MSc Thesis - Environmental Engineering

Deep Learning based Mass-Flux Estimation of floating (plastic) litter in waterways

F. Boccacci

TU Delft

NORIA

Deep Learning based Mass-Flux Estimation of floating (plastic) litter in waterways

by

F. Boccacci

to obtain the degree of Master of Science

at the Delft University of Technology,

to be defended publicly on Monday December 15, 2015 at 11:00 AM.

Student number: 6066836

Project duration: May 15, 2025 – December 15, 2025

Thesis committee:

Dr. R. Taormina,	TU Delft, Chair
Dr. M. Hrachowitz,	TU Delft, Supervisor
J. van Wijk,	Noria - Sustainable Innovators (Advisor)

TU Delft

Acknowledgment

This thesis represents a challenging, but very rewarding final milestone of my academic journey at TU Delft. Over the past seven months I have trained, what felt like an endless amount of computer vision models, I have counted more trash than I really wanted to and spent many late nights bringing it all together in this document. During this time I have felt doubt, excitement, exhaustion, curiosity and accomplishment. But most importantly I felt gratitude towards everyone that helped me along this way.

First and foremost I would like to thank my supervisors Riccardo Taormina, Markus Hrachowitz and Jur van Wijk, who provided me all with support, guidance and encouragement.

Jur, you were the first person I met and talked to at Noria. Even though you acted as my supervisor at Noria, our relationship felt more like a partnership to me. I'm thankful for all the time you dedicated to this project and the almost weekly feedback you gave me. I was always anticipating which new Jur I would get this week. This is something I will certainly miss!

Riccardo, I thank you for taking me on as a thesis student even though you probably had one too many other commitments already. I'm grateful for your passion and knowledge about AI and Deep Learning, even though I left some meetings with more confusion than clarity. I appreciate your ideas, enthusiasm and ultimately the freedom you gave me, which allowed me to write the thesis I wanted. I also thank you for your honesty when I needed to hear it the most.

Markus, thank you for getting on board for this project without any hesitation. I value that you held me accountable to a certain academic standard and gave me many ideas, which definitely improved my analysis and critical thinking. Thank you also for giving me the chance and trust to present this research in the Colloquium.

I also would like to thank the Noria family for their support and trust. There were never questions, when I asked for certain equipment, tools or workshop space for my research. The focus was always on how to make it happen, which I am very thankful for. I appreciate all the discussions we had on and off the project. Thank you Arnoud and Rinze for giving me the chance join such a tight team, thank you Parshva for getting the GPS trackers and batteries ready, thank you Adriaan for helping with the camera mounts, thank you Joris for getting in touch with the contractors in Den Helder and thank you Fokke for exchanging many ideas about the fieldwork.

I would also like to express my gratitude to all my amazing volunteers that helped me out during fieldwork experiments and waste processing. I would like to thank Sean Paul Scott for joining me on multiple fieldwork days as well as in the workshop, I could rely on your more than once! We mounted cameras as well as sorted and classified a lot of litter. I'm happy that you are so passionate about waste, which made you provide your costly time during your own research. Thank you Emil Sigmann Engh for being a reliable boat captain during my fieldwork experiments. I also thank Hrishyank Shetty for helping out on fieldwork days. The classification and quantification of the large amounts of trash would not have been possible without the help of David Jacob and Richard Teschl, thank you both!

Lastly, I would like to thank my family, my parents and my brother. Your support allowed me to fully focus on this research in the past months and you allowed me to chase any ambition I had during my studies. I hope I made you proud!

This topic is very important to me, which is why I can proudly say that I gave it my all. People say that you will lose blood, sweat and tears during your thesis. Well, I can say they were right! While I haven't cried during this research, maybe I will get emotional when I think back to this time and my studies in general. I hope you enjoy reading this report!

Abstract

Plastic debris in aquatic environments has become an increasing concern for wildlife and human health. Rivers and canals serve as major sources of plastic transport to the oceans, yet global estimates of plastic flux remain highly uncertain, partly due to non-uniform and sparse field measurements. This poses a significant challenge for effective mitigation strategies, as the contribution of individual streams can vary by multiple orders of magnitude.

This thesis develops and tests deep-learning models that use images from fixed cameras to detect and classify floating litter based on material types. Following these detections, a conversion into a mass estimate is made based on the size of the litter object. A combination of real-world datasets and synthetically generated images was used to train a YOLOv11 architecture for a single-class task and two multi-class setups. The performance and generalization capability was evaluated on in-domain test sets and out-of-domain locations, while mass estimates were validated against physical sampled litter at two locations.

The results indicate robust performance for single-class models, while the introduction of material classes caused a performance decrease of 12% in mAP@50-95, due to the visual ambiguity of heterogeneous plastic types. The integration of synthetic data improved generalization to unseen locations through higher recall, albeit at the cost of higher false positive rates. Field validation showed that mass estimates based on traditional methods, such as human visual counting, can cause an uncertainty of up to an order of magnitude. Object detection models tended to underestimate the total mass due to heavy outliers in the ground-truth and a detection bias towards plastic categories. However, the estimate at the urban site was within 20% of the recovered plastic load. Overall, results indicate that deep learning can provide conservative and reviewable estimates of plastic mass flux from camera data. Detailed material classification is feasible for visually distinct categories, but remains data-limited for more heterogeneous materials.

List of Abbreviations

AI	Artificial Intelligence
AP	Average Precision
AUC	Area Under the Curve (of the precision–recall or ROC curve)
CNN	Convolutional Neural Network
COCO	Common Objects in Context dataset
DETR	DEtection TRansformer object detection architecture
DL	Deep Learning
EPS	Expanded Polystyrene
F1	F1-score (harmonic mean of precision and recall)
FN	False Negative
FP	False Positive
GPS	Global Positioning System
GPU	Graphics Processing Unit
IMR	IJssel–Meuse–Rhine riverbank dataset
IoU	Intersection over Union
MAE	Mean Absolute Error
mAP	mean Average Precision
ML	Machine Learning
MLP	Multilayer Perceptron
OD	Object Detection
OOD	Out-of-Domain
OSPAR	Oslo–Paris convention for marine litter
PET	Polyethylene Terephthalate
PO-hard	Hard polyolefin plastic
PO-soft	Soft (flexible) polyolefin plastic
PR	Precision–Recall
PS	Polystyrene
R-CNN	Regions with Convolutional Neural Networks object detector (Faster R-CNN)
ReLU	Rectified Linear Unit activation function
RoIs	Regions of Interest
RPD	RiverPlasticDataset
RPN	Region Proposal Network
RT-DETR	Real-Time DEtection TRansformer
SQ	Sub-question
SSL	Semi-Supervised Learning
TP	True Positive
TU Delft	Delft University of Technology
TUD-GV	TU Delft Green Village dataset
TUD-V	TU Delft Vietnam dataset
WUR	Wageningen University and Research
YOLO	You Only Look Once object detection architecture

Contents

1 Introduction

1.1. Monitoring Riverine Plastic Pollution

Plastic pollution is one of the most pressing global environmental challenges, particularly in aquatic ecosystems such as rivers and oceans (Rochman et al., 2013). Annually, more than 450 million tonnes of plastic are produced (Ritchie et al., 2023), while estimated emissions of plastic waste from land-based sources into marine environments range from 4.8 to 12.7 million tonnes annually (Jambeck et al., 2015).

Rivers and canals serve as major pathways for transporting plastic waste to the oceans, contributing to the accumulation of plastic debris in coastal regions and open water (Meijer et al., 2021). Around 80 percent of ocean plastic comes from land sources, making intervention necessary to effectively mitigate this problem (Jambeck et al., 2015). Due to the persistence of plastics in the environment, health concerns arise for wildlife and humans. Fragmentation occurs due to weathering effects, chemical degradation, wave mechanics as well as interaction with marine animals, which cause plastics to split into smaller particles (Goldstein and Goodwin, 2013). These processes make it difficult to remove and trace its origin. To address this issue, there is an increasing focus on technologies that allow real-time detection and data collection on riverine plastic litter, focusing on identifying and quantifying this waste, allowing proper mitigation strategies. Currently, in situ measurements include observation-based methods such as human visual counting or interception-based sampling with drifting nets or existing floating surface booms (Gasperi et al., 2014). Although these methods allow site-specific measurements, monitoring longer time periods is labor intensive and is often constrained by extensive requirements for sampling equipment (Van Emmerik and Schwarz, 2020).

These quantification methods are often used in modeling approaches, which can range from simple temporal extrapolation to numerical models, where in situ assessments are combined with secondary data such as population density and geographical and hydrological parameters (Lebreton et al., 2017; Schmidt et al., 2017). However, a limiting factor that remains is the reliance on a small number of isolated short-time measurements, taken at a single location along the length of the river (Roebroek et al., 2022). Therefore, observations and models of riverine plastic fluxes are subjected to high uncertainties, especially in global annual estimations. While global estimates of river-to-ocean fluxes differ significantly, individual contributions of smaller rivers vary by up to five orders of magnitude (González-Fernández et al., 2023).

Differences in modeling techniques and the question whether small drainage basins contribute significantly to global estimations are among the main reasons behind the flux uncertainty. In recent years, research has shown that the inclusion of smaller rivers explains the majority of the plastic load entering the ocean (González-Fernández et al., 2021; Meijer et al., 2021). Another reason is the lack of standardized monitoring and observation techniques, which lay the foundation of every modeling approach. Often, observations are limited in their spatial and temporal resolution, due to the dependence of people performing visual counting observations. In addition, measurement campaigns often only include items larger than 2.5 cm and sometimes cover only part of the width or depth of the river (Roebroek et al., 2022). Although plastic fluxes are usually reported in tonnes per year, observations mostly measure in number of items per time unit (van Calcar and van Emmerik, 2019). Therefore, the required item to mass conversion introduces uncertainty by up to two orders of magnitude (González-Fernández et al.,

2023). To reduce uncertainty in estimations of river plastic emissions, comparability and consistency between data sources are needed. A visual observation approach introduced by (González-Fernández and Hanke, 2017) introduced a monitoring app, which allowed for consistent item categories. However, visual counting is still affected by observer bias, which can increase during high flow velocities and increased plastic fluxes. Furthermore, the level of qualification or training of the observer itself plays a crucial role in correctly assessing litter items (Van Emmerik et al., 2018). Yet, visual counting still represents a popular method in the field of plastic monitoring, even with the current low degree of automation (Castro-Jiménez et al., 2019; Crosti et al., 2018).

Therefore, observation methods are required, which allow for simple deployment to capture a larger spatial and temporal window. Accurate long-term data on the type and quantity of litter is essential for targeted interventions, policymaking, and public awareness initiatives. In recent years, the field of deep learning (DL) has presented a promising solution to automate the monitoring of floating litter. Camera-based systems mounted on infrastructure like bridges can capture a continuous stream of images, while DL models can be trained on different computer vision tasks such as image classification, object detection or instance segmentation to automatically detect debris in this imagery (Jia, Kapelan, et al., 2023). Advanced architectures such as Convolutional Neural Networks (CNNs) have demonstrated promising performance in object detection and classification tasks, offering the potential to create scalable and cost-effective monitoring networks (Wolf et al., 2020).

However, the application of DL to riverine plastic detection is not without its challenges. Many existing models struggle with generalization, performing poorly when deployed at new locations or under variable environmental conditions such as changing light, weather, and water levels (Jia et al., 2024). These challenges are often rooted in a lack of large-scale, diverse, and accurately annotated training datasets. Although studies have achieved success in detecting floating litter, they often focus on general litter detection, which features a single universal litter class (Jia, Kapelan, et al., 2023). Therefore, specific information about the type, shape and size of the detected or observed item is neglected. A recent study by Kataoka et al. (2024) achieved an average precision of 55.3% across seven categories (e.g. bottle, cup), highlighting both the potential and the existing performance limitations of detailed classification (Kataoka et al., 2024). Previous attempts to bridge the gap between camera monitoring and mass quantification have explored area-to-mass conversion techniques. For example, Kataoka et al. (2020) calculated an area-flux, which was later converted to mass-flux using empirically derived mass-per-unit-area coefficients (Kataoka and Nihei, 2020). Although this approach established a significant correlation between image-derived metrics and physical mass during flood events, the authors highlighted the significant variance between different types of debris, which remains a major source of uncertainty. For accurate mass-flux estimation, a classification based on material type (e.g., PET, Polystyrene, etc.) could be advantageous, since the material is often a critical factor for an item's weight. Thus, the inclusion of such information in observation methods could contribute towards a more nuanced mass estimation.

1.2. Knowledge Gaps

Despite advances in computer vision and riverine plastic monitoring, several knowledge gaps persist that hinder the operational deployment of automated mass-flux monitoring systems.

Limitations in Multi-Class and Material-Based Detection: Current research mostly focuses on binary detection (litter vs. background) or general item-based classification. While recent studies, such as Kataoka et al. (2024), have applied deep learning for detailed classification across specific item categories, there is a lack of models designed to classify litter based on material composition (Kataoka et al., 2024). Furthermore, the generalization capabilities of multi-class models in out-of-domain environments with varying environmental conditions remain largely unexplored.

Scarcity of Annotated Datasets for Diverse Litter Types: The development of robust Deep Learning (DL) models is currently restrained by the lack of large-scale, detailed datasets. While general floating litter datasets exist, they often lack annotations for specific material classes. A limited number of datasets provide detailed labels (Jia, Vallendar, et al., 2023; Saddi et al., 2025; Jia et al., 2025), yet they often suffer from class imbalance. Data regarding underrepresented litter types such as glass, paper or textile items are scarce, hindering the ability of models to learn and detect these categories in real-world scenarios.

Validation of Long-Term Monitoring and Mass Estimation in Uncontrolled Environments: A significant gap exists between model performance on test sets and operational utility in real-world environments. Most DL studies evaluate performance using in-domain or out-of-domain test sets but rarely

assess efficacy under the continuous, uncontrolled constraints of long-term monitoring (e.g., variable weather). While methods for area-flux estimation exist (e.g., Kataoka and Nihei, 2020), no study has yet validated a DL-based material classification approach against collected physical litter data in a real-world case study. Consequently, the suitability of camera-based systems as a direct substitute for physical sampling for mass quantification remains unverified.

1.3. Research Questions and Contributions

This thesis aims to address the identified knowledge gaps by developing a deep learning framework capable of material-based classification for mass flux estimation. To achieve this, the research is guided by the following central research question:

How can a deep learning model enable material-based classification of floating litter in waterways for the purpose of accurate mass-flux estimation?

To answer this central question, several sub-questions (SQ) address the knowledge gaps outlined above.

1. *To what extent does expanding the training dataset with synthetic data improve model robustness and detection performance?*
2. *How accurately can a deep learning model classify floating litter into material-based categories compared to traditional item-based approaches?*
3. *How does image based object detection compare to traditional visual counting methods in terms of temporal coverage and consistency during long-term deployment?*
4. *Can material-based classification be reliably converted into mass flux estimates, and what is the accuracy of these estimates when validated against physically collected litter data?*

In addressing these questions, this thesis aims for the following contributions:

Alleviation of Data Scarcity via Synthetic Data Integration (SQ1): To overcome the challenge of limited training data, this study explores the generation and integration of synthetic data. By creating 3D-rendered images of diverse litter items within simulated riverine environments, this work expands the training distribution, particularly for rare litter items.

Material-Based Classification (SQ2): This research proposes a classification scheme based on the Oslo-Paris (OSPAR) convention, which was refined into different material categories Tasseron et al., 2020. By training the model to recognize material types, the detection output can be linked to specific weight characteristics, facilitating a more physics-based approach to mass quantification.

Real-World Validation and Mass Flux Correlation (SQ3 & SQ4): Bridging the gap between theory and practice, the developed framework is deployed at two case study locations in the Netherlands: an urban setting (Amsterdam) and a rural setting (Den Helder). This field study provides comparisons between DL-estimated mass fluxes and weights of litter items obtained from physical removal systems, offering insights into the suitability for long-term deployment.

1.4. Thesis Outline

This thesis is structured as follows. Chapter 2 provides a comprehensive review of the relevant literature on riverine plastic monitoring, computer vision techniques for object detection, and data augmentation strategies. Chapter 3 details the methodology, including data acquisition and preprocessing, the generation of synthetic data, the selection and finetuning of deep learning models, and the design of the case study. Chapter 4 presents the results of the model training and validation, including performance metrics and the mass flux estimation from the test locations. Finally, Chapter 5 discusses the implications of these results, and Chapter 6 summarizes the key findings, outlines the limitations of the study, and provides recommendations for future research.

This research is supported by Noria Sustainable Innovators, a tech-scale up based in Delft that monitors and removes litter from Dutch waterways. Contributions include the provision of camera and fieldwork equipment, computational hardware for training models and the permission to use their automated removal systems.

2

Theoretical Background

2.1. Plastic Pollution in Riverine Environments

2.1.1. Riverine Litter and Macroplastics

Riverine litter encompasses any persistent, manufactured, or processed anthropogenic material discarded, disposed of, or abandoned in the marine and coastal environment (Van Emmerik et al., 2018). A significant portion of this litter is plastic, which is often categorized by size. This research focuses on macroplastics, which are plastic debris items typically larger than 2.5 cm in at least one dimension (Castro-Jiménez et al., 2019). These items constitute a major fraction of the visible pollution transported by rivers and are the primary targets for monitoring with camera-based systems due to their detectability in visual-spectrum imagery (Van Emmerik and Schwarz, 2020).

2.1.2. The OSPAR Framework for Litter Classification

To ensure data consistency and comparability across studies, a standardized classification protocol is necessary. This thesis adopts the methodology developed under the Oslo-Paris (OSPAR) convention, which is a key international agreement for the protection of the marine environment in the North-East Atlantic (Tasseron et al., 2020). Originally developed for beach litter, the OSPAR guidelines have been adapted for riverine environments (van Emmerik et al., 2020).

The OSPAR guidelines define over 120 item based categories across several main material groups (OSPAR Commission, 2010). This research focuses on a material-based classification scheme derived from the OSPAR protocol (Vriend et al., 2020). Instead of categorizing litter by item type (e.g., bottle or cup), items are classified by their constituent polymer material. The main categories used in this study include specific plastic polymers as well as other common debris materials based on the original OSPAR format, as detailed in Table 2.1 (Tasseron et al., 2020).

Table 2.1: Proposed material-based classification based on OSPAR guidelines.

Main Category	Sub-category	Example
Plastic	PET	Bottles
	PS	Plastic cutlery, straws
	EPS	Cups, foamed meat trays
	PO-hard	Shampoo bottles, lunch boxes
	PO-soft	Shopping bags
	Multilayer	Food wrappers
Metal	-	Cans, foil
Glass	-	Glass bottles
Wood	-	Sticks
Paper	-	Newspaper, Cigarette Packaging
Textile	-	Clothing
Rubber	-	Tire

The full list with individual OSPAR items per material class can be found in Appendix A. For the purpose of this research, any reference to OSPAR classification, especially in case of model training, refers to this adapted material based approach.

2.2. Deep Learning in Computer Vision

2.2.1. Machine Learning Paradigms

Within machine learning (ML), Deep Learning (DL) has emerged as a dominant paradigm, characterized by its use of deep artificial neural networks to model complex, high-level abstractions in data (LeCun and Bengio, 1998).

Within ML, this research operates primarily under the *Supervised Learning* paradigm. In a supervised task, the model learns from a dataset where each data point (e.g., an image) is accompanied by a ground-truth label (e.g., "PET" or a set of bounding box coordinates). The model's goal is to learn a mapping function that can correctly predict the labels for new, unseen data.

In a supervised setting, the model is provided with a training dataset composed of input-data-and-label pairs, (X, y). For plastic litter detection, the input X is an image of a waterway, and the label y is the ground-truth annotation (e.g., a bounding box and class label for each piece of litter).

The training process is iterative:

1. **Forward Pass:** An input X is passed through the network, which produces a prediction, $\hat{y}$.
2. **Loss Calculation:** A *loss function* $L(\hat{y}, y)$ quantifies the dissimilarity between the prediction $\hat{y}$ and the true label y.
3. **Backpropagation:** The gradient of this loss with respect to every weight in the network is computed using the backpropagation algorithm. This gradient indicates the direction and magnitude of change needed for each weight to reduce the loss.
4. **Weight Update:** An *optimizer* uses these gradients to update the network's weights, $w \leftarrow w - \eta \nabla L(w)$, where η is the learning rate.

This cycle is repeated for many epochs (passes through the entire training dataset) until the model's performance on a separate validation dataset ceases to improve. To prevent overfitting, a regularization term $\Omega(f)$ is often added, leading to the minimization of a regularized risk function.

While other paradigms exist, such as unsupervised learning (finding patterns in unlabeled data) or reinforcement learning (learning through trial and error), they are less suitable for the specific task of litter detection. The primary challenge in this field is the scarcity of labeled data, which might suggest an unsupervised approach. However, the task of identifying and classifying specific litter types (e.g., OSPAR categories) is a high-precision, multi-class problem. Therefore, despite data scarcity, supervised learning remains the chosen approach for this research, with the data scarcity challenges being addressed through other methods like data augmentation, transfer learning and synthetic data generation.

2.2.2. Convolutional Neural Networks

The deep learning revolution in computer vision was largely initiated by the development and application of the Convolutional Neural Network (CNN) (LeCun and Bengio, 1998). While recent years have seen the emergence of transformer-based architectures, CNNs remain a primary methodology for processing data such as images. Their architecture is inspired by the hierarchical organization of the human visual cortex, which allows them to automatically learn spatial features from high-dimensional inputs without the need for manual feature engineering.

A CNN processes an image by passing it through a sequence of mathematical operations, often organized into repeated units known as convolutional blocks. The fundamental components of these blocks include:

- **Convolutional Layers:** These layers act as the core building blocks of the network. They employ a set of learnable filters (kernels) that slide across the input image to perform convolution operations. This process detects specific features ranging from low-level edges and corners in earlier layers to complex patterns and object parts in deeper layers (LeCun and Bengio, 1998).

- **Activation Functions:** Following the convolution operation, a non-linear activation function is applied to the output. Common functions, such as the Rectified Linear Unit (ReLU) or Swish, introduce non-linearity into the model, which enables the network to learn more complex patterns that a simple linear function could not represent.
- **Pooling Layers:** These layers serve to downsample the feature maps, reducing their spatial dimensions (width and height) while retaining the depth (number of features). Techniques, such as max-pooling, extract the maximum value from a small window of the feature map. This makes the feature representation more robust to small translations in the image and reduces the computational load.
- **Batch Normalization:** This technique is applied to normalize the inputs of a layer, typically by re-centering and re-scaling the data. It stabilizes the learning process, accelerates convergence, and acts as a regularizer, which reduces the model's sensitivity to initialization parameters (Ioffe and Szegedy, 2015).

The sequential stacking of these convolutional blocks constitutes the feature extractor or backbone of the network. This section of the model is responsible for transforming raw pixel data into a rich, compressed numerical representation of the image content.

The architecture of the head, which follows the feature extractor, varies depending on the specific downstream task. In standard image classification, the extracted features are typically flattened and fed into a Multilayer Perceptron (MLP) to produce class probabilities. In semantic segmentation, the head often remains fully convolutional to preserve spatial dimensions for pixel-level predictions. For object detection, the head design is more complex. It must predict both the spatial coordinates of objects and their specific categories. Therefore, the structure of the head is directly influenced by the number of target classes. The final classification layer within the head consists of N output channels (or neurons), where N corresponds to the total number of classes defined in the dataset. Therefore, increasing the detail of the classification scheme (e.g., from Binary to 13 OSPAR based material classes) requires a corresponding expansion of the model's output dimension to accommodate the probability scores for each additional category.

2.2.3. Computer Vision Tasks

The features extracted by a CNN backbone can be used to solve several computer vision tasks, which differ in their objective and output. For the context of this research, the most relevant tasks are classification, object detection, and segmentation.

Image Classification

Image Classification is one of the most fundamental tasks in computer vision. The objective is to assign a single category or class to an entire image (Lu and Weng, 2007). The model analyzes an image and outputs a single label, such as 'Litter Present' or 'No Litter'. However, this approach does not provide information about where the litter is located or how many items are present. It is a high-level summary of the entire scene.

Object Detection

Object Detection (OD) is a more advanced task. It goes beyond simple classification by identifying and localizing multiple objects within a single image. The goal is to produce a list of detections, where each detection consists of two parts:

1. **Class Label:** The category of the object (e.g., 'Bottle', 'Plastic Bag', 'Wood').
2. **Bounding Box:** A set of coordinates (e.g., $x, y, width, height$) that define a rectangular box tightly enclosing the detected object.

OD requires image labeling, so that objects are annotated with class labels and bounding boxes. This process can be time consuming but is essential for effective training (Zou et al., 2023).

Image Segmentation

Image Segmentation provides the most detailed approach. Instead of just drawing a rectangular box around an object, segmentation models classify every single pixel in the image. There are two main types. First, in *semantic segmentation* class labels are assigned to each label. However, there is no distinction between different instances of the same class. This means a model cannot differentiate between

two separate bottles. In *instance segmentation* the goals of object detection and semantic segmentation are combined. The model identifies every distinct object instance in an image and generates a pixel-wise mask for each one (Minaee et al., 2021). Therefore, the exact dimensions of a litter item could be estimated.

Although segmentation provides a high level of detail, it is also the most computationally intensive and requires significantly more detailed (and thus more expensive) pixel-level annotations for training. Given the research goals of counting and classifying distinct items, object detection presents an optimal balance between informative output and computational feasibility.

2.2.4. Architectures for Object Detection

For object detection tasks, specialized architectures are required. This research employs three distinct types:

- **YOLO (You Only Look Once):** YOLO is a state of the art single stage detection model (Jocher and Qiu, 2024). It is renowned for its exceptional speed and real-time capabilities. It frames object detection as a single regression problem, directly predicting bounding boxes and class probabilities from the full image in one pass. This research uses YOLOv11, which uses a highly efficient backbone and a decoupled head to process classification and regression tasks separately, improving accuracy. Benchmarking results on the official Common Objects in Context (COCO) 2017 dataset for the YOLOv11m (medium sized) model show that it outperforms previous YOLO iterations (mAP@50-95: 51.5%, validation set) (Khanam and Hussain, 2024). YOLO models have been commonly used in the field of plastic litter detection due to their ease of use and real time capabilities, which are beneficial for automated monitoring networks (Kylili et al., 2021; Maharjan et al., 2022; Kataoka et al., 2024; Lin et al., 2021)
- **Faster R-CNN (Regions with CNNs):** The Faster RCNN architecture is a two-stage detector and known for its high accuracy, which is why it presents a popular choice in the field (van Lieshout et al., 2020; Jia et al., 2024). In the first stage, a Region Proposal Network (RPN) scans the image and proposes a set of candidate regions of interest (RoIs) that might contain an object. In the second stage, these proposals are individually passed to a classification and regression head to be refined **Ren2015**. This two-stage approach is typically more computationally extensive but can be more precise. The initial COCO validation set performance with a ResNet-101 Backbone achieved an mAP@50-95 of 48.4%, which was the highest upon its release (Ren et al., 2015).
- **RT-DETR (Real-Time DETR):** This model is based on the DEtection TRansformer (DETR) architecture, which was the first to apply the Transformer (the basis of models like ChatGPT) to object detection Carion et al., 2020. DETR removes the need for many hand-designed components (like region proposals) and instead uses an encoder-decoder structure with attention mechanisms to directly output a set of predictions. RT-DETR is a recent variant optimized for real-time performance while maintaining high accuracy. The COCO mAP@50-95 performance of the RT-DETR (ResNet Backbone) on the validation set reached 53.1%, outperforming similar sized YOLO models at the time (Zhao et al., 2024).

Although benchmark results on official datasets give good indications of overall peak performance, results in different detection tasks can vary. Taking hardware limitations into account, peak performance might not be reproducible as declared in official benchmarking publications (Asselin et al., 2025). Therefore, training different architectures on a new task can generate new insights, even if official results indicate a potential best performing model.

2.2.5. Optimization Strategies for Model Performance

Transfer Learning

Transfer learning uses knowledge from previous tasks and transfers it to a new one. Instead of training a model from scratch (with random initial weights), which requires a large dataset, transfer learning leverages a model that has already been pre-trained on a general-purpose dataset (Pan and Yang, 2009). Common object detection models like the ones mentioned in 2.2.4 are typically trained on large datasets like the COCO dataset. The core principle is that the features this model has learned (e.g., edges, textures, shapes) are useful for many different vision tasks. The pre-trained backbone is then fine-tuned on the smaller, specific dataset (e.g. litter data), allowing it to adapt its general knowledge to the new task (Yosinski et al., 2014). Finetuning involves several steps of adjustments, although unfreezing all

layers and fine-tuning them, instead of only the classifier, seems to improve performance the most for the specific task of litter detection (Jia, Vallendar, et al., 2023).

Data Augmentation
Data augmentation is used to combat over- and underfitting and expand the effective size of a small dataset (Shorten and Khoshgoftaar, 2019). This involves applying random, realistic transformations to the training images during training (online augmentation) or prior to the training, where images are augmented and saved locally (offline augmentation). Each time the model sees an image, it sees a slightly different version. Common augmentations include:

- **Geometric Augmentations:** rotation, scaling, translation, shear, and flipping (horizontal and vertical).
- **Color Augmentations:** adjustments to hue, saturation, and value (brightness).
- **Object-Level Augmentations:** Mosaic (stitching four images together) and MixUp (linearly blending two images), which forces the model to learn from complex and partially occluded scenes.

Data augmentation has been widely used in the field of litter detection and monitoring to improve model performance (Lin et al., 2021; van Lieshout et al., 2020). While some techniques seem more effective than others, such as flipping, data augmentation can lead to insignificant performance gains (Jia, Vallendar, et al., 2023). In some rare cases, the inclusion of data augmentation can also lead to a decrease in performance (Musić et al., 2020).

Hyperparameter Optimization
Finding the best hyperparameter settings for the training process, such as the learning rate, weight decay, and probabilities for each augmentation, is a major challenge (Raiaan et al., 2024). These parameters are not learned during training but are set beforehand. Different methods can be applied for a systematic approach. The definition of a parameter grid is a simple approach, that defines the search space for a selection of hyperparameters. However, with an increased amount of parameters to test and a wide search space, testing each combination becomes computationally expensive (Alibrahim and Ludwig, 2021).

A possible solution is to apply a random search, in which a subset of possible combinations is explored (Bergstra and Bengio, 2012). This requires expert knowledge for efficient results, otherwise an optimal setup might not be found. Alternatively, the search space can be explored through a probabilistic model, called Bayesian Optimization (Alibrahim and Ludwig, 2021). It uses the results from past trials to make informed decisions about which new combination of hyperparameters to try next, focusing its search on promising areas of the hyperparameter space. This allows it to find better configurations in fewer trials than other methods.

3

Methodology

This chapter details the systematic approach taken to develop and validate deep learning models for mass-flux estimation of floating litter. The methodology is structured into several key phases: (1) data acquisition and preparation, including the generation of a novel synthetic dataset; (2) a multi-stage experimental design to train, optimize and evaluate the model, starting with a binary classification task, progressing to a detailed multi-class OSPAR setup and leading to a benchmark against other model architectures; and (3) final field tests to estimate mass-fluxes on two different sites. An overview of the experimental framework is presented in Figure 3.1.

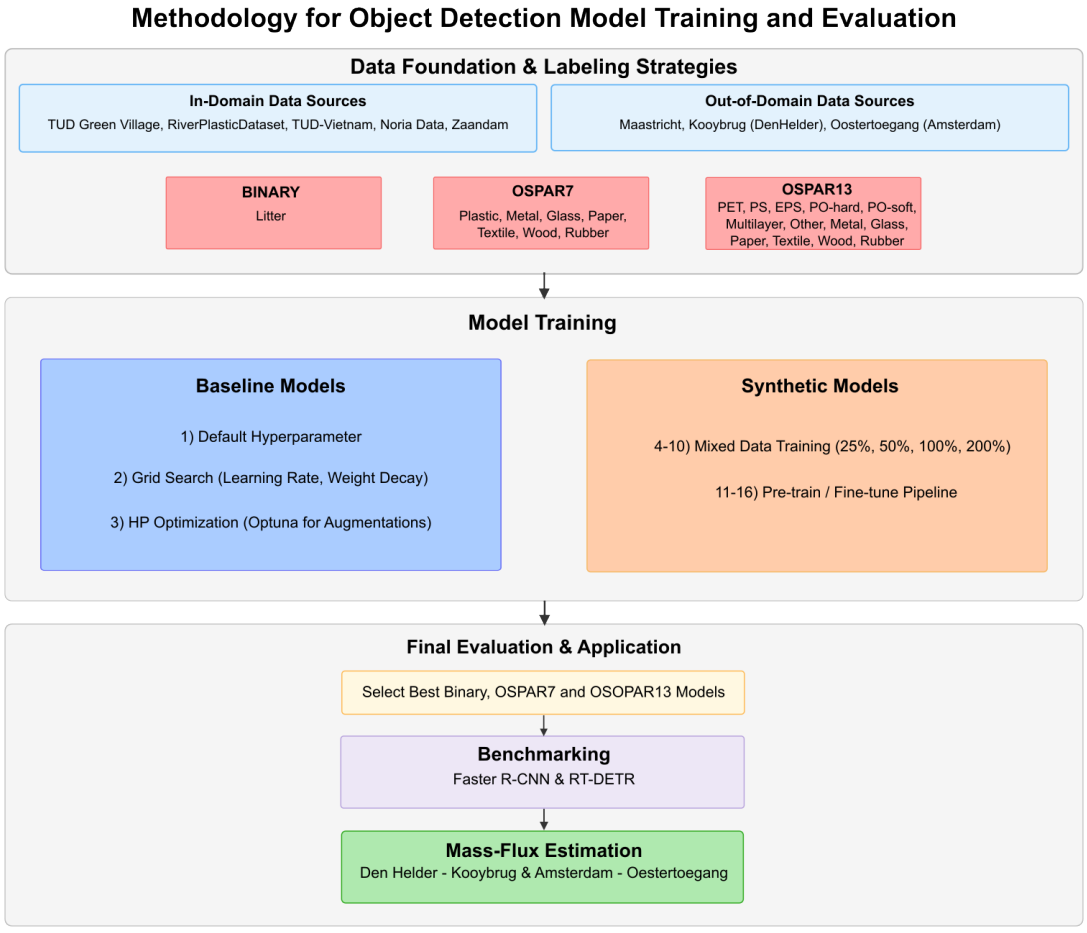


Figure 3.1: Overview of the research methodology, from data preparation to final model validation.

3.1. Data Acquisition and Preparation

3.1.1. Real-World Image Datasets

The real-world image dataset used in this research is a composite of several sources to ensure variability in location, environmental conditions, and litter types. The initial set of training images was provided by Noria from existing camera monitoring stations in the Netherlands. Furthermore, publicly available datasets are incorporated to enhance diversity, including the TU Delft Green Village dataset (TUD-GV) (Jia, Vallendar, et al., 2023), the TUD-Vietnam dataset (Jia et al., 2025), and data from (Saddi et al., 2025) which aggregates data from Jakarta, Vietnam (WUR) and Amsterdam under the RiverPlastic-Dataset (RPD). Additionally, more than 800 images were manually annotated from a catch and release experiment in Zaandam (Netherlands). This experiment was conducted by Noria prior to this research. Noria also provided data from Maastricht after the completion of a monitoring assignment, which was included for testing purposes. Finally, this research deployed cameras in Den Helder (Kooybrug) and Amsterdam (Oostertoegang) for mass-flux validation, where subsets of unlabeled data were manually annotated. Table 3.1 shows the details of the real world datasets.

Table 3.1: Summary of datasets used in the study, including unseen test locations.

Name	Location	Camera Device	Device Height (m)	Image Resolution	Number of Images	Number of Annotations	Label Strategy
TUD-GV	Delft, the Netherlands	GoPro Hero 4 GoPro MAX 360 Huawei P30 Pro	2.7	1920x1080	1501	8,195	Binary, OSPAR
RPD	Amsterdam, the Netherlands Jakarta, Indonesia Ho Chi Minh City, Vietnam	GoPro Hero 10/11, DJI Phantom 4 pro, Dhaka Easy4ip	1-2 4.5 7.4-8.6 (cameras) 11-14 (drones)	4000x3000 1920x1080 2795x2095	918	14,244	Binary, OSPAR
TUD-V	Ho Chi Minh City, Vietnam	Pentax K-serie		6016x4000	449	1,243	Binary, OSPAR
Zaandam	Zaandam, the Netherlands	Obscape HQ	5	4056x3040	814	1,384	OSPAR
Noria Dataset	Rotterdam, the Netherlands Groningen, the Netherlands Delft (Oostport), the Netherlands	Obscape HQ GoCam3 GoPro MAX 360	4-8	640x640 4056x3040	1974	4,069	Binary
Unseen Locations (Not Included in Training)							
Maastricht	Maastricht, the Netherlands	Obscape HQ	7	4056x3040	160	176	Binary
Kooybrug	Den Helder, the Netherlands	Obscape HQ	9	4056x3040	80	97	Binary, OSPAR
Oostertoegang	Amsterdam, the Netherlands	Obscape HQ	1.35	4056x3040	76	101	Binary, OSPAR

The open datasets contain their own individual labeling scheme instead of one uniform label strategy. These datasets were selected because of their detailed, object-based annotations for items such as plastic bag or bottle rather than a broader litter only label. This level of detail makes a straightforward conversion possible to a new label scheme. A standard 70%/20%/10% split was employed to divide the data into training, validation, and test sets, respectively. To avoid model bias and avoid the results being skewed by a single dominant data source, care was taken to ensure that each original dataset contributed equally to the image count in each subset. To evaluate the generalization capacity of models, data from specific locations were intentionally excluded from all training sets. These unseen locations were used for testing the out-of-domain (OOD) capabilities. This included data from Maastricht (160 images), Zaandam (814 images), Kooybrug in Den Helder (80 images), and Oostertoegang in Amsterdam (76 images).

3.1.2. Synthetic Dataset Generation

To mitigate the limitations of data scarcity and the inherent class imbalance of real-world datasets, an automated workflow was established, which uses the open-source 3D creation suite Blender. After the initial setup, synthetic data was generated automatically. Unlike generative AI or manual composition methods, this 3D-rendering approach allows for the programmability of scene parameters, enabling the generation of large-scale datasets with automatically generated annotations.

The synthetic environment was constructed to replicate the visual variability of riverine systems. The base of the scene consisted of an adjustable water layer to simulate varying turbidity levels, surface movements and wave patterns. To ensure diverse lighting conditions, the illumination of the scene was randomized using a collection of parameters to simulate different times of day, cloud cover, and solar angles. To populate the scene with litter items, a library of over 60 unique 3D models was compiled and sourced from open repositories. To improve realism, the library included non-litter objects such as

floating leaves, organic vegetation and foam patches, which are frequent sources of false positives in operational monitoring.

Object placement within the scene was automated. The position, rotation, and relative scale of items was randomized to mimic the chaotic nature of floating debris. During the rendering process, the bounding box coordinates for each object were calculated by projecting the object's 3D mesh onto the 2D camera plane, ensuring that labels were generated without error or manual intervention.

To specifically address the class imbalance observed in the real-world training data, two distinct object population strategies were implemented. The first, a random method, assigned an equal selection probability to all assets, creating a baseline dataset. The second, a weighted method, used an inverse frequency strategy to preferentially use underrepresented classes. The spawn weight $w(c_i)$ for a specific class i was derived from its frequency in the combined real-world datasets. Given the probability of occurrence $P(c_i)$, calculated as the count of class instances c_i divided by the total dataset size N, the spawn weight was defined as the inverse square root of the class probability:

$$w(c_i) = \frac{1}{\sqrt{P(c_i)}} = \frac{1}{\sqrt{\frac{c_i}{N}}} \tag{3.1}$$

This weighting function suppresses the generation of over-abundant classes while prioritizing rare categories. In total, this workflow generated 3,593 images using the random logic and 3,403 images using the weighted logic, resulting in a synthetic corpus of nearly 7,000 fully annotated images to supplement the real-world training sets. Figure 3.2 displays representative examples of both methods with varying backgrounds and litter items.

Figure 3.2: Examples of generated synthetic images

3.1.3. Labeling Strategies

Three different labeling methodologies were implemented to evaluate the effects on model performance and mass-flux estimation. These strategies form a hierarchy of difficulty, ranging from general litter detection to detailed material characterization.

1. **Binary Classification:** All litter types are consolidated into a single universal class labeled litter. This setup establishes a baseline for general detection performance, focusing solely on distinguishing debris from the background without classification.
2. **Material-Based Classification:** The original litter class is split into different classes based on their material. While the OSPAR framework has listed more than 120 different litter types, the available data does not sufficiently support this level of detail for an object detection task. Instead, different litter items from the OSPAR guidelines are aggregated according to their material (Vriend et al., 2020). For the purpose of this research this classification approach is referred to as OSPAR with the addition of a number to indicate the amount of classes. While this method is only based on the original framework, which does not necessarily define all of these material types, the terminology is adapted for simplicity. Therefore, material-based classification was examined at two levels of detail:
    - **OSPAR7:** The general litter class is divided into 7 main material types, which are Plastic, Metal, Glass, Paper, Textile, Wood and Rubber. Each category contains different items, although some share the same shape, such as a plastic or glass bottle.
    - **OSPAR13:** The Plastic class is further split into specific polymer types (PET, PS, EPS, Multilayer, PO-Soft, PO-Hard, Other Plastics) alongside the major material categories, which expands the class count to 13.

A standardization process was performed to unify the different datasets under the proposed labeling scheme. As the original datasets used varying classes, a relabeling step converted all native labels to the target classes. It is important to note that this process involved only the modification of class labels. The original bounding boxes were retained without manual verification of their positional accuracy. Table 3.2 lists the original labels matched to the new scheme.

Table 3.2: Mapping of original dataset labels to the unified OSPAR classification scheme.

TUD-GV	Vietnam	RPD	Zaandam	New OSPAR Class
Plastic bottles	Plastic bottles	Plastic bottle	Bottles	**PET**
Cups	Cups	Plastic straw	-	**PS**
Cutlery/trays/straws	Cutlery/trays/straws	Polystyrene		
Styrofoam	Styrofoam	-	Foam big	**EPS**
			Foam small	
Caps and lids	Caps and lids	Food container & cutlery	Caps and lids	**PO_hard**
	Floats/buoys	Bottle cap & container		
Bags	Bags	Plastic bag	Plastic film small	**PO_soft**
Food packaging	Food packaging	Transparent bag	Plastic film big	
Plastic films	Plastic films			
Food wrapping	Food wrapping	Food wrapper	Food wrapping small	**Multilayer_plastic**
			Food wrapping big	
Others	Nets	Unclassified Plastic Object	-	**Other_plastic**
	Others			
Drink cans	Drink cans	Metallic can	-	**Metal**
-	-	Glass bottle	-	**Glass**
-	-	-	-	**Wood**
-	-	Drink carton	Cigarette butts	**Paper**
-	Shoes/sandals	-	-	**Textile**
-	-	Rubber ball	-	**Rubber**
-	Vegetation	-	-	*Non-litter*

The datasets used in this research contain an inherent class imbalance, where plastic items disproportionately outnumber materials such as Rubber, Wood, Paper, Glass, and Textile. To address this issue, the RPD and Vietnam subsets were manually audited to annotate previously unlabeled instances

of these underrepresented materials. While only a few objects were present in the existing images, the synthetic data provides more examples for training. Due to the remaining low representation with real examples, a data curation step was implemented for these rare classes. A standard 70/20/10 data split would have resulted in a statistically insignificant number of instances (e.g., <5 objects) in the validation and test sets. To guarantee a robust evaluation, images containing these rare class instances were partitioned using a 50%/25%/25% split (Train/Val/Test). This modification ensures that performance metrics for underrepresented classes are derived from a more meaningful sample size.

All annotations were standardized to the YOLO format, comprising a text file per image with normalized bounding box coordinates (center-x, center-y, width, height) and the corresponding class index. Figure 3.3 illustrates the final distribution of object instances per class for the OSPAR13 setup utilized in the training phase.

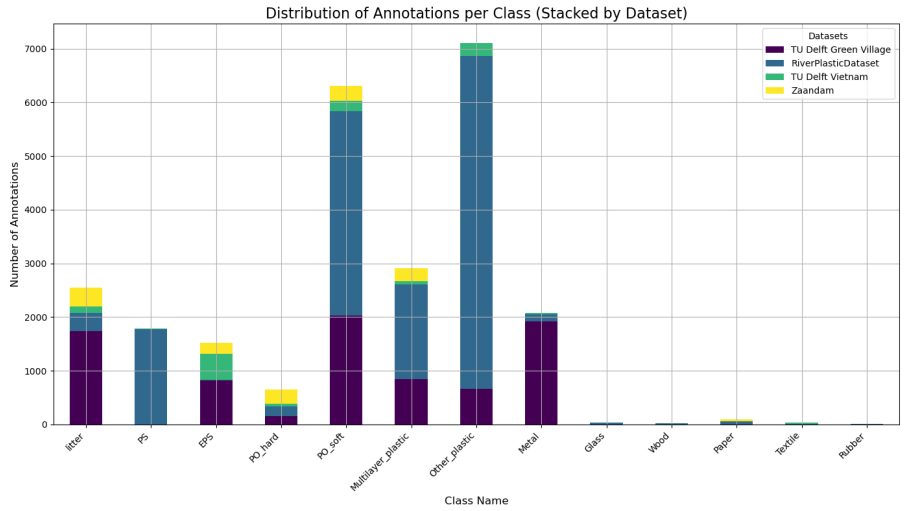


Figure 3.3: Annotations per class for OSPAR13 setup

3.1.4. Weight Datasets for Litter

To enable the conversion from litter count into mass estimates, three distinct sources of empirical weight data were used. The primary dataset is from a study, which sampled over 14,000 items from riverbanks at the IJssel, Meuse and Rhine, in the following referred to as the IMR dataset (de Lange et al., 2023). In addition to the weight of the items, the physical length and width were measured. Furthermore, Noria provided a weight database, which consists of data from sampled litter in Groningen and Rotterdam in 2023 Noria, 2023. The dataset contains over 2,700 individual items with a total weight of 26.8 kg. Lastly, litter was sampled at the Oostertoegang location in Amsterdam during this study. A total of 659 items were collected, weighing just over 5.7 kg. Next to this batch, more litter was sampled during this research at the Oostertoegang location as well as in Den Helder as part of the mass-flux field experiments. As this data acts as the ground-truth for the mass estimation, it is only used for the purpose of validation.

3.2. Model Training

The YOLO framework by Ultralytics was selected as the base model for this research Jocher and Qiu, 2024. This choice was motivated by its established state-of-the-art performance on benchmark datasets such as COCO, its relative ease of use, and its real-time detection capabilities, which are important for practical environmental monitoring. At the beginning of this study, YOLOv11 was the latest fully refined version available. The medium-sized model (YOLOv11m) was chosen since it balances performance with computational efficiency. Transfer learning was applied by initializing YOLOv11m with weights pre-trained on the COCO dataset and then fine-tuning all layers on the specific litter dataset. The detection head was reconfigured to match the class set of each label setup. No layers were frozen, which allowed both backbone and head to adapt to the new feature representations of the litter data. Due to hardware constraints, all training runs used a consistent batch size of 16 and an input resolution of 640x640 pixels, with the AdamW optimizer employed for effective weight decay regularization. All models were trained on a single NVIDIA GeFroces RTX 3090 GPU. The following training procedure was conducted and repeated for all three label setups.

3.2.1. Baseline Training and Optimization

The goal of model training was to maximize performance with different optimization techniques. To measure the impact of each step, changes in configuration were isolated and final models were trained for 100 epochs each time. First, a model was trained with default hyperparameters (learning rate 0.01, weight decay 0.0005) and no data augmentation. Second, a 5x5 grid search was performed to test 25 combinations of different learning rate and weight decay settings. While each combination was trained for a reduced duration of 50 epochs, the optimal configuration was trained for 100 epochs to ensure comparability. Third, a Bayesian optimization study, based on the Optuna Framework, was used to tune the data augmentation parameters (Akiba et al., 2019). This study optimized geometric transformations (`degrees`, `translate`, `scale`, `shear`, `perspective`, `flipud`, `fliplr`), color space augmentations (`hsv_h`, `hsv_s`, `hsv_v`) and object-level augmentations (`mosaic`, `mixup`, `copy_paste`) to maximize model robustness against environmental variability. In total 25 trials were conducted to establish an optimized set of hyperparameters, with which a third final model was trained. In total, 3 baseline models are trained to indicate the performance range of using real data with optimized model configurations.

3.2.2. Synthetic Integration Strategies

The determined set of hyperparameters of the baseline training was kept consistent during the addition of synthetic data. To evaluate the utility of this data, two distinct integration strategies were explored. A total of 13 synthetic model variants were trained per label setup to capture the variability of these approaches. A list of all trained models per setup can be found in Appendix B.

The first strategy, *Mixed-Data Training*, added synthetic images to the real-world training set. Various mixing ratios were tested, ranging from 25%, 50%, 100% to 200% of the real data volume, using the random and class-weighted placement logic to assess the impact of balancing rare classes.

The second strategy, *Pretraining and Fine-Tuning*, leveraged a two stage approach. First, models were trained on a purely synthetic dataset to learn general feature representations. This pre-trained state was then fine-tuned on the real-world dataset using the optimized hyperparameters. Variations of this strategy included using synthetic validation sets versus real-world validation sets during the pretraining phase to monitor domain generalization.

For statistical robustness in the final analysis, the performance metrics of these individual strategies were aggregated. In the Results chapter, these are presented as a collective Baseline and Synthetic category, with performance reported as a range (minimum to maximum) to illustrate the stability and potential gain of hyperparameter optimization and synthetic data integration.

3.2.3. Benchmarking

While the primary analysis focused on the YOLOv11m architecture, a benchmarking study was conducted to validate performance against alternative state-of-the-art detectors. Two additional architectures were selected: the RT-DETR and the two-stage Faster R-CNN. To ensure a fair comparison of model capacity, a ResNet-18 backbone was chosen to match the YOLOv11m parameter count of approximately 20 million. Due to time constraints only the best performing model (out of 16) on the in domain test set of the Binary and OSPAR13 approach is replicated on the other architectures. This included a repeat of the optimization steps (grid search and optuna study) to determine the optimal hyperparameters, which ensures a fair comparison between all architectures.

3.2.4. Evaluation Metrics

Model performance was evaluated using standard object detection metrics. All final performance comparisons were conducted exclusively on the held-out test set and the unseen out-of-domain locations to ensure an unbiased assessment of generalization capabilities.

3.2.5. Performance Evaluation Metrics

To quantitatively assess model performance across all experiments, a standard set of object detection metrics is employed. All final performance comparisons were conducted on the in-domain test set and the unseen out-of-domain locations to ensure an assessment of generalization capabilities. Unless explicitly stated otherwise, all validation results were estimated using an Intersection over Union (IoU) threshold of 0.5. In addition, calculated error rates were determined using a confidence threshold of 0.7. The performance metrics are defined in the following.

Intersection over Union (IoU)
Intersection over Union quantifies the spatial overlap between a predicted bounding box B_p and the corresponding ground-truth box B_{gt}. It is defined as:

$$\mathrm{IoU}(B_p, B_{gt}) = \frac{|B_p \cap B_{gt}|}{|B_p \cup B_{gt}|}. \tag{3.2}$$

A detection is considered a True Positive (TP) if $\mathrm{IoU} \geq \tau$, where τ is the chosen IoU threshold. Predictions with insufficient overlap are counted as False Positives (FP), while missed objects contribute to False Negatives (FN).

Confidence Score
Object detectors assign each predicted bounding box a confidence score $c \in [0, 1]$ representing the estimated probability that: 1) an object is present, and 2) the predicted class label is correct.

Formally, for a predicted bounding box i:

$$c_i = P(\text{objectness}) \cdot P(\text{class} \mid \text{object}). \tag{3.3}$$

Varying the confidence threshold controls the precision–recall trade-off: higher thresholds reduce FP but typically increase FN, while lower thresholds increase recall at the cost of more FP. This study evaluates performance across multiple confidence levels for the final selected models.

Precision, Recall, and F1-Score

- Precision quantifies the accuracy of positive predictions:

$$\mathrm{Precision} = \frac{TP}{TP + FP}. \tag{3.4}$$

- Recall measures the proportion of actual objects that are successfully detected:

$$\mathrm{Recall} = \frac{TP}{TP + FN}. \tag{3.5}$$

- F1-Score is the harmonic mean of Precision and Recall:

$$\mathrm{F1} = 2 \times \frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}. \tag{3.6}$$

Average Precision (AP) and mean Average Precision (mAP)
Average Precision summarizes the area under the Precision–Recall (PR) curve. For a given class i, AP is computed as:

$$\mathrm{AP}_i = \int_0^1 \mathrm{Precision}_i(\mathrm{Recall})\, d\mathrm{Recall}. \tag{3.7}$$

The mean Average Precision over N classes is:

$$\mathrm{mAP} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{AP}_i. \tag{3.8}$$

Two mAP variants are used in this study:

- **mAP@50**: AP computed at a single IoU threshold of $\tau = 0.50$.
- **mAP@50–95**: The COCO-style metric, averaging AP over 10 IoU thresholds:

$$\mathrm{mAP}_{50:95} = \frac{1}{10} \sum_{\tau \in \{0.50, 0.55, \ldots, 0.95\}} \mathrm{mAP}(\tau), \tag{3.9}$$

providing a stricter and more comprehensive measure of detection quality, especially for precise localization.

Class-Averaging Methods: Micro, Macro, and Weighted
Since the dataset is highly imbalanced, multiple averaging schemes are considered. The index i denotes the class label, such that TP_i, FP_i, and FN_i correspond to the true positives, false positives, and false negatives for class i.

- Micro Average aggregates TP, FP, and FN across all classes before computing the metric. Precision is given as an example:

$$\text{Precision}_{micro} = \frac{\sum_i TP_i}{\sum_i (TP_i + FP_i)}.$$

  Micro averaging emphasizes performance on frequent classes and is dominated by majority categories.

- **Macro Average** Computes the unweighted mean of per-class metrics:

$$\text{Precision}_{macro} = \frac{1}{N} \sum_{i=1}^{N} \text{Precision}_i.$$

  Macro averaging treats all classes equally but becomes unreliable when many classes have few or zero samples, artificially lowering the score. Here, N represents the total number of classes.

- **Weighted Average** Each class is weighted by its support (number of instances):

$$\text{Precision}_{weighted} = \frac{\sum_i n_i \cdot \text{Precision}_i}{\sum_i n_i},$$

  where n_i is the number of ground-truth instances for class i.

  Weighted averaging preserves the influence of common classes while still incorporating information from rare ones. This makes it the most appropriate choice to evaluate the highly imbalanced OSPAR models, which enables a fair comparison with the single-class Binary setup.

3.3. Mass-Flux Estimation

The developed models were used and validated in the field for mass-flux estimation. This enabled an assessment of the practical applicability of OSPAR-based models in operational environments. This study prioritized in-situ validation using real-world data streams over controlled catch-and-release experiments. Artificial tests often fail to replicate the complex morphological characteristics of riverine litter, which typically exhibits significant environmental degradation, fragmentation, and biofouling.

Two monitoring locations were selected to represent contrasting environmental characteristics: a rural canal system in Den Helder and an urban canal in Amsterdam. Both sites featured automated litter removal systems operated by Noria, providing the necessary infrastructure to collect ground-truth retrieval data.

3.3.1. Location 1: Kooybrug (Den Helder)

Site Characteristics
The Kooybrug location represents a rural setting. The removal system is positioned in front of a pumping station, which induces a directional flow towards the retrieval point when active. The monitoring point for observations was established at the Kooybrug, a bridge, where a permit already existed for mounting cameras. At a mounting height of approximately nine meters, two Obscape cameras, equipped with 6mm focal length lenses, were installed to monitor the two canal sections passing beneath the bridge.

Figure 3.4: Overview of the Den Helder mass-flux test location, illustrating the spatial separation between the monitoring bridge and the removal point.

A critical challenge at this site was the spatial decoupling of the monitoring and retrieval points. As illustrated in Figure 3.4, the camera systems were separated from the removal system by a canal section of approximately 3.8 to 4.0 kilometers. This distance introduces significant uncertainty regarding mass flux continuity. Litter observed at the bridge is not guaranteed to reach the removal system due to potential retention along canal banks or the influence of external factors such as wind shear and vessel traffic. Furthermore, the transport dynamics are influenced by the intermittent activation of the pumping station and nearby sluice operations. The system was emptied on three occasions by the local operator and the retained litter was stored for ground-truth analysis.

Monitoring Setup

To address the uncertainty introduced by the spatial separation, field experiments were conducted on the 2nd and 4th September 2025 to characterize the transport behavior. Three GPS trackers were deployed at the bridge to log hourly positional data, which enables the estimation of average flow velocity, travel duration, and the identification of retention zones. The results of this transport study were used to define the valid temporal observation window for the mass flux correlation.

Next to the automated monitoring, visual counting surveys were conducted on two separate days. Dedicated observers were assigned to each canal section to classify passing debris according to the OSPAR guidelines (survey sheets can be found in Appendix A). For the automated system, the solar-powered cameras were operated with progressively extended monitoring durations to validate system stability and energy consumption under varying field conditions.

3.3.2. Location 2: Oostertoegang (Amsterdam)

Site Characteristics

The Oostertoegang location, located next to the Amsterdam Central Station, represents a highly urbanized environment. In contrast to the Den Helder site, this location benefits from close spatial proximity between the observation point and the removal system. The Noria system retains floating litter within a confined canal section, minimizing transport losses. Consequently, for the purpose of this study it was assumed that the flux observed at the monitoring point is effectively conserved upon reaching the retrieval system. The field validation at this site was conducted over a 72-hour period from September 12th to September 15th, 2025. The system was manually cleaned before and after the testing period to ensure an unbiased ground-truth sample.

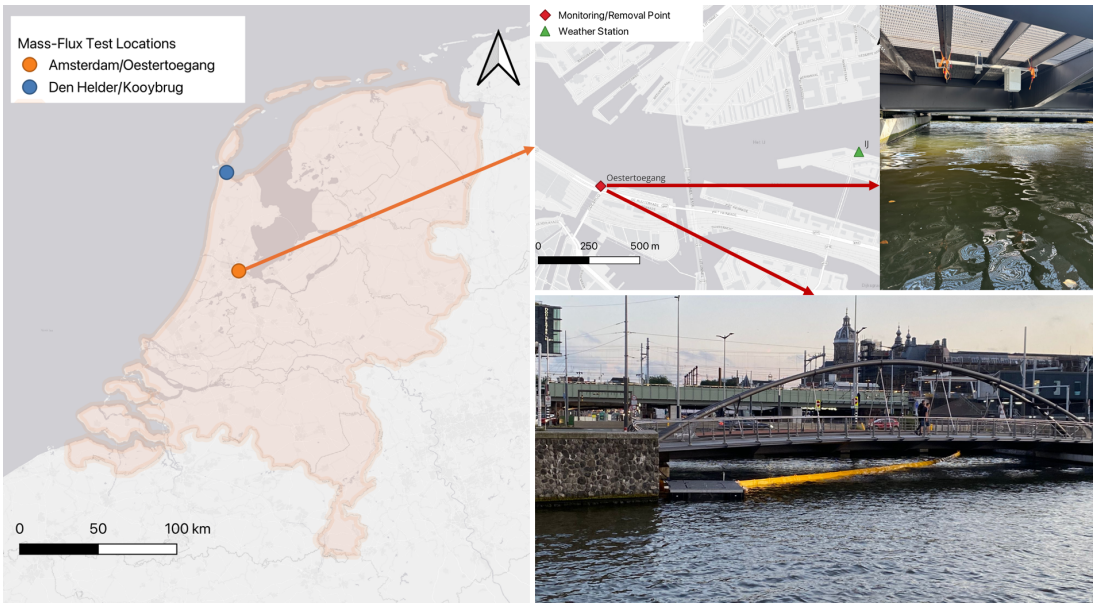


Figure 3.5: Overview of the Amsterdam mass-flux test location, highlighting the proximity of the observation point to the removal system.

Monitoring Setup
The monitoring equipment was installed on a low clearance bridge, resulting in a camera mounting height of approximately 1.5 meters above the water surface. This setup utilized an Obscape camera equipped with a 4mm focal length lens. The observation window was set at 10 hours a day, which utilized maximum daylight hours. Visual counting was performed by a single observer positioned to monitor the full inlet width, ensuring a complete coverage of surface debris during the survey intervals.

3.3.3. Flux Quantification
To translate raw model detections into a reliable estimate of litter transport, multiple post-processing steps were implemented. First, image bursts were used to enhance detection reliability. Rather than processing single static images, the monitoring system recorded three consecutive frames at a frequency of 4 Hz for every detection interval. A valid detection was defined as an object that appeared with consistent bounding box dimensions and overlapping spatial coordinates across all three frames of the burst. This enabled the exclusion of common FP sources, such as sun glint or wind induced ripples, since they typically shift position or shape between frames.

To address the challenge of slow flow velocities, which can cause a single item to remain in the camera's field of view across multiple recording intervals, a Kalman filter was implemented for object tracking. This algorithm estimated the velocity vector of valid detections to predict their trajectory in subsequent frames. If a detection in a following interval spatially aligned with the predicted position of a previously detected object, assuming the same class in case of multi-class models, it was matched rather than counted as a new flux event. Finally, the net flux calculation incorporated flow directionality. The velocity vector derived from the burst analysis determined the movement of the object relative to the removal system. Items moving towards the collection point contributed positively to the total count, while objects moving in the opposite direction were subtracted from the flux. This logic ensured that the final count represented the net transport rather than a simple sum of detections.

3.3.4. Temporal Extrapolation Scenarios
Visual surveys and camera monitoring were limited to specific observation windows (e.g., daylight hours), while the Noria removal systems operated continuously. Therefore, extrapolation models are required to estimate the total load for the full testing period. To ensure methodological consistency, identical extrapolation logic was applied to both visual and automated datasets where applicable.

The most fundamental approach employed for both locations was a Linear Extrapolation. This model calculated the mean hourly flux observed during the active monitoring periods and applied this average rate to the entire duration of the field test. This method assumes a constant background flux, which

averages out any variation observed during specific conditions.

However, for the Kooybrug location, the extrapolation was dependent on the flow direction. Litter flux was assumed to be non-zero only during hours when hydraulic conditions favored downstream transport. Valid transport intervals were defined as periods when the pumping station was active or when wind direction data from the nearby De Kooy meteorological station indicated south-west winds. In addition to the first linear scenario, a sensitivity analysis was conducted to address spatial separation at this site. A probabilistic model based on Newbould et al., 2021 was applied to the Kooybrug location to account for possible retention of litter items between the observation and removal point. This resulted in a transport coefficient of 0.48, implying that approximately 48% of the litter passing the bridge reaches the removal system, while the rest is retained by canal banks or vegetation for the specific duration of field tests. The calculation of the retention factor can be found in Appendix C

At the Oostertoegang location, a second extrapolation method was defined based on local wind conditions. This scenario restricted the application of observed flux rates to hours where wind direction supported transport into the collection system. This model filtered out periods with unfavorable wind conditions for litter transport. This possibly depicts a more realistic estimate since it tries to establish a relation between the physical drivers of the specific environment.

3.3.5. Mass Conversion

Bootstrapping for Visual Counting

The conversion of visual count data into mass estimates presents a statistical challenge due to the high variance in litter weights and is therefore one of the biggest sources of uncertainty (González-Fernández et al., 2023). Standard parametric approaches, which assume a normal distribution of item weights, are often unsuitable for riverine debris, since the weight distribution is typically dominated by a heavy tail. To address this, a bootstrapping method was employed to quantify the total mass and its associated uncertainty without relying on assumed probability density functions.

For a given material class c, the total extrapolated count N_c was converted into a mass distribution through iterative resampling. In each of the $k = 10,000$ iterations, a random sample of size N_c was drawn with replacement from the empirical weight datasets $W_c = \{w_1, w_2, ..., w_m\}$. The total mass estimate $M_{c,k}$ for iteration k is calculated as:

$$M_{c,k} = \sum_{i=1}^{N_c} w_{i,k}^{*} \tag{3.10}$$

Where $w_{i,k}^{*}$ represents a randomly selected weight from the source dataset. This process generates a distribution of 10,000 potential total mass values for each class. From this distribution, the median (50^{th} percentile) was extracted as the central estimate, while the 2.5^{th} and 97.5^{th} percentiles defined the 95% confidence interval. This procedure was repeated using the different weight datasets mentioned in 3.1.4.

Bounding Box Regression

A limitation of count-based quantification is the loss of information about the size of litter items. A small fragment and a large item are both registered as single counts despite their varying size and masses. To overcome this, a regression-based approach was used to estimate mass directly from the bounding box dimensions generated by the object detection model. The mass estimation consists of two stages. First, the pixel dimensions of the bounding box are converted into real-world units (centimeters) using the known camera specifications and height above the water surface. Second, these physical dimensions serve as inputs for a regression model trained to predict item mass.

The regression models were developed using the IMR dataset (de Lange et al., 2023), which contains length, width, and mass measurements for approximately 14,000 items. To establish the area-mass relationship, the dataset was partitioned into an 80% training set and a 20% testing set. Three regression architectures were evaluated: Linear Regression (predicting the mean), Random Forest (predicting the mean), and Quantile Regression (predicting the median).

4

Results

4.1. Model Training

4.1.1. In Domain Performance

Comparison of Label Setups

To evaluate the impact of added classes on model performance, the weighted mAP50-95 was compared between the three label setups. Figure 4.1 illustrates the performance of all trained models within each setup, with error bars representing the minimum and maximum scores achieved across different training strategies.

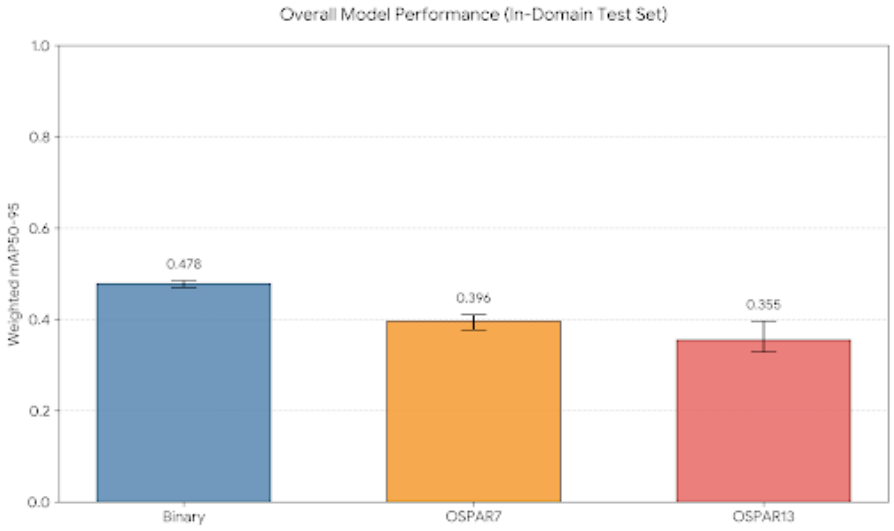


Figure 4.1: Comparison of overall model performance (Weighted mAP50-95) on the in-domain test set. The bars represent the mean performance, while the error bars indicate the range (min-max) across all trained model variations.

The Binary models achieved the highest mean weighted mAP of 0.478. The introduction of material-based categorization resulted in a performance drop, with the OSPAR7 setup decreasing to a mean mAP of 0.396. The OSPAR13 models further subdivide plastics into polymer types, which resulted in the lowest mean score of 0.355.

Beyond the mean performance, the spread of results (indicated by the error bars) highlights the variation in model stability. The Binary models show a narrow performance range of 1.5%, suggesting that variations in training data or hyperparameters have minimal effect. In contrast, the OSPAR13 models displayed a significantly wider spread of 6.7%, indicating that complex multi-class models are more sensitive to training strategies, such as the inclusion of synthetic data or HP improvement techniques.

Impact of Synthetic Data Integration
Figure 4.2 illustrates the Weighted F1-Confidence curves across all label setups.

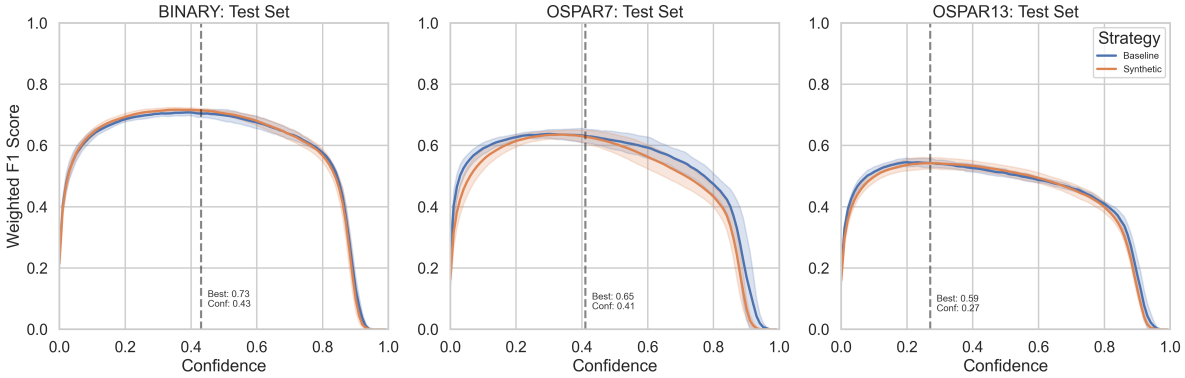


Figure 4.2: Comparison of Weighted F1-Confidence curves between Baseline models (Real data only) and Synthetic models (Real + Synthetic data) on the in-domain test set. The curves represent the aggregated average of all trained models.

The results on the in-domain test set show a high degree of similarity between the Baseline and Synthetic models. The F1 trajectories are nearly identical across the entire confidence threshold range. While the synthetic models achieved a marginally higher peak F1-score compared to the baseline, this difference is minimal. Given the variance observed in model training, this slight improvement is likely explained by stochastic variations in weight initialization or batch sampling rather than a fundamental information gain from the synthetic domain. This suggests that for the available real-world test set, the existing real training data was sufficient to learn the necessary features, or that the domain gap between the synthetic and real imagery limited the transfer of learnable patterns for this specific test environment.

Further analysis was conducted to determine if specific integration strategies, such as varying the ratio of synthetic data in the training set (25%, 50%, 100%) or employing a pretraining (synthetic data) and finetuning strategy (real data). The comparison indicates no distinct trend since all synthetic variations clustered around the baseline performance. This reinforces that increasing quantity of synthetic data does not translate to in-domain accuracy gains without addressing the underlying domain gap.

Error Analysis: Detection vs. Classification
To further investigate the performance trade-offs between single and multi-class detections as well as the impact of synthetic data, the error counts of the three label setups are presented in figures 4.3 and 4.4.

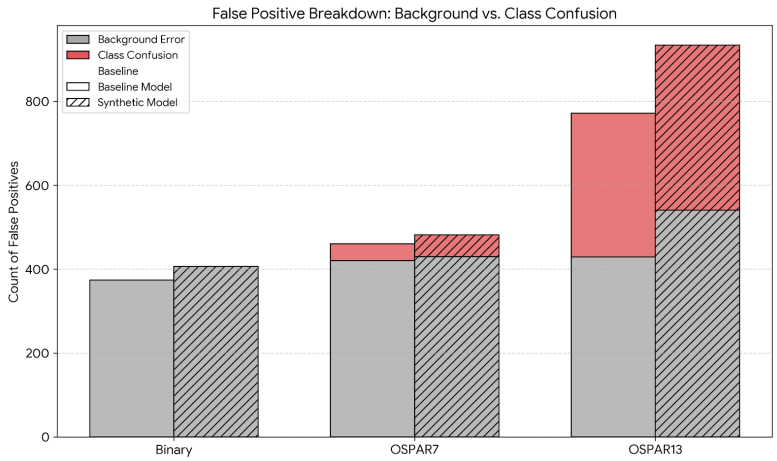


Figure 4.3: Breakdown of FP by error type across label setups. Background Error refers to non-litter objects detected as litter, while Class Confusion refers to correctly detected litter assigned the wrong material label. Hatched bars indicate models trained with synthetic data.

For the Binary setup, all FPs are inherently background errors since only one class exists. However, as the amount of classes increases, a significant portion of FPs is attributed to class confusion (red segments). OSPAR7 models maintain low class confusion, mostly due to low representation of the rare material classes (Wood, Paper, etc.) in the test set. Therefore, the OSPAR7 models mainly have to detect the plastic class, which closely resembles a binary detection task and is reflected in the low performance decrease compared to the single-class task. In contrast, OSPAR13 models struggle to distinguish between specific polymer types, as the amount of miss-classified objects increases. Nevertheless, the performance drop is mostly driven by the confusion of classes (wrong labels) instead of missing objects altogether, as the number of background errors remains relatively consistent between single and multi-class models. Furthermore, the integration of synthetic data consistently resulted in a slight increase in Background Errors across all setups. This suggests that while synthetic data introduces new features, it may also cause the model to detect background textures (e.g., reflections or foam) as debris more frequently than the baseline.

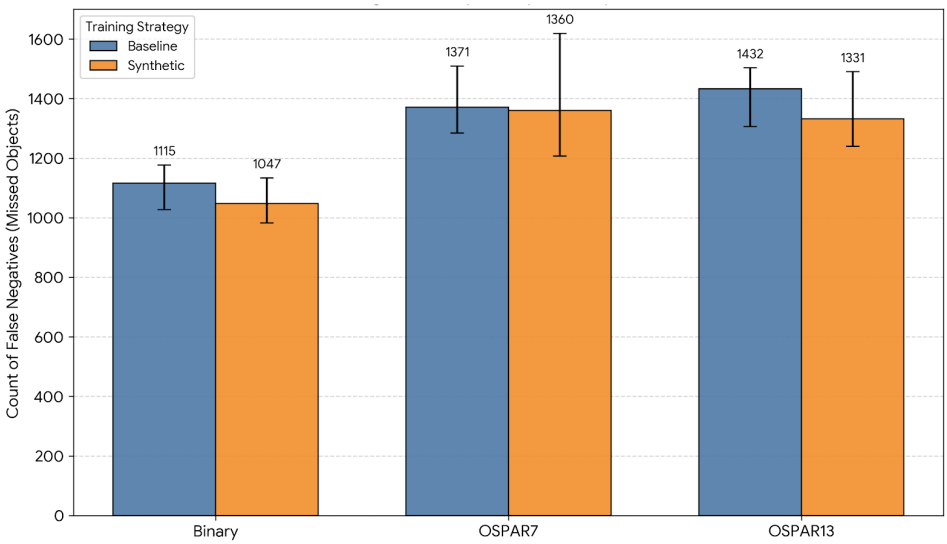


Figure 4.4: Comparison of FNs between Baseline and Synthetic models. The error bars represent the range of minimum to maximum of missed objects across all trained models. The number above bars indicate the median FN count.

In contrast, figure 4.4 highlights the behavior of FNs or missed detections among trained models. Across all setups, models trained with synthetic data (orange bars) consistently missed fewer objects than their baseline counterparts. For example, in the OSPAR13 setup, the average number of missed objects dropped from 1433 (Baseline) to 1332 (Synthetic). This explains the neutral impact of synthetic data observed in the overall F1 and mAP scores. While it seems that synthetic data improves the model's ability to detect more objects (lowering FNs), the simultaneous increase in FPs cancels out any overall performance gain. Thus, while the overall metric remained static, the underlying behavior of the model shifted towards higher recall at the cost of precision.

Material-Specific Performance and Confusion

While aggregated metrics indicate a general performance decline in the OSPAR13 setup, this degradation is not uniform across all materials. Figure 4.5 presents the Precision-Recall curves for the individual OSPAR13 categories.

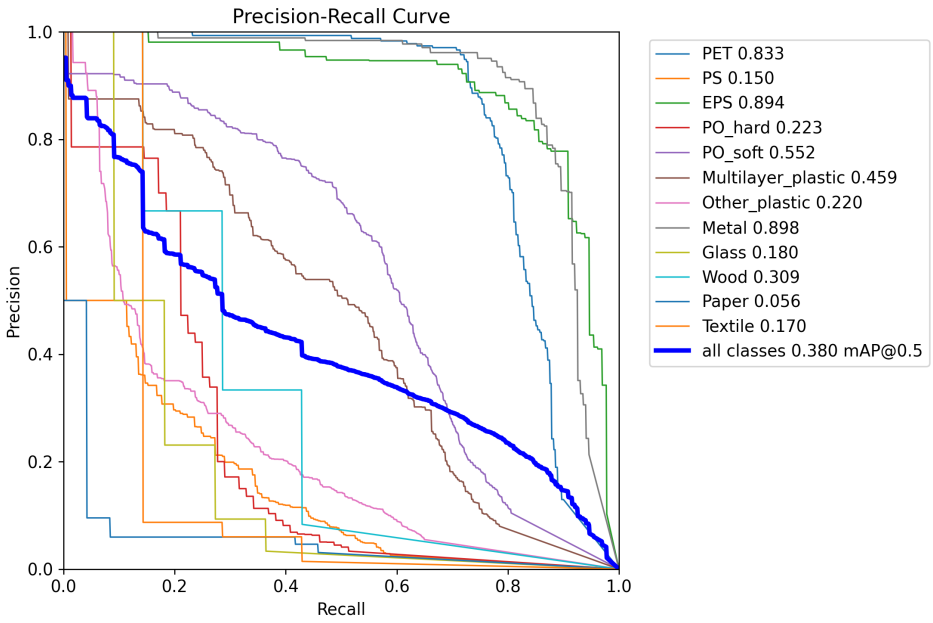


Figure 4.5: Class-wise Precision-Recall curves for the OSPAR13 setup at IoU=0.5

Homogeneous classes with consistent items, such as PET (mostly bottles) and Metal (mostly drinking cans), achieve a high Area Under the Curve (AUC), maintaining high precision even at higher recall levels. Heterogeneous categories such as PS (straws, cutlery, cups) and PO-hard (fragments, containers) show significantly lower performance. These curves plateau early, indicating that OSPAR13 models struggle to consistently identify these items without incurring a high rate of FPs or FNs. Medium performance is observed in semi-homogeneous classes like PO-soft and Multilayer. Figure 4.6 represents a normalized confusion matrix and indicates that these classes are often confused with each another, most likely due to similar appearances. This matrix visualizes the probability of a true class (y-axis) being predicted as a certain output class (x-axis).

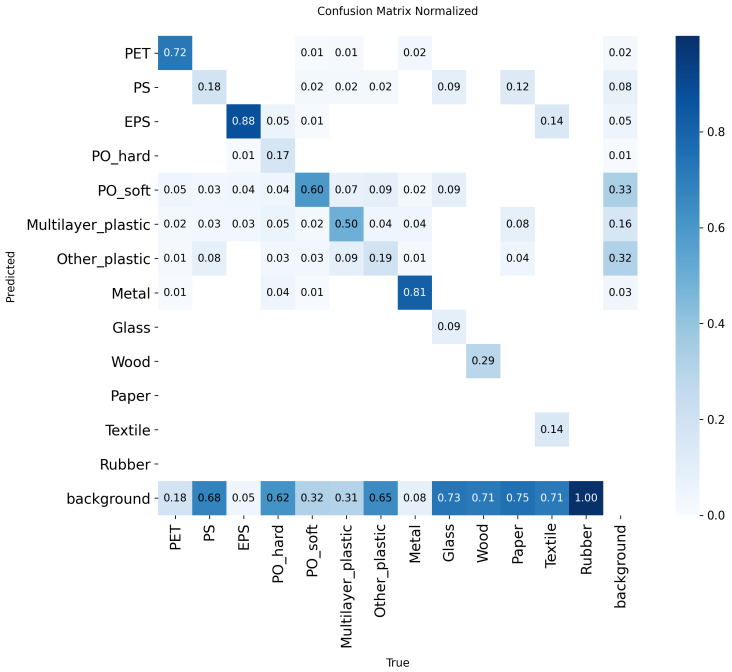


Figure 4.6: Normalized Confusion Matrix for the OSPAR13 model. Values represent the proportion of ground-truth instances (rows) predicted as each category (columns). The diagonal represents correct classification, while off-diagonal elements indicate confusion.

A significant cluster of confusion is observed among the polyolefins. A high percentage of PO-soft, Multilayer and Other items are misclassified. This is attributable to the visual ambiguity of these classes. Furthermore, the classes PO-soft and Other plastic struggle with FNs and often predict objects (around 30%) when none are present. Especially light reflections and smaller natural debris can look similar to actual litter in static frames. The lower performing classes such as PO-hard and PS are less confused, but a significant amount remains completely undetected. This can be partly explained by an under-representation in annotation quantity. More labeled examples could boost performance, although a detection and classification limitation will probably remain due to different shapes. Objects from other underrepresented classes such as Glass, Wood, Paper, Textile and Rubber are mostly not detected at all. In comparison to PS and PO-hard, there were even less examples in the training set for the model to learn from. Consequently, not enough examples are provided in the test set, which limits the informative value of the results for these classes. However, a few examples got already detected with no intra-class confusion. Here, synthetic data was also not able to improve class-wise detection. More real annotations examples are necessary to evaluate the suitability of these classes for litter detection. A detailed overview of class-wise performance of synthetic and baseline models is presented in Appendix D

Location-Dependent Performance Variability

The previous analysis demonstrated that increasing the amount of classes generally leads to a performance tradeoff. However, this tradeoff is not uniform across all environments. The in-domain test set is composed of different locations with varying camera heights, water conditions, and object distances. Figure 4.7 illustrates the weighted PR curves split by test location.

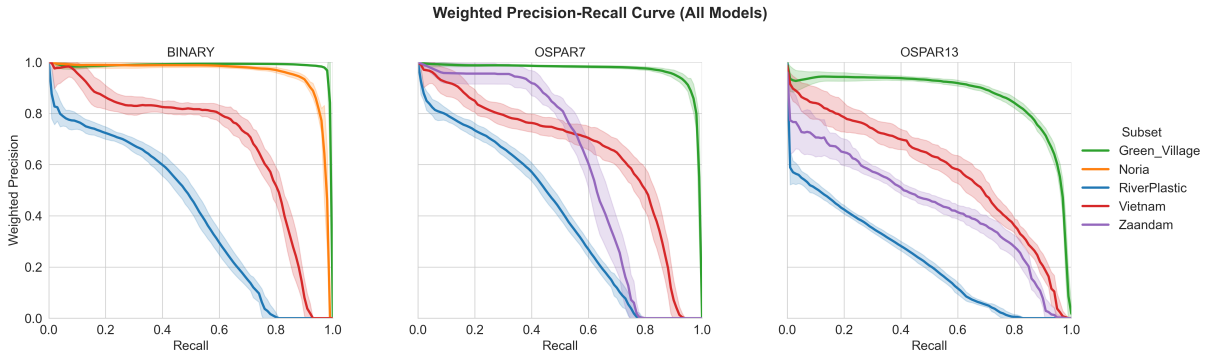


Figure 4.7: Weighted PR curves across model setups differentiated by test location.

In locations such as Vietnam (red lines) and Green Village (green lines), the PR curves for the detailed OSPAR13 setup maintain a high AUC in comparison to Binary models. In these datasets, the camera is typically positioned closer to the water surface, resulting in a higher pixel density per object. This suggests that when spatial resolution is sufficient, detailed classification is more feasible. Notably, litter items in these datasets are less fragmented and retain their original appearance, which contributes towards an easier identification. The Noria dataset was only used for the Binary models, since detailed labels were not available. However, it is expected that a performance behavior similar to the Green Village dataset can be achieved between Binary and OSPAR models since the single-class performance shows already a high AUC.

In contrast, the Zaandam dataset (excluded from Binary) shows a more significant decrease between OSPAR7 and OSPAR13. This location is characterized by an intermediate distance between the water surface and the camera, which was deployed in two configurations. Next to the use of a standard lens for this distance, an ultra wide lens was also used, which causes litter items to be distorted at the edges. Furthermore, a variety of smaller and fragmented items are present, which adds difficulty. The RPD dataset is the most difficult location overall for all setups. Images typically exhibit a high quantity of small litter items, which makes general detection challenging. Therefore, a more detailed detection and classification as in OSPAR13 is difficult and highly sensitive to environmental constraints. For monitoring from high bridges, a Binary or simplified OSPAR7 approach is preferable to maintain reliability. However, for setups with lower camera positioning, the OSPAR13 model can better distinguish between material types.

4.1.2. Out of Domain Performance

Figure 4.8 illustrates the performance drop observed when transitioning from the In-Domain test tet to the Out-of-Domain (OOD) locations.

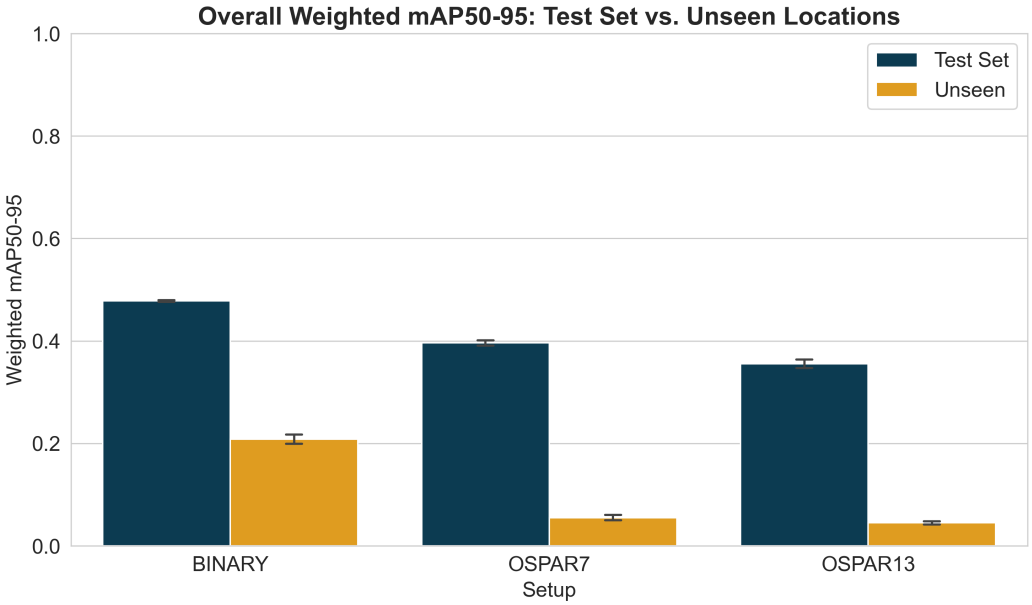


Figure 4.8: Mean weighted mAP50-95 results for trained locations (blue) and unseen locations (yellow). The error bars indicate the range of results achieved across all trained models.

A performance decrease is evident across all label setups. For the Binary models, the mean mAP drops from ≈ 0.48 to ≈ 0.20. The decline is similar for the multiclass setups, where a performance drop of around 30% is observed. This indicates that the models have learned to detect litter in familiar settings, but struggle to transfer this knowledge to novel environments. While Binary models were tested across 4 locations (Zaandam, Maastricht, Oostertoegang, Kooybrug), the OSPAR models were only evaluated for 2 locations (Oostertoegang, Kooybrug). Furthermore, the number of images per unseen location varies from over 800 images for Zaandam to around 80 images for both Oostertoegang and Kooybrug. Therefore, results in figure 4.8 only give an indication about true performance since they can vary depending on the sample size. This is especially a concern for the OSPAR locations since not all classes are present, which introduces bias. More images are needed for testing to achieve full representativeness.

Synthetic data yielded negligible improvements on the In-Domain test set. However, figure 4.9 shows a trend in the Out-of-Domain analysis when comparing Baseline and Synthetic models on unseen locations.

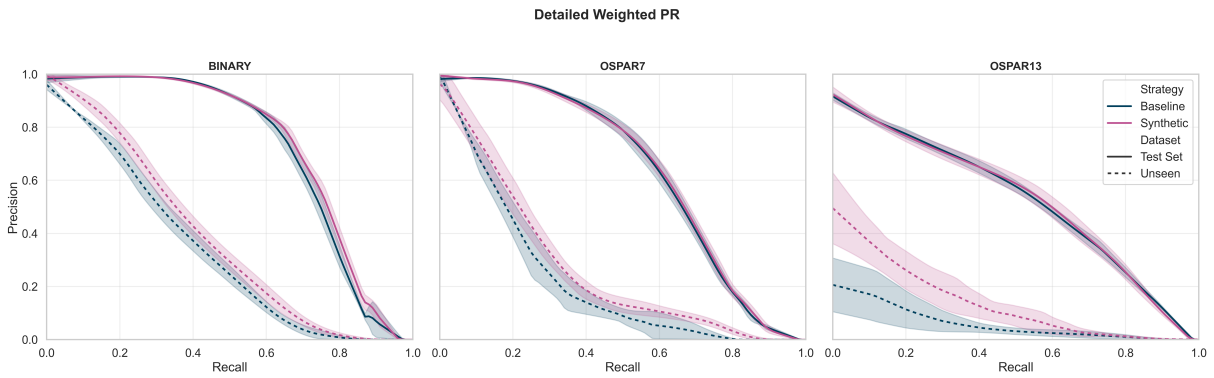


Figure 4.9: Weighted PR curves comparing In-Domain performance (Solid lines) vs. Out-of-Domain performance (Dashed lines).

In contrast to the in-domain results, the synthetic models consistently perform better than the baseline models across all three label setups in the OOD environments. Although the absolute performance gain is modest, the trend is uniform. The integration of synthetic data shifts the PR curve upwards, improving both precision and recall. This suggests that the synthetic dataset helps the model learn more generalized feature representations that are less overfitted to the specific characteristics of the training locations.

The weighted PR curves hide the variability between the individual sites. As mentioned, the difficulty of a location can be influenced by physical factors such as camera height as well as the sample size used for model evaluation. Figure 4.10 details the performance of Binary models for the Maastricht and Zaandam location which were only used for the Binary setup.

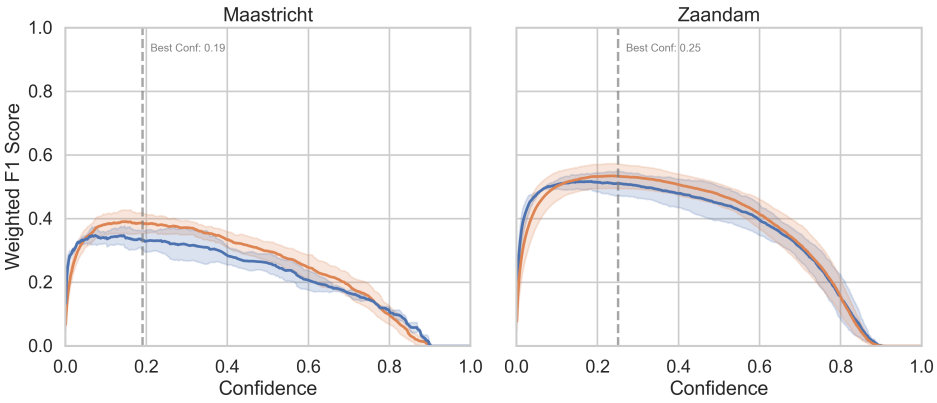


Figure 4.10: Weighted F1 curves for baseline and synthetic models for the Binary setup

The results highlight the variance between locations. The Zaandam dataset comprises the largest sample size and offers the most statistically robust assessment. Here, the models achieve a moderate peak F1 score of close to 60%, while the Maastricht location seems more difficult for models since the peak F1 score is significantly lower at around 40%. The images in Zaandam were taken from a lower camera height, which makes detection in general easier and could be one of the reasons for the performance difference. Another explanation for the performance difference could be the type of litter that is observed. The images at the Zaandam location were taken from a catch and release experiment, which featured different types and sizes of litter. However, some items, such as the PET bottles, were deliberately placed in the canal and did not experience any environmental degradation or bio-fouling. In contrast, the Maastricht images contain litter items that entered the environment naturally and likely have been present for a longer duration, thus making processes such as degradation or fragmentation more likely.

The other unseen locations (Kooybrug and Oostertoegang) also feature natural litter and were evaluated for all label setups. Figure 4.11 shows the performance behavior at these locations for all models.

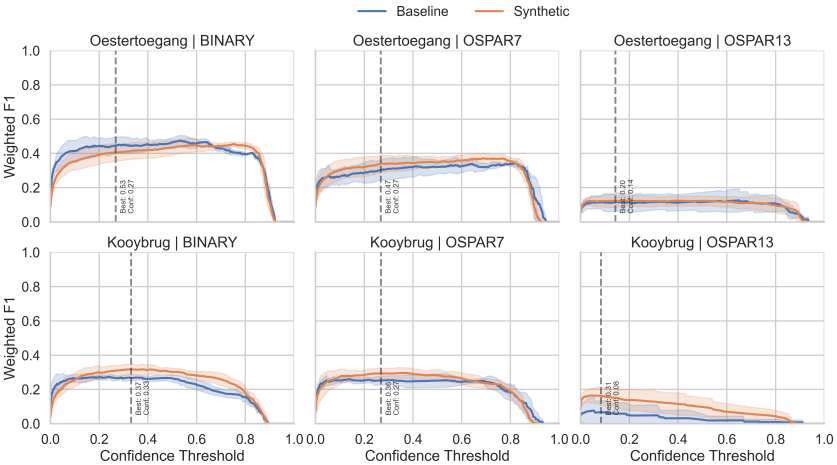


Figure 4.11: Weighted F1 performance of Binary, OSPAR7 and OSPAR13 models on unseen locations.

Even though both datasets are comprised of images with natural litter, the models on the Kooybrug location achieve a performance of around 10 to 20% lower than on the Oostertoegang location for the Binary and OSPAR7 setups. For the OSPAR13 models, the performance for the Kooybrug location is initially higher for low confidence values, but is overall more unstable and declines faster with an increasing confidence threshold compared to the Oostertoegang results. Here, the camera height above the water surface plays a role, since the Kooybrug images were taken from a height of around 9m, while the camera at the Oostertoegang was placed less than 2m above water surface. The latter makes objects much bigger and in theory easier to identify, though models have never been trained on such a short distance. Therefore, natural items, such as leaves and foam, appear bigger and could be mistaken for litter. Lastly, different camera lenses (including an ultra wide lens) were used to capture as much as possible of the canals cross-section. This distorts objects especially at the edge of the images. All models likely struggle with this perspective since only a few examples of images with similar camera settings were included in the original training set. The Kooybrug location features a more known camera setup for the models, however litter items not only appear much smaller on the images due to the height, but they usually are much more fragmented due to the more rural setting. In addition, litter items are often trapped in larger tree branches, adding difficulty to the detection and classification task. Lastly, the number of images used for the evaluation on these two locations was limited, due to the manual labeling effort. Consequently, not all class for the OSPAR models were represented in the evaluation. It is expected that the performance of OSPAR models increases with more examples, since easier classes like PET and EPS (as depicted in Figure 4.5) are not present. Figure 4.12 shows example images of each location including predictions of each model. The ground truth is given with OSPAR13 labels, while the number next to predictions indicate the confidence score of that specific detection.

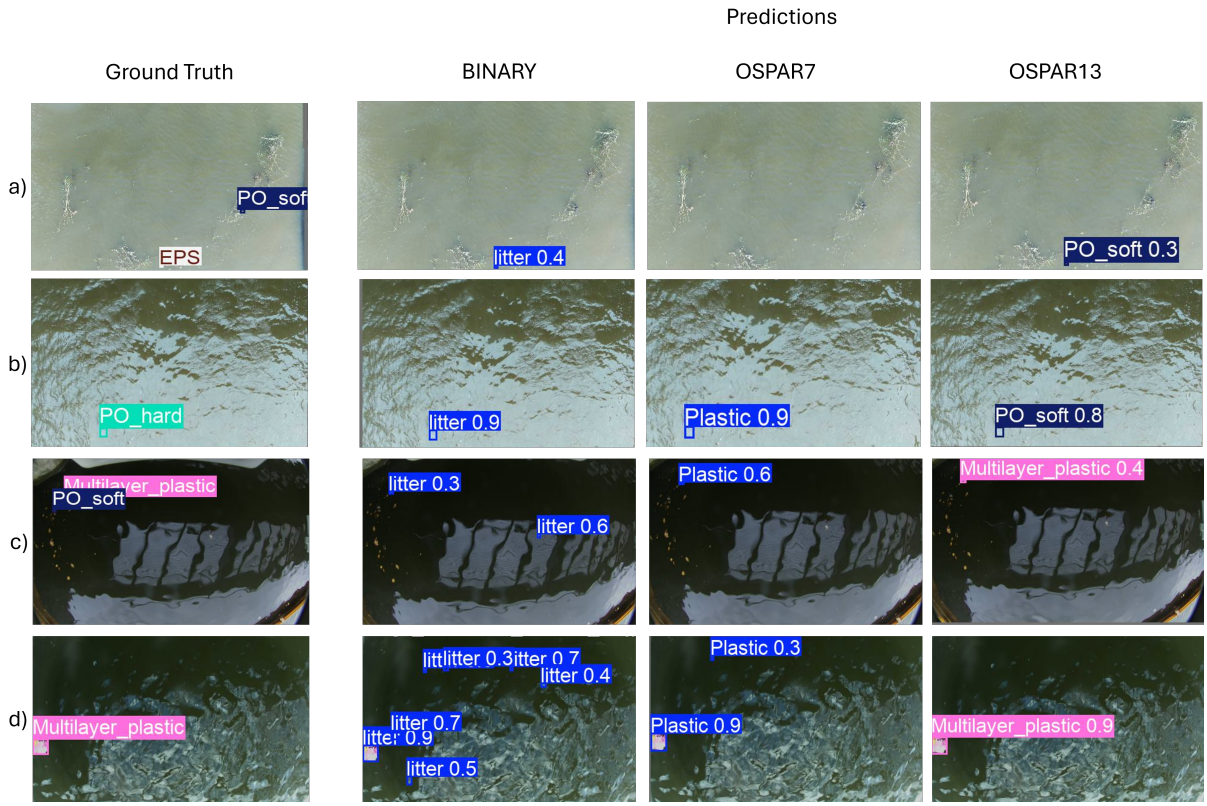


Figure 4.12: Example images from the OOD test set for the Kooybrug (a, b) and Oostertoegang (c, d) location. Example a) shows natural vegetation with small trapped litter items, which are only partly detected with low confidence scores. Image b) shows a bigger item in difficult lighting conditions, where all models detect the item with a higher confidence, but the OSPAR13 model miss-classifies example b). Image c) shows an example of the ultra wide lens used, where only one item at the edge is detected with a low confidence. Example d) shows a bigger and eaiser to detect ite„ though reflections cause FPs for Binary and OSPAR7 models.

The images from Kooybrug are characterized by small, low-contrast objects. As observed in the sample images, the primary error is omission. Small plastic fragments that blend into the dark water or are partially occluded by organic debris are frequently missed by both baseline and synthetic models. However,

the models rarely produce false detections and are generally able to distinguish water reflections as well as vegetation from actual litter. This results in low FP rates but overall lower mAP and F1 due to poor recall.

In contrast, the Oostertoegang samples reveal that models are much more active in predicting litter items. The higher resolution allows the model to successfully detect varied litter, including those with complex shapes. However, this comes with a trade-off. The examples show errors, where sun glint, foam patches, and ripples are confidently bounded as plastic. Furthermore, for the multi-class OSPAR13 models, objects are usually correctly localized, the specific label often fluctuates due to ambiguous litter items.

Following the findings from visual inspection, error dynamics are quantified in figure 4.13 with FP rates per ground truth object across the test set and the unseen locations.

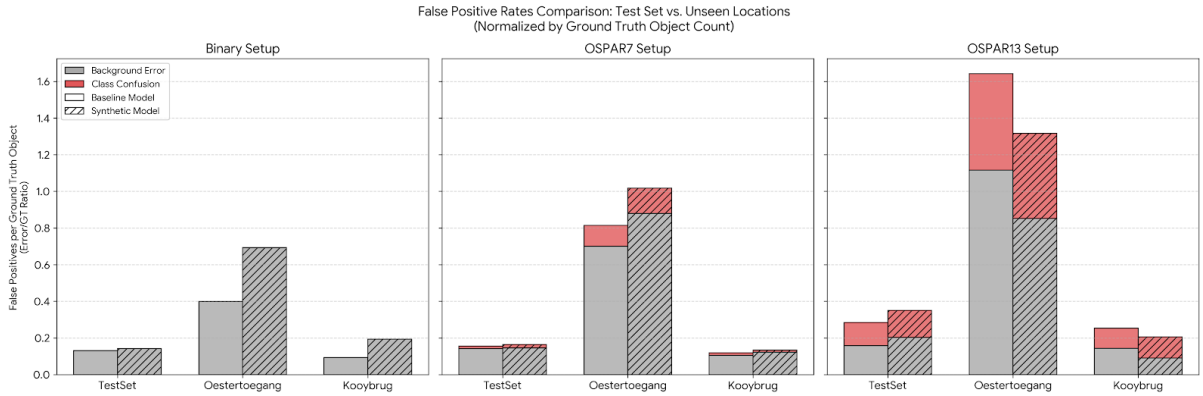


Figure 4.13: Comparison of False Positive Rates normalized by Ground Truth object count.

All models maintain a low error rate (< 0.25 errors per object) at the Kooybrug location, due to general detection difficulty induced by the distance. In contrast, models at the Oostertoegang location suffer from noise, with OSPAR13 models generating nearly one background error for every real object detected. This indicates that the better F1 performance at Oostertoegang is driven by recall, despite the detection output being significantly noisier than at Kooybrug. Another trend observed in figure 4.13 is the behavior of the synthetic models. Across all setups at Oostertoegang (except OPAR13), the synthetic models exhibit a markedly higher FP rate than the baselines (e.g., increasing from ≈ 0.4 to ≈ 0.7 in the Binary setup). This indicates that synthetic models are more aggressive than baseline models. Although this behavior leads to more FPs, the gain in TPs is enough to score a higher overall mAP and F1 scores. For operational monitoring, this implies that synthetic models provide higher noise, which is superior for ensuring less litter is missed.

4.1.3. Benchmarking

The benchmarking step was prioritized for the Binary and OSPAR13 setups, since the performance behavior between single and multi-class detection can be sufficiently evaluated for these two setups (OSPAR7 results would likely fall in between). Due to time constraints, only one model per setup was replicated on the different architectures. Therefore, possible performance ranges cannot be shown. Figure 4.14 presents the PR curves for the three architectures on the test set. The results indicate a superior detection and classification capability of the YOLO architecture across both task setups. In the Binary task, YOLOv11m maintains high precision even as recall increases, indicated by the largest AUC. While Faster R-CNN achieves competitive precision at lower recall levels, its performance degrades slightly faster at higher recall levels than the YOLO model. RT-DETR exhibits a drop in precision even earlier in the recall range, suggesting increased FPs. The OSPAR13 setup further amplifies these disparities, with the YOLOv11m model exhibiting a more robust trade-off, whereas both comparative models struggle to achieve high recall without a severe penalty in precision.

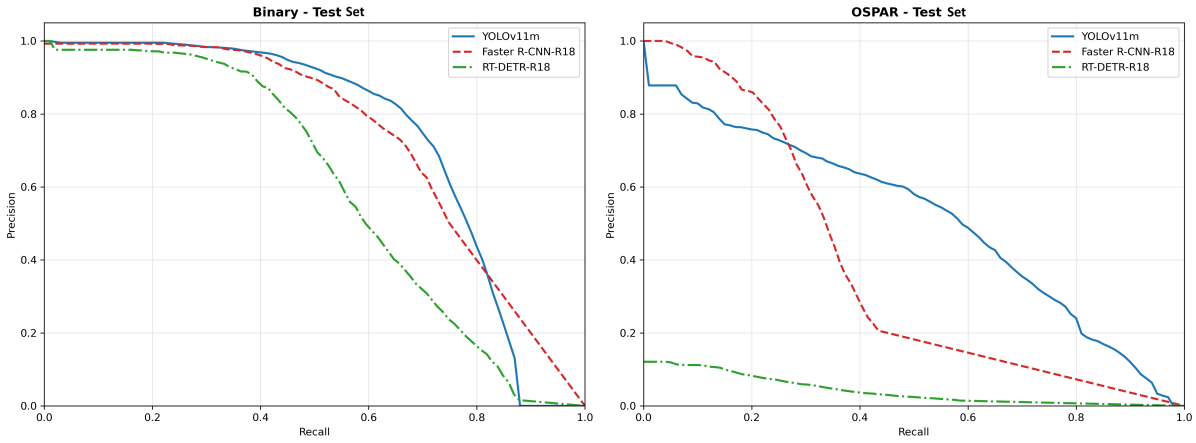


Figure 4.14: Weighted PR curves for YOLOv11m, Faster R-CNN-R18, and RT-DETR-R18 on the In-Domain Test Set. The YOLOv11m architecture achieves a superior area under the curve for both Binary and OSPAR tasks compared to the alternative architectures.

The evaluation on the unseen dataset, illustrated in Figure 4.15, highlights the significant challenges of domain generalization. All architectures show a performance attenuation compared to the test set, though the degree of degradation varies substantially. The YOLOv11m model maintains a curve that, while lower than in the in-domain test, still demonstrates a responsive trade-off between precision and recall.

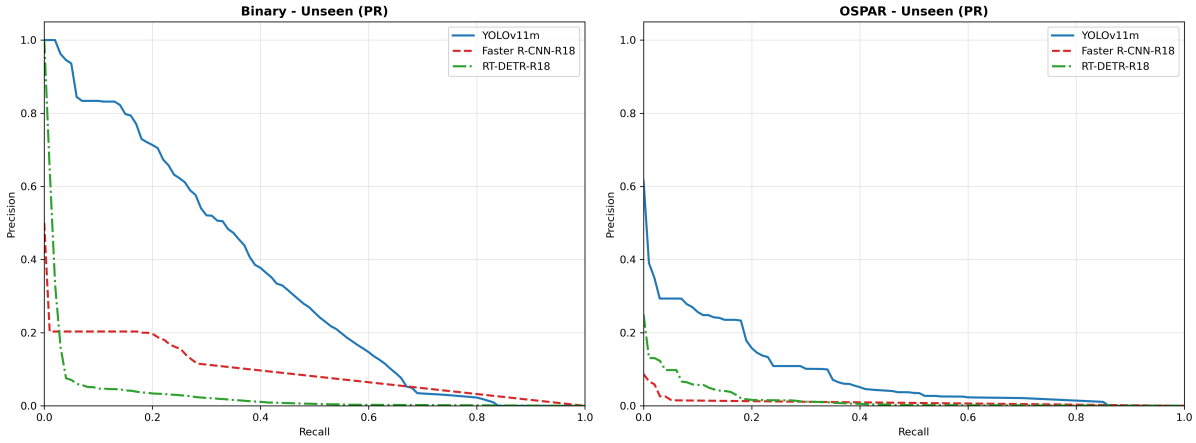


Figure 4.15: Precision-Recall curves evaluated on the Unseen dataset. While all models suffer performance drops due to domain shift, the comparative architectures (Faster R-CNN and RT-DETR) exhibit a collapse in precision, whereas YOLOv11m retains a functional detection capability.

In contrast, the alternative architectures exhibit a significant collapse in performance on the unseen data. The Faster R-CNN model flattens at a very low precision level (approximately 0.2) in the Binary task, indicating that the model produces a high volume of false positives regardless of the recall threshold. Similarly, the RT-DETR model fails to generalize, yielding negligible precision across the recall range for both tasks. This stark contrast underscores that while the YOLOv11m model is affected by the domain shift, it possesses a higher degree of robustness compared to the Faster R-CNN and RT-DETR implementations tested.

Further analysis of the F1 scores across confidence thresholds (shown in Appendix E) reveals characteristics regarding stability. The Faster R-CNN architecture demonstrates high stability across a broad range of confidence thresholds, with a relatively flat F1 curve. This suggests that its probability output is well-calibrated despite the lower overall peak performance. In contrast, the RT-DETR model exhibits significant instability, where F1 scores peak sharply around specific confidence values and decay rapidly elsewhere. This behavior implies a narrow optimal operating window, making deployment in variable environments challenging. The benchmarking confirms that the limitations observed in previous sections,

specifically the difficulty of OSPAR13 classification and the performance drop in OOD locations, are data-centric challenges (resolution, class ambiguity) rather than architectural deficiencies. Ultimately, the YOLOv11m architecture provides the highest generalization performance while avoiding the sensitivity of the confidence threshold observed in the transformer-based alternative.

4.2. Mass-Flux Estimation

4.2.1. Ground Truth Composition of Recovered Litter

To evaluate the mass flux estimations of the models, the physical debris collected by the removal systems at both locations was analyzed. This ground-truth data serves as the benchmark against the extrapolations based on visual and camera observations.

Den Helder (Kooybrug)

The Kooybrug location represents a rural environment with overall low item fluxes. All items collected between September 2nd and September 12th, 2025, were manually sorted, cleaned, and categorized according to the OSPAR monitoring guidelines. To ensure consistency with the deep learning framework, the 133 OSPAR sub-categories were aggregated into the 13 primary material classes used during training. Mass measurements were conducted on dried items using a precision scale with a resolution of 0.01 g.

The composition of the catch is detailed in Table 4.1. A total of 320 items were collected over the 10-day period, with a cumulative mass of 444.20 g. This yields an average item weight of approximately 1.38 g, indicating a high degree of fragmentation.

Table 4.1: Ground-truth data from the Noria removal system, Den Helder. (Total Items: 320, Total Mass: 444.20 g).

Material Class	Number of Items	Percentage (Count)	Weight (grams)	Percentage (Weight)
PO hard	124	38.8%	121.89	27.4%
PO soft	120	37.5%	12.52	2.8%
Multilayer	49	15.3%	22.46	5.1%
EPS	11	3.4%	4.32	1.0%
Metal	5	1.6%	78.64	17.7%
PS	3	0.9%	5.12	1.2%
Rubber	3	0.9%	0.28	0.1%
Wood	2	0.6%	196.89	44.3%
Other Plastic	2	0.6%	1.68	0.4%
Paper	1	0.3%	0.40	0.1%
PET	0	0.0%	0.00	0.0%
Textile	0	0.0%	0.00	0.0%
Glass	0	0.0%	0.00	0.0%
Total	**320**	**100.0%**	**444.20**	**100.0%**

The breakdown reveals a disparity between item count and mass contribution. In terms of abundance, the catch was dominated by plastics, specifically hard polyolefins (38.8%) and soft polyolefins (37.5%). Combined with multilayer items (15.3%), these three plastic categories accounted for 91.6% of all collected items. However, their contribution to the total mass was low. In contrast, the Wood category, which consisted of only two individual items (0.6% of the count), was the largest contributor to mass, accounting for 44.3% of the total weight. Similarly, metal items represented only 1.6% of the count but contributed 17.7% of the mass. This underscores the challenge of using item counts as a proxy for plastic load, as the vast majority of plastic items collected were lightweight fragments rather than intact objects.

Oostertoegang (Amsterdam)

In contrast to the rural setting of Den Helder, the Oostertoegang location represents a high-flux urban environment. The ground-truth data was collected over a shorter timeframe, from September 12th to September 15th, 2025, yet yielded a significantly larger volume of debris. The processing methodology remained identical to the Den Helder case study.

As presented in Table 4.2, a total of 1,490 items were retrieved, with a combined mass of 27.07 kg. The average item weight in this urban setting was approximately 18.1 g, more than an order of magnitude

higher than in Den Helder. This indicates the presence of larger, more intact items such as bottles, packaging, and wooden logs.

Table 4.2: Ground-truth data from the Noria removal system, Oostertoegang (Amsterdam). (Total Items: 1,490, Total Mass: 27,067.35 g).

Material Class	Number of Items	Percentage (Count)	Weight (grams)	Percentage (Weight)
EPS	398	26.7%	1385.28	5.1%
Paper	385	25.8%	1214.68	4.5%
PO hard	219	14.7%	1003.51	3.7%
PO soft	180	12.1%	429.02	1.6%
Multilayer	92	6.2%	291.03	1.1%
Metal	55	3.7%	949.42	3.5%
PET	53	3.6%	1296.46	4.8%
Wood	42	2.8%	17339.19	64.0%
PS	34	2.3%	180.35	0.7%
Glass	10	0.7%	2363.81	8.7%
Other Plastic	9	0.6%	4.10	0.0%
Rubber	7	0.5%	4.13	0.0%
Textile	6	0.4%	606.37	2.2%
Total	**1,490**	**100.0%**	**27,067.35**	**100.0%**

The data from Amsterdam amplifies the count-mass discrepancy observed in Den Helder. The numerical composition was dominated by expanded polystyrene (EPS, 26.7%) and paper (25.8%), reflecting urban consumption and packaging waste. While plastics accounted for 66.1% of the total count (985 items), their combined mass contribution was low at just 16.9% (4.59 kg). The mass distribution was instead heavily skewed by rare, high-density items. The Wood category accounted for 64.0% of the total mass, a number driven largely by a single log weighing 11.5 kg. Similarly, glass items, representing less than 1% of the count, contributed nearly 9% of the mass. This reinforces the finding that in both rural and urban contexts, the plastic flux is numerically dominant but in terms of weight secondary to denser anthropogenic materials. Consequently, monitoring systems that rely solely on item counts risk misrepresenting the true material transport in the river system.

4.2.2. Validation of Area-to-Mass Conversion Models

To enable the conversion of bounding box detections into mass estimates, the relationship between object area (cm^2) and mass (grams) was analyzed using three regression techniques: Linear Regression, Quantile Regression (Median), and Random Forest. These models were trained and evaluated (80, 20 split) on a dataset of approximately 14,000 items with measured dimensions and weights. The aggregated performance on the overall class (all classes combined) reveals the challenges of establishing a universal area-mass proxy. Figure 4.16 compares the coefficient of determination (R^2) for the three model architectures on both Train and Test sets.

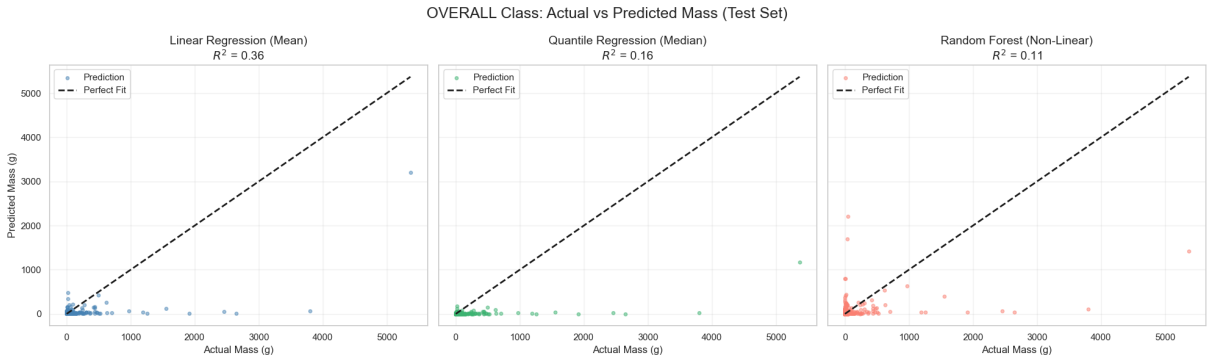


Figure 4.16: Comparison of R^2 performance for the overall class across three regression models.

The analysis indicates that a universal model generally correlates poorly ($R^2 < 0.2$) due to the extreme

density variations between material types (e.g., Styrofoam vs. Glass). The Random Forest model exhibited severe overfitting, achieving an R^2 of 0.62 on the training set but dropping to 0.11 on the test set. The Quantile Regression model (Median) proved to perform slightly better on unseen data, achieving a test R^2 of 0.163 and the lowest Mean Absolute Error (MAE) of 14.58 g. While the Linear Regression model achieved the highest Test R^2 of 0.36, suggesting it captures the variance driven by outliers (e.g., heavy logs) effectively. However, this sensitivity to outliers resulted in a higher MAE of 19.64 g. This indicates that for the majority of typical litter items, the median-based approach provides a more accurate weight estimate than the mean-based linear model. Consequently, the Quantile Regression model was selected for the estimation of the mass flux to provide conservative median weight estimates and prediction intervals (10th–90th percentiles).

While the overall correlation is weak, the area-mass relationship is highly material-dependent. Table 4.3 details the Test Set R^2 scores for individual material classes.

Table 4.3: Test set R^2 scores across material classes.

Material Class	Linear Reg	Quantile Reg	Random Forest
High Correlation			
Textile	0.84	0.84	0.19
Glass	0.79	0.77	0.14
Rubber	0.44	0.52	0.48
Paper	0.41	0.47	0.49
Moderate Correlation			
EPS	-0.04	0.06	**0.44**
PS	0.35	0.17	-2.39
PO hard	0.11	**0.40**	-0.74
PO soft	0.25	0.17	-0.67
Low/Negative Correlation			
Multilayer	0.05	0.02	0.13
PET	-1.79	-0.09	-9.51
Metal	-0.62	-0.14	-3.32
Wood	-0.06	-0.84	-3.26

The results demonstrate that materials with consistent densities and shapes, such as textile ($R^2 = 0.84$) and glass ($R^2 = 0.79$), can be accurately estimated based on area. In contrast, classes that are either highly deformable (e.g., multilayer) or have variable thickness (e.g., wood) show negligible or negative correlations. For these classes, the regression model effectively defaults to predicting the median weight of the training distribution rather than an instance-specific value.

4.2.3. Validation of Monitoring Methods

To evaluate detection performance of the full unlabeled data from the entire testing duration, model predictions are compared with human visual observations for the same monitoring windows at both locations.

Kooybrug

Table 4.4 summarizes the model predictions at the Kooybrug during the overlapping observation windows on September 2nd and 4th, 2025.

Table 4.4: Direct comparison of Visual Counting vs. Object Detection counts at Kooybrug.

Date	Time Slot	Visual Count	VS Count per Class	Binary Detections	OSPAR Detections (Total & Class-wise)
02/09	13:00-14:00	**3**	PO-soft: 2 Metal: 1	**2**	**0** (OSPAR7) **2** (OSPAR13) (Other: 1, PO-hard: 1)
02/09	14:30-15:30	**1**	PS: 1	**13**	**7** (OSPAR7) (Plastic: 7) **40** (OSPAR13) (PO-soft: 10, Other: 16 Multilayer: 6, PS: 4, PET: 3, PO-hard: 1)
04/09	11:48-12:48	**3**	Paper: 1, PO-soft: 1 Other: 1	**58**	**45** (OSPAR7) (Textile: 15, Plastic: 25, Paper: 2, Glass: 3) **101** (OSPAR13) (Other: 51, PET: 26 Multilayer: 8, PO-soft: 11, PO-hard: 3, PS: 2)
04/09	13:50-14:50	**5**	PO-soft: 2, Paper: 1 EPS: 1, PO-hard: 1	**13**	**53** (OSPAR7) (Textile: 8, Glass: 1, Plastic: 44) **5** (OSPAR13) (PO-soft: 3, PET: 1, PS: 1)

The comparative analysis reveals over-predictions relative to the low visual flux. The Binary model mostly overestimated the litter count, especially on September 4th (11:48-12:48), where it detected 58 items against a visual count of only 5. While the Binary model over predicted, the OSPAR models show a more unstable behavior, fluctuating between high and low detection counts. Furthermore, class wise predictions are unreliable for the 7 and 13 class models, since classes were recognized which were not present as confirmed by the human visual observations. A universal increase in detections for all models on 04/09 (11:48-12:48) coincided with strong observed sun reflections on the water surface. This lead to misinterpretations of sun glints as specific litter types. Additionally, low observed flow rates at the Kooybrug can cause items to drift slowly. While a Kalman filter is used to prevent double counting during postprocessing, this step can fail as it relies on consistent object recognition. For OSPAR models, if the predicted class of a single object fluctuates between frames, the tracker fails to match the instances, registering them as new items. This also happens if the object is not detected at all in one interval, but is again detected in the consecutive frame (e.g. due to better lighting conditions). This could be one of the reasons for the high estimates across all models. Figure 4.17 displays multiple examples of detection across label setup during the total monitoring period.

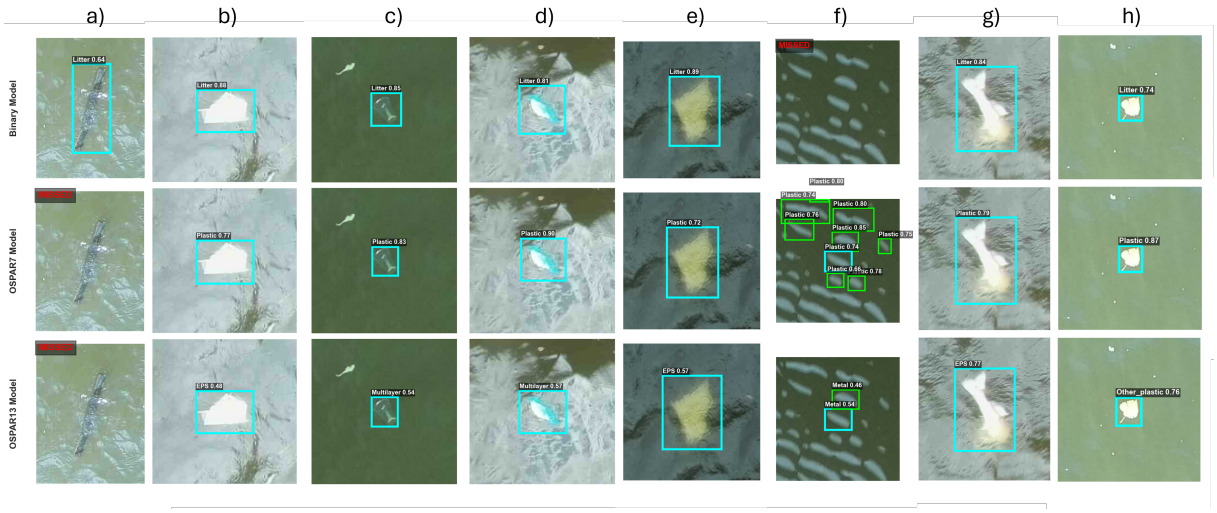


Figure 4.17: Example detections at the Kooybrug during the full monitoring period. Example a) shows a missed Wood item by the OSPAR models, images b) to e) show the correct detection of various litter items, though frequent miss-classification for OSPAR13. Examples f), g) and h) show frequent FP sources, such as water reflections, wildlife and vegetation/leaves.

Oostertoegang

The comparative monitoring at Oostertoegang, detailed in Table 4.5, shows lower estimated counts by the models in comparison to human visual observations. This trend is expected and can be attributed to the physical constraints of the monitoring setup, as the camera's field of view did not span the entire width of the canal section monitored by the human observer.

Table 4.5: Direct comparison of Visual Counting vs. Object Detection counts at Oostertoegang.

Date	Time Slot	Visual Count	VS Count per Class	Binary Detections	OSPAR Detections (Total & Class-wise)
13/09	11:45-12:45	**17**	PO-soft: 8, Paper: 6 PO-hard: 3	**10**	**19** (OSPAR7) (Plastic: 19) **3** (OSPAR13) (Multilayer: 2, Other: 1)
13/09	13:30-14:30	**35**	Paper: 12, Multilayer: 7 PO-soft: 5, PET: 5 EPS: 4, PO-hard: 1 Metal: 1	**86**	**58** (OSPAR7) (Plastic: 58) **7** (OSPAR13) (EPS: 4, Multilayer: 2 PET: 1)
14/09	12:00-13:00	**157**	Paper: 53, PO-soft: 51 Multilayer: 18, Other: 8 PO-hard: 6, Wood: 6 Metal: 5, PET: 5 EPS: 4, PS: 1	**128**	**88** (OSPAR7) (Plastic: 88) **16** (OSPAR13) (PET: 7, EPS: 5 PO-soft: 2, Other: 1 PO-hard: 1)
14/09	13:45-14:45	**32**	Paper: 12, PO-soft: 6 Multilayer: 4, PO-hard: 3 EPS: 2, Other: 2 Metal: 1, PET: 1 Wood: 1	**21**	**18** (OSPAR7) (Plastic: 18) **5** (OSPAR13) (PO-soft: 2, EPS: 1 Multilayer: 1, PET: 1)

Within the model setups, a clear pattern of detection sensitivity is observable. The Binary model consistently registered the highest number of detections, approaching the visual count during the high-flux event on September 14th (128 detections vs. 157 visual counts). The OSPAR7 model followed with moderate estimates, while the OSPAR13 model proved to be the most conservative, recording significantly fewer detections than its counterparts. Despite the low total counts, the class-specific detections of the OSPAR13 model aligned qualitatively with the visual composition, correctly identifying the presence of key categories such as EPS, PET and Multilayer, albeit at lower absolute frequencies.

The results from Kooybrug and Oostertoegang highlight the complex relationship between location specific conditions. At the Kooybrug location, models tended to overpredict relative to visual counts, likely driven by the misinterpretation of sun glint and surface reflections as litter. In contrast, the models provided conservative estimates at Oostertoegang that systematically undercount the true flux, primarily due to spatial coverage limitations. These findings diverge partially from the initial evaluation of OOD performance as presented in 4.1.2, where OSPAR13 initially showed tendencies towards overprediction at Oostertoegang and conservative behavior at Kooybrug. Several factors contribute to this discrepancy. First, the OOD assessment was based on an average of 16 models per setup, whereas the final field monitoring utilized a single, high-performing model selected based on metrics from the in-domain test set. It is possible that this specific model behaves differently than the average when exposed to novel environments. Second, the initial OOD evaluation was constrained to a small sample size (approximately 80 images per location), which likely failed to capture the full range of environmental variability, such as changing lighting angles, wind conditions and litter, which fluctuates more in the over 30 hours of monitoring. Consequently, the initial performance metrics provided only a partial indication of model behavior. The direct comparison of visual counting and camera based object detection indicates that model results from the Kooybrug are affected by noise and are therefore mostly inaccurate. Any extrapolated estimations require cautious interpretation. For the Oostertoegang location, model detections are more aligned with visual observations. The models generally demonstrate higher confidence in their detections. Figure 4.18 shows multiple detection examples at the Oostertoegang location.

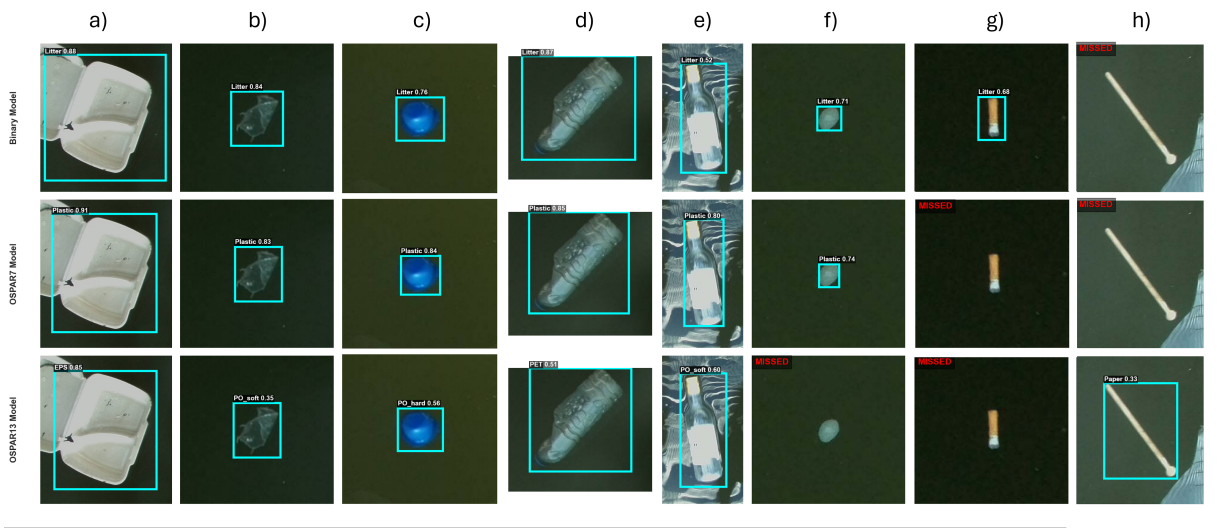


Figure 4.18: Example detections at the Oostertoegang during the full monitoring period. Example a) through d) show correctly detected and classified objects of all models, while the bounding box in example a) only covers part of the litter item. Example e) shows a correct detection of a glass bottle but a common class confusion ofr the OSPAR13 model. Example f) shows a FP for Binary and OSPAR7 models, while OSPAR13 correctly ignores the foam patch. Example g) shows a cigarette filter, which is detected by the Binary model, but missed as a Paper item by the OSPAR models. Example h) shows missed detection by Binary and OSPAR7 models but a correct classification by the OSPAR13 model for an underrepresented class.

4.2.4. Extrapolation

Location Dynamics and Extrapolation Parameters

To convert detections and observations into a mass estimate, the spatial and temporal dynamics characterizing each monitoring location were evaluated. This defined the valid observation windows and the specific extrapolation parameters applied to the model outputs.

A significant challenge for the Den Helder case study was the separation distance of approximately 4 km between the monitoring bridge and the removal system. To quantify the temporal lag introduced by this distance, field experiments using GPS trackers were conducted on September 4th, 2025. The sensor pathways presented in Figure 4.19 reveal a longer travel duration than initially expected. Of the three deployed trackers, two successfully reached the pumping station, recording travel times of 7.9 days (188.4 hours) and 8.1 days (193.4 hours), respectively. The third tracker ceased transmission mid-journey but was later recovered on Texel, suggesting eventual passage. These results establish a characteristic travel time of 7 to 8 days for floating objects in this canal section, confirming that litter transport is not instantaneous but subject to significant retention and re-mobilization cycles. Complementary tracer experiments utilizing biodegradable floating items yielded no recoveries in the final removal batch, further indicating high retention rates along the canal banks.

The established travel time required a temporal adjustment of the analysis period. Litter passing the Kooybrug monitoring point after September 4th would plausibly not reach the removal system before the final cleaning on September 12th. Consequently, the valid observation window was defined as August 26th to September 4th. Figure 4.20 illustrates the alignment of this observation window with the environmental conditions and monitoring activities. Image data collected outside this period were excluded from the mass flux correlation to ensure temporal consistency with the ground-truth batch.

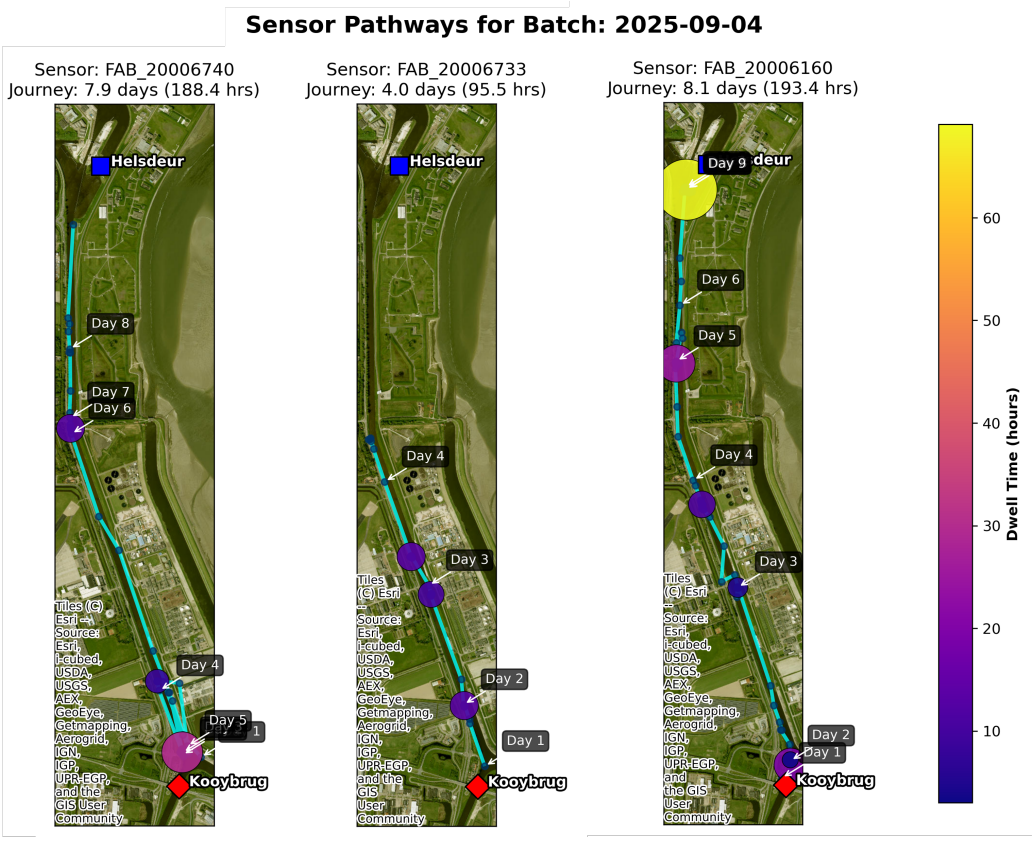


Figure 4.19: Sensor pathways for GPS trackers deployed from the Kooybrug monitoring point.

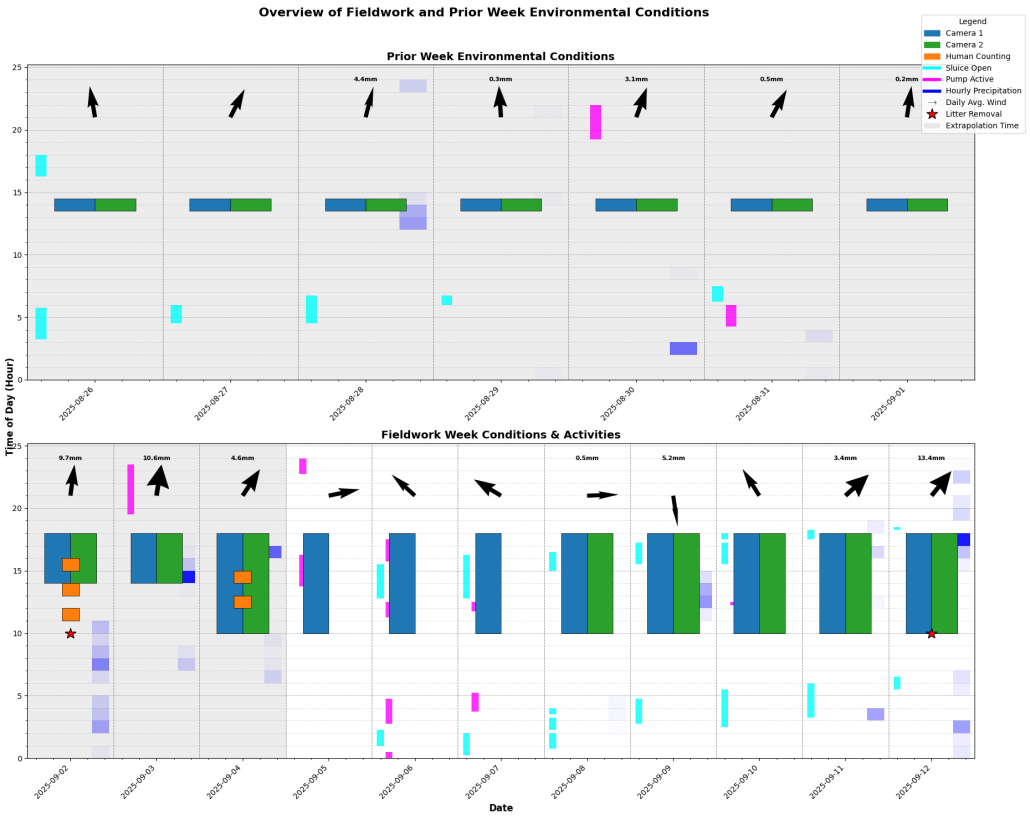


Figure 4.20: Overview of fieldwork activities and environmental conditions at Kooybrug. The valid analysis period is defined as August 26th to September 4th, accounting for the identified travel lag.

Extrapolation of the count data over this 10-day period incorporated hydraulic and meteorological conditions. The average flux based on observations was only extrapolated when the pumping station (Helsdeur) was active or when the wind direction favored downstream transport (South/South-West). To account for the losses identified in the tracer experiments, two retention scenarios were applied: a baseline scenario assuming ideal transport, and a retention-corrected scenario applying a 48% factor based on a model from a study in the UK (Newbould et al., 2021). The exact calculation is detailed in Appendix C.

Oostertoegang

The Oostertoegang site is defined by the close proximity between the observation point and the removal system, minimizing uncertainty for removal and retention. The primary variable in this environment was the temporal fluctuation of litter. A linear extrapolation model applied the mean detection rate from observation hours to the entire 72-hour test duration. To account for the wind dependence observed during visual counting, a wind-based model was also implemented. This scenario applied the average item flux rate only to hours with supportive wind direction. All observed litter (through visual counting and model detections) was assumed to reach the removal system and remain contained, as the system covers the entire canal section with floating lines.

Total Count Estimates

Following the establishment of field dynamics, the detection rates were extrapolated over the full relevant monitoring periods to estimate the total litter count. Figure 4.21 and Figure 4.22 present the comparison between these extrapolated values and the actual number of items retrieved by the removal systems.

At the Kooybrug location, the linear extrapolation of both visual surveys and model detections consistently exceeded the physical ground truth collected by the removal system. The visual survey extrapolation yielded the highest estimates, likely because the limited observation window failed to capture the full temporal variability of the flux, resulting in an upward bias from the sampled hours. The object detection models, which benefited from a longer monitoring duration, provided slightly lower estimates (except OSPAR7) but still surpassed the physical catch in the linear scenarios.

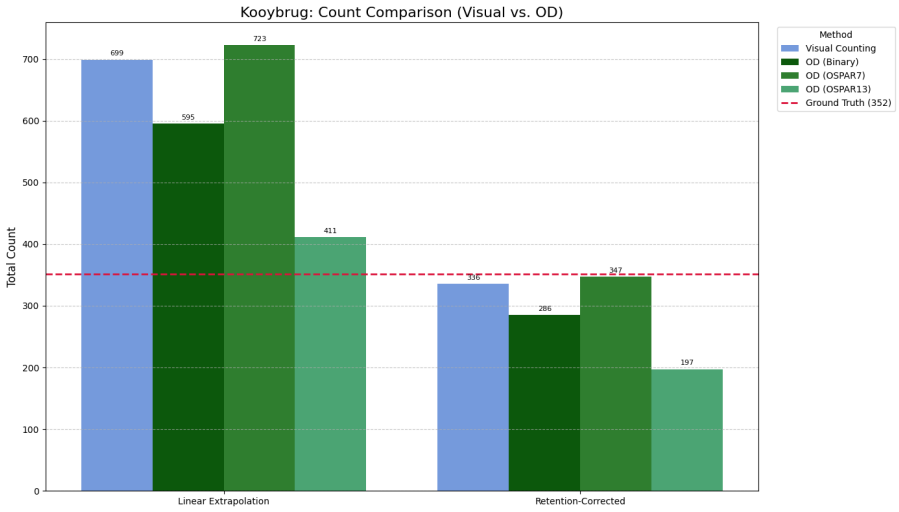


Figure 4.21: Estimation of final counts after extrapolation at Kooybrug

The application of the retention-corrected scenario brought the extrapolated counts significantly closer to the ground truth, resulting in slight underpredictions in some cases. However, this convergence must be interpreted with caution. As identified in the comparative monitoring analysis, the camera detections at this location were heavily influenced by false positives. The physical ground truth consisted largely of fragmented and small scale items that are likely not detectable by the camera system or even human observers at the bridge height. Furthermore, it is plausible that a portion of the captured litter traveled below the surface, evading detection entirely. Therefore, the apparent accuracy of the retention-corrected estimates is likely a result of error cancellation, where the inflation caused by false positives was balanced by the inclusion of the retention factor.

A similar trend of overestimation was observed at the Oostertoegang location, although the driving mechanisms differed. The linear extrapolation of the visual survey data yielded a significant overprediction.

This was primarily driven by a single high-flux event recorded during the visual counting, where 158 items were observed in one hour. This outlier skewed the linearly extrapolated average, demonstrating the sensitivity of short-term monitoring to sporadic flux variations.

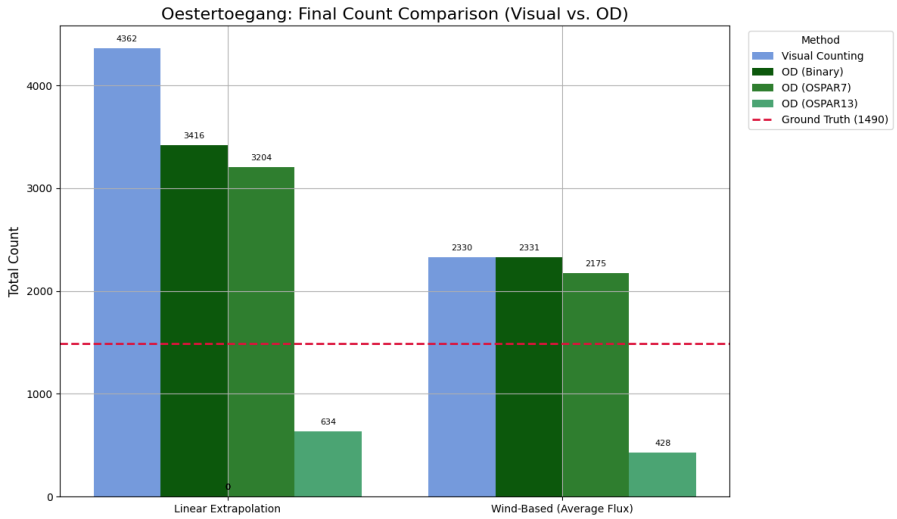


Figure 4.22: Estimation of final counts after extrapolation at Oostertoegang

Among the automated systems, the Binary and OSPAR7 models also predicted total counts, which are significantly higher than the physical catch. Given that the camera's field of view did not cover the entire width of the canal, a degree of underprediction would be expected if detection were more accurate. The fact that these models instead overestimated the count suggests that the rate of false positives outweighed the items missed due to the field-of-view limitation.

In contrast, the OSPAR13 model produced a conservative estimate that underpredicted the total count. This behavior aligns with the results from the head-to-head comparison, where OSPAR13 consistently filtered out ambiguous detections that triggered other models (see Figure 4.18. This underprediction is consistent with the incomplete camera coverage, implying that the OSPAR13 model provided the most reliable estimation. While items outside the frame are missed and potentially legitimate but ambiguous items are filtered out, it avoided the inflation through FPs that compromised the Binary and OSPAR7 estimates. This suggests that in an environment like the Oostertoegang, multi-class detection is feasible and can be accurate. The class wise estimates in comparison to the ground truth count for OSPAR7 and OSPAR13 models are presented in Appendix F.

Total Mass Estimates

The final step of the analysis involved converting the extrapolated item counts into total mass flux estimates. This conversion utilized two distinct methodologies: for visual surveys, a bootstrapping approach resampling from external weight datasets (10,000 iterations); and for object detection models, the instance-specific bounding box regression model validated in Section 4.2.2. Figure 4.23 and Figure 4.24 present the median mass estimates with their associated uncertainty intervals compared against the physical collected ground truth data.

At the Kooybrug location, the mass estimation results diverge significantly from the count extrapolations. The visual survey extrapolation yielded a median mass estimate of approximately 5 kg (aggregated across weight sources), which is an order of magnitude higher than the actual ground truth of 0.44 kg. The uncertainty interval for this estimate is notably wide, reflecting the significant variability introduced when applying literature-based average weights to a specific local context. This overestimation persists even though the initial visual item counts were relatively close to the physical reality, suggesting that the average litter item in this rural canal is significantly lighter and more fragmented than the standard items found in general weight datasets.

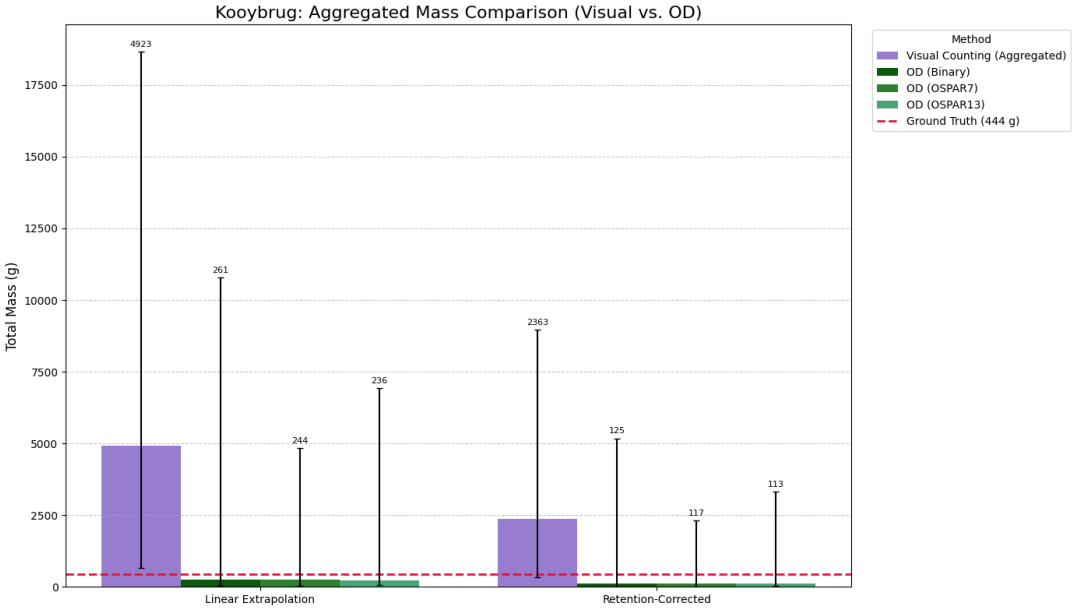


Figure 4.23: Extrapolated total mass estimates for Kooybrug. Visual survey estimates (left) vary by an order of magnitude depending on the reference weight dataset. Model estimates (right) consistently underpredict but provide constrained uncertainty intervals.

In contrast, all object detection models consistently underpredicted the total mass. The retention-corrected Binary and OSPAR models estimated between 0.11 kg and 0.13 kg, corresponding to approximately 25-30% of the actual load. This underprediction highlights a critical stabilizing feature of the bounding box regression method. Although the models overestimated the item count, which consisted mostly of false detections, the resulting mass estimates did not increase proportionally. This indicates that the majority of false positives were characterized by small bounding box areas, which the regression model assigned negligible weights. Consequently, the area-based weighting acted as an effective filter, dampening the impact of count-based hallucinations that would otherwise skew a median-weight multiplication approach.

The results at Oostertoegang further illustrate the comparative strengths of the monitoring approaches. As shown in Figure 4.24, the aggregated visual survey extrapolation overestimated the load for both scenarios, predicting a median mass of approximately 70.6 kg and 37.8 kg compared to the ground truth of 27.07 kg. While the physical ground truth falls within the lower bounds of the visual uncertainty interval in the wind-based scenario, the general tendency to overestimate reaffirms the sensitivity of visual counting to the assumed average item weight, particularly when using different weight databases. The detailed estimations are presented in Appendix G.

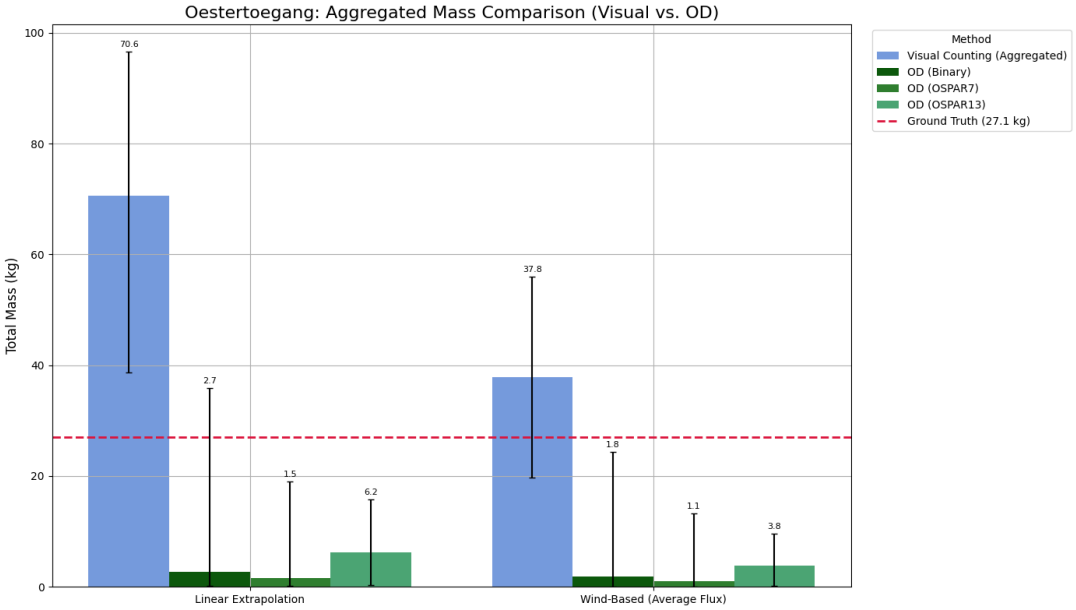


Figure 4.24: Extrapolated total mass estimates for Oostertoegang. Note that while OSPAR13 (far right) predicted the lowest item count, it yielded the highest and most accurate median mass estimate among the automated models.

Among the automated models, a reversed trend was observed compared to count-based extrapolations. While the Binary and OSPAR7 models predicted significantly higher counts compared to OSPAR13, their mass estimates were lower. The Binary model estimated only 1.80 kg, an underprediction of 93% relative to the ground truth, though the ground truth falls within the uncertainty interval. This indicates that despite the high detection count, the model likely detected small features which contributed minimally to the calculated mass. This is partly due to low performance of the regression model for the overall litter class and general under estimation when predicting mass based on litter size. In contrast, the OSPAR13 model, despite its conservative count, yielded the highest median mass estimate of 6.2 and 3.8 kg among the detection models, with a narrower uncertainty interval. As established in the ground truth analysis, 83% of the total mass at Oostertoegang consisted of non-plastic classes such as Wood, Glass and Textile. The object detection models, which were biased towards plastic categories due to training class imbalances, likely missed these heavy non-plastic items or misclassified them as lighter plastic equivalents. However, when considering only the plastic fraction of the ground truth (4.59 kg), the OSPAR13 estimate of 3.79 kg is accurate, capturing 82.5% of the relevant plastic load. This implies that the multi-class model was accurate under the present restrictions at this location. Additionally, the bounding box regression for different classes showed more nuanced estimations, which are closer to the ground truth.

The field validation demonstrates distinct operational trade-offs between manual and automated mass quantification. Visual surveys, while providing a more comprehensive capture of total debris including non-plastic items, are prone to overestimation due to the reliance on extrapolation and generalized average weights that often fail to reflect local characteristics. In contrast, the automated system utilizing object detection models and bounding box regression acts as a conservative estimator for litter mass. While it currently overestimates total item flux due to low performance in OOD environments, it offers benefits in quantifying the plastic load. The area-to-mass conversion method proved effective at filtering out noise, ensuring that false positive detections did not translate into significant mass errors.

5

Discussion

5.1. Model Performance and Classification Strategy

5.1.1. Comparability with Literature

To validate the efficacy of the optimized training pipeline, the model performance was first compared against studies using similar datasets. The RPD datset, characterized by high image complexity, served as a common evaluation ground. Saddi et al., 2025 reported a binary F1-score of 0.46 on the RPD dataset. In comparison, the optimized binary model developed in this thesis achieved a superior F1-score of approximately 0.52. A similar performance advantage was observed for multi-class detection, whereas Saddi et al. reported an F1-score of approximately 0.25 (likely a macro average score) for their 13-class model, the OSPAR13 model developed here achieved a macro F1-score of 0.25 as well (0.38 for weighted F1). This suggests that the inclusion of other external datasets, such as Green Village and Noria Data, alongside a new classification scheme with the integration of synthetic data, did not enhance the multi-class performance beyond the baseline established in the current literature. Furthermore, the study by Saddi et al. used YOLOv7/v8 models, while this research used a YOLOv11 model. The official ultralytics benchmark shows a performance increase of up to 2% of mAP50-95 on the COCO val set between similar sized models (Jocher and Qiu, 2024), while other studies reported similar performance increases on other detection tasks (Hung and Rodriguez, 2025; Khanam and Hussain, 2024). Therefore, the marginal binary performance increase on the RPD dataset is a result of multiple changes in data, training strategies and model choice. However, this did not result in an improvement for the multi-class models, which can be attributed to limitations in the the data itself, see 5.1.2.

Regarding class-specific performance, the results also relate to the findings of Kataoka et al., who utilized a 7-class YOLOv8 model. Both studies report high detection accuracy for visually distinct categories, such as bottles and shopping bags, while performance degrades significantly for heterogeneous categories like other plastics and other bags. This confirms that the difficulty in detecting classes like PO-hard and PS is not unique to this study but represents a universal challenge in differentiating classes with high internal variance (Kataoka et al., 2024).

A similar performance behavior from existing literature was observed regarding the relationship between class detail and model performance. Previous studies, generally report that reducing the number of classes improves performance, implying that a binary or 7-class model should outperform a 13-class model (Saddi et al., 2025; Kataoka et al., 2024). This thesis reports similar findings, where the Binary and OSPAR7 outperformed the OSPAR13 model. The performance decrease is mostly explained by the confusion of classes instead of a general inferior detection capability. This also relates to the findings of Kataoka et al., where consistent detection performance is reported, but decreasing classification scores. Furthermore, their estimated FP rates per 1000 images show a higher rate for a single-class detection model than for one of their multi-class models. A similar behavior, though not fully quantified, was observed at the Oostertoegang location when evaluating the full unlabeled data from the field tests. There, Binary and OSPAR7 models produced significantly more detections than the OSPAR13 model, which likely correlates to an increased FP rate as well for models with fewer classes. This could suggest, that a single-class model has difficulty to distinguish litter from ambiguous reflections or vegetation objects, since all shapes and textures of litter are merged into one class.

Despite the strong in-domain results, the challenge of generalizing to new, unseen riverine environments remains a persistent barrier. The OOD tests conducted in this study revealed a substantial performance decline when models were applied to novel environments such as Kooybrug and Oostertoegang. This phenomenon is consistent with the work of van Lieshout et al., who observed a precision drop from 68.7% (in-domain) to a range of 20-54% on unseen locations. The consistency of these findings across different architectures (Faster R-CNN vs. YOLO) and geographic regions underscores that the generalization gap is a fundamental characteristic of the high variance in riverine environments (van Lieshout et al., 2020).

This limitation is further contextualized by the work of Jia et al., 2025, who investigated domain shifts using the TUD-Vietnam dataset. Their baseline supervised model achieved a low average precision (AP50) of approximately 13.1% on unseen data. However, they demonstrated that a Semi-Supervised Learning (SSL) approach, utilizing 100,000 unlabeled in-domain images, increased this performance to 20.6%. While the absolute performance in both studies remains low, the comparison highlights a methodological limitation of this thesis. The models developed here relied exclusively on supervised learning. The relative performance uplift reported by Jia et al. implies that future iterations of the OSPAR models could benefit significantly from incorporating SSL pre-training pipelines to leverage vast quantities of unlabeled monitoring data, potentially mitigating the steep performance drop observed in the OOD case studies.

5.1.2. Limitations of Open Data sets

While the aggregation of diverse open-source datasets provided the necessary volume of training data for deep learning, this methodological choice introduced significant label noise that constrained model performance. As evidenced by the subset-specific performance metrics, the RPD dataset, comprising the Jakarta, Vietnam, and Amsterdam subsets, consistently yielded the lowest detection scores. Qualitative inspection suggests that this performance deficit is driven less by the model's inability to extract features and more by the degradation of ground-truth quality within the training images.

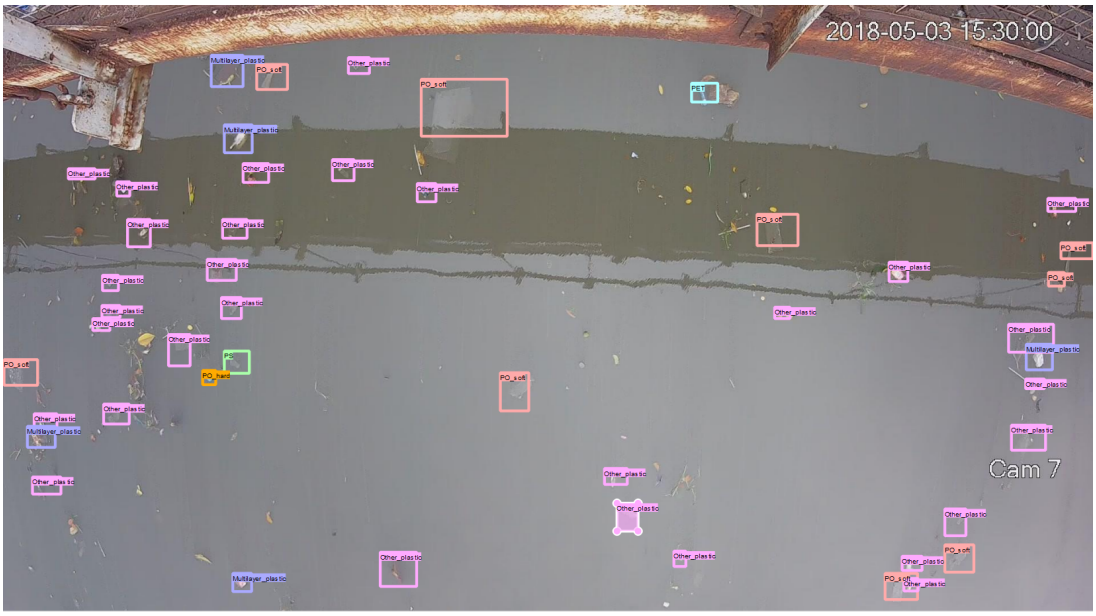


Figure 5.1: Example of inaccurate ground-truth labels from the RPD dataset. Several pink boxes (e.g., PO-soft) are offset from the visible object or do not accurately capture its boundaries.

A primary mode of failure in the annotated data is spatial inaccuracy. As illustrated in Figure 5.1, numerous bounding boxes in the training set are offset from the visible objects or fail to encompass the entire item. This introduces contradictory signals during training, as the model attempts to minimize localization loss against a coordinate system that does not align with the visual reality of the object features. This spatial noise forces the model to learn an averaged or loosened representation of object boundaries, potentially reducing precision during inference.

Figure 5.2: Example of ”missing labels” in the RPD dataset (Amsterdam). The model’s predictions (blue boxes) correctly identify numerous litter items that are missing from the ground truth (green boxes).

A more critical limitation is the prevalence of incomplete annotations, where visible litter items remain unlabeled. Figure 5.2 demonstrates a scene from an Amsterdam canal where the ground truth annotations account for only a fraction of the visible debris. The trained YOLO model successfully detected these omitted items; however, in a standard evaluation pipeline, these valid detections are penalized as false positives. This creates a divergence between quantitative metrics and qualitative utility, artificially deflating the reported mAP and F1-scores. It punishes the model for high-recall behavior, essentially treating superior detection capabilities as errors.

Beyond annotation mechanics, the open datasets exhibit a distinct material bias. The vast majority of available training images focus on plastic debris, resulting in a scarcity of labeled examples for other anthropogenic materials (such as paper and glass). This limitation had direct impact for the mass flux estimation in the Oostertoegang case study, where the model likely failed to detect the heavy non-plastic items that constituted the majority of the physical catch mass. By relying on datasets that prioritize plastic, the model effectively learns a plastic-centric representation of riverine debris, limiting its applicability for total waste load monitoring.

Finally, the RPD dataset introduces extreme domain variance that diverges from the operational reality of automated monitoring (especially for the Amsterdam subset). Unlike the Noria and Green Village datasets, which utilize fixed camera angles typical of bridge-mounted systems, the RPD includes handheld and drone imagery from widely varying perspectives, scales, and lighting conditions. While visual diversity is generally beneficial for generalization, the combination of unstable camera angles with high label noise creates a difficult learning environment. The high variance effectively treats every test image as an out-of-domain sample. Future iterations of such datasets would benefit significantly from a rigorous re-labeling effort to rectify missing annotations and a strategic expansion of class categories to include non-plastic riverine debris.

5.1.3. Material-based Classification

While the results indicate that increased class detail does not decrease general detection capability, the reliance on the OSPAR material-based classification scheme introduces a fundamental challenge for computer vision. Deep learning models, rely primarily on shapes and surface textures to catego-

rize objects. However, in a material-based scheme, the relationship between shape and class is often decoupled. Common items such as cups may share identical geometric profiles yet be composed of distinct materials such as PET, Glass, Paper, or PS. Consequently, a model trained to recognize the shape of a bottle must infer the material properties solely from secondary visual cues like transparency. This limitation is location dependent. As observed in the Oostertoegang case study, the close-range imagery provided sufficient textural resolution to differentiate between the transparency of a plastic cup and the opacity of a paper cup. However, in the images from Kooybrug with increase in distance of camera to water surface, these textural cues can be lost, leaving the model to rely on shape alone. In such scenarios, the distinction between material classes becomes arbitrary for an RGB sensor, leading to the misclassification of visually similar but materially different items.

Furthermore, categories such as PO-hard, Textile, or Other Plastic represent heterogeneous classes with high intra-class variance. Unlike a specific object class (e.g., bottle, like used in other studies), which has a finite range of visual representations, a PO-hard item can manifest as a crate, a fragment, a toy, or an industrial part and more. This results in an effectively infinite feature space that the model must learn to encompass. Even with manual inspection, determining the specific polymer type often requires physical handling or spectroscopic analysis, a level of discrimination that is physically impossible for a standard camera system.

To render this material-based approach robust for operational monitoring, the inherent ambiguity of these classes must be addressed through a data-centric strategy. The current labeled datasets are insufficient to capture the long-tail distribution of possible items within these material categories. Future developments must therefore focus on the expansion of annotated datasets or the generation of high quality synthetic data that explicitly models the textural nuances of different materials. Without such data to bridge the gap between visual appearance and material composition, the classification of generic items into polymer and material specific categories will remain a source of uncertainty in RGB-based monitoring.

5.1.4. Synthetic Data Domain Gap

To interpret the mixed operational impact of synthetic data, specifically the trade-off between improved recall and increased false positives, a separate analysis was conducted to assess the qualities of the generated dataset. By training models exclusively on synthetic data and evaluating them across domains, the characteristics of the transferability of the synthetic to the real environment was assessed.

The training behavior of the synthetic only models revealed limitations in the generated dataset, which are a lack of diversity and complexity. When trained and validated solely on synthetic images, the models achieved very high performance metrics. As illustrated in Figure 5.3, the OSPAR13 synthetic only models achieved a mAP0.5-0.95 of over 0.92.

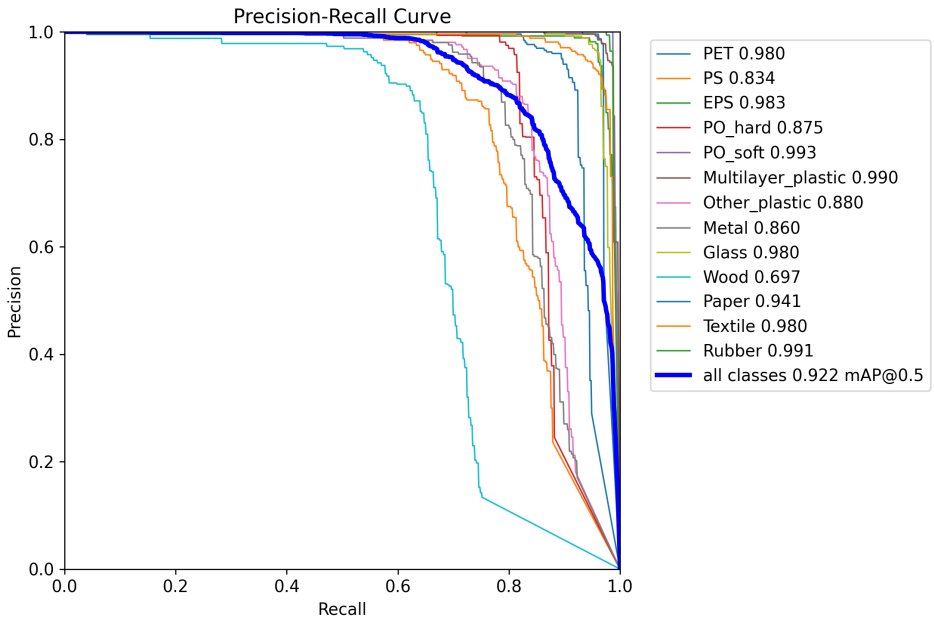


Figure 5.3: Precision-Recall curve for the OSPAR model trained and validated on synthetic data.

This convergence indicates that the synthetic domain is not diverse enough. The use of approximately 70 unique 3D assets to represent 13 material classes resulted in a feature space that was easily memorized by the models. Furthermore, the synthetic rendering pipeline produced objects that were geometrically perfect and texturally clean, lacking the fragmentation, biofouling, and deformation characteristic of riverine debris. Consequently, the model learned to identify idealized representations of litter, creating a disparity when applied to the noisy real world.

The domain gap was found to be highly dependent on the complexity of the location and classification scheme. Figure 5.4 compares the performance of the synthetic only models against the baseline real-data models when evaluated on the in-domain test set.

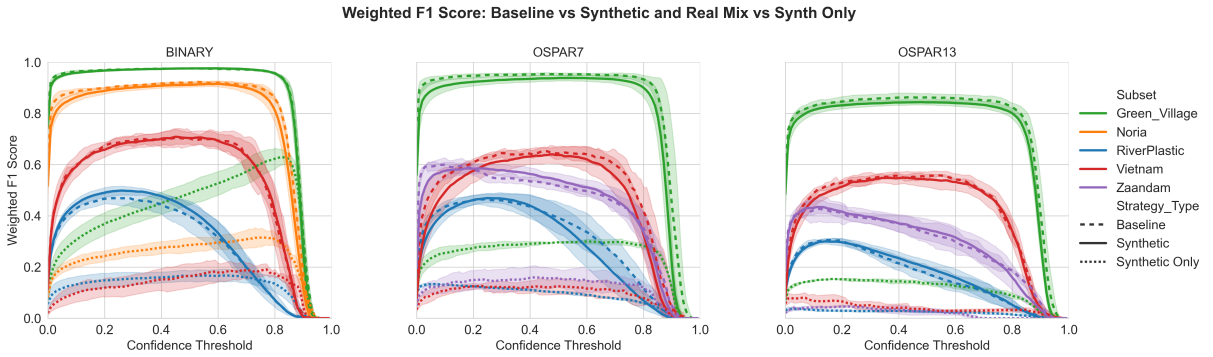


Figure 5.4: Weighted Precision-Recall curves comparing models trained on Real data (solid lines) versus models trained exclusively on Synthetic data (dashed lines) when evaluated on the real-world test set. Note the significantly higher transferability for the Binary task compared to the OSPAR tasks.

For the Binary task, the synthetic only models demonstrated some degree of transferability, achieving an F1-score of approximately 0.63 on the Green Village subset. This suggests that the synthetic generation pipeline successfully captured the fundamental visual cues of litter, such as high contrast against the water surface and unnatural geometric edges. The model learned a generalized concept, allowing it to identify debris in real images without ever having seen a real photo. Furthermore, the Green Village dataset was constructed through controlled catch and release experiments, where a specific catalog of litter items was repeatedly deployed into the water. In this sense, the GV dataset represents a real world environment with artificial characteristics. While the lighting and physics are real, the item distribution is

derived from a closed catalog, mirroring the procedural generation logic of the synthetic dataset. Therefore, synthetic data generation has the potential to simplify future data collection strategies. With some improvements, synthetic data could be equivalent for controlled field experiments and data campaigns. If a model can learn the features of a closed catalog just as well from a renderer as from a staged experiment, the labor intensive process of manual data collection becomes redundant. Future efforts can thus pivot entirely to a hybrid data regime, which leverages synthetic data to learn the controlled features, while reserving real world data collection for monitoring locations which capture the variability of natural litter that simulation cannot yet replicate.

However, this is at the current stage only feasible for a Binary approach since the performance for the OSPAR multi-class tasks collapsed. As shown by the lower curves in Figure 5.4, the features learned from synthetic textures did not generalize to real materials. The visual distinction between a synthetic paper cup and a plastic cup in the renderer was often based on simple opacity, which did not align with the complex and weathered appearance of these objects in the field.

These limitations in transferability can be traced to specific constraints within the process of generating synthetic data in this study. First, on a class wise scale, the synthetic data did not have enough variability of items, which limited transferability. Second, prioritizing automation to achieve dataset scale (generating over 7,000 images) required a placement logic that lacked physics-based interaction. While the rendering engine recreated the optical appearance of floating objects, it did not simulate physical buoyancy or structural deformation through interaction with the environment. Consequently, objects remained rigid and occasionally appeared in physically implausible configurations, such as tires floating with unrealistic orientation or items lacking the appropriate wetting artifacts. This focus on volume over detail introduces a trade-off. The randomization process, while efficient for scaling, produced a subset of training examples with unrealistic lighting or unrealistic placements of items. The experimental results indicated that increasing the quantity of synthetic examples beyond a certain threshold yielded no additional performance gain. This suggests that the domain gap is dependent on quality rather than data quantity. It remains to be determined whether a smaller, manually curated dataset of high quality, physics-compliant renders would outperform the current large-scale approach. Future optimization should therefore pivot from data expansion to data refinement, ensuring that synthetic instances respect the physical constraints of the riverine environment.

5.1.5. Benchmarking and Architectural Differences between models

To validate that the observed performance trends were not artifacts of the specific YOLO architecture, a comparative benchmark was conducted using the Faster R-CNN and RT-DETR architectures. Due to the significant computational resources required, the total amount of trained models used for the primary YOLO analysis was not replicated for these alternative benchmarks. Instead, the best single model setup based on in-domain performance was established for each label configuration to facilitate a direct head-to-head comparison.

Observations regarding computational efficiency showed a behavior, which the performance metrics do not capture. While alternative architectures often exhibited earlier convergence in terms of epoch count compared to the YOLO pipeline, the actual time required for training was significantly higher on the used hardware. The YOLOv11m architecture showed an advantage in inference latency. In the context of developing fully automated monitoring networks, detection speed is a parameter of equal importance to accuracy. Future operational frameworks will likely rely on decentralized processing—where inference occurs on edge devices directly at the monitoring site, or high throughput centralized systems processing streams from multiple locations simultaneously. In both scenarios, the low-latency capabilities of the YOLO family allow for true real-time processing, a capability that heavier two-stage detectors like Faster R-CNN or computationally intensive transformers struggle to maintain without significant hardware acceleration. Consequently, the YOLO architecture offers a more favorable balance of speed and performance, making it the most viable candidate for scalable, continuous riverine monitoring.

In terms of detection accuracy, the experimental results demonstrated a consistent performance advantage for the YOLOv11m pipeline across all test scenarios, including Binary, OSPAR and unseen locations. The RT-DETR model, in particular, exhibited significant stability issues. Despite achieving decent mAP scores, the model only produced competitive F1-scores within a specific confidence range in nearly all evaluations. This discrepancy indicates a critical lack of precision at standard operating thresholds. This behavior suggests that the RT-DETR output, at least within this specific implementation, requires specific threshold tuning to become operationally viable.

The Faster R-CNN model, using a ResNet-18 backbone, was similarly outperformed by the YOLOv11m baseline. A primary hypothesis for this performance gap is the limited capacity of the ResNet-18 backbone relative to the dual-task demands of a two-stage detector. In the Faster R-CNN architecture, the backbone must generate feature maps sufficiently to support both the Region Proposal Network (RPN) and the subsequent classification and regression heads. While the ResNet-18 was selected to maintain parity in parameter count with the YOLOv11m (approximately 20 million parameters), this lightweight backbone appears to act as a bottleneck, limiting the quality of region proposals and the final classifier. This interpretation aligns with established literature, where competitive results for Faster R-CNN are typically reported using significantly deeper backbones such as ResNet-50 or ResNet-101 (van Lieshout et al., 2020; Jia et al., 2025).

Furthermore, the analysis suggests that the training pipeline, which was optimized for the YOLO architecture, likely introduced a bias against the alternative models. Hyperparameter tuning revealed fundamentally different training dynamics. For example, the grid search for Faster R-CNN and RT-DETR identified a higher optimal learning rate than what was effective for YOLO, indicating a difference in convergence behavior. Moreover, the models exhibited varying sensitivities to data augmentation. The Optuna studies demonstrated that the YOLO model improved on augmentations, such as high HSV shifts, Mosaic, and vertical flips, to achieve peak performance. In contrast, the other models failed to improve when trained with these aggressive alterations. Consequently, the final benchmarks for the alternative architectures utilized simpler augmentation techniques. This indicates that the YOLOv11m model is more capable out-of-the-box for this specific application and more robust to aggressive training strategies, which enable benefits from regularization techniques that destabilize lighter architectures. Future research could focus more on the hyperparameter search, ensuring a wider search space and exploring larger backbone variants.

5.2. Comparability between Visual Counting and Camera-based Object Detection

The experimental design of this research facilitated a direct evaluation of two distinct monitoring methodologies: traditional Visual Counting and automated Object Detection. The design of the field study used the two methods for different durations. However, this contrast of four hours of manual observation compared to over thirty hours of automated monitoring per location highlights the fundamental operational differences between the approaches.

In terms of detection accuracy, human observers typically outperform automated models, particularly in complex environments. As noted by Jia et al., 2025, human cognition is more robust to occlusion and environmental noise, allowing for higher precision during short observation windows. This was supported by the comparative results at Oostertoegang, where the observer accurately recorded a high-flux event that the models undercounted, partly due to field-of-view limitations. This is in line with experiments conducted by van Lieshout et al., 2020, which observed a similar trend in their experiments. However, visual counting is constrained by physical and physiological factors. At the Kooybrug location, the considerable vertical distance of nine meters likely impaired human classification capabilities similarly to the camera system, making the distinction between small material fragments unreliable. Furthermore, manual counting is susceptible to observer bias, fatigue, and concentration loss, particularly during long-duration surveys, which introduces non-systematic errors that are difficult to quantify. Furthermore, in high-flux scenarios, as observed in one of the hours at Oostertoegang, it can be overwhelming for human observers to keep track of multiple objects at the same time, taking records/notes of the items as well as classifying them accurately.

A primary advantage of the visual counting method remains its low barrier to entry. It requires no specialized infrastructure, power supply, or permitting, making it a highly effective tool for rapid assessments in resource-constrained environments or for initial site scoping. However, its scalability for long-term monitoring is limited by labor costs and logistical constraints. As demonstrated by the extrapolation analysis at Oostertoegang, reliance on sparse manual observations can lead to significant errors when events, such as wind-driven flux spikes, are either missed or disproportionately weighted. The automated system, conversely, offers high temporal resolution, capturing variations that manual observations miss.

Beyond scalability, a critical distinction lies in data verifiability. Visual counting produces a singular numerical value that lacks an audit trail. It is impossible to retrospectively validate a count or analyze the specific composition of the observed flux if the litter was not sampled. In contrast, the object detection

pipeline generates a transparent, reviewable dataset where every detection is linked to a source image, bounding box, and confidence score. As demonstrated in the mass flux quantification results, the retention of spatial information (bounding box size) allows for instance-specific weight estimation. This approach acts as a conservative filter, mitigating the impact of false positives on the final mass estimate. Conversely, visual counting relies on applying generalized average weights from external databases to raw counts since any information about the size of the item is typically lost. As shown in the Kooybrug case study, this method can result in an order-of-magnitude overestimation when local litter fragmentation differs from the reference database.

Therefore, while visual counting remains a valuable tool for first order estimates of fluxes and calibration, the results suggest drawbacks when trying to quantify a reliable mass. In addition, the high uncertainty inherent in extrapolating data from sparse visual counting observations can be reduced through continuous and automated monitoring networks. Although camera-based systems currently struggle with specific classification tasks and environmental noise, their ability to provide consistent, auditable, and temporally dense data offers a pathway to reduce the uncertainty bounds that global flux models currently struggle with. Ideally, future monitoring frameworks should integrate both approaches, using periodic visual surveys to validate model performance while relying on automated systems for the continuous quantification. However, it must be noted that both methods are limited to the surface layer. A total mass budget will require integration with physical sampling methods to account for the submerged fraction of the plastic load.

5.3. Mass Conversion and Flux Estimation

The mass quantification method used in this research relies on the fundamental assumption that the two-dimensional area of a bounding box correlates sufficiently with the mass of the object. While the regression analysis on the IMR dataset (de Lange et al., 2023) established a relationship, the application of this proxy in dynamic riverine environments introduces limitations. A bounding box, by definition, simplifies complex morphologies into a rectangle. Consequently, objects with vastly different densities and aspect ratios, such as a compact food container versus a straw, may present identical bounding box areas. In such cases, the regression model assigns a similar mass estimate, ignoring the actual surface area and volumetric properties of the item.

This limitation is further reinforced by the optical constraints of the monitoring setup. The system quantifies mass based solely on the visible surface area. Objects that are partially submerged present a reduced cross-sectional area relative to their total mass. This phenomenon creates a systematic bias towards underestimation, particularly for high-density items that sit lower in the water column. Furthermore, the quantile regression model itself is an approximation derived from a specific training distribution. As observed in the results, the fit is imperfect for outliers, leading to potential inaccuracies when the model encounters litter classes that were underrepresented in the training data, such as rubber or wood items.

Despite these geometric inaccuracies, the bounding box to mass approach demonstrated a unique behavior when applied to the full detection set. A primary finding of the field study was that the area-based weighting acts as a potential filter against detection errors. As evidenced by the generalization tests at Kooybrug and Oostertoegang, the object detection models were prone to generating a high volume of false positives. However, these errors were characteristically triggered by small features such as foam patches, leafs or water reflections.

In a traditional count-based extrapolation, these false positives introduce linear errors, where a misidentified ripple contributes the same as a genuine plastic crate without manual verification. However, within the mass estimation framework, this impact is reduced. The quantile regression model correctly assigns these small, incidental bounding boxes a low mass value. Consequently, a high amount of FPs with low areas contribute a less significant amount to the total aggregate mass as the detection of larger litter items. This effect explains the result observed at the Kooybrug and Oostertoegang location, where overestimation in item counts still result in conservative mass estimations. Thus, while the 2D proxy is inaccurate for individual item characterization, it can provide an alternative for long-term mass estimations in riverine environments.

Furthermore, the current system utilizes a Kalman filter to associate detections between frames to prevent double-counting. However, the qualitative analysis suggests that tracking failures contributed to the overestimation of counts, particularly for the multi-class models. For example, if the same object appears in consecutive images, but the multi-class models assigns a different class, the filter won't match these

detections even if the predicted location overlays with the actual position. Therefore, the same litter item could have been detected multiple times, which results in a higher estimated mass and count.

To further refine the model outputs and reduce uncertainty, the detection logic can be further calibrated. First, the Kalman filter could be adjusted, so that predicted and actual litter positions are closer together and result in more correct matches. Second, the precision scores obtained from the location specific test sets (such as the OOD subsets for Kooybrug and Oostertoegang) could serve as a correction factor, discounting the raw detection count by the expected false positive rate. However, this approach assumes that the test set is perfectly representative of the full monitoring period. A more robust calibration method involves the integration of periodic visual surveys. A correction factor can be derived by calculating a ratio between the detected objects of the models and the human visual count during overlapping periods. This approach uses the high accuracy of human observation to calibrate the models output to the actual flux. This could mitigate a systematic over- or underestimation, while still benefiting from the advantages of an automated system in terms of temporal coverage.

5.4. Case Study Limitations

While the field deployments provided valuable data on the operational viability of the monitoring models, the conclusions drawn from these case studies must be interpreted within the context of experimental limitations. These constraints, imposed by the physical infrastructure and logistical realities of the monitoring sites, introduced sources of uncertainty that complicate the direct correlation between model predictions and physical ground truth.

5.4.1. Uncertainty in the Kooybrug Location

The selection of the Kooybrug location in Den Helder was primarily driven by the operational availability of a removal system and existing permit for cameras rather than scientific suitability. This compromise resulted in a separation distance of approximately four kilometers between the camera monitoring point and the litter removal point. This introduced a substantial variable of transport uncertainty that proved difficult to quantify.

The results from the ground-truth analysis indicated that the litter retrieved from this location was dominated by small, fragmented items, such as pieces of hard and soft polyolefins. Given the camera mounting height of nine meters, reliable detection of such small scale debris is physically improbable for both human observers and computer vision systems. Furthermore, the transport dynamics within the canal introduced significant temporal lag and retention. The deployment of GPS trackers confirmed a travel time of seven to eight days and they were retained multiple times at the canal banks or vegetation. The trackers are characterized by uniform shapes and high buoyancy, which results in high surface area above the water. Therefore, transport is strongly influenced by wind, which could result in a faster travel time than actual litter. Moreover, real litter items are susceptible to sinking, entrapment or fragmentation that the trackers did not replicate. While the trackers gave good insights into the local conditions during the testing period, they are less ideal as proxies for true litter behavior. As a consequence, it is impossible to decouple the error contribution of the model from the uncertainty of the physical transport. The seemingly accurate total mass estimate produced by the models at this location should therefore be viewed with caution, as it likely results from the cancellation of errors, where high false positive rates were offset by the conversion into a mass, rather than precise detection of fragments.

5.4.2. Physical Constraints at Oostertoegang

While the location at the Oostertoegang solved the issue of uncertainty induced by distance between monitoring and removal point, the experimental setup presented a different set of methodological challenges. To mitigate security risks, the camera equipment was installed beneath the bridge deck, resulting in a short distance to the water surface. This perspective deviated from the standard training data distribution, which typically features higher views. This could have potentially confused the models, leading to errors despite encountering mostly intact litter items.

A critical limitation was introduced by the optical configuration of the system. Due to a hardware failure of the intended ultra wide lens, a narrower four millimeter lens was used. This lead to a shortened field of view, which failed to capture the entire width of the canal section between the floating lines. Therefore, a portion of the litter flux could bypass the camera while still being captured by the removal system.

This limitation is evident in the direct comparison between visual surveys and automated detections. In

three of the four synchronized observation windows, the automated models recorded lower counts than the human observer. While this behavior is not reflected in the final extrapolated counts, it is offset by a high FP rate for most models and simplified extrapolation scenarios. Only the OSPAR13 model showed a conservative detection behavior in which every visual observation hour was under predicted, including the overall ground truth count. To fully isolate the model performance from the field-of-view limitation, a complete annotation of the data from these specific hours would be required. Without such ground truth, it remains difficult to quantify exactly what proportion of the discrepancy is due to model failure versus optical exclusion. Ideally, the removal system would have been cleared and sampled on an hourly basis to match the visual surveys, allowing for a precise validation of the count and weight estimation. However, a continuous inflow of litter items as well as logistical constraints made such detailed sampling infeasible, necessitating reliance on the extrapolated aggregate data.

5.5. Recommendations for Future Work

The findings of this thesis highlight several possibilities for advancing the capability of automated riverine litter monitoring. Future research should focus on addressing the data limitations identified in the training phase, refining operational deployment strategies and exploring novel architectural and hardware solutions to overcome the semantic ambiguities of RGB-based monitoring.

5.5.1. Advancing Synthetic Data Generation

While the integration of synthetic data successfully improved detection recall, the current synthetic dataset remains limited. Future iterations of the synthetic data tool should expand the asset library from the current set of approximately seventy items to a minimum of a few hundreds. This expansion is at the moment still constrained by a manual import of 3D objects. With more funding, more detailed objects can be used that improve upon open source models. The inclusion of more objects could prevent the model from memorizing items. Furthermore, the generation logic should be improved to simulate environmental variance more detailed. By randomizing camera characteristics, such as focal length, sensor height and angle, the synthetic dataset can better approximate the diverse conditions of real deployments. Future work should incorporate physics-based deformation to simulate the interaction of objects with the water surface, which would add realism to the current established baseline. After improving on these aspects, the relationship between synthetic data quantity and quality should be investigated in regards to improvement of model performance.

5.5.2. Data Acquisition and Fine-Tuning

To resolve the performance deficits observed in the OSPAR multi-class models, a targeted campaign to acquire real-world training data for underrepresented classes is essential. The reliance on general open datasets resulted in a class imbalance that biased mass estimates. Future collection efforts should prioritize non-plastic categories to ensure the model learns a representative distribution of riverine debris.

Parallel to data expansion, the effects of location specific fine-tuning should be further investigated. The results indicated that models struggle to generalize to new environments. However, it has been demonstrated in the field that labeling a small quantity of site specific images (50 to 200 examples) can yield precision improvements of over 20% (Van Emmerik and Schwarz, 2020; van Lieshout et al., 2020). Future research should systematically quantify the stagnation point for fine-tuning in diverse environments. Determining the minimum number of local annotations required to stabilize model performance would establish a standard operating procedure for rapid camera deployment, balancing the cost of manual labeling against the benefit of automated accuracy. Different strategies should be explored for this goal, as fine-tuning with new data on existing models and completely re-training come with their specific advantages and drawbacks.

5.5.3. Extended Field Validation

Due to limitations of the current case studies, more rigorous field studies are required to fully establish decisive results. Future deployments must ensure setups that guarantee full-width coverage of the water body to eliminate the uncertainty of missed detections. Furthermore, monitoring campaigns should span across multiple seasons to assess the model's robustness against changing vegetation cover, lighting angles, and weather conditions such as fog or heavy rain, which were only partially represented in this study.

A particularly promising opportunity for validating mass flux estimation capabilities lies in controlled re-

lease experiments, such as the initiative conducted at the Themsebrug in Belgium. Unlike the monitoring at Kooybrug or Oostertoegang, this experimental design involves the release of a known quantity and mass of specific litter items into a monitored section. Testing the Binary and OSPAR models in such a controlled environment would provide a ground-truth benchmark free from the transport lag, retention and coverage uncertainties that complicated the analysis in this research. This would enable a precise calibration of the area-to-mass regression logic.

5.5.4. Architectural and Hardware Innovations

Beyond improvements in data and deployment, methodological innovations could address the fundamental limitations of the current workflow. One potential approach is a two-stage detection framework. Rather than training a single end-to-end detector, a binary model could first localize potential debris, after which high-resolution crops of the detections are passed to a dedicated image classifier. This would allow the system to leverage extensive land-based waste classification datasets, such as TrashNet, to improve material identification. However, this approach is contingent on high-resolution imagery, as the cropping process requires sufficient pixel density to resolve material textures.

Alternatively, end-to-end mass estimation could be explored through multi-modal data fusion. By annotating training images with the exact mass of specific items, models could be trained to regress weight directly from visual features, bypassing the intermediate bounding box proxy. Techniques such as late fusion, where visual feature vectors are concatenated with metadata before the regression head, offer theoretical promise. However, the creation of such a dataset, requiring the individual weighing and image matching of thousands of floating items, presents a substantial logistical barrier.

Finally, the inherent semantic ambiguity of RGB imagery suggests that hardware solutions may offer the most definitive solution. Models using standard cameras cannot reliably distinguish between a transparent PET bottle and a glass bottle based on visual appearance alone. The integration of hyperspectral or multispectral sensors, which capture data beyond the visible spectrum, could resolve this limitation. Polymers exhibit distinct spectral signatures in the near-infrared region that are invisible to RGB sensors. Incorporating spectral data would shift the classification task from shape based inference to material analysis, potentially improving upon the class confusion that currently limits the accuracy of OSPAR-based monitoring.

6

Conclusion

This thesis conducted a comprehensive investigation into the development, optimization, and real-world application of deep learning models for monitoring floating litter in waterways. Through various training methods, different labeling strategies and model architectures were evaluated. The best-performing models were then deployed in two real-world case studies to assess their capability for mass-flux estimation against ground-truth data. This final chapter revisits the research questions posed at the beginning of this study, providing an answer to each based on the experimental results and discussion.

6.1. Sub-Questions

To what extent does expanding the training dataset with synthetic data improve model robustness and detection performance?

The addition of synthetically generated data provided a negligible or marginal increase in performance on the in-domain test set when compared to real-data baselines. While the overall best performing models for each label setup included synthetic data to a capacity, no real trend was observed. The addition of synthetic data in various approaches yielded rather random results, where one specific strategy did not show constant improvement.

However, synthetic models exhibited a shift towards higher recall when compared with baseline models at the same confidence thresholds, which generally resulted in a slight decrease in False Negatives but a corresponding increase in False Positives. This detection behavior was amplified when evaluating on out-of-domain and unseen locations. On these unseen sets, synthetic models often yielded higher F1 scores, indicating that their improved ability to detect more objects slightly outweighed the cost of increased detection errors.

It is concluded that synthetic data is currently effective for inducing a high recall behavior in models, which may be prioritized in monitoring scenarios where detecting more items is important. Baseline models trained solely on real data proved more suitable when high precision and a lower FP rate are required. The synthetic strategies did not significantly improve the detection of underrepresented material classes, suggesting that increased variety and realism in synthetic data are necessary to effectively address class imbalance.

How accurately can a deep learning model classify floating litter into material-based categories compared to traditional item-based approaches?

Classification accuracy under the OSPAR-based material approach was highly variable and directly dependent on the visual homogeneity of the target class. Overall detection performance, remained relatively consistent across single-class and multi-class setups. However, the introduction of more detailed material classes led to an increase in class confusion, which was the dominant source of error and the main driver for decreased mAP.

Models demonstrated high and reliable performance on visually distinct and homogenous classes, such as PET, Metal, and EPS. These categories achieved the cleanest PR curves and highest F1 scores. In contrast, performance degraded significantly for heterogeneous classes like PO-Hard, PS and Other

Plastic, which represent collections of items with widely varying shapes and appearances. This heterogeneity introduced ambiguity that models struggled to resolve, resulting in low precision and recall.

For scarce material classes, including Wood, Glass, and Textile, the model's performance was inconclusive due to insufficient data representation in the test set, resulting wide performance for the F1-Score. However, the detection of a few true positives by the YOLOv11m architecture indicates that material-based classification is feasible, provided that dedicated and diverse annotated data is collected for these classes.

How does image based object detection compare to traditional visual counting methods in terms of temporal coverage and consistency during long-term deployment?

The comparison of flux estimates revealed systematic location-specific biases. Automated models at the Kooybrug site generally showed a tendency for overprediction relative to the visual counts. At the Oostertoegang site, models systematically underpredicted the total flux, which was the expected outcome due to the known limitation of the camera's restricted field of view. The OSPAR13 model exhibited the most conservative counting behavior across the field sites. A full conclusion and quantification of model behavior can be drawn through a full manual review and annotation of the field data, which has not been completed at this point.

In the context of long-term monitoring, camera-based object detection offers a more scalable methodology. Automated systems provide continuous observation, which yields a more representative flux estimate than sparse visual counting data, which is prone to high or low bias variations observed in limited windows. Furthermore, camera detections are reviewable, while visual counting is non-reviewable and highly susceptible to observer bias and fatigue. Therefore, it is concluded that reliance on automated camera systems is preferred for future long-term monitoring efforts, with visual counting best utilized as an efficient, low-cost method for frequent calibration checks to account for systematic model biases or first order estimates at new potential deployment locations.

Can material-based classification be reliably converted into mass flux estimates, and what is the accuracy of these estimates when validated against physically collected litter data?

Mass conversion was assessed using two distinct approaches. For visual counting, the bootstrapping method converts counts by randomly sampling empirical weight distributions, yielding a mass estimate alongside an uncertainty interval. This approach revealed high uncertainty, with confidence intervals often exceeding an order of magnitude, as visual cues about item size are lost. The use of local, site-specific weight data was found to produce the most accurate estimates relative to the ground truth catch. For camera-based detections, the Quantile Regression model converts object bounding box area into a mass prediction. While the regression relationship between area and mass showed variable per-class performance, this method offers an advantage as it can serves as a filter against environmental noise. Small False Positives, such as reflections or background debris, are assigned a low mass by the model due to small size, which dampens their impact on the overall flux calculation. This resulted in conservative mass estimates compared to the ground truth, mitigating the effect of detection based overpredictions.

At the Kooybrug location, visual counting overpredicted the mass by over an order of magnitude, while object detection methods were more accurate. However, all results are significantly influenced by the uncertainty surrounding this location. At the Oostertoegang location, the ground truth mass falls within the confidence bounds of the visual counting method, although mass is generally overpredicted. The estimation from object detection models significantly underpredicts up to an order of magnitude lower. Considering the exclusion of outliers and the bias towards the plastic classes, the estimation comes close to the actual retrieved plastic load. Nevertheless, the physical retrieval data confirmed the major challenge in litter quantification. The mass of debris is often dominated by a few heavy outliers (e.g., a few items of Wood, Metal or Glass), while item count is dominated by small and light plastic. The choice between quantifying mass flux (kg/day) or count flux (items/day) depends entirely on the monitoring objective, and both methods were shown to be feasible, each with its own specific characteristics regarding uncertainty.

6.2. Central Research Question

With the sub-questions answered, the main research question can be addressed:

How can a deep learning model enable material-based classification of floating litter in waterways for the purpose of accurate mass-flux estimation?

In response to the central research question, deep learning models can enable material-based classification for the purpose of mass-flux estimation, provided that the monitoring objectives and data constraints are clearly understood.

- **Feasibility and Classification:** A material-based (OSPAR) classification is reliable for visually distinct, homogenous materials but requires substantial data improvement or a remapping of classes for heterogeneous and underrepresented categories. The consistent performance of the Binary approach confirmed that high-level detection is robust, even when detailed classification is less precise.
- **Generalization:** Model performance remains susceptible to location-specific parameters. Generalization to new, unseen locations is limited, confirming that systematic biases in training data prevent reliable out-of-the-box deployment without site-specific calibration.
- **Mass Flux Quantification:** The combination of flux quantification and the Quantile Regression mass estimation model constitutes a functional and auditable approach. In most cases, this automated method proved to be superior to visual counting due to its continuous temporal coverage and its ability to conservatively filter detection noise through mass assignment. While the correlation between area and mass is imperfect, the final estimate can come close to the actual plastic litter load present.

In conclusion, this research validates the current feasibility for deploying automated, material-based litter monitoring. The primary focus for future work should be on data curation and collection. A greater diversity of annotated real-world data as well as the improvement of realism of synthetic data can potentially improve upon the current performance plateau for ambiguous and scarce material classes.

Bibliography

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. https://arxiv.org/abs/1907.10902

Alibrahim, H., & Ludwig, S. A. (2021). Hyperparameter optimization: Comparing genetic algorithm against grid search and bayesian optimization. *2021 IEEE congress on evolutionary computation (CEC)*, 1551–1559.

Asselin, P.-L., Coulombe, V., Guimont-Martin, W., & Larrivée-Hardy, W. (2025). Replication study and benchmarking of real-time object detection models. https://arxiv.org/abs/2405.06911

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *The journal of machine learning research*, *13*(1), 281–305.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. *European conference on computer vision*, 213–229.

Castro-Jiménez, J., González-Fernández, D., Fornier, M., Schmidt, N., & Sempéré, R. (2019). Macrolitter in surface waters from the rhone river: Plastic pollution and loading to the nw mediterranean sea. *Marine Pollution Bulletin*, *146*, 60–66.

Crosti, R., Arcangeli, A., Campana, I., Paraboschi, M., & González-Fernández, D. (2018). 'down to the river': Amount, composition, and economic sector of litter entering the marine compartment, through the tiber river in the western mediterranean sea. *Rendiconti Lincei. Scienze Fisiche e Naturali*, *29*(4), 859–866.

de Lange, S. I., Mellink, Y., Vriend, P., Tasseron, P. F., Begemann, F., Hauk, R., Aalderink, H., Hamers, E., Jansson, P., Joosse, N., et al. (2023). Sample size requirements for riverbank macrolitter characterization. *Frontiers in Water*, *4*, 1085285.

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

Goldstein, M. C., & Goodwin, D. S. (2013). Gooseneck barnacles (lepas spp.) ingest microplastic debris in the north pacific subtropical gyre. *PeerJ*, *1*, e184.

González-Fernández, D., Cózar, A., Hanke, G., Viejo, J., Morales-Caselles, C., Bakiu, R., Barceló, D., Bessa, F., Bruge, A., Cabrera, M., et al. (2021). Floating macrolitter leaked from europe into the ocean. *Nature Sustainability*, *4*(6), 474–483.

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

González-Fernández, D., Roebroek, C. T., Laufkötter, C., Cózar, A., & Van Emmerik, T. H. (2023). Diverging estimates of river plastic input to the ocean. *Nature Reviews Earth & Environment*, *4*(7), 424–426.

Hung, G., & Rodriguez, I. F. (2025). A comparative study of yolov8 to yolov11 performance in underwater vision tasks. *arXiv preprint arXiv:2509.12682*.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *International conference on machine learning*, 448–456.

Jambeck, J. R., Geyer, R., Wilcox, C., Siegler, T. R., Perryman, M., Andrady, A., Narayan, R., & Law, K. L. (2015). Plastic waste inputs from land into the ocean. *science*, *347*(6223), 768–771.

Jia, T., de Vries, R., Kapelan, Z., van Emmerik, T. H. M., & Taormina, R. (2024). Detecting floating litter in freshwater bodies with semi-supervised deep learning. *Water Research*, *266*, 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.

Jia, T., Taormina, R., de Vries, R., Kapelan, Z., van Emmerik, T. H., Vriend, P., & Okkerman, I. (2025). A semi-supervised learning-based framework for quantifying litter fluxes in river systems. *Water Research*, 124833.

Jia, T., Vallendar, A. J., de Vries, R., Kapelan, Z., & Taormina, R. (2023). Advancing deep learning-based detection of floating litter using a novel open dataset. *Frontiers in Water*, *5*, 1298465.

Jocher, G., & Qiu, J. (2024). *Ultralytics yolo11* (Version 11.0.0). https://github.com/ultralytics/ultralytics

Kataoka, T., & Nihei, Y. (2020). Quantification of floating riverine macro-debris transport using an image processing approach. *Scientific reports*, *10*(1), 2198.

Kataoka, T., Yoshida, T., & Yamamoto, N. (2024). Instance segmentation models for detecting floating macroplastic debris from river surface images. *Frontiers in Earth Science*, *12*, 1427132.

Khanam, R., & Hussain, M. (2024). Yolov11: An overview of the key architectural enhancements. https://arxiv.org/abs/2410.17725

Kylili, K., Artusi, A., & Hadjistassou, C. (2021). A new paradigm for estimating the prevalence of plastic litter in the marine environment. *Marine Pollution Bulletin*, *173*, 113127.

Lebreton, L. C., Van Der Zwet, J., Damsteeg, J.-W., Slat, B., Andrady, A., & Reisser, J. (2017). River plastic emissions to the world's oceans. *Nature communications*, *8*(1), 15611.

LeCun, Y., & Bengio, Y. (1998). Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*.

Lin, F., Hou, T., Jin, Q., & You, A. (2021). Improved yolo based detection algorithm for floating debris in waterway. *Entropy*, *23*(9), 1111.

Lu, D., & Weng, Q. (2007). A survey of image classification methods and techniques for improving classification performance. *International journal of Remote sensing*, *28*(5), 823–870.

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.

Meijer, L. J., Van Emmerik, T., 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.

Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., & Terzopoulos, D. (2021). Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, *44*(7), 3523–3542.

Musić, J., Kružić, S., Stančić, I., & Alexandrou, F. (2020). Detecting underwater sea litter using deep neural networks: An initial study. *2020 5th International Conference on Smart and Sustainable Technologies (SpliTech)*, 1–6.

Newbould, R. A., Powell, D. M., & Whelan, M. J. (2021). Macroplastic debris transfer in rivers: A travel distance approach. *Frontiers in Water*, *3*, 724596.

Noria. (2023, May). Kentallen dataset rivieren [In opdracht van I&W]. https://www.afvalcirculair.nl/zwerfafval-microplastics/kennisbibliotheek/rivier/kentallen-dataset-drijvend-zwerfafval/

OSPAR Commission. (2010). *Guideline for monitoring marine litter on the beaches in the OSPAR maritime area* (Agreement No. 2010-02) (Prepared by B. Wenneker and L. Oosterbaan). OSPAR Commission. London, United Kingdom. https://www.ospar.org/ospar-data/10-02e_beachlitter%20guideline_english%20only.pdf

Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, *22*(10), 1345–1359.

Raiaan, M. A. K., Sakib, S., Fahad, N. M., Al Mamun, A., Rahman, M. A., Shatabda, S., & Mukta, M. S. H. (2024). A systematic review of hyperparameter optimization techniques in convolutional neural networks. *Decision Analytics Journal*, *11*, 100470.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, *28*.

Ritchie, H., Samborska, V., & Roser, M. (2023). Plastic pollution [https://ourworldindata.org/plastic-pollution]. *Our World in Data*.

Rochman, C. M., Browne, M. A., Halpern, B. S., Hentschel, B. T., Hoh, E., Karapanagioti, H. K., Rios-Mendoza, L. M., Takada, H., Teh, S., & Thompson, R. C. (2013). Classify plastic waste as hazardous. *Nature*, *494*(7436), 169–171.

Roebroek, C. T., Laufkötter, C., González-Fernández, D., & van Emmerik, T. (2022). The quest for the missing plastics: Large uncertainties in river plastic export into the sea. *Environmental pollution*, *312*, 119948.

Saddi, K. C., van Emmerik, T. H. M., Miglino, D., Poggi, M., Isgrò, F., Tasseron, P., Daniele, L., & Manfreda, S. (2025). River plastic dataset.

Schmidt, C., Krauth, T., & Wagner, S. (2017). Export of plastic debris by rivers into the sea. *Environmental science & technology*, *51*(21), 12246–12253.

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of big data*, *6*(1), 1–48.

Tasseron, P., Zinsmeister, H., Rambonnet, L., Hiemstra, A.-F., Siepman, D., & van Emmerik, T. (2020). Plastic hotspot mapping in urban water systems. *Geosciences*, *10*(9), 342.

van Calcar, C. v., & van Emmerik, T. v. (2019). Abundance of plastic debris across european and asian rivers. *Environmental Research Letters*, *14*(12), 124051.

van Emmerik, T., Vriend, P., & Roebroek, J. (2020). *An evaluation of the river-ospar method for quantifying macrolitter on dutch riverbanks*. Wageningen University Wageningen.

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.

Van Emmerik, T., Kieu-Le, T.-C., Loozen, M., Van Oeveren, K., Strady, E., Bui, X.-T., Egger, M., Gasperi, J., Lebreton, L., Nguyen, P.-D., et al. (2018). A methodology to characterize riverine macroplastic emission into the ocean. *Frontiers in Marine Science*, *5*, 372.

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

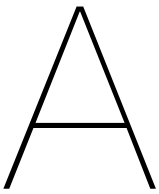
Vriend, P., Roebroek, C. T., & Van Emmerik, T. (2020). Same but different: A framework to design and compare riverbank plastic monitoring strategies. *Frontiers in water*, *2*, 563791.

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.

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? *Advances in neural information processing systems*, *27*.

Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., & Chen, J. (2024). Detrs beat yolos on real-time object detection. https://arxiv.org/abs/2304.08069

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.



Litter Survey with detailed OSPAR classes

OSPAR Floating Macro-Litter Survey Sheet

1. General Information

Survey Date:_____

Observer(s):_____

Location Name:_____ (e.g., Station Bridge, Delft)

River/Canal Name:_____

GPS Coordinates:

Lat:_____ Long:_____

Observation Point:_____

(e.g., Center of bridge)

Start Time (HH:MM):_____ End Time (HH:MM):_____

Total Duration (min):_____

Observed Width (m):_____ (Width of canal section monitored)

2. Environmental Conditions

Condition	Details
Weather:	☐ Sunny ☐ Partly Cloudy ☐ Overcast ☐ Drizzle ☐ Rain ☐ Fog
Wind Direction:	(Direction wind is coming **FROM**) N / E / S / W / Other: _____
Wind Speed (Beaufort):	☐ 0 (Calm) ☐ 1–2 (Light) ☐ 3–4 (Moderate) ☐ 5+ (Strong)
Water Flow Direction:	*(e.g., North to South)* _____
Water Flow Speed:	☐ Stagnant ☐ Slow ☐ Moderate ☐ Fast
Recent Weather:	*(e.g., Heavy rain yesterday, dry spell for 3 days)* _____ _____

3. Litter Count

Use tally marks (e.g., ||||) for counting during observation.

Item Description	Material	Tally Marks	Total
— PLASTICS —			
Bottle (>= 0.5 L)	PET		
Bottle (< 0.5 L)	PET		
Six pack ring	PO soft		
Plastic bag	PO soft		
Garbage bag	PO soft		
Small bag	PO soft		
Net bag	PO soft		
Bottle label	PO soft		
Cleaning glove	PO soft		
Rope D>1cm	PO soft		
Rope D<1cm	PO soft		
Fishing gear	PO soft		
Fishing wire	PO soft		
Tape	PO soft		
Industrial packaging	PO soft		
Other medical	PO soft		
Glove	PO soft		
Cleaning product packaging	PO hard		
Cosmetics packaging	PO hard		
Motor oil packaging (< 50 cm)	PO hard		
Motor oil packaging (>= 50 cm)	PO hard		
Jerrycan	PO hard		
Caulking nozzle	PO hard		
Hair brush	PO hard		
Lollipop stick	PO hard		
Glowstick	PO hard		
Bucket	PO hard		
Plastic plant pot	PO hard		
Cable tie	PO hard		
Helmet	PO hard		
Rifle cartridge case	PO hard		
Firework	PO hard		
Cotton swab	PO hard		
Toilet refresher	PO hard		
Syringe	PO hard		
Water filter	PO hard		
Food packaging	PS		
Caps and lids	PS		
Lighter	PO hard		

Pen	PO hard
Toy	PS
Cup	PS
Cutlery	PS
Straw	PS
Swizzle stick	PS
Plastic plate	PS
Foam food packaging	EPS
Foam cup	EPS
Food wrapping	Multilayer
Medical packaging	Multilayer
Hard fragment (>= 2.5 cm)	PO hard
Soft fragment (>= 2.5 cm)	PO soft
Hard fragment (> 50 cm)	PO hard
Soft fragment (> 50 cm)	PO soft
Hard fragment (< 2.5 cm)	PO hard
Soft fragment (< 2.5 cm)	PO soft
Foam fragment (>= 2.5 cm)	EPS
Foam (> 50 cm)	EPS
Foam fragment (< 2.5 cm)	EPS
Other	Other plastic
— RUBBER —	
Balloon	Rubber
Tire	Rubber
Condom	Rubber
Other rubber	Rubber
— TEXTILE / CLOTH —	
Clothing	Textile
Carpet	Textile
Shoeware	Textile
Sanitary towel	Textile
Tampon (applicator)	Textile
Wet tissue	Textile
Toilet paper	Textile
Other sanitary	Textile
Other textile	Textile
— PAPER / CARDBOARD —	
Paper bag	Paper
Cartboard	Paper
Drink carton	Paper
Cigarette pack	Paper
Cigarette filter	Paper
Cartboard cup	Paper
Newspaper	Paper
Carton cotton swab	Paper
Other paper	Paper
— WOOD —	
Cork	Wood
Pellet	Wood
Popsicle stick	Wood
Paintbrush	Wood
Other wood (< 50 cm)	Wood
Other wood (>= 50 cm)	Wood
— METAL —	
Spray can	Metal
Metal bottle cap	Metal
Drink can	Metal
Food can	Metal
Electrical wire	Metal
Fish lead	Metal

Aluminium foil	Metal
Metal capsule	Metal
Old iron scrap	Metal
Oil drum	Metal
Paint can	Metal
Barbed wire	Metal
Single use grill	Metal
Other metal (< 50 cm)	Metal
Other metal (>= 50 cm)	Metal
— GLASS —	
Glass bottles and ceramics	Glass
Tube lamp	Glass
Other glass	Glass
— OTHER —	
Other	

(Specify “Other” items in the notes below.)

4. Notes & Observations

Record any relevant information, such as potential litter sources (e.g., overflow, nearby event), presence of litter traps, photos taken, or specific details about “Other” items.

B

Full List of Trained Models for each Label Setup

Table B.1: Overview of experimental runs, strategy groups, and descriptions.

Run Name	Strategy Group	Description
01_Baseline	Baseline (real Data)	Baseline model with no optimized HPs
02_Improved_lr_wd	Baseline (real Data)	Model with optimal learning rate and weight decay
03_Improved_HP	Baseline (real Data)	Model with optimized augmentation parameters after Optuna study
04_Random_25%	Synthetic	Model with 25% added synthetic data to the training set.
05_Random_50%	Synthetic	Model with 50% added synthetic data to the training set.
06_Random_100%	Synthetic	Model with 100% added synthetic data to the training set.
07_Weighted_25%	Synthetic	Model with 25% added synthetic data to the training set.
08_Weighted_50%	Synthetic	Model with 50% added synthetic data to the training set.
09_Weighted_100%	Synthetic	Model with 100% added synthetic data to the training set.
10_Mixed_200%	Synthetic	200% of synthetic images were added combining both random and weighted logics.
11_Finetuned_Random	Synthetic	Pretraining on a synthetic data model, validation against real data; Finetuning with real data.
12_Finetuned_Weighted	Synthetic	Pretraining on a synthetic data model, validation against real data; Finetuning with real data.
13_Finetuned_Mixed	Synthetic	Pretraining on a synthetic data model, validation against real data; Finetuning with real data.
14_Finetuned_Weighted_synth_split	Synthetic	Pretraining on a synthetic data, validation against synthetic data; Finetuning with real data.
15_Finetuned_Random_synth_split	Synthetic	Pretraining on a synthetic data, validation against synthetic data; Finetuning with real data.
16_Finetuned_Weighted_synth_split	Synthetic	Pretraining on a synthetic data, validation against synthetic data; Finetuning with real data.

C

Calculation of Retention Factor for Kooybrug/Den Helder

This section details the probabilistic modelling approach used to estimate the retention capacity of the study reach. The estimation relies on the stochastic transport framework developed by Newbould et al. (2021), which conceptualizes the downstream movement of macroplastic debris as a series of discrete events.

The model divides the river reach into a series of cells of length L (10 m). For a particle entering a cell i, the probability of becoming trapped, $p(T)_i$, is defined by the interaction of three independent trapping mechanisms: meander bends (M), channel banks (CB), and vegetation (V). The total probability of trapping is given by Equation C.1:

$$p(T) = 1 - ((1 - p(M)) \cdot (1 - p(CB)) \cdot (1 - p(V))) \tag{C.1}$$

Once trapped, a particle has a daily probability of release, $q(i)$, which determines the duration of the rest phase. A value of $q(i) = 0.5$ was applied in this study, consistent with the baseline assumption established by Newbould et al. (2021) in the absence of trap-specific release data.

The general model was adapted to the specific morphological characteristics of the study site:

1. **Meanders ($p(M) \approx 0$):** As the canal is characterized by a straight channel geometry, the sinuosity ratio S approaches unity ($S \approx 1$). According to the function $p(M) = 1 - S^{-a}$, this yields a negligible trapping probability for meanders.
2. **Channel Banks ($p(CB) \approx 0$):** The study reach features reinforced artificial banks (e.g., sheet piling/stone) rather than natural rough banks. Consequently, hydraulic trapping due to bank roughness was assumed to be negligible compared to vegetation trapping.
3. **Vegetation ($p(V)$):** Riparian vegetation was identified as the dominant trapping mechanism. For cells containing vegetation, a trapping probability of $p(V) = 0.2$ was applied, adopting the value from the reference study (Newbould et al., 2021).

The density of effective vegetation traps was determined through inverse modeling, using in-situ tracer observations. Field data indicated that macroplastic items required approximately 7 days to traverse the 3.8 km reach ($v \approx 540$ m day^{-1}). A Monte Carlo simulation ($N = 5000$) was performed to calibrate the effective vegetation cover. The simulation indicated that an effective trap density of approximately 1% (reflecting overhanging vegetation) was required to reproduce the observed 7-day travel time. Under these calibrated conditions, the model yielded a retention factor 0.52. This implies that, approximately 52% of the floating litter load is expected to be retained within the canal reach for a duration exceeding the 7-day observation window, while 48% is exported to the removal point. This calculation does not take the influence of the pumping station into account, which likely affects retention and re-mobilization behavior of litter items. Further field work is needed to further investigate the local characteristics.

D

Class-wise OSPAR7 and OSPAR13 PR curves for in-domain test set

The following figures show the class wise PR curves on the in-domain test set for OSPAR7 and OSPAR13 models. In general, synthetic data does not improve class wise performance. Especially, underrepresented classes do not seem to improve with the addition of synthetic images.

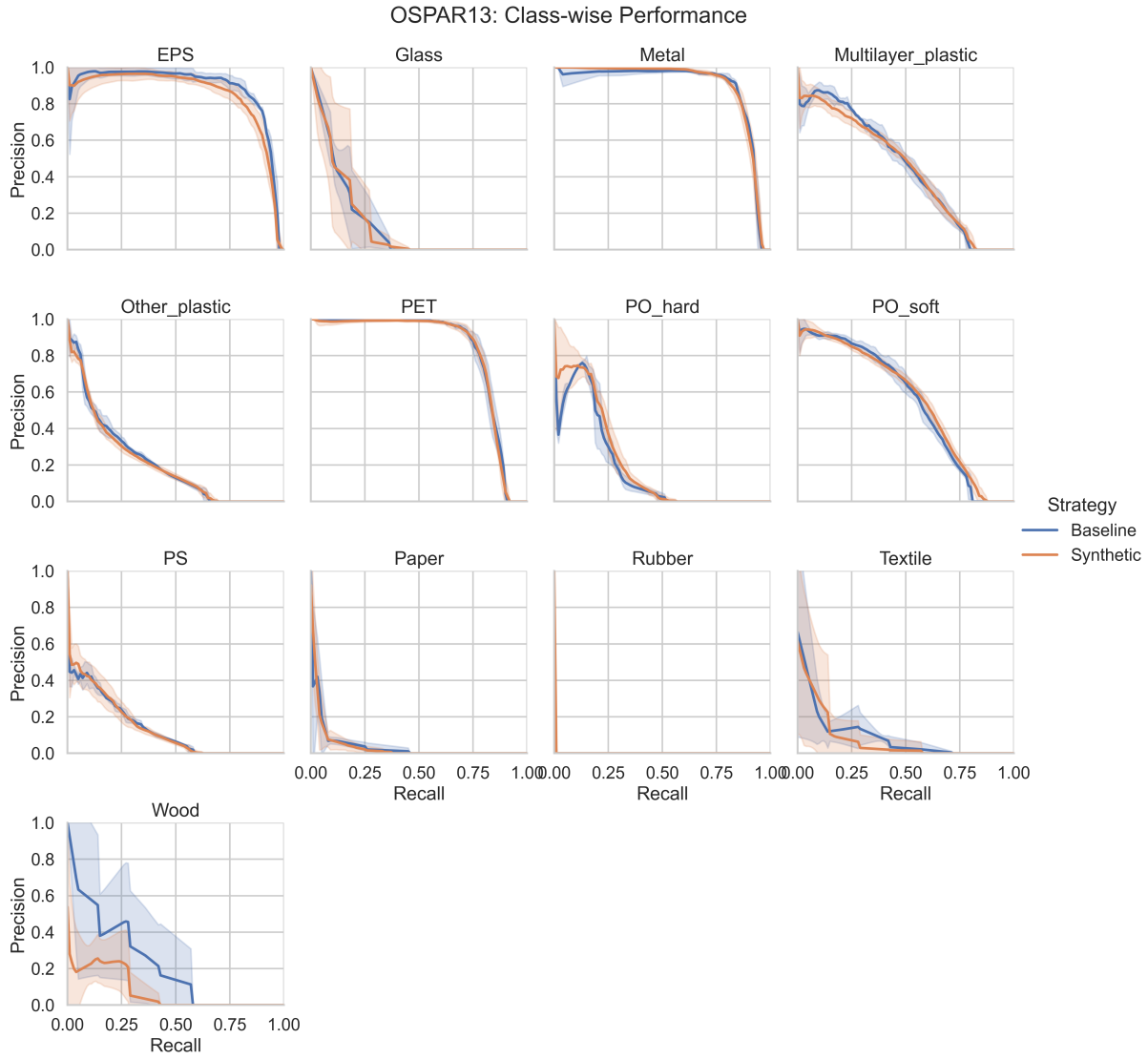


Figure D.1: Class wise PR curves for OSPAR13 on in-domain test set

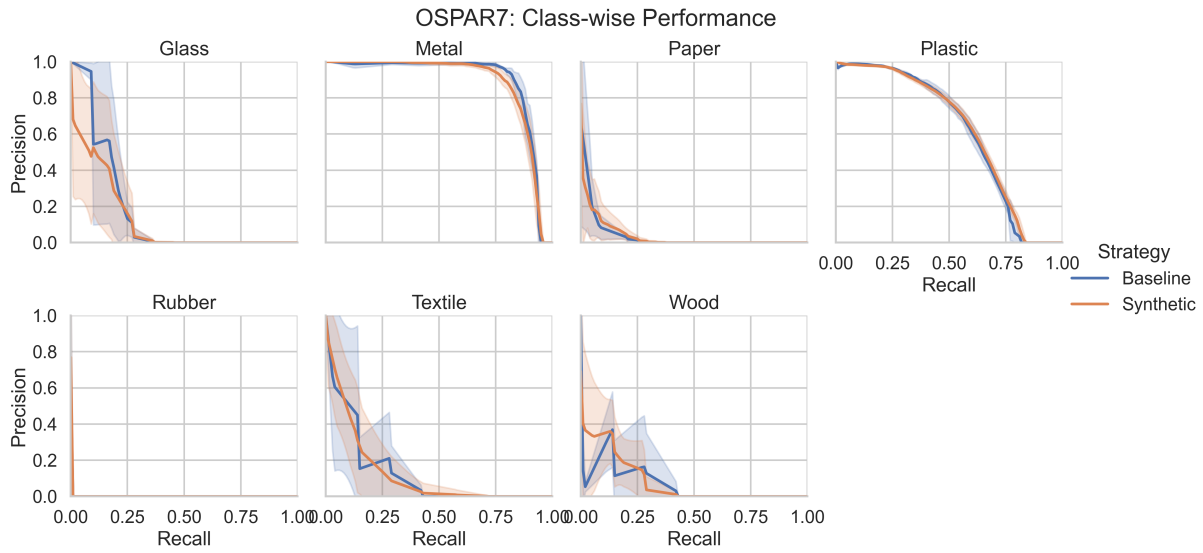


Figure D.2: Class wise PR curves for OSPAR7 on in-domain test set

E

Benchmark F1 curves for YOLOv11m, RT-DETR-R18 and Faster RCNN-R18

The following figures show the weighted F1 curves across three tested architectures. Faster RCNN shows stable performance across the whole confidence range for both Binary and OSPAR13 setups. The RT-DETR model seems to peak around a confidence threshold of 0.3, but shows overall instability. Especially, the multi-class setup on the RT-DETR is not competitive with the other architectures. Generalization for both architectures is inferior compared to YOLOv11.

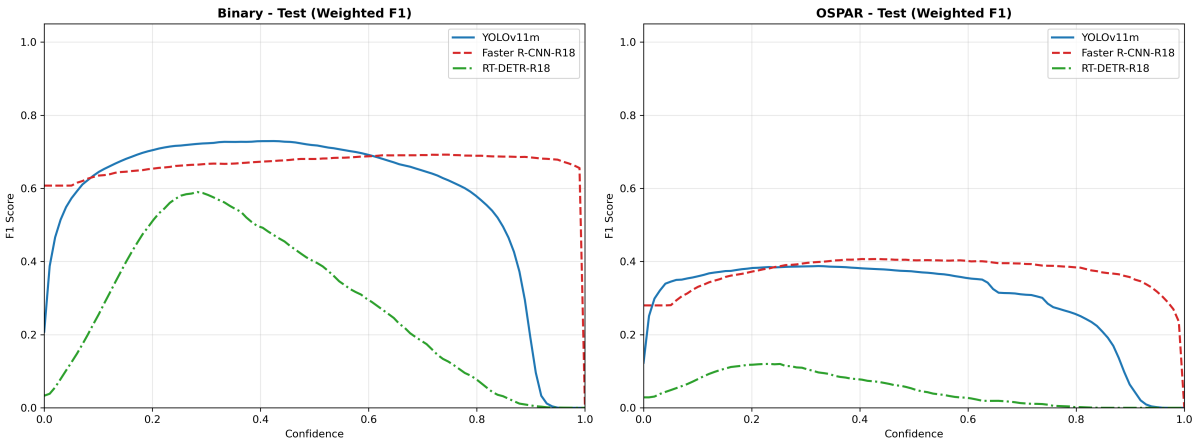


Figure E.1: Benchmark F1 curves for in-domain test set

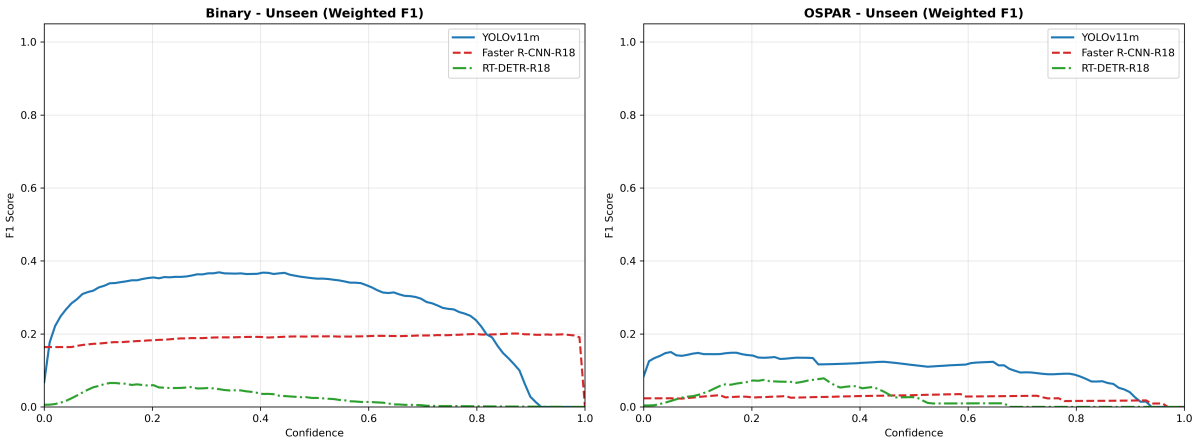


Figure E.2: Benchmark F1 curves for out-of-domain test sets

Class Wise Detection Counts and Mass Estimates for mass-flux tests

These are the detailed class-wise estimations for the test period in Den Helder for both OSPAR7 and OSPAR13. The following plots show significant differences from the ground-truth composition.

Kooybrug

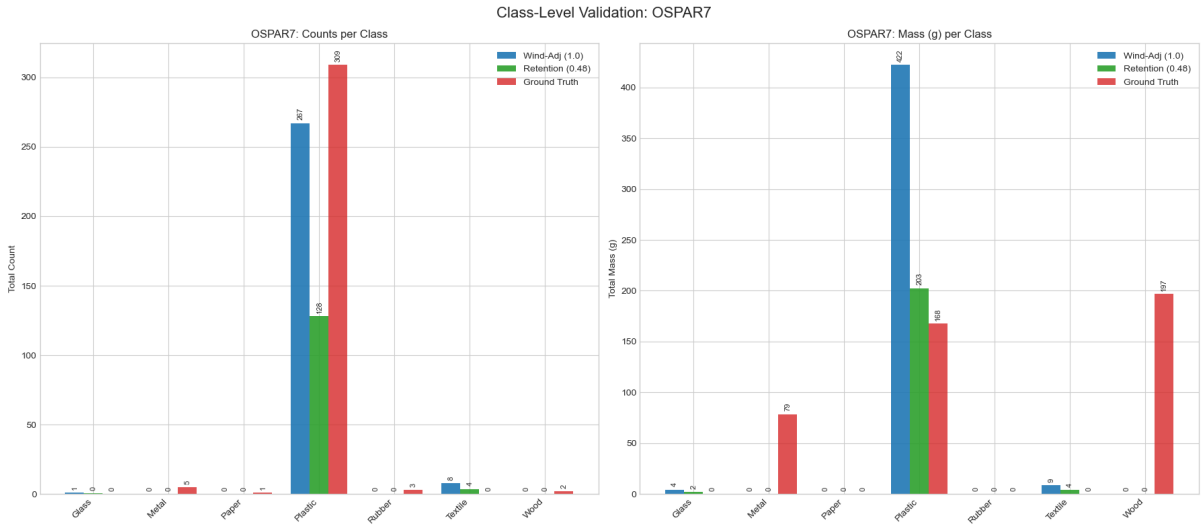


Figure F.1: Class wise counts and mass predictions OSPAR7 at Kooybrug

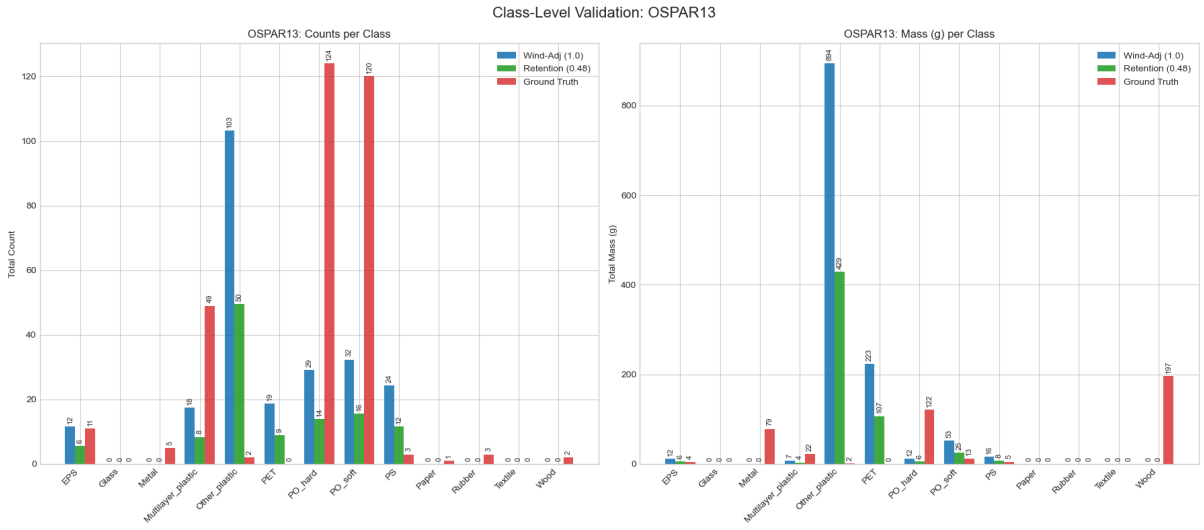


Figure F.2: Class wise counts and mass predictions OSPAR13 at Kooybrug

Oostertoegang

The following plots show the detailed class-wise estimations for the test period in Amsterdam for both OSPAR7 and OSPAR13 models. For both models, the majority of items from underrepresented classes are not detected, especially Paper items (mostly cigarette filters). The plastic fraction is more accurate.

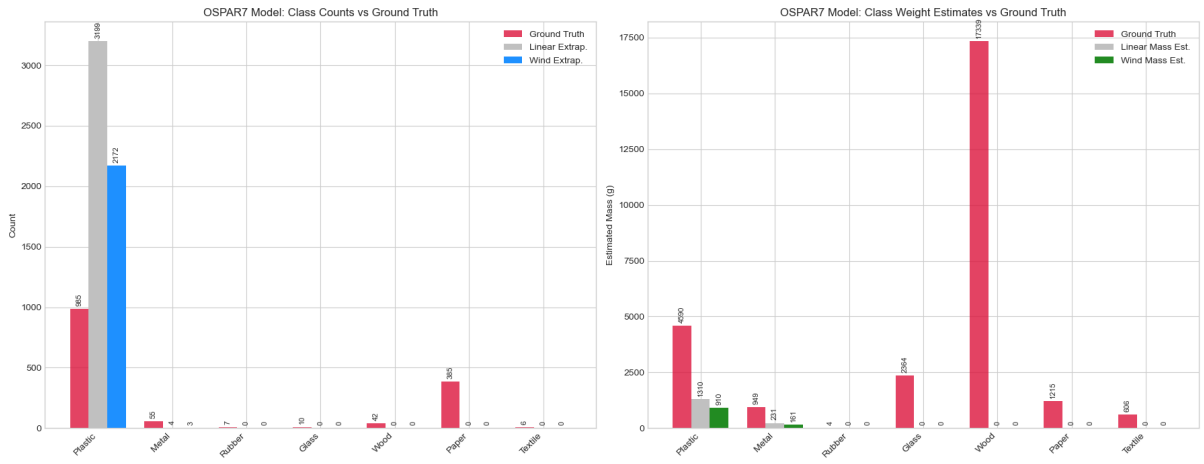


Figure F.3: Class wise counts and mass predictions OSPAR7 at Oostertoegang

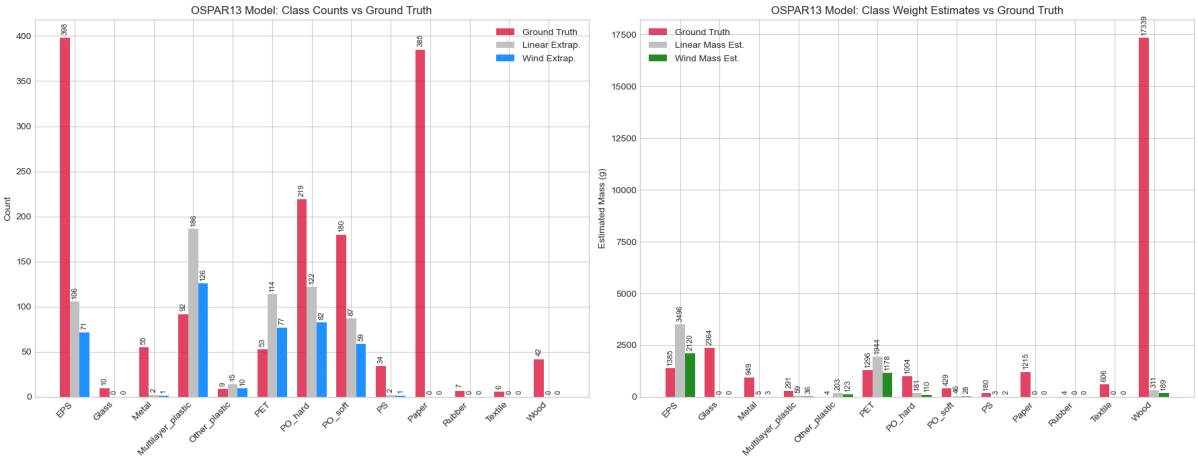


Figure F.4: Class wise counts and mass predictions OSPAR13 at Oostertoegang

G

Detailed Weight Estimations for Visual Counting

The results section indicated weight estimates for Visual Counting for all weight datasets combined. The following plots show the more detauked estimates. For Den Helder, all weight sources lead to a significant overestimation of the weight. The site specific data in Den Helder is comprised of the actual ground truth data as no data from prior batches was available.

At Oostertoegang, the site specific data from a prior batch resulted in the closest estimates as well. Using data from other locations still results in significant overestimation for the visual counting method. Site specific data might also not always be available.

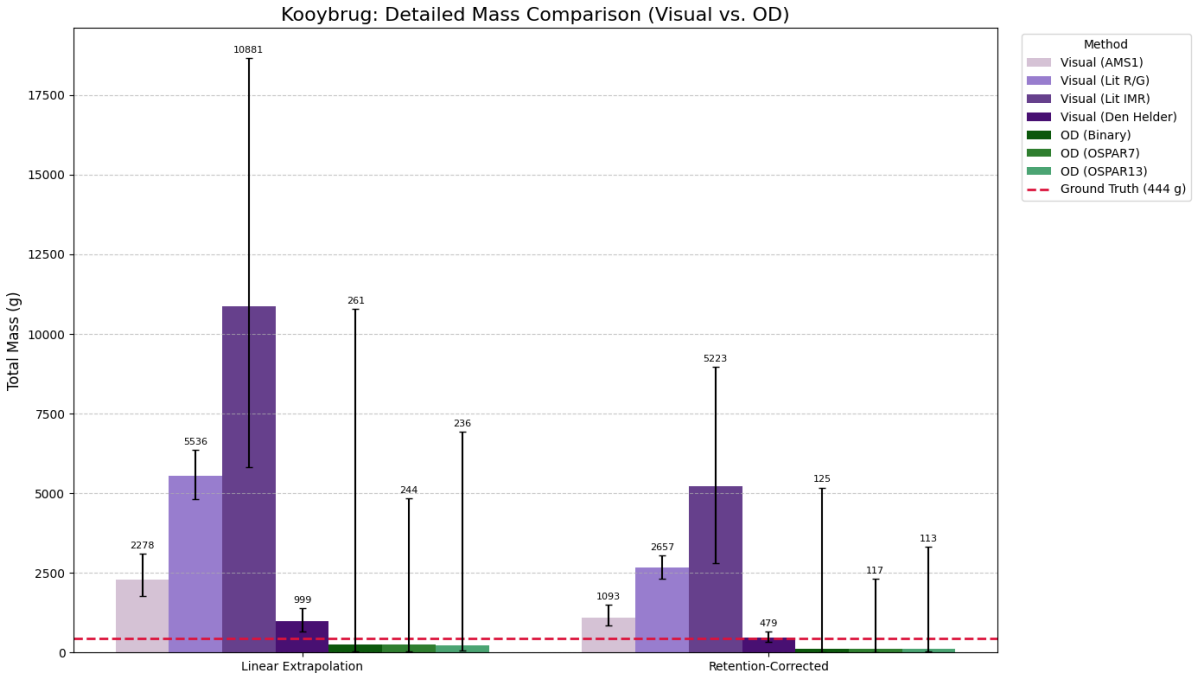


Figure G.1: Detailed Weight Estimates for Visual Counting at Kooybrug/Den Helder

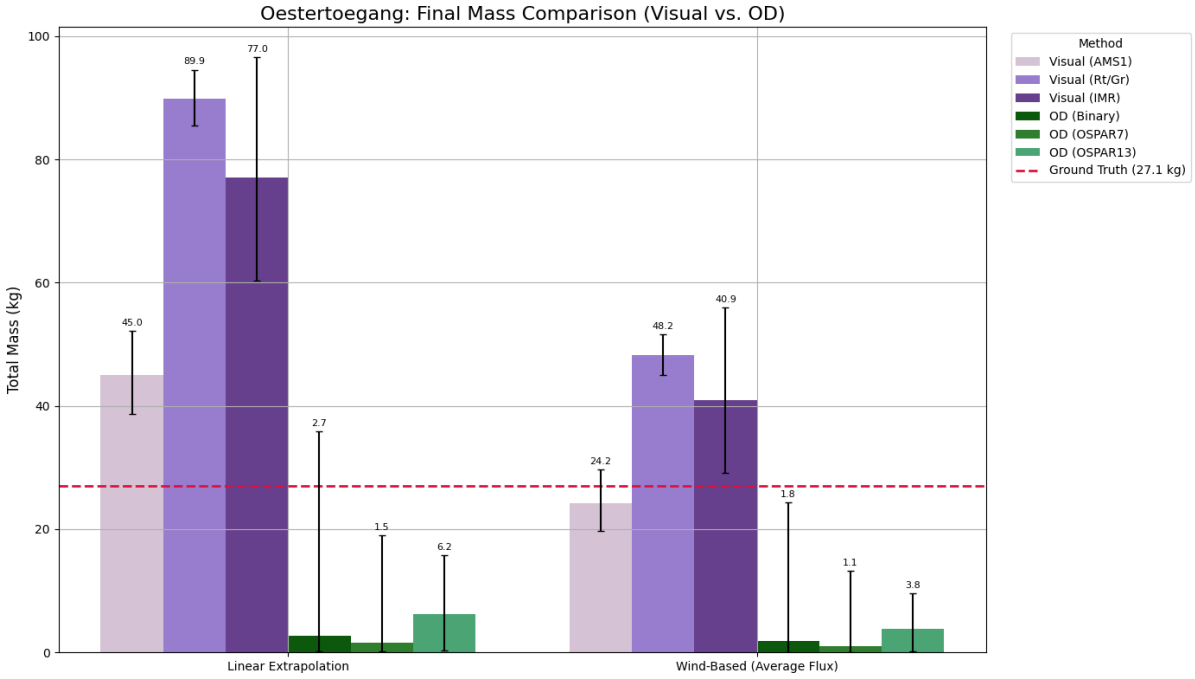


Figure G.2: Detailed Weight Estimates for Visual Counting at Oostertoegang/Amsterdam