system), especially when tiering progressed (>Tier 0). These findings highlight the need for *transferability-focused metrics* and *data set-merging frameworks* that will enable future studies to measure transferability performance while maintaining a small mixed data set sizes to enhance capability of global river plastic monitoring. In fact, the mechanisms driving transferability improvements remain partially critical and warrant additional investigation.

Furthermore, we also observed strong class-specific data requirements. Some classes appear clearly data-limited —e.g., Food container and cutlery, Plastic straw, Polystyrene—where low annotation counts (A < 1,000) coincide with low performance. Other classes show less predictable behavior: despite higher annotation counts (A > 1,000), additional data are still needed to characterize performance saturation and variability (e.g., UPO, Plastic bag, Transparent bag). For certain items (e.g., Plastic bottle, Transparent bag, Food wrapper), performance seems to depend strongly on environmental diversity: they perform well in The Netherlands but less consistently in Indonesia or Vietnam, suggesting that broader background and lighting variability may reduce annotation demand. Conversely, Drink carton, Glass bottle, Metallic can remain relatively robust across conditions and may require fewer annotations. These distinctions can guide future annotation strategies toward the classes where additional data yield the highest marginal gains.

In the future, we will further investigate which data features will help improve our own data set and transferability performance using XAI, and also the data fusion potential of Tier 0 and Tier 4. We also plan to explore RGB, multispectral, and hyperspectral sensing, particularly for challenging scenarios (e.g., high plastic density) where object-based detection may be unreliable. While the annotation thresholds required for long-term deployment remain uncertain, our results demonstrate that small, diverse mixed data sets can substantially advance river plastic detection, reinforcing that data set quality—potentially supported by citizen science—is central to scaling monitoring efforts globally.

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Availability Statement

Data and codes used in this research are available in these in-text data citation references: Saddi et al. (2025), (ShareAlike 4.0 International). Original source of Indonesia: cyliesho (2020), (GNU GPLv3). Original source of Vietnam: L. Schreyers et al. (2022), (ShareAlike 4.0 International).

Acknowledgments
This paper and related research work has been conducted during and with the support of the Italian national inter-university PhD course in Sustainable Development and Climate change (www.phd-sdc.it), Cycle XXXVIII, with the support of a scholarship financed by the Ministerial Decree no. 351 of 9 April 2022, based on the NRRP—funded by the European Union—NextGenerationEU—Mission 4 "Education and Research," Investment 4.1 "Extension of the number of research doctorates and innovative doctorates for public administration and cultural heritage," the projects "OurMED: Sustainable water storage and distribution in the Mediterranean," which is part of the PRIMA Programme supported by the European Union's Horizon 2020 Research and Innovation Programme under Grant 2222; and the PRIN project 2022MMBA8X—"RiverWatch: A Citizen-Science Approach to River Pollution Monitoring" (CUP: J53D23002260006) funded by the Italian Ministry of University and Research. Open access publishing facilitated by Universita degli Studi di Napoli Federico II, as part of the Wiley - CRUI-CARE agreement.

References

Adnan, M. M., Rahim, M. S. M., Khan, A. R., Alkhayyat, A., Alamri, F. S., Saba, T., & Bahaj, S. A. (2023). Automated image annotation with novel features based on deep ResNet50-SLT. *IEEE Access*, *11*, 40258–40277. https://doi.org/10.1109/access.2023.3266296

Ahmed, M., Hashmi, K. A., Pagani, A., Liwicki, M., Stricker, D., & Afzal, M. Z. (2021). Survey and performance analysis of deep learning based object detection in challenging environments. *Sensors*, *21*(15), 5116. https://doi.org/10.3390/s21155116

Al-Jarrah, O. Y., Yoo, P. D., Muhaidat, S., Karagiannidis, G. K., & Taha, K. (2015). Efficient machine learning for big data: A review. *Big Data Research*, *2*(3), 87–93. https://doi.org/10.1016/j.bdr.2015.04.001

Alshalali, T., & Josyula, D. (2018). Fine-tuning of pre-trained deep learning models with extreme learning machine. In *2018 international conference on computational science and computational intelligence (CSCI)* (pp. 469–473). IEEE.

Armitage, S., Awty-Carroll, K., Clewley, D., & Martinez-Vicente, V. (2022). Detection and classification of floating plastic litter using a vessel-mounted video camera and deep learning. *Remote Sensing*, *14*(14), 3425. https://doi.org/10.3390/rs14143425

Braioni, M. G., Villani, M. C., Braioni, A., Salmoiraghi, G., Locascio, A., & Cannata, G. (2009). The restoration of some stretches of the Sarno River (Southern Italy): From canalized environment to fluvial corridor. In *River basin management V. International conference on river basin management* (pp. 293–304). Wessex Institute of Technology: Ashurst.

Chui, K. T., Arya, V., Band, S. S., Alhalabi, M., Liu, R. W., & Chi, H. R. (2023). Facilitating innovation and knowledge transfer between homogeneous and heterogeneous datasets: Generic incremental transfer learning approach and multidisciplinary studies. *Journal of Innovation & Knowledge*, *8*(2), 100313. https://doi.org/10.1016/j.jik.2023.100313

Cook, J., & Ramadas, V. (2020). When to consult precision-recall curves. *STATA Journal*, *20*(1), 131–148. https://doi.org/10.1177/1536867x20909693

Cortesi, I., Masiero, A., De Giglio, M., Tucci, G., & Dubbini, M. (2021). Random forest-based river plastic detection with a handheld multispectral camera. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *43*, 9–14. https://doi.org/10.5194/isprs-archives-xliii-b1-2021-9-2021

CrowdWater. (2026). CrowdWater. Retrieved from https://crowdwater.ch/en/start/

cyliesho. (2020). *colinvanlieshout/riverplasticdetection: Automated river plastic monitoring using deep learning and cameras (version 3)*. Zenodo. https://doi.org/10.5281/zenodo.3817117

Day, O., & Khoshgoftaar, T. M. (2017). A survey on heterogeneous transfer learning. *Journal of Big Data*, *4*, 1–42. https://doi.org/10.1186/s40537-017-0089-0

de Carvalho, A. R., Garcia, F., Riem-Galliano, L., Tudesque, L., Albignac, M., Ter Halle, A., & Cucherousset, J. (2021). Urbanization and hydrological conditions drive the spatial and temporal variability of microplastic pollution in the Garonne River. *Science of the Total Environment*, *769*, 144479. https://doi.org/10.1016/j.scitotenv.2020.144479

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248–255). IEEE.

Djolonga, J., Yung, J., Tschannen, M., Romijnders, R., Beyer, L., Kolesnikov, A., & Lucic, M. (2021). On robustness and transferability of convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 16458–16468).

Dwyer, B., Nelson, J., Hansen, T., et al. (2026). Roboflow (version 1.0) [Software]. https://roboflow.com.Computervision

EEA. (2024). *Surface water ecological status map*. RBD, European Environment Agency (EEA). Retrieved from https://water.europa.eu/freshwater/europe-freshwater/maps/ecological-status-map-by-rbd?activeTab=d6af4137-9d5b-4f92-9387-3d9ece248634

European Union. (2024). *Copernicus Land Monitoring Service 2018*. European Environment Agency (EEA).

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, *88*(2), 303–338. https://doi.org/10.1007/s11263-009-0275-4

Fallati, L., Polidori, A., Salvatore, C., Saponari, L., Savini, A., & Galli, P. (2019). Anthropogenic Marine Debris assessment with Unmanned Aerial Vehicle imagery and deep learning: A case study along the beaches of the Republic of Maldives. *Science of the Total Environment*, *693*, 133581. https://doi.org/10.1016/j.scitotenv.2019.133581

Feng, C., Zhong, Y., Jie, Z., Chu, X., Ren, H., Wei, X., & Ma, L. (2022). Promptdet: Towards open-vocabulary detection using uncurated images. In *European conference on computer vision* (pp. 701–717). Springer Nature Switzerland.

Fouquet, L., Maggio, S., & Dreyfus-Schmidt, L. (2023). Transferability metrics for object detection. *arXiv preprint arXiv:2306.15306*.

Fulton, M., Hong, J., Islam, M. J., & Sattar, J. (2019). Robotic detection of marine litter using deep visual detection models. In *2019 International conference on robotics and automation (ICRA)* (pp. 5752–5758). IEEE.

Galgani, F., Ruiz, O. S. P. L., Ronchi, F., Tallec, K., Fischer, E., Matiddi, M., & Hanke, G. (2013). Guidance on the monitoring of marine litter in European seas.

Gashler, M., Giraud-Carrier, C., & Martinez, T. (2008). Decision tree ensemble: Small heterogeneous is better than large homogeneous. In *2008 seventh international conference on machine learning and applications* (pp. 900–905). IEEE.

Gnann, N., Baschek, B., & Ternes, T. A. (2022). Close-range remote sensing-based detection and identification of macroplastics on water assisted by artificial intelligence: A review. *Water Research*, *222*, 118902. https://doi.org/10.1016/j.watres.2022.118902

Gonçalves, G., Andriolo, U., Pinto, L., & Duarte, D. (2020). Mapping marine litter with Unmanned Aerial Systems: A showcase comparison among manual image screening and machine learning techniques. *Marine Pollution Bulletin*, *155*, 111158. https://doi.org/10.1016/j.marpolbul.2020.111158

Hao, Y., Pei, H., Lyu, Y., Yuan, Z., Rizzo, J. R., Wang, Y., & Fang, Y. (2023). Understanding the impact of image quality and distance of objects to object detection performance. In *2023 IEEE/RSJ international conference on Intelligent Robots and Systems (IROS)* (pp. 11436–11442). IEEE.

Huang, K., Lei, H., Jiao, Z., & Zhong, Z. (2021). Recycling waste classification using vision transformer on portable device. *Sustainability*, *13*(21), 11572. https://doi.org/10.3390/su132111572

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis*, *6*(5), 429–449. https://doi.org/10.3233/ida-2002-6504

Jia, T., de Vries, R., Kapelan, Z., Van Emmerik, T. H., & Taormina, R. (2024). Detecting floating litter in freshwater bodies with semi-supervised deep learning. *Water Research*, *266*, 122405. https://doi.org/10.1016/j.watres.2024.122405

Jia, T., Kapelan, Z., de Vries, R., Vriend, P., Peereboom, E. C., Okkerman, I., & Taormina, R. (2023). Deep learning for detecting macroplastic litter in water bodies: A review. *Water Research*, *231*, 119632. https://doi.org/10.1016/j.watres.2023.119632

Jocher, G., Chaurasia, A., & Qiu, J. (2023). Ultralytics YOLOv8 (version 8.0.0) [Computer software]. Retrieved from https://github.com/ultralytics/ultralytics

Kachouri, R., Djemal, K., & Maaref, H. (2010). Adaptive feature selection for heterogeneous image databases. In *2010 2nd international conference on image processing theory, tools and applications* (pp. 26–31). IEEE.

Kataoka, T., Iga, Y., Baihaqi, R. A., Hadiyanto, H., & Nihei, Y. (2024). Geometric relationship between the projected surface area and mass of a plastic particle. *Water Research*, *261*, 122061. https://doi.org/10.1016/j.watres.2024.122061

Khan, S., Yin, P., Guo, Y., Asim, M., & Abd El-Latif, A. A. (2024). Heterogeneous transfer learning: Recent developments, applications, and challenges. *Multimedia Tools and Applications*, *27*, 1–37. https://doi.org/10.1007/s11042-024-18352-3

Krawczyk, B. (2016). Learning from imbalanced data: Open challenges and future directions. *Progress in artificial intelligence*, *5*(4), 221–232. https://doi.org/10.1007/s13748-016-0094-0

Kuwata, K., & Shibasaki, R. (2015). Estimating crop yields with deep learning and remotely sensed data. In *2015 IEEE international geoscience and remote sensing symposium (IGARSS)* (pp. 858–861). IEEE.

Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., et al. (2020). The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, *128*(7), 1956–1981. https://doi.org/10.1007/s11263-020-01316-z

Lebreton, L., & Andrady, A. (2019). Future scenarios of global plastic waste generation and disposal. *Palgrave Communications*, *5*(1), 6. https://doi.org/10.1057/s41599-018-0212-7

Li, G., Li, X., Wang, Y., Wu, Y., Liang, D., & Zhang, S. (2022). Pseco: Pseudo labeling and consistency training for semi-supervised object detection. In *In European conference on computer vision* (pp. 457–472). Springer Nature Switzerland.

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13* (pp. 740–755). Springer International Publishing.

Lin, W. (2021). Yolo-green: A real-time classification and object detection model optimized for waste management. In *2021 IEEE international conference on big data* (pp. 51–57). IEEE.

Lusher, A. L., & Primpke, S. (2023). Finding the balance between research and monitoring: When are methods good enough to understand plastic pollution? *Environmental Science & Technology*, *57*(15), 6033–6039. https://doi.org/10.1021/acs.est.2c06018

Maharjan, N., Miyazaki, H., Pati, B. M., Dailey, M. N., Shrestha, S., & Nakamura, T. (2022). Detection of river plastic using UAV sensor data and deep learning. *Remote Sensing*, *14*(13), 3049. https://doi.org/10.3390/rs14133049

Manfreda, S., Miglino, D., Saddi, K. C., Jomaa, S., Etner, A., Perks, M., et al. (2024). Advancing river monitoring using image-based techniques: Challenges and opportunities. *Hydrological Sciences Journal*, *69*(6), 657–677. https://doi.org/10.1080/02626667.2024.2333846

Marye, A. T., Caramiello, C., De Nardi, D., Miglino, D., Proietti, G., Saddi, K. C., et al. (2025). Remote sensing for monitoring macroplastics in rivers: A review. *Wiley Interdisciplinary Reviews. Water*, *12*(2), e70020. https://doi.org/10.1002/wat2.70020

Meijer, L. J., van Emmerik, T. H. M., van Der Ent, R., Schmidt, C., & Lebreton, L. (2021). More than 1000 rivers account for 80% of global riverine plastic emissions into the ocean. *Science Advances*, *7*(18), eaaz5803. https://doi.org/10.1126/sciadv.aaz5803

Miao, J., & Zhu, W. (2022). Precision–recall curve (PRC) classification trees. *Evolutionary intelligence*, *15*(3), 1545–1569. https://doi.org/10.1007/s12065-021-00565-2

Moore, C. J., Moore, S. L., Leecaster, M. K., & Weisberg, S. B. (2001). A comparison of plastic and plankton in the North Pacific central gyre. *Marine Pollution Bulletin*, *42*(12), 1297–1300. https://doi.org/10.1016/s0025-326x(01)00114-x

Morales-Caselles, C., Viejo, J., Martí, E., González-Fernández, D., Pragnell-Raasch, H., González-Gordillo, J. I., et al. (2021). An inshore–offshore sorting system revealed from global classification of ocean litter. *Nature Sustainability*, *4*(6), 484–493. https://doi.org/10.1038/s41893-021-00720-8

OECD. (2024). Policy scenarios for eliminating plastic pollution by 2040. Retrieved from https://www.oecd.org/en/publications/policy-scenarios-for-eliminating-plastic-pollution-by-2040_76400890-en.html#:~:text=Global%20production%20and%20use%20of,further%20increase%20in%20OECD%20countries

Oksuz, K., Cam, B. C., Akbas, E., & Kalkan, S. (2018). Localization recall precision (LRP): A new performance metric for object detection. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 504–519).

Oksuz, K., Cam, B. C., Kalkan, S., & Akbas, E. (2020). Imbalance problems in object detection: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *43*(10), 3388–3415. https://doi.org/10.1109/tpami.2020.2981890

Padilla, R., Netto, S. L., & Da Silva, E. A. (2020). A survey on performance metrics for object-detection algorithms. In *2020 international conference on systems, signals and image processing (IWSSIP)* (pp. 237–242). IEEE.

Pérez, R., Schubert, F., Rasshofer, R., & Biebl, E. (2019). Deep learning radar object detection and classification for urban automotive scenarios. In *2019 Kleinheubach conference* (pp. 1–4). IEEE.

Pérez-García, Á., van Emmerik, T. H., Mata, A., Tasseron, P. F., & López, J. F. (2024). Efficient plastic detection in coastal areas with selected spectral bands. *Marine Pollution Bulletin*, *207*, 116914. https://doi.org/10.1016/j.marpolbul.2024.116914

Pham, T. V., & Smeulders, A. W. (2005). Object recognition with uncertain geometry and uncertain part detection. *Computer Vision and Image Understanding*, *99*(2), 241–258. https://doi.org/10.1016/j.cviu.2005.01.006

Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

Ryan, P. G. (2015). A brief history of marine litter research. In *Marine anthropogenic litter* (pp. 1–25). Springer International Publishing. https://doi.org/10.1007/978-3-319-16510-3_1

Saddi, K. C., Miglino, D., Isgro, F., van Emmerik, T. H. M., & Manfreda, S. (2024). Balancing accuracy and efficiency for river plastic monitoring. In *IGARSS 2024-2024 IEEE international geoscience and remote sensing symposium* (pp. 3836–3840). IEEE.

Saddi, K. C., van Emmerik, T. H. M., Miglino, D., Poggi, M., Isgrò, F., Tasseron, P., et al. (2025). River plastic dataset. *Github*. Retrieved from https://github.com/khmSDDi/RiverPlasticDataset/

Saini, M., & Susan, S. (2023). Tackling class imbalance in computer vision: A contemporary review. *Artificial Intelligence Review*, *56*(Suppl 1), 1279–1335. https://doi.org/10.1007/s10462-023-10557-6

Salari, A., Djavadifar, A., Liu, X., & Najjaran, H. (2022). Object recognition datasets and challenges: A review. *Neurocomputing*, *495*, 129–152. https://doi.org/10.1016/j.neucom.2022.01.022

Sánchez, F. L., Hupont, I., Tabik, S., & Herrera, F. (2020). Revisiting crowd behaviour analysis through deep learning: Taxonomy, anomaly detection, crowd emotions, datasets, opportunities and prospects. *Information Fusion*, *64*, 318–335. https://doi.org/10.1016/j.inffus.2020.07.008

Sanness Salmon, H. R., Baker, L. J., Kozarek, J. L., & Coletti, F. (2023). Effect of shape and size on the transport of floating particles on the free surface in a natural stream. *Water Resources Research*, *59*(10), e2023WR035716. https://doi.org/10.1029/2023wr035716

Savoca, M. S., Kühn, S., Sun, C., Avery-Gomm, S., Choy, C. A., Dudas, S., et al. (2022). Towards a North Pacific Ocean long-term monitoring program for plastic pollution: A review and recommendations for plastic ingestion bioindicators. *Environmental Pollution*, *310*, 119861. https://doi.org/10.1016/j.envpol.2022.119861

Schreyers, L., Bui, K., & van Emmerik, T. (2022). Drone images over the Saigon river—Hyacinth & Plastic patches. https://doi.org/10.4121/21648152.v1

Schreyers, L. J., van Emmerik, T. H., Huthoff, F., Collas, F. P., Wegman, C., Vriend, P., et al. (2024). River plastic transport and storage budget. *Water Research*, *259*, 121786. https://doi.org/10.1016/j.watres.2024.121786

Schwarz, A. E., Ligthart, T. N., Boukris, E., & Van Harmelen, T. (2019). Sources, transport, and accumulation of different types of plastic litter in aquatic environments: A review study. *Marine Pollution Bulletin*, *143*, 92–100. https://doi.org/10.1016/j.marpolbul.2019.04.029

Sohn, K., Zhang, Z., Li, C. L., Zhang, H., Lee, C. Y., & Pfister, T. (2020). A simple semi-supervised learning framework for object detection. *arXiv preprint arXiv:2005.04757*.

Tang, P., Ramaiah, C., Wang, Y., Xu, R., & Xiong, C. (2021). Proposal learning for semi-supervised object detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 2291–2301).

Thiombane, M., Martín-Fernández, J. A., Albanese, S., Lima, A., Doherty, A., & De Vivo, B. (2018). Exploratory analysis of multi-element geochemical patterns in soil from the Sarno River Basin (Campania region, southern Italy) through compositional data analysis (CODA). *Journal of Geochemical Exploration*, *195*, 110–120. https://doi.org/10.1016/j.gexplo.2018.03.010

Thomson, T. J., & Uddin, S. (2023). Contemporary ways of seeing: Exploring how smartphone cameras shape visual culture and literacy. *Journal of Visual Literacy*, *42*(4), 269–286. https://doi.org/10.1080/1051144x.2023.2281163

Thrun, S., & Pratt, L. (1998). Learning to learn: Introduction and overview. In *Learning to learn* (pp. 3–17). Springer US.

UNEP. (2024). UNEP plastics initiative. Retrieved from https://wedocs.unep.org/bitstream/handle/20.500.11822/44503/One_plastics_initiative.pdf?sequence=3https://wedocs.unep.org/bitstream/handle/20.500.11822/44503/One_plastics_initiative.pdf?sequence=3https://wedocs.unep.org/bitstream/handle/20.500.11822/44503/One_plastics_initiative.pdf?sequence=3

van Emmerik, T., Mellink, Y., Hauk, R., Waldschläger, K., & Schreyers, L. (2022). Rivers as plastic reservoirs. *Frontiers in Water*, *3*, 786936. https://doi.org/10.3389/frwa.2021.786936

van Emmerik, T., Strady, E., Kieu-Le, T. C., Nguyen, L., & Gratiot, N. (2019). Seasonality of riverine macroplastic transport. *Scientific Reports*, *9*(1), 13549. https://doi.org/10.1038/s41598-019-50096-1

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*(8), e2019EA000960. https://doi.org/10.1029/2019ea000960

Veettil, B. K., Quan, N. H., Hauser, L. T., Van, D. D., & Quang, N. X. (2022). Coastal and marine plastic litter monitoring using remote sensing: A review. *Estuarine, Coastal and Shelf Science*, *279*, 108160. https://doi.org/10.1016/j.ecss.2022.108160

Wang, C. Y., Bochkovskiy, A., & Liao, H. Y. M. (2023). YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 7464–7475).

Wenneker, B., & Oosterbaan, L. (2010). *Guideline for monitoring marine litter on the beaches in the OSPAR maritime area* (1st edn.). OSPAR Commission.

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. https://doi.org/10.1088/1748-9326/abbd01

Ying, X. (2019). An overview of overfitting and its solutions. In *Journal of Physics: Conference Series* (Vol. 1168, p. 022022). https://doi.org/10.1088/1742-6596/1168/2/022022, IOP Publishing.

Zhang, R., Zhang, W., Fang, R., Gao, P., Li, K., Dai, J., & Li, H. (2022). Tip-adapter: Training-free adaption of clip for few-shot classification. In *European conference on computer vision* (pp. 493–510). Springer Nature Switzerland.

Zhu, X., Vondrick, C., Fowlkes, C. C., & Ramanan, D. (2016). Do we need more training data? *International Journal of Computer Vision*, *119*(1), 76–92. https://doi.org/10.1007/s11263-015-0812-2

Zou, Z., Chen, K., Shi, Z., Guo, Y., & Ye, J. (2023). Object detection in 20 years: A survey. *Proceedings of the IEEE*, *111*(3), 257–276. https://doi.org/10.1109/jproc.2023.3238524