

ARCAM

Domain Adaptation for Camera-Based River Waste
Detection in Durban, South Africa

RO57035: Master Thesis

Korneel Somers



ARCAM

Domain Adaptation for Camera-Based River Waste Detection in Durban, South Africa

by

Korneel Somers

to obtain the degree of Master of Science

at the Delft University of Technology,

to be defended publicly on Thursday March 13, 2025 at 11:00 AM.

Student number:	4851730	
Faculty:	Cognitive Robotics, Mechanical Engineering	
Project Duration:	March, 2024 - March, 2025	
Supervisors:	J. Kooij	TU Delft
	H. Kolvenbach	ETH Zurich
	E. Elbir	ETH Zurich
	J. Tkaczuk	ETH Zurich
External Assessor:	J. Alonso-Mora	TU Delft

An electronic version of this thesis is available at <http://repository.tudelft.nl/>.

Contents

Preface	v
Abstract	vi
Symbols	vii
1 Introduction	2
1.1 Plastic Pollution	2
1.2 River Waste Monitoring	3
1.3 Thesis Outline	4
2 Related Work	6
2.1 Vision-Based Waste Detection	6
2.2 Domain Adaptation	8
2.3 Previous Thesis Work	9
2.3.1 Bridge Mounted Camera in Switzerland	10
2.3.2 Physical Sampling in South Africa	13
2.4 Thesis Contributions	14
3 Methods	16
3.1 System Design	17
3.1.1 Detection Box	17
3.1.2 Power Management	19
3.1.3 Mounting System	19
3.2 Data Annotation	22
3.3 Object Detector	23
3.3.1 Baseline Model	24
3.3.2 Few-Shot Fine-Tuning	25
3.3.3 Pseudolabel Self Training	25
3.4 Multimodal Foundation Models	26
3.5 Evaluation Metrics	26
3.5.1 Object Detection	27
3.5.2 Scalability	28
4 Data Acquisition	29
4.1 Imaging Parameters	29
4.2 Image Collection in Switzerland	30
4.2.1 ARCHE	30
4.2.2 Street Parade	31
4.3 Image Collection in South Africa	31
4.3.1 Green Corridors	32
4.3.2 uMhlangane River	32
4.3.3 Weather-Waste Relationship	33

4.4	Datasets Overview	34
5	Results & Discussion	37
5.1	Baseline Detector	37
5.2	Few-Shot Fine-Tuning	42
5.3	Pseudolabel Self Training	42
5.4	Comparison	46
5.5	Multimodal Foundation Models	47
6	Conclusions & Outlook	48
	Bibliography	55
A	Related Work Overview	57
B	Design Evolution	59
C	Stakeholder Interviews	64
C.1	Nick Swan - Green Spaces Programme Manager Green Corridors . .	64
C.2	Siphiwe Rakgabale - Coordinator Litter Boom Project	66
D	Experiments	69
D.1	Baseline Model Hyperparameters	69
D.2	Fewshot Finetuning Results	69

Preface

I would like to start by thanking my supervisors. Emre, Hendrik, and Jakub thank you for the guidance and for enabling me to have a master thesis project that was incredibly diverse. It allowed me to learn far beyond than what is documented in this report. An incredible thank you to Julian as well, I have appreciated your forever critical and spot-on feedback on my work.

The thesis had many phases, and along the way, I was supported by some wonderful people. It started with the prototyping at the university workshop, where I could always spar with Wiebe about the design choices. Then came the trips to Switzerland for initial testing, where I was welcomed and supported by the ARC team. But the highlight must have been the time spent in South Africa. Thank you, Marc, for welcoming Emre and me to Durban, making sure we were all set to go, and checking in throughout the whole experience. It was an epic journey where hands-on problem-solving became part of my daily routine. Jonathan, your support there was invaluable.

A special thank you to Sips, who joined me in the field time and again and also introduced me to South African culture, it was a pleasure! I'd also like to thank Musa and his team from Green Corridors for their unwavering support.

Lastly, this master's thesis was part of a collaboration between the Autonomous River Cleanup (ARC) Project, housed in the Robotic Systems Lab, and the Global Health Engineering Lab. It would not have been possible without the support of these fantastic labs at ETH Zurich. A big thank you to Prof. Elizabeth Tilley and Prof. Marco Hutter for supporting such fascinating and insightful projects.

Abstract

Plastic pollution in rivers is a growing environmental issue with widespread impacts. Monitoring the movement of plastic waste across different river systems is challenging due to environmental variability and the limited availability of labeled data. This thesis investigates camera-based methods for detecting floating macroplastics in rivers and explores ways to adapt detection systems to new locations with minimal data. Collecting data from the Limmat River in Zurich and the uMhlangane River in Durban, South Africa, the study assesses the impact of domain shifts on detection performance and proposes a semi-automated annotation pipeline to improve labeling. Furthermore, it tests techniques like few-shot learning and pseudolabeling to address the performance dip. The results show that model performance decreases significantly when applied to new locations but that even with minimal data, camera-based monitoring can provide useful insights for understanding waste movement and informing plastic waste management strategies.

Symbols

Symbols

D_S	Source domain dataset (Zurich)
D_{T1}	Target domain dataset 1 (Durban)
D_{T2}	Target domain dataset 2 (Jakarta)
M_B	Baseline YOLO model
M_F	Few-shot fine-tuned model
M_P	Pseudolabel-based model
θ	Confidence threshold for pseudolabeling
$\mathcal{L}_{\text{YOLO}}$	YOLO loss function
$\mathcal{L}_{\text{bbox}}$	Bounding box loss
\mathcal{L}_{cls}	Classification loss
\mathcal{L}_{df}	Distribution focal loss
$\mathcal{L}_{\text{soft}}$	Soft-label training loss
IoU	Intersection over Union
mAP@50	Mean Average Precision at IoU 0.5
$o_{i,j}$	Oriented bounding box for object j in image i
$b_{i,j}$	Regular bounding box for object j in image i
x_i	Input image
$C(y_i)$	Confidence score of detection y_i

Indices

i	Image index
j	Object index within an image
N_S	Number of images in the source dataset
N_T	Number of images in the target dataset
N_t	Number of images in a subset of the target dataset

Acronyms and Abbreviations

ARC	Autonomous River Cleanup
ARCAM	Autonomous River Cleanup Camera
ETH	Eidgenössische Technische Hochschule

EPR	Extended Producer Responsibility
GSD	Ground Sampling Distance
HFOV	Horizontal Field of View
OBB	Oriented Bounding Box
RBB	Regular Bounding Box
ROS	Robot Operating System
SAM2	Segment Anything Model 2
UAV	Unmanned Aerial Vehicle
VFM	Vision Foundation Model
NGO	- Non-Governmental Organization

Chapter 1

Introduction

Plastic pollution is an increasingly urgent environmental crisis with far-reaching consequences for ecosystems, economies, and human health. Rivers act as major transport pathways for mismanaged waste, carrying plastic from land to sea. Understanding how plastic moves through these waterways is crucial for effective mitigation but is challenging due to the large amount of data required across diverse locations. This chapter provides an overview of the problems of plastic pollution, the need for river waste monitoring, and the challenges in implementing scalable detection systems.

1.1 Plastic Pollution

Materials have always defined human development. In the last century, a new type of material was invented that is now indispensable from our daily lives. Just like stone and bronze did in the past, plastic defines the way we create and consume the world around us. It has seen a growing increase in production from 3 million metric tons (Mt) in 1950 to over 450 Mt per year nowadays [1, 2]. More than a third of this amount is packaging, which is often single-use, which in itself would not be as big of a problem if the end-of-life would be managed well. However, the reality is that only 9% of plastic waste is recycled, and the large majority ends up as landfill or is discarded into the environment [3].

Estimations about the amount of mismanaged waste from land-based activities that ends up in the ocean through rivers range between 0.8 and 13 Mt annually [2, 4]. The magnitude of this range indicates the difficulty of properly monitoring this problem. Research by Meijer et al. (2021) indicates that approximately 1% of the world's rivers, totaling around 1000, account for 80% of the plastic transfer from land to sea [5]. While the number of rivers contributing to the majority of riverine plastic emissions seems to be relatively small and manageable, the challenge lies in implementing consistent and accurate measurement techniques for global analysis. The road from user to ocean can be long, and many more plastics strand along the way. Some will sink, and others will get stuck or land onshore again before reaching the ocean [6]. The variety of factors like weather and material type that influence their movements create high uncertainties about how much plastic will end up where and when. This, together with the complex and global scope of the issue, makes it hard to measure the problem and model the solutions.

One thing that all studies agree on is that mismanaged plastic waste (MPW) has found its way into the hydrosphere and brought negative impacts with it. We like plastics because they are durable and resistant, but that also makes them degrade slowly in the environment. On land, the waste can block drainage systems or hydro-

electric power plants, which can cause flooding and power shortages, respectively [7]. In the ocean, wildlife is directly impacted through entanglement and ingestion of plastic debris, potentially leading to injury or death [8]. The ingestion also leads to plastics ending up in the human food chain. This is concerning because plastics may act as a transport medium for toxic chemicals, posing health risks to humans [9, 10]. Additionally, the mingling of plastics with marine ecosystems interferes with their major role in carbon sequestration, thus affecting the climate [11]. Economically, the presence of plastics in the ocean comes with a considerable cost. Deloitte has estimated the economic impact of marine plastic pollution to range between 6 to 19 billion USD annually, accounting for cleanup efforts and the loss of economic value in sectors such as marine tourism and fisheries [12].

The growing awareness regarding plastic pollution has stimulated various initiatives aimed at mitigating its impact. The Netherlands, amongst others, implemented PET deposit schemes in 2021 to incentivize the return of plastic bottles [13]. Even more important are efforts from middle-income countries, the biggest contributors to MPW, where fast growth in plastic consumption has outpaced the capacity of existing waste management systems [14]. An example is South Africa, which introduced Extended Producer Responsibility payment schemes in 2021 [15]. These initiatives require companies to take responsibility for the end-of-life of their products and pay fees based on the materials they use. Internationally, the United Nations has expressed its commitment to tackling plastic pollution through the 2022 resolution "End plastic pollution: Towards an international legally binding instrument." This resolution marks the start of negotiations for a global agreement and requires a comprehensive understanding of the magnitude and origins of the problem to ensure its effectiveness [16].

Overall, to better understand the problem and assess the effectiveness of mitigation strategies, there is a strong need to gain a better grasp of the scale and impact of mismanaged plastic waste in the hydrosphere. This can be achieved by establishing monitoring projects to track plastic movements within our water networks.

1.2 River Waste Monitoring

River waste monitoring can generally be divided into low-tech and high-tech approaches. Low-tech methods include visual counting and physical sampling, while high-tech methods involve bridge-mounted imagery, UAVs, and remote sensing.

While low-tech methods are valuable because they can be quickly deployed, they have several limitations. They are labour-intensive and offer limited spatiotemporal information. For example, physical sampling typically captures only a small subset of the plastic travelling through rivers, and visual counting depends on the observer's ability to focus and register individual items. The latter gets increasingly difficult in very polluted and fast-flowing rivers. Additionally, low-tech methods are subject to observer bias and fatigue, creating inconsistencies and inaccuracies in the collected dataset [17, 18, 19]. In contrast, high-tech methods offer great potential to overcome these limitations and address the significant spatiotemporal variability required to understand plastic transportation across global water networks, but each approach comes with distinct operating boundaries [20].

Remote sensing via satellites, like the Sentinel-2 by the European Space Agency, is effective for studying large-scale movements of floating debris due to its ability to cover vast areas [21]. However, with a spatial resolution of 10 meters in its highest spectral bands and temporal resolution of 2–5 days, it struggles with detecting smaller, fast-changing debris patterns in rivers [22]. For more detailed waste flux analysis in rivers, UAVs and bridge-mounted cameras are better suited. UAVs provide precise spatiotemporal monitoring and can detect individual waste items

in remote areas, but they are constrained by short flight durations and weather conditions [23, 24]. In contrast, bridge-mounted cameras enable continuous, long-term monitoring without interfering with wildlife or river flow [20].

While each approach has its advantages, scalability remains a significant challenge. To achieve a comprehensive understanding of plastic waste transport, monitoring must extend beyond single rivers to cover broader spatial and temporal scales. However, the current lack of harmonization between monitoring techniques across the globe and the fact that detection algorithms struggle to generalize across diverse locations and conditions make this hard.

Scalability requires monitoring systems to be adaptable both in terms of hardware and software. Physical setups must be cost-effective and deployable in various settings, and algorithms must overcome domain shifts caused by variations in river characteristics, environmental conditions, and data availability. Since there is usually limited labeled data available, exploring unsupervised domain adaptation techniques like image-to-image translation, pseudolabeling, and adversarial feature learning holds great potential to overcome the domain shift.

1.3 Thesis Outline

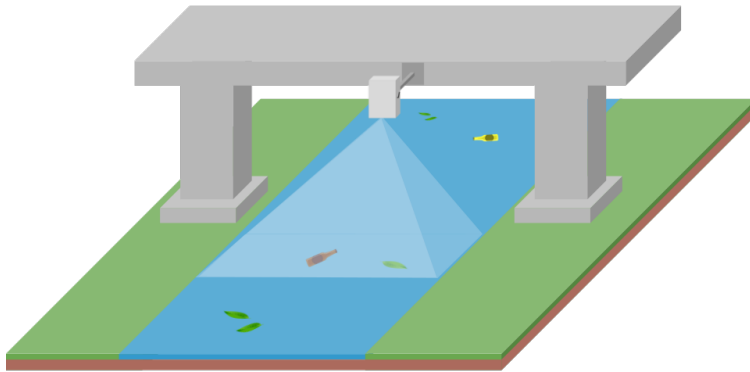
In summary, there is a clear problem of plastic pollution in the hydrosphere and a willingness to do something about it, but the scope and complexity make it hard to model both the problem and solutions. Monitoring waste flux in rivers can help to gain a better understanding, but it is important to be able to cover a large spatiotemporal variability. The latter can be done with vision-based detection strategies, but only if the detections can overcome domain shifts and generalize well to new locations, where limited to no labeled data is available. This is visualized in Figure 1.1 and leads to the following research question:

How can minimal data and effort be leveraged to train a detector for floating macroplastics at new locations, providing actionable insights for policymakers?

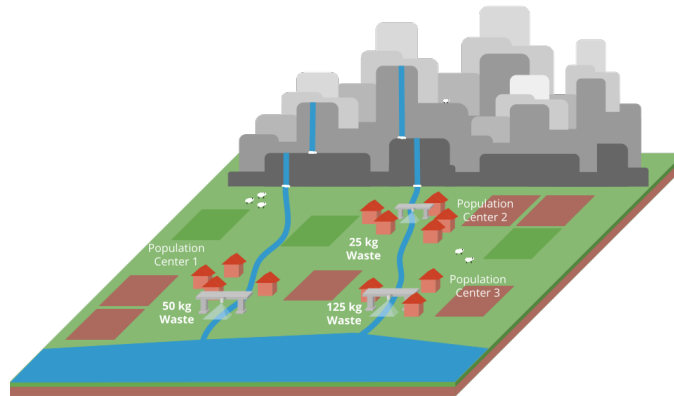
To find an answer to this question, several sub-questions are posed:

1. How are datasets currently processed and analyzed in vision-based monitoring systems, and which detection frameworks and taxonomies provide the most actionable insights for policymakers?
2. What is the magnitude of the domain shift that occurs when moving a system from one river to another?
3. Which domain adaptation techniques could address the domain shift across diverse riverine environments?

As a part of this study, fieldwork will first be conducted in Switzerland with the support of the Autonomous River Cleanup (ARC) project, followed by fieldwork in South Africa in collaboration with a local NGO, Green Corridors. The thesis is structured as follows: Chapter 2, Related Work, reviews the state of the art in vision-based riverine waste detection and outlines the contributions of this thesis. Chapter 3, Methods, describes the setup of the bridge-mounted system for data collection and the methods used to analyze the datasets. Chapter 4, Data Acquisition, details the data collection processes in Zurich and Durban. Chapter 5, Experiments, presents the experimental results and provides an analysis of the findings. Finally, Chapter 6 summarizes the study’s outcomes and offers an outlook for future research.



(a) Monitoring system deployed over a single river.



(b) Monitoring multiple rivers within a network (*adaptation of figure by Emre Elbir*)

Figure 1.1: Scaling waste detection from single river monitoring to waste flux assessment in waternetwork

Chapter 2

Related Work

This chapter will address subquestion one by reviewing existing research on vision-based waste detection and domain adaptation to determine challenges and advancements relevant to riverine waste monitoring (Section 2.1). Current detection models, primarily based on RGB imaging, rely on pre-trained networks fine-tuned on waste-specific datasets. However, their scalability is limited due to the scarcity of labeled data and poor generalization across different environments. Therefore, various domain adaptation techniques from other fields are explored to address these challenges (Section 2.2).

Section 2.3 outlines the previous work that this thesis builds upon. It describes the evolution of the bridge-mounted real-time monitoring system that was developed in Switzerland and the physical waste sampling studies in South Africa that were used for policy implementation.

2.1 Vision-Based Waste Detection

Depending on the purpose of the data outcome, different monitoring techniques must be deployed. For general feedback on the movement of larger waste patches, satellite hyperspectral imagery can be used, but it lacks the precision needed for local policy-making [22]. More targeted monitoring is done using cameras mounted on drones, ships, or bridges. Drones offer flexibility and can reach remote areas without infrastructure but are limited by weather and energy constraints [24]. Ships provide longer monitoring durations but from a ship-view perspective, which is less suitable for waste flux monitoring compared to the general top-view overview that drones offer. Bridge-mounted cameras, though requiring the right infrastructure, can do the longest continuous monitoring and provide a full view of the river if the camera and bridge specifications allow it, making them very suitable for waste flux analysis. All camera-based techniques can use either hyperspectral or RGB imaging, but RGB is the most common due to its affordability and compatibility with existing datasets.

A detailed overview of recent floating waste detection research based on RGB imagery can be found in Appendix A. They often start with pre-trained weights from general-purpose datasets like COCO [25] or ImageNet [26]. These models are then fine-tuned on waste-specific datasets, such as TACO [27], TrashNet [28], and FloW-Img [29]. However, since waste-specific data in riverine environments are limited, researchers tend to create tailored regional datasets, like van Lieshout et al. [30] did at the Grogol River in Jakarta and Jia et al. [31] at the canals in Delft. Figure 2.1 shows examples from each of these waste datasets. Still, many of these custom-collected datasets are not publicly available, limiting accessibility and reproducibil-

ity. Additionally, most are captured from a "ship-view" perspective, whereas a "top-view" approach would be more suitable for waste flux monitoring. Furthermore, many studies are focused on specific locations, making it difficult to generalize findings across different global contexts. This issue is further amplified by the absence of harmonized methodologies or global coordination between studies, making cross-study comparisons challenging [18, 32]. Data augmentations can address these limitations to some extent by increasing dataset variability and accounting for the many variations objects can exhibit in riverine environments. Especially techniques like flipping [31], scaling, and cropping are effective [33], which is logical given that a Coca-Cola bottle can appear in different orientations, sizes, and levels of compression.



Figure 2.1: Examples of waste detection datasets

To detect waste objects, studies typically employ two-stage detectors (e.g., Faster R-CNN, Mask R-CNN, or Cascade R-CNN) or single-stage detectors like those from the YOLO family. Pu et al. [34] compared several models for detecting floating objects and concluded that YOLO models often outperform R-CNN variants both in performance and speed, making them particularly suitable for real-time applications. However, comparing across studies remains difficult due to varying experiment setups, datasets, and evaluation criteria. Reported mean average precision (mAP, further explained in Section 3.5) values can range from 65% [35] to 98% [36], but these figures are highly dependent on camera quality, dataset size, and taxonomy complexity.

A finer taxonomy is more useful for policy-making but increases complexity as it requires sufficient data from each class and appropriate instruments. For instance, South Africa’s EPR framework uses material-based classifications to levy fees on companies based on the type of material their products are made of [15]. Material-based detection is best achieved with hyperspectral cameras in the short-wave infrared (SWIR) range, which captures unique absorption features of plastics around 1215 nm and 1410 nm [37]. RGB imaging can do object-based detection at best, making material classification indirect. It relies on the assumption that an object is typically made of a specific material, as illustrated in Figure 2.2.

For object-based taxonomy, image resolution plays a big role. Schreyers et al. [20] found that a resolution of approximately 0.1 cm/pixel, such as UAV imagery at 5 meters elevation, is sufficient to detect plastic categories and estimate object sizes. However, imagery taken from higher elevations, such as from bridges of 12 to 16

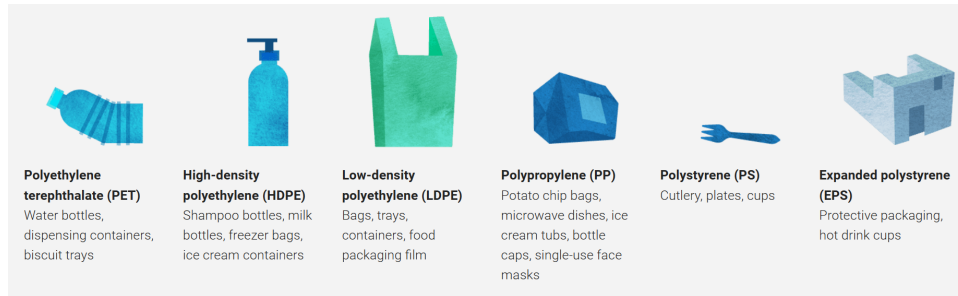


Figure 2.2: Common plastic materials and their typical applications [38]

meters, often has a lower resolution of approximately 0.3 cm/pixel, which allows for detecting items but not distinguishing their categories. Additionally, the more classes, the higher the complexity, making it harder to get satisfactory detection results. As an alternative, a binary taxonomy can be used (e.g., waste vs. no-waste). This requires lower resolution images with less detailed labeling, but is sufficient for understanding overall waste fluxes perform targeted cleanups, see if they are effective and understand better how waste generally moves through their region, rather than targeting specific companies upstream

Nevertheless, even simple binary classification models struggle with domain shifts. For example, a detection model trained in Jakarta performed poorly when applied to other locations within the same water network [30]. This issue gets worse when attempting to scale systems globally, stressing the importance of domain adaptation strategies and harmonized protocols to support widespread applicability. This leads to the most pressing challenge: making river waste detection algorithms more scalable to use their outcomes on a city, country, or global level for policymaking. The main issue is the limited availability of real-world labeled data and the focus on case studies of specific locations with little exploration of domain adaptation techniques.

2.2 Domain Adaptation

From the previous section, we established that limited labeled river waste data is available and that current detection models often perform poorly in new locations. This happens because of the domain shifts that can arise due to differences in backgrounds, lighting, camera equipment, and other factors that alter data distributions between the source and target domains. Extensive manual labeling of a new dataset every time the monitoring system is moved would be a tedious task, for which unsupervised domain adaptation (UDA) techniques can offer a solution. Unsupervised domain adaptation addresses domain shift by aligning the source and target domains without relying on labeled target data [39].

Oza et al. [40] conducted a survey on unsupervised domain adaptation for object detection and identified key challenges and advancements. Three of the most common methods are image-to-image translation, pseudolabeling with self-training, and adversarial feature learning. Each approach operates at different stages of the detection pipeline, offering complementary techniques to overcome the domain shift.

- Image-to-image translation works in the input space and minimizes visual differences between the source and target domains. For example, when lighting or color schemes of the river strongly differ across domains, a model like Cycle-GAN [41] can transform source domain images to resemble the target domain. It does so by introducing generators that map images between the

source and target domains and discriminators that try to distinguish between the translated images and the real ones in the target domain. Guided by an adversarial loss, the generators try to make that distinction impossible to detect. This approach is often combined with other methods to prevent that errors during image translation propagate through the pipeline [40].

- Pseudolabel self-training operates further down the pipeline by using the model trained on the source domain to generate pseudolabels for the target domain. These pseudolabels will then act as ground truth during the next training iterations to improve detection on the target domain. However, to prevent noisy pseudolabels from degrading performance, only high-confidence labels are retained. This filtering can be expanded by generating soft labels, assigning greater weight to highly confident labels [42], or incorporating auxiliary classifiers [43].
- Adversarial feature learning operates in the latent space and aims to make features domain-invariant. It can do so by using Gradient Reversal Layers (GRL) [39] to align the feature representations between domains. The GRL passes data unchanged during forward propagation but inverts gradients during back-propagation to make the feature extractor generate features that confuse the domain classifier. This setup trains the model to, at the same time, minimize task-specific loss for detection performance and maximize domain classification loss to ensure domain invariance. This approach can be extended by, for example, implementing multi-level alignment [44] or strong/weak alignment [45]. They align both global and object-specific features, but they differ in how and where the alignment is applied in the network.

An alternative approach is exploring the potential of vision foundation models (VFMs) to bypass domain adaptation altogether. VFMs are trained on extensive and diverse datasets and may inherently handle domain shifts. For waste detection, it becomes particularly interesting when the semantic understanding of visual scenes is combined with other modalities like natural language, which enables open-vocabulary detection. Multimodal models like Grounding DINO [46] and OWL-ViT [47] accept both an image and text prompts as input (e.g. a picture of a river with trash and the prompt "plastic bottle") and output bounding boxes with confidence scores indicating where the text prompt is detected in the image. Ideally, this approach could support finer taxonomies without requiring retraining. However, it is unsure how well these models perform out-of-the-box in specific scenarios like riverine waste detection. For now, they are still outperformed by lightweight models such as YOLO in terms of inference speed.

A foundation model that could be used in addition rather than as an alternative is the Segment Anything Model 2 (SAM2) by Meta [48]. This model takes images as input and can either automatically segment entire images or specified parts. While it can automate full segmentation, a human is still needed to assign the classes for each segment. When combined with a detection model pre-trained on COCO classes, even labeling could be automated to some extent. However, due to the specific need for waste item identification, human involvement would still be required. This opens the possibility for a combined pipeline where loose human labeling, assisted by SAM2, could speed up the tedious labeling process and improve the precision of labels.

2.3 Previous Thesis Work

This thesis builds on the work of two departments at ETH Zurich. Therefore, it is important to give an overview of the detection system from the Autonomous River

Cleanup (ARC) project that has been through several iterations so far, as well as the manual waste analysis done in Durban by previous students.

2.3.1 Bridge Mounted Camera in Switzerland

The Autonomous River Cleanup (ARC) is an initiative that is part of the Robotic Systems Lab (RSL) at ETH Zurich with the mission to use robotics and machine learning to combat riverine plastic pollution. This goal is pursued through several projects: waste monitoring, waste analysis, and robotic sorting [49]. This thesis is part of the waste monitoring division and will build upon previous work that was done to create a reliable bridge-mounted waste detection camera. The long-term vision is that the camera setup can be simplified and placed at multiple bridges across multiple cities to get a precise and real-time understanding of plastic waste flux. This can lead to targeted action and policies to mitigate plastic pollution in rivers (see Figure 1.1).

At first, ARC performed a preliminary study to explore the possibility of using a bridge-mounted RGB camera with a Mask R-CNN network to detect floating river waste [50]. Figure 2.3 shows the simple setup at 2 meters above the river with detection examples that served as a proof of concept and inspired further development.

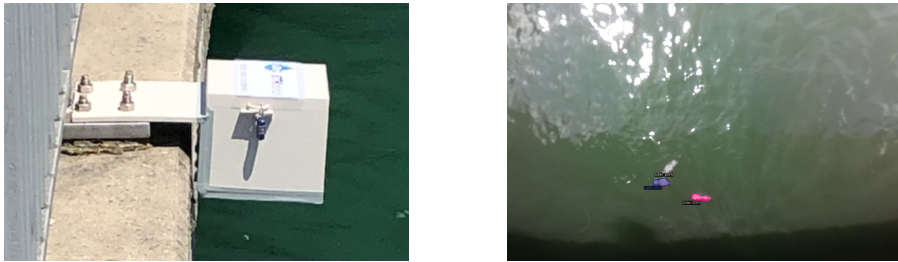


Figure 2.3: Preliminary bridge mounted camera and detection at river Limmat, Zurich

The next step focused on developing a bridge-mounted camera system capable of processing RGB images on-site and in real time, eliminating the need for time-consuming post-processing. This system was named ARCAM (Autonomous River CAMera), and three different students worked on the prototype’s development [51, 52, 53]. The casing and mounting system remained the same across iterations, and the hardware components had some updates, but most change happened with respect to the detection algorithm.

Initially, the setup consisted of a camera and lens for image capture, a computer for real-time inference, a router for wireless data transmission, and a solar panel with a battery and a solar charger to make the system autonomous [51]. At first, the router was not yet integrated into the detection system, that happened later when the IoT capabilities were added together with a new router. Over time, the camera and lens were upgraded several times to improve image resolution. By the last iteration, the hardware components had evolved into the configuration that can be seen in Table 2.1. The biggest drawbacks of the setup are the relatively high cost, which hinders the vision of deploying it at multiple locations, and the difficulty of installing it. Currently, mounting it requires at least 3 people.

The waste detection algorithm progressed in each iteration as shown in Figure 2.5. The first implementation of the waste detection algorithm used the Mask R-CNN instance segmentation model for object detection, outputting segmentations and bounding boxes. For tracking, it employed the Simple Online and Realtime Tracker

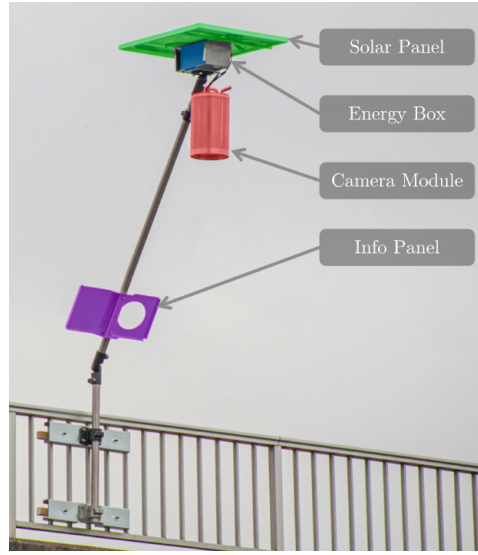


Figure 2.4: ARCAM Bridge Mounted Camera [51]

Component Type	Features and Specifications
Camera	Arducam IMX477 12.3 MP with 1/2.3" sensor
Lens	Arducam 2.8-12 mm, F1.6
Computing unit	Jetson AGX Orin 32 GB RAM, 200 TOPS
Router	Teltonika RUT955 4G, Wi-Fi, GPS
Solar panel	Swaytronic 110 W peak
Battery	LiFePO4 12 V, 18 Ah
Solar charger	SmartSolar MPPT 75/10

Table 2.1: Hardware setup [53]

(SORT) with a Kalman Filter for multi-object tracking [51]. However, Mask R-CNN struggled with cluttered scenes, leading to missed or double detections. The computational demands of the model also limited real-time performance and the absence of wireless data transfer prevented scalability.

Building on the initial efforts, Marco's system introduced a filtering step using HSV (Hue, Saturation, Value) thresholds to identify relevant regions, followed by SORT for tracking. A fine-tuned ResNet-50 model was then used for object detection at intervals rather than every frame. Additionally, IoT capabilities were incorporated which enabled remote data transfer. This iteration addressed issues like double detections by prioritizing existing trackers and filtering objects based on river flow [52]. Nevertheless, reflections and cluttered waste remained challenging and limited real-world data for training restricted the system's robustness.

The last iteration integrated optical flow techniques for pre-filtering, combining ORB feature detection and Lucas-Kanade optical flow with density-based clustering (DBSCAN) to isolate moving objects. A modified version of SORT was used for tracking and minimizing redundant detections. The object detection was done with YOLOv8, which could improve confidence through repeated detections and is very suitable for real-time monitoring. Despite advancements, the datasets remain limited and consist of staged "waste" items that are relatively clean and undamaged, failing to imitate the real-world degradation expected in actual waste. Besides, the system has never been extensively tested outdoors for more than a few hours.

Thus far, ARC has been developing and testing its technologies in Zurich, using the

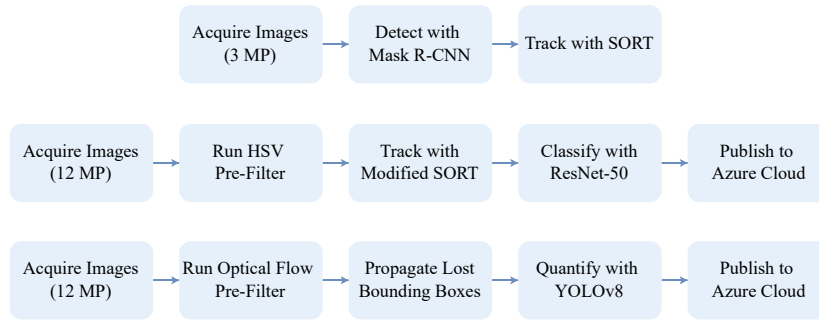


Figure 2.5: Previous iterations of the waste detection algorithm

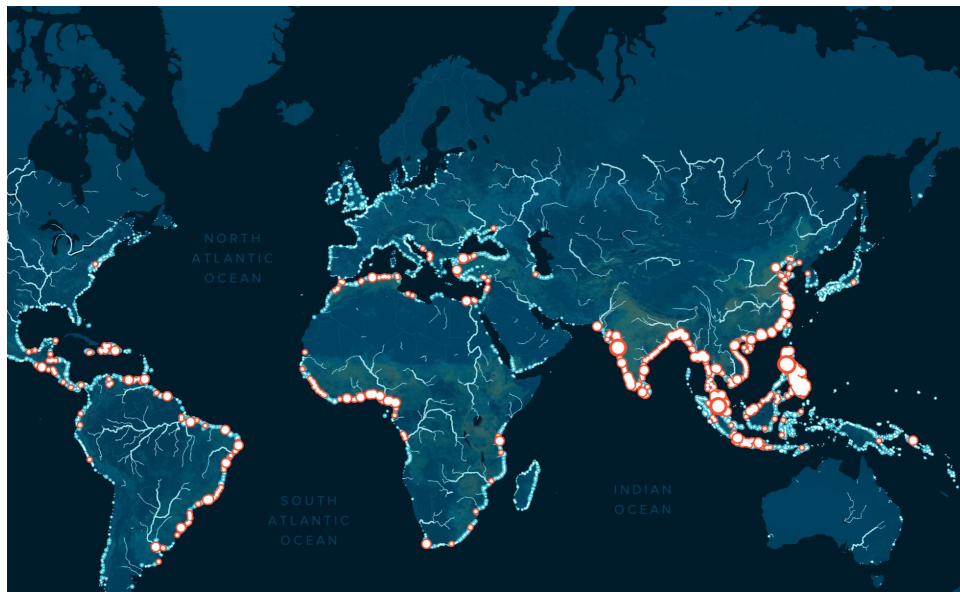


Figure 2.6: World map with 1000 rivers accounting for about 80% of global riverine plastic emissions into the ocean marked in red [54]

Limmat River as a testing ground. In reality, very little waste flows through the river, which requires researchers to manually introduce waste for data collection. Only on exceptional occasions like the Zurich Street Parade, which attracts almost a million visitors, there might be a noticeable increase in plastic waste in the rivers. This highlights the importance of identifying locations where such a system could be most impactful. A good starting point is the map of the 1,000 most polluting rivers by Meijer et al. [5]. Figure 2.6 shows that South-East Asia has the highest density of polluting rivers, but small urban rivers in Africa and Central America are also among the large contributors.

For this research, a river in Durban, South Africa, was chosen because of the already existing collaborations between Green Corridors and the Global Health Engineering (GHE) Department at ETH Zurich. This department focuses on using systems and technologies that can help improve health in over-exploited countries [55]. By working in both Zurich and Durban, the same system can be tested in diverse environments, facilitating domain adaptation experiments. The following section will discuss the physical sampling that has been done by previous students in collaboration with Green Corridors.

2.3.2 Physical Sampling in South Africa

In Autumn 2022, two complementary studies were conducted by Raúl Bergen [56] and Chiara Meyer-Piening [57] in Durban, South Africa. They investigated the composition, sources, and recyclability of plastic pollution in the uMngeni catchment area to inform waste management initiatives and strategies for leveraging EPR returns.

Raúl assessed the impact of South Africa’s mandatory EPR scheme on small waste collection enterprises. Over two months, 906.5 kg of plastic waste was characterized by type, application, and brand to determine who should be targeted for financing or partnership opportunities. Figure 2.7 shows an overview of the waste categories that were found in the litter booms, almost a third of which are soda bottles. This gives an idea about the objects that the detection algorithm will be dealing with. Findings on a material level showed that 54% of PET waste originated from Coca-Cola Bottling South Africa and HDPE/PP plastics mostly came from United National Breweries (33% of the total).

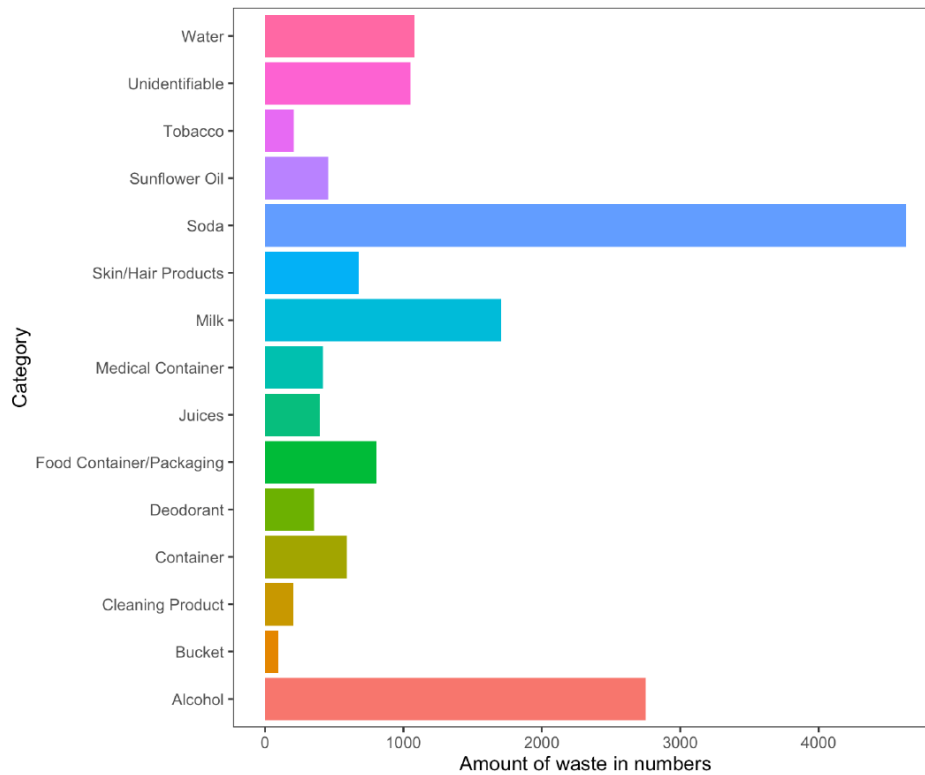


Figure 2.7: Object-based categorization and their respective amount of waste in numbers [56]

Where Raúl focused on what was collected by the litterbooms, Chiara investigated the composition of plastic pollution on Durban’s beaches. By identifying the types of waste and their condition, she aimed to assess their recyclability and determine actionable strategies under South Africa’s EPR policy. Over 1,000 kg of waste was categorized, showing that plastics constituted approximately 58% of the total beachfront waste. The most common plastic types were PET (12.7%), PE (10.5%), PP (12.4%), and PS (11.4%). In particular, PS was often found in a highly degraded state, broken into small pieces, which made it difficult to recycle. Likely because it is a light material that is easily fragmented, PS was rarely found in the litterbooms.

This shows that the litterbooms are mostly effective for capturing larger and heavier floating plastics.

Plastic Type	Price per kg (ZAR)
PET clear	8.0
PET green	6.8
PET brown	6.8
HDPE White	6.3
HDPE mix	6.0
LDPE clear	4.1
PP	3.8

Table 2.2: Listing of the prices per kg of plastic [57]

The outcomes of these studies provide a good starting point for enhancing South Africa’s EPR policy, particularly to better support informal and small-scale waste collectors. The brand and material audits can help to identify partnership opportunities and beneficiation possibilities, but Raúl mentions the lack of consistency. Their efforts, though valuable, covered only a portion of the larger issue as the uMgeni river is estimated to emit about 380,000 kg of plastic waste into the ocean annually [54]. Therefore, there is a need for more quantitative data on plastic waste management. This is where the vision-based monitoring solutions come in. They can provide a more comprehensive view of the waste flux in rivers and complement the existing measures, including all floating plastics. For example, polystyrene, which is often missed by litter booms, could be identified through camera detection.

Other challenges with the EPR fees is that they are spread across numerous projects, reducing their direct impact on collectors. Additionally, personal conversations with stakeholders indicate that, although EPR fees are being paid, up until today, funds are often held up at the Producer Responsibility Organizations (PROs) and do not always seem to flow to relevant initiatives as intended. Even though vision-based monitoring will not directly solve these systemic issues, it can create awareness and provide visual evidence of the problem and the work being done by organizations like Green Corridors. On top of that, the footage can also serve to educate and mobilize the communities to take preventive measures.

Both Raúl and Chiara emphasized the need for empowering informal waste collectors through targeted funding and data-driven strategies. Greater investment is required in education, enhanced waste interception technologies, and adaptive EPR policies that directly benefit small-scale collectors and recyclers. When funding from EPR is not accessible, informal waste pickers rely on the value of the plastics they collect, as shown in Table 2.2. This dependence can then result in selective cleanups, leaving less valuable plastics uncollected.

2.4 Thesis Contributions

Summarizing the gaps in the state of the art: limited scalability of vision-based systems, poor domain generalization across rivers, and inadequate datasets for robust training and evaluation. This thesis addresses these gaps through the following contributions:

- A real-world dataset of riverine waste images from Switzerland and South Africa, addressing the lack of top-view imagery for waste flux analysis.

- A semi-automated annotation pipeline using SAM2 for riverine waste datasets to speed up and improve human labeling.
- An expanded YOLO-based waste detection pipeline integrating domain adaptation techniques and a comparative evaluation of their effectiveness to establish a scalable monitoring strategy.

Beyond those contributions, this thesis also contributed to the ARC project through extensive testing, leading to iterative improvements based on field experience at various locations. This included redesigning the housing and mounting system of ARCAM for better functionality.

Additionally, this work has a societal impact. Tackling plastic pollution also requires consumer awareness. The advantage of this system is its visual output, and since a picture is worth a thousand words, it can support Green Corridors' mission to educate and change the mindset of local communities regarding waste. It also provides concrete evidence of detected waste, which can be used as leverage towards policymakers for support.

Chapter 3

Methods

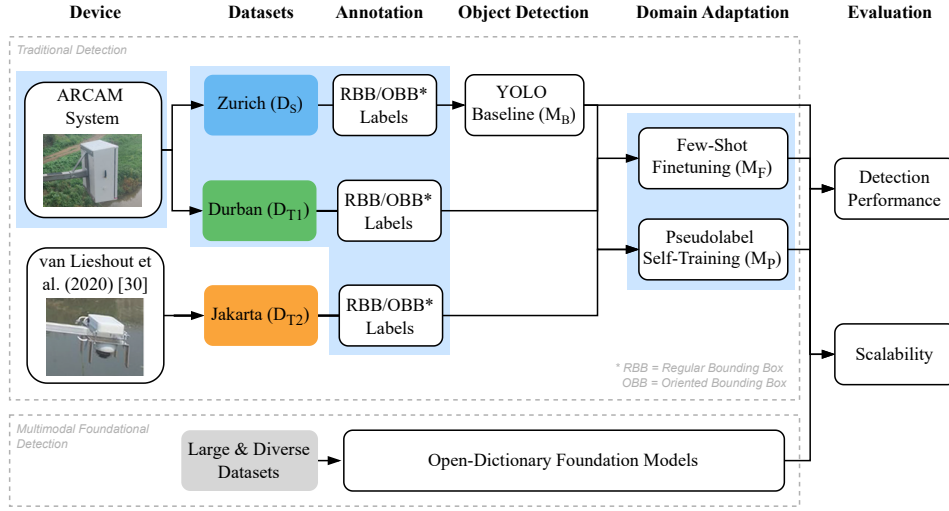


Figure 3.1: Schematic Overview of Methods (contributions marked in blue)

This chapter details the design and implementation of the bridge-mounted river waste detection system for data collection and real-time monitoring. The goal is to develop a robust detector and adaptation strategy that allows the transfer of the system to new locations with minimal effort and data while maintaining consistent performance. A river waste dataset from Zurich serves as the source dataset and one from Durban is used as the target dataset. Both were acquired using the same camera system named ARCAM, which is described in Section 3.1. Furthermore, a dataset from Jakarta that was recorded by van Lieshout et al. [30] is used as a second target dataset to evaluate the consistency of the proposed methods to datasets from a new location with a different camera setup. More details on the data acquisition will be explained in Chapter 4.

With the source dataset, a river waste detector can be developed and enhanced with an annotation improvement pipeline based on the Segment Anything Model 2 (SAM 2) by Meta [48] (see Section 3.2). This improvement would generate better-fitting and oriented bounding boxes to boost detection performance. For the new locations, three approaches are proposed for utilizing the detector: (1) applying the Zurich YOLO detector directly to assess its performance without modification, (2) performing few-shot fine-tuning to determine the minimal amount of new location data required for effective adaptation, and (3) implementing pseudolabeling for self-

training as an unsupervised domain adaptation technique. These approaches are elaborated on in Section 3.3 and will later be analyzed to understand how well the detection system adapts to new environments with varying levels of intervention. Lastly, two multimodal foundation models are introduced as a complete alternative to the traditional pipeline of gathering data and training models to recognize specific objects: OwL-ViT [47] and Grounding DINO [46]. Assessing the performance and suitability of these open-dictionary models in comparison with the proposed pipelines can show if they are ready to replace them. Figure 3.1 shows a schematic overview of how all these proposed methods are linked together.

3.1 System Design

This section presents the redesign of the ARCAM system, introduced earlier in Section 2.3.1. Tables 3.2, 3.3 and 3.4 show the requirements and wishes that guided the process with compactness and ease of installation being particularly important to improve on the previous design while maintaining weather resistance and the capability of continuous monitoring for a longer period of time. The final designs differ slightly for Switzerland and South Africa because the change in circumstances required some adaptations regarding security. Every round of tests and data collection, including those conducted in Geneva and Zurich (see Section 4.2), provided feedback that informed the next iterations and improvements. Details regarding the design evolution are provided in Appendix B.

The entire system is shown in Figure 3.2, with a schematic overview in Figure 3.3. It consists of two subsystems, the energy box and the detection box. They are held together by a mounting system that can be attached to a bridge. The key difference between the system setup for the two locations where it was used is that the Switzerland version includes a solar panel with a solar charger and battery, while the South African version only uses a battery that has to be recharged manually. Additionally, a protective cage was made for security, and the mounting system was changed from a single-axis setup to a multi-screw mount for stability.



Figure 3.2: The ARCAM system deployed in different locations: Geneva, Switzerland (left), Zurich, Switzerland (middle), and Durban, South Africa (right).

3.1.1 Detection Box

The detection box houses all the components that are needed to acquire images of the river surface, execute detection pipelines and transfer relevant information to other devices. These functions are executed respectively by a camera with a lens, a computer and a router. Additionally, a temperature sensor is included to address potential overheating, along with an AirTag for traceability. All components are mounted on a customized baseplate inside an off-the-shelf electricity box to guarantee initial waterproofing.

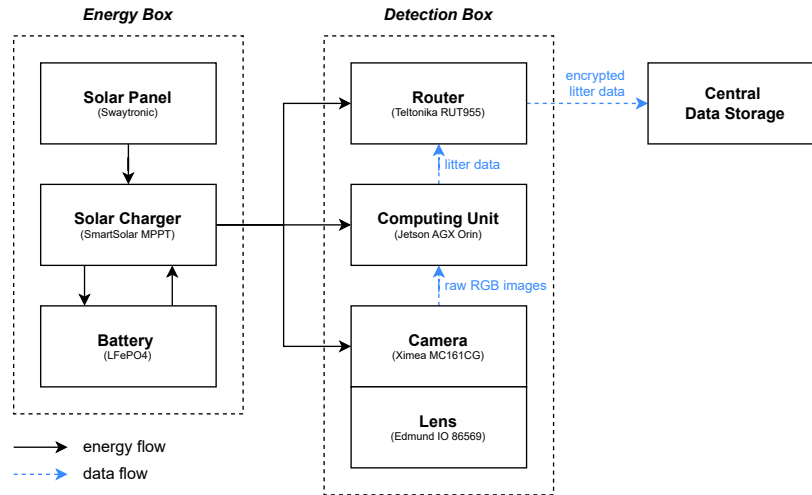


Figure 3.3: Schematic Overview of ARCAM

Different pipelines can be run in a ROS architecture within a docker container. Two primary launch files are used: one for data collection and the other for real-time monitoring. The data collection pipeline saves images at a predefined rate, as recording continuously at 24fps would quickly fill up storage and is unnecessary for training the model. The real-time monitoring is described in Section 2.3.1. Figure 3.4 shows the inside of the detection box and its two pipelines.

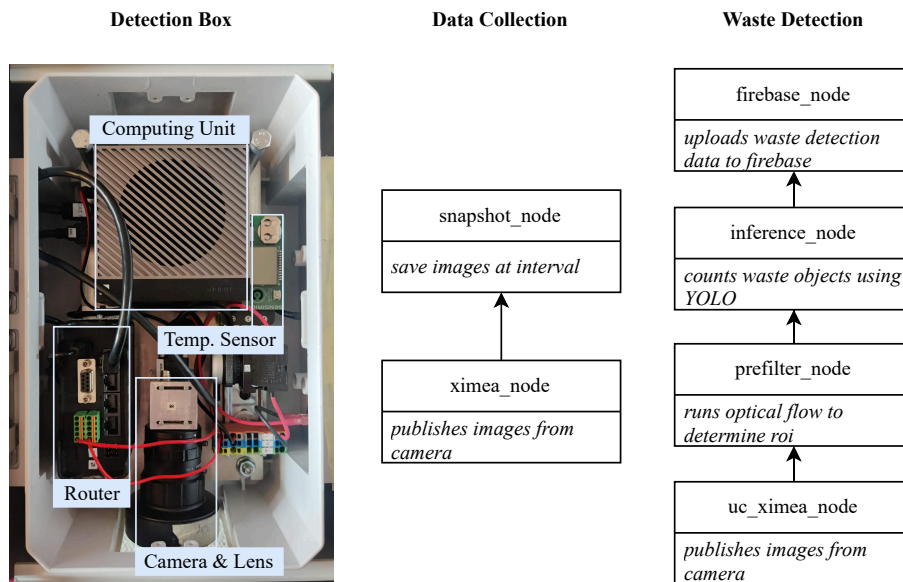


Figure 3.4: Detection Box and Software Pipelines

3.1.2 Power Management

The so-called energy box contains the power management system. In Switzerland, the system used a solar panel with a solar charger for continuous operation, falling back on the battery only during periods without sufficient sunlight. In Durban, the system was adapted to use a rechargeable battery exclusively. The main reason for this change was the fact that a solar panel would attract too much attention, increasing the risk of vandalism or theft.

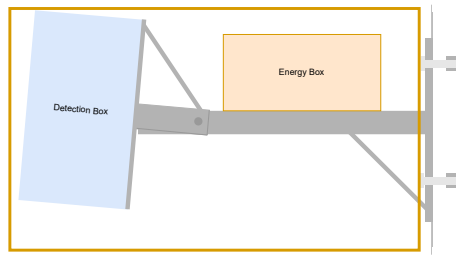
The power consumption of the main components is summarized in Table 3.1. The total power consumption typically ranges from 20–50 W, depending on the workload. Testing in Geneva (see Section 4.2) showed that running the real-time detection pipeline required an average of 23 W. The battery capacity is 18 Ah (230.4 Wh), which supports approximately 10 hours of operation without sunlight, making it sufficient for one full day of daylight measurements.

Component	Power Consumption (W)
Computing Unit	15–40
Router	2–7
Camera	ca. 3.4
Total Typical Consumption	20–50

Table 3.1: Power consumption of the ARCAM system components.

3.1.3 Mounting System

The mounting system is divided into several parts. It starts with the bridge mount, which consists of two plates clamped to the bridge (see Figure 3.5a). Initially, this was done with a single-axis configuration, but in South Africa, it was upgraded to a four-axis design to improve stability and security. The outer plate has a beam attached to it with a screw at the end to adjust the tilt of the detection box (see Figure 3.5b). Lastly, the box mount is screwed to the detection box. It slides over the beam and is secured in place with a pin, as illustrated in Figure 3.5c.



(a) Bridge clamp securing the mount.



(b) Adjustment screw for box tilt.



(c) Locking pin to secure the box mount.

Figure 3.5: Components of the mounting system.

Table 3.2: Requirements (R) and Wishes (W) for system functionality and performance. Results and status indicate the outcome: ✓ fulfilled, * partially fulfilled, and ✗ unfulfilled. Blue rows highlight contributions made in this thesis.

ID	Description	Results	Status
System Requirements (R)			
R1.1	Continuously acquire RGB imagery of floating riverine waste.	Ximea camera captures RGB images, and Snapshot Taker ROS node saves them at set intervals.	✓
R1.2	Capture and transmit data of macro-plastics ($5\text{cm}^2 - 25\text{cm}^2$), including size, type, image, time, and location.	Built on prior work, with scalable detection pipeline proposals detailed in Section 3.3.	✓
R1.3	Maintain a log of GPS coordinates.	Static location ensures consistent GPS logging.	✓
R1.4	Allow for remote access and operation of the system.	Systemd service startup removes the need for remote access, but improved housing and antennas enhance connectivity for checkups.	*
R1.5	Provide real-time data transfer with encryption and secure connection.	Azure Cloud by [52] ensures encryption and secure connectivity.	✓
R1.6	Operate for 8 hours/day.	System consumes 24W on average (see Section 4.2.1), and a 230.4Wh battery suffices (see Section 3.1.2).	✓
R1.7	Capture a large river surface area (at least 10m FOV at 10m height).	At 10m bridge height, the field of view is 18m with a GSD of 0.34 cm/pixel (see Section 4.1).	✓
R1.8	Send notifications and perform emergency shutdown procedures in case of overheating or water intrusion.	Integrated temperature and humidity sensors trigger alerts.	✓
System Wishes (W)			
W1.1	Operate as energy self-sufficient.	Solar power is sufficient in Zurich but was adapted for one-day use in South Africa due to security concerns.	✗
W1.2	Allow data storage (e.g., 1TB) for uninterrupted operation during limited connectivity.	External hard drive on Jetson Orin AGX meets storage needs.	✓
W1.3	Achieve high detection performance.	Performance depends on image resolution and model choice. Benchmarks in Chapter 5.	✓
W1.4	Ensure system safety during operation.	A steel cable secures the system to the bridge, with warning signs and an Failure Modes, Effects, and Criticality Analysis (FMECA).	✓

Table 3.3: Requirements (R) and Wishes (W) for the system and environmental specifications. Results and status indicate the outcome: ✓ fulfilled, * partially fulfilled, and ✗ unfulfilled. Blue rows highlight contributions made in this thesis.

ID	Description	Results	Status
System Requirements (R)			
R2.1	House all components.	Successfully integrated all parts on baseplate.	✓
R2.2	Protect against water and heavy rain (IP55 or higher).	Off-the-shelf electricity box guarantees IP67 compliance.	✓
R2.3	Ensure impact resistance (IK08: 1.7kg dropped from 300mm).	Turned out to be of low priority and therefore not extensively tested.	✗
R2.4	Keep internal temperature below 50°C.	Temperature sensors monitored internal temperature; ventilation vents, a high-albedo case, and optional reflective foil were sufficient (see Chapter 4).	✓
R2.5	Operate in varying lighting conditions.	Functions well during the day with an optional polarizing filter to reduce sunlight glare but struggles at dusk and dawn.	*
System Wishes (W)			
W2.1	Lightweight design.	Weight of detection box was reduced from 13.5 kg (previous design) to 7 kg.	✓
W2.2	Compact design.	Volume of the detection box was reduced from 16,600cm ³ to 13,000cm ³ .	✓
W2.3	Low production costs.	Standardized parts reduce material cost of housing and mount. However, electronics make up for the large majority of the costs and remained unchanged.	*
W2.4	Allow for heat dissipation.	Added air vents for passive cooling.	✓
W2.5	Adapt to different river/bridge heights.	The single-axis mounting system adapts to different bridges without modifications, while the multi-axis version requires different screws but no other changes. The software strategy is detailed in Section 3.3.	*

Table 3.4: Requirements (R) and Wishes (W) for installation and lifecycle considerations. Results and status indicate the outcome: ✓ fulfilled, * partially fulfilled, and ✗ unfulfilled. Blue rows highlight contributions made in this thesis.

ID	Description	Results	Status
System Requirements (R)			
R3.1	Protected against vandalism and theft.	Tamper-proof screws, a welded cage (South Africa), an AirTag, and a weathered exterior provide protection.	✓
R3.2	Ensure a lifespan of at least 6 months.	System sustained throughout a research period of 5 months, although not continuously deployed.	*
W3.3	Allow mounting by one person in under 15 minutes.	Practiced setup time is under ~10 minutes, including the cage. The systemd service automates startup without manual intervention.	✓
System Wishes (W)			
W3.1	Easy mounting on common bridge types.	Successfully mounted on 5 different bridge and railing types with small adjustments.	✓
W3.2	Shield from direct sunlight.	Since it operates throughout the day, avoiding sunlight is challenging. A polarizing filter can reduce glare effects.	*
W3.3	Be accessible for bi-weekly maintenance.	Due to circumstances, maintenance was almost daily, which was feasible given the low setup time.	✗

3.2 Data Annotation

After collecting the data with the ARCAM system, it has to be annotated to train and evaluate the detector. A problem is that human bounding box annotations are prone to inaccuracies due to the often small size of objects in the image, making it a tedious task of zooming in for every label where a mistake is easily made. Examples of poorly annotated bounding boxes are shown in Figure 3.7. These inaccuracies can negatively impact the model during training and validation. During training, imprecise ground truths can lead to suboptimal weight updates by wrongly increasing the bounding box loss $\mathcal{L}_{\text{bbox}}$ and the distribution focal loss \mathcal{L}_{dfl} . During validation, the discrepancies between the predicted bounding boxes and the inaccurate ground truth might penalize the model’s performance unfairly.

To address this, an improved annotation pipeline is proposed using the Segment Anything Model 2 (SAM2) by Meta [48], a foundation model for promptable segmentation. SAM2 can encode the imprecise human labels together with the image and generate precise segmentation masks of the object, which can then be converted into fitting oriented bounding boxes (OBBs) as shown in Figure. The benefit is not only that it should improve training and validation but also that the bounding box becomes more informative as it becomes a better representation of the object size.



Figure 3.6: Examples of inaccurate bounding boxes from Zurich dataset (left), Durban dataset (middle), and Jakarta dataset (right).

$$o_{i,j} = \text{MinAreaBox}(m_{i,j}), \quad \text{where} \quad m_{i,j} = \text{SAM2}(b_{i,j}, I_i). \quad (3.1)$$

where:

- $o_{i,j} = \text{MinAreaBox}(m_{i,j})$: oriented bounding box computed as the minimum-area box around the extracted contour of the segmentation mask.
- $m_{i,j} = \text{SAM2}(b_{i,j}, x_i)$: segmentation mask of the object inside the initial bounding box.
- $b_{i,j}$: initial bounding box.
- x_i : input image.



Figure 3.7: Examples of improved oriented bounding boxes from Zurich dataset (left), Durban dataset (middle), and Jakarta dataset (right).

3.3 Object Detector

The ARCAM system introduced in the previous section will be used to collect datasets from multiple locations. These can be used to explore the scalability of the detection models, which is a critical factor for effective waste monitoring. This section will propose three approaches, starting with a baseline model (M_B) trained on the source dataset (D_S) for riverine waste detection. This baseline will give an idea about the initial performance that can be achieved for one location. Given that M_B is likely to perform poorly on unseen target datasets (D_{T1} , D_{T2}), two adaptation strategies are proposed: *few-shot fine-tuning* (M_F) and *pseudolabel self-training* (M_P). These strategies aim to mitigate the performance drop caused by domain shifts. Besides detection performance, they will be assessed on scalability, for which the metrics are detailed in Section 3.5.

3.3.1 Baseline Model

The detection architecture for M_B will be the latest version of the YOLO family (YOLOv11 developed by Ultralytics). Previous work [53, 34] has shown that YOLO models, which are pre-trained on the COCO dataset, achieve promising results for riverine monitoring in terms of prediction performance and inference speed. To optimize the baseline, a custom source dataset of riverine waste will be trained with YOLOv11 variations with different model sizes to identify the best trade-off. Additionally, data augmentations like flipping and scaling will be applied to diversify the training dataset. During training the YOLO loss function, $\mathcal{L}_{\text{YOLO}}$, is minimized. It contains three factors: distributed focal loss, bounding box regression loss, and class probability loss. The latter is less relevant for this case as the model is only trained as a binary detector:

$$\mathcal{L}_{\text{YOLO}} = \mathcal{L}_{\text{bbox}} + \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{dff}} \quad (3.2)$$

where:

- **Bounding box loss ($\mathcal{L}_{\text{bbox}}$):** Penalizes misalignment between predicted and ground-truth boxes using Intersection over Union (IoU).
- **Classification loss (\mathcal{L}_{cls}):** Measures how well the model predicts object presence.
- **Distribution focal loss (\mathcal{L}_{dff}):** Improves bounding box predictions by modeling them as probabilistic distributions.

The baseline model M_B can be trained on two versions of the labeled source dataset: the initial regular bounding box dataset (D_S^{RBB}) and the refined bounding box dataset (D_S^{OBB}):

$$D_S^{\text{RBB}} = \{(x_i, b_i)\}_{i=1}^{N_S}, \quad b_i = \{b_{i,j}\}_{j=1}^{n_i} \quad (3.3)$$

$$D_S^{\text{OBB}} = \{(x_i, o_i)\}_{i=1}^{N_S}, \quad o_i = \{o_{i,j}\}_{j=1}^{n_i} \quad (3.4)$$

where:

- x_i : input image
- b_i : set of loose bounding boxes for all objects in image x_i
- o_i : set of oriented bounding boxes for all objects in image x_i
- N_S : Number of images in the source dataset
- n_i : Number of objects in image x_i

To establish a benchmark, the performance of the baseline model (M_B) will be compared to similar detectors used in the previous studies on the Limmat River in Zurich [51, 52, 53]. To evaluate the domain shift, the trained model (M_B) will be tested on two target datasets, D_{T1} (Durban) and D_{T2} (Jakarta). This evaluation will help to determine the effectiveness of the proposed adaptation strategies.

3.3.2 Few-Shot Fine-Tuning

After establishing the performance drop due to domain shift, the effort required to adapt the model to new locations has to be assessed. This raises the question of how much labeled data from the target domain is needed for effective fine-tuning. Since the model is already trained on river images from Zurich, it is expected that fewer images from Durban or Jakarta will be required to achieve a similar performance. To explore this hypothesis, the model will be fine-tuned by incrementally adding labeled instances from the target dataset, starting with the baseline model M_B and progressively refining it into M_F using a small, growing subset of the target dataset (D_t) that contains a certain amount of images (N_t):

$$D_t \subset D_T, \quad |D_t| = N_t, \quad N_t \ll N_S \quad (3.5)$$

Similar to the baseline model, the objective of the few-shot fine-tuned model M_F is to minimize the YOLO loss function (see Equation 3.2). The performance of M_F is then evaluated on an unseen subset of the target dataset, using the mean Average Precision (mAP, see Section 3.5).

3.3.3 Pseudolabel Self Training

Where few-shot fine-tuning still requires some manual data labeling, unsupervised domain adaptation techniques eliminate that entirely. This section will explore an approach where only an unlabeled dataset from the new location is needed: self-training with pseudolabels. The idea is that the baseline model M_B runs inference on the unseen target dataset D_T and that the resulting detections are used for retraining to obtain model M_P . However, these pseudolabels may contain false detections, which can introduce noise and degrade the performance of M_P . Therefore, high-confidence filtering can be applied to exclude low-confidence predictions from retraining. This can be repeated to gradually improve the performance. Alternatively, images with low-confidence pseudolabels can be used to augment the dataset with synthetic data.

Before applying filtering, baseline performances have to be established. As discussed before the direct application of M_B on D_T quantifies the domain shift. Secondly, training M_P on the entire dataset with pseudolabels from M_B without filtering will help to quantify the effect of noise reduction in the next steps.

High-confidence Filtering

To only retrain with high-confidence labels, a confidence threshold θ is applied to remove unreliable predictions. A pseudolabel is accepted if the confidence score of a prediction $C(y_i)$ is above this threshold:

$$C(y_i) > \theta, \quad y_i \in M_B(x_i), \quad x_i \in D_T \quad (3.6)$$

The confidence threshold θ can be set directly or dynamically by determining the number of instances to be used for retraining. In the dynamic case, θ is adjusted to select exactly the top N_p highest-confidence pseudolabels. If θ is set directly, the number of selected pseudolabels (N_p) is:

$$N_p = \sum_{y_i \in M_B(D_T)} C(y_i) > \theta \quad (3.7)$$

A challenge in using hard pseudolabels is the fact that some correct pseudolabels will obtain confidences below the threshold. In that case, the model may learn that those objects are not waste and reinforce incorrect negative predictions. It is possible

that filtering leads to significant improvement but not optimal results. Then, the process can be repeated so that more labels receive high-confidence predictions in every round, which allows for iterative improvement to bridge the domain gap.

Low-confidence Filtering

Low-confidence predictions (or no predictions at all) indicate that the model has not detected meaningful features in those images. These images are considered "empty" and can be used for data augmentation. The goal is to filter out these images with low-confidence predictions and use them to add synthetic waste objects. For this alternative approach, pseudolabels are the means to determine which images to use for augmentation rather than using the pseudolabels directly for fine-tuning.

The low-confidence images are defined as those where the maximum confidence score for any object in the image is below a threshold θ_{low} . To generate synthetic data, the TrashNet [28] dataset will be utilized.

3.4 Multimodal Foundation Models

Vision Foundation Models have brought significant advancements to computer vision by offering tools that generalize well across tasks like object detection and tracking. They are often built using transformer architectures [58] and typically consist of a backbone encoder, which extracts visual features, multi-modal embeddings, and a decoder that is tailored for downstream tasks. The key strength is that they are trained on very large and diverse datasets to find universal visual features. This makes them particularly useful for tasks that require flexibility like monitoring river waste.

Multimodal foundation models combine the semantic understanding of visual scenes with other modalities like natural language which enables open-vocabulary detection. This allows them to generate bounding boxes based on any text prompt, potentially eliminating the need for task-specific detection models.

To explore their potential, two open-dictionary models, OwL-ViT [47] and Grounding DINO [46], will be compared against the traditional detection pipelines discussed earlier. A small subset of D_S , D_{T1} , and D_{T2} will be tested using various prompts related to waste in the river, followed by more specific categories such as "plastic bottle." The results (see Section 5.5) will be analyzed in terms of detection performance and computational cost to make suggestions on the feasibility of replacing traditional YOLO pipelines.

3.5 Evaluation Metrics

For the evaluation and discussion of the proposed approaches, it is important to focus on the long-term goal: achieving high-quality detection with spatiotemporal variability to make sure that the outputs of riverine waste monitoring efforts can be put to use in policy making. The evaluation is, therefore, divided into two main categories: (1) detection performance and (2) scalability.

Detection performance is important because the results need to be reliable, especially when decisions or actions are based on the system's output. Secondly, scalability is key to addressing spatial variability. The system and algorithm should be robust and easily adaptable to different locations with minimal effort. The goal is to achieve a balanced trade-off, providing accurate and useful predictions at a reasonable computational cost while maintaining real-time performance and scalability for deployment across multiple locations.

3.5.1 Object Detection

These metrics evaluate the ability of a model to identify and localize waste items. In the formulas, TP, TN, FP and FN represent true positives, true negatives, false positives, and false negatives.

- **Precision and Recall:** Precision measures the proportion of correctly identified waste items among all detections, while recall represents the proportion of actual waste that was detected:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (3.8)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3.9)$$

- **F1-Score:** The harmonic mean of precision and recall, balancing false positives and false negatives:

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3.10)$$

Precision and recall depend on two key thresholds: the intersection over union (IoU) threshold, which determines the required overlap between predicted and ground truth bounding boxes (fixed at 0.5 in this study), and the confidence threshold, which sets the minimum confidence for a prediction to be considered valid. The confidence threshold can be optimized to maximize the F1-score and find the trade-off between precision and recall, but its optimal value will vary across test sets. As a result, a single confidence threshold does not fairly represent overall performance and must be adjusted for each new environment, making direct model comparisons difficult. Instead, the mean average precision at constant IoU threshold 0.5 (mAP50) provides a more reliable comparison because it evaluates precision and recall across all confidence levels by determining performance as the area under the Precision-Recall (PR) curve shown in Figure 3.8.

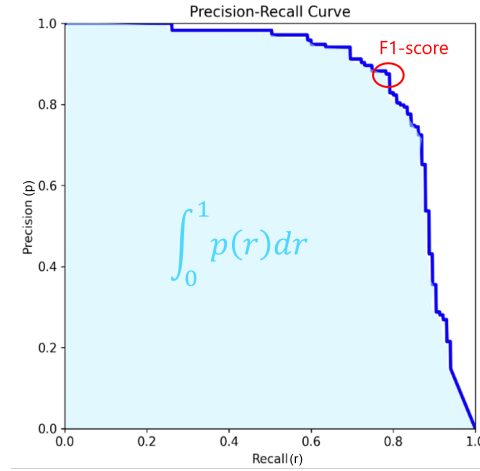


Figure 3.8: Example of Precision-Recall Curve

- **Average Precision (AP):** The integral of the precision-recall curve over all recall values:

$$\text{AP} = \int_0^1 p(r) dr \quad (3.11)$$

- **Mean Average Precision (mAP):** The mean of AP across n classes. Since this study uses only one class, mAP is equivalent to AP:

$$\text{mAP} = \frac{1}{n} \sum_{i=1}^n \text{AP}_i \quad (3.12)$$

3.5.2 Scalability

Scalability is about assessing the system's feasibility for deployment across multiple locations. It is difficult to quantify precisely as it entails hardware, software, and operational considerations with many uncertainties and approximations. Nevertheless, the following aspects have to be considered:

- **Software Adaptation:** This includes the effort of data annotation, defined as the number of new instances that must be manually labeled to retrain the model. It also involves minimizing computational costs, as resource-heavy training pipelines may limit scalability in environments with limited access to high-performance hardware.
- **Hardware Adaptation:** This is optimized by reducing production costs and simplifying the installation process to eliminate the need for specialized expertise. Additionally, ensuring the adaptability of the bridge mount allows for it to be installed at different locations.
- **Maintenance:** Computational costs should remain low to facilitate real-time monitoring. The real-time aspect is needed to capture high temporal variability without requiring extensive storage or frequent maintenance for offline analysis. Additionally, the system should withstand outdoor weather conditions to allow for long-term deployment.

Chapter 4

Data Acquisition

Data acquisition was conducted across multiple locations to test the system setup and ensure hardware consistency, with location being the only variable. At the end of this chapter, an overview is provided of the datasets collected during this thesis and additional datasets from other studies used in this research. This allows for evaluating whether the proposed domain adaptation strategies generalize not only across locations within the same system but also across different systems in different locations. Section 4.1 provides an overview of the key imaging parameters and their impact on image quality. Afterwards, Section 4.2 and 4.3 dive into the image collection that took place in Switzerland and South Africa.

4.1 Imaging Parameters

The most important imaging parameters that influence the detection pipeline are the field of view (FOV), ground sampling distance (GSD), and storage requirements in combination with the collection rate. These parameters depend on the characteristics of the camera and lens that are used. The FOV refers to the visible area captured by the camera and is determined by the sensor size and focal length. The GSD is the real-world distance represented by a single pixel in an image. It is influenced by the camera's resolution, focal length, sensor dimensions, and the distance to the object. These parameters are defined as follows:

Field of View (FOV):

$$\text{HFOV} = 2 \cdot \arctan\left(\frac{\text{Sensor Width}}{2 \cdot \text{Focal Length}}\right), \quad \text{VFOV} = 2 \cdot \arctan\left(\frac{\text{Sensor Height}}{2 \cdot \text{Focal Length}}\right) \quad (4.1)$$

Field of View at Distance D (cm):

$$\text{HFOV}_{\text{cm}} = 2 \cdot D \cdot \tan\left(\frac{\text{HFOV}}{2}\right), \quad \text{VFOV}_{\text{cm}} = 2 \cdot D \cdot \tan\left(\frac{\text{VFOV}}{2}\right) \quad (4.2)$$

Ground Sampling Distance (GSD):

$$\text{Horizontal GSD} = \frac{\text{HFOV}_{\text{cm}}}{\text{Image Width}}, \quad \text{Vertical GSD} = \frac{\text{VFOV}_{\text{cm}}}{\text{Image Height}} \quad (4.3)$$

Pixels per Square Centimeter:

$$\text{Pixels per cm}^2 = \frac{1}{\text{Horizontal GSD} \cdot \text{Vertical GSD}} \quad (4.4)$$

The system used to record the datasets contains a camera with parameters summarized in Table 4.1, providing a horizontal field of view (HFOV) of approximately 84.82° and a vertical field of view (VFOV) of approximately 54.78° . Other parameters like the field of view in centimeters and the ground sampling distance depend on the bridge's height where the system is mounted and will be mentioned for the specific data collection in the following sections.

Table 4.1: Camera (Ximea MC161CG-SY-UB-HDR) Specifications

Parameter	Value
Image Width	5320 pixels
Image Height	3032 pixels
Sensor Width	14.5 mm
Sensor Height	8.3 mm
Focal Length	8 mm
Distance from Camera to Water (D)	Variable

4.2 Image Collection in Switzerland

ARCAM was installed in Switzerland on two occasions. To get familiarized and test its functionalities, the system was set up at ARCHE, a week-long event with the entire Robotic Systems Lab (RSL) held at a military test base near Geneva. Later, it was deployed at the Kornhausbrücke during the Street Parade 2024 for real-world data collection at the Limmat River.

4.2.1 ARCHE

At ARCHE in Geneva, the ARCAM system underwent hardware validation tests in a controlled environment to establish its baseline functionality and readiness for deployment in South Africa. The network reliability and power management were assessed through several tests. Additionally, both the data collection and detection pipeline were tested by dragging plastic objects through the water, see Figure 4.1. This resulted in a list of improvements for the detection box, energy box, mount system, and software, as well as the creation of a small dataset at a height of about 3 meters with a resolution of 0.10 cm/pixel.



Figure 4.1: ARCAM at ARCHE

4.2.2 Street Parade

The second deployment took place during the 2024 Street Parade in Zurich, a large-scale electronic music festival that attracts nearly twice the city’s population in a single day. The Limmat River is usually well-maintained, and past research has often required manually introducing waste items for detection experiments [51, 53, 52]. However, major public events like the Street Parade increase the amount of mismanaged plastic waste entering the river [59]. The deployment of ARCAM on that day (see Figure 4.2) provided an opportunity to observe river pollution under natural conditions and gather a real-world dataset from Zurich. Additionally, the continuous sunlight throughout the day allowed for the evaluation of potential overheating issues. Positioned at a height of approximately 10 meters above the river, the system captured a horizontal field of view (HFOV) of about 18 meters and a ground sampling distance (GSD) of 0.34 cm/pixel. At this resolution, some waste items were clearly identifiable, while others were visibly waste but difficult to classify. Examples are shown in Figure 4.3 where one object is quite certainly a Coca-Cola bottle and the other is unknown.

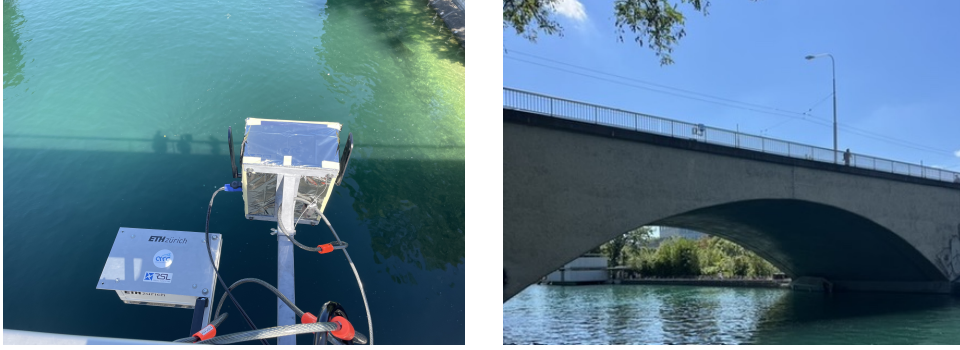


Figure 4.2: ARCAM at Kornhausbrücke during the Zurich Street Parade 2024.



Figure 4.3: Examples of detected waste in the Limmat River during the Zurich Street Parade 2024.

4.3 Image Collection in South Africa

Durban is the third most populated city in South Africa and is home to the largest harbor in Africa. It is also where the uMgeni river, one of the 1000 most polluting globally [57], flows into the ocean. With a catchment area of approximately 4,500 km², this river network is responsible for an estimated 380,300 kg of plastic emissions into the ocean annually. In collaboration with the local NGO Green Corridors (see Section 4.3.1), ARCAM was deployed at the uMhlangane River, a tributary of the

uMgeni River. The location was selected based on safety and accessibility. Section 4.3.2 provides details on the uMhlangane River and ARCAM's operations there. Furthermore, a strong correlation was observed between precipitation and waste accumulation. These preliminary findings, discussed in Section 4.3.3, suggest the need for a more extensive evaluation.

4.3.1 Green Corridors

Green Corridors is an organization dedicated to addressing the environmental and social challenges of riverine ecosystems. Operating in the eThekweni metropolitan area (Durban), Green Corridors focuses on transformative riverine management, including mitigating the impact of invasive alien plants and solid waste accumulation. Through partnerships with local communities, they focus on nature-based solutions that promote environmental restoration, sustainable development, and community well-being. By maintaining and improving riverine landscapes, Green Corridors highlights the cost-effectiveness of proactive environmental management, where 1.8 ZAR spent on such efforts can save 3.8 ZAR in infrastructure repairs.

The Litter Boom Project is a collaborative initiative led by Siphwe Rakgabale, founder of TriEcoTours, in partnership with Green Corridors. The project aims to intercept waste flowing down rivers before it reaches the ocean while educating local communities about sustainable waste management. So far, the project has installed 21 litter booms in the uMgeni area, with each boom monitored by assistants from nearby communities. The project's vision extends beyond waste collection. The goal is to establish multiple receiving stations where waste can be sorted, processed, and beneficiated. In these operations, the biggest challenges are theft, limited infrastructure, and unpredictable waste composition. Vision-based monitoring could help evaluate the litter booms, optimize their placement, and create educative material. For the full interviews with Nick and Siphwe, go to Appendix C.

Besides having a profound understanding of the operating context, collaborations like these are extremely valuable to further develop waste monitoring strategies that ensure a link with the implementation step.

4.3.2 uMhlangane River

ARCAM was deployed over the uMhlangane River on a bridge at Riverhorse Valley. Figure 4.4 shows a map of the region, along with the mounted system and its protective cage. The bridge has a height of about 8 meters which provides an HFOV of 15 meters. This allows the entire river width to be captured in the images. The GSD is 0.27 cm/pixel, which is slightly better than the resolution achieved at the Limmat River in Zurich. However, while some waste items can be classified as specific objects, many remain only identifiable as general waste. Examples from the dataset are shown in Figure 4.5.

The goal of the deployment was to gather river waste footage and top-view waste data, as well as to field test the ARCAM system and assess how it could contribute to Green Corridors' efforts to mitigate the impact of solid waste in the riverine environment. The system was mounted at the side of a highway with little passage, reducing the risk of vandalism or theft during operations. For the data collection, a downside was the ongoing informal waste picking in Riverhorse Valley, which resulted in a relatively clean river with longer periods of little to no waste floating by. Over the period of October to December 2024, the system was deployed regularly for testing and data collection. The final Durban dataset contains images from seven days in November and December. While the system was deployed more frequently, especially during longer periods without rain (see Section 4.3.3), many recordings did not contain any waste and were therefore left out. Weather conditions during

the deployment ranged from sunny and cloudy to heavy rains (more than 4mm/h), providing a good test for the system’s resistance to harsh weather. Unlike the Zurich dataset, direct sunlight and different times of day had less effect on the dataset’s color variations due to the murkiness of the river water.



Figure 4.4: Map of the deployment site at the uMhlangane River, Riverhorse Valley, Durban.

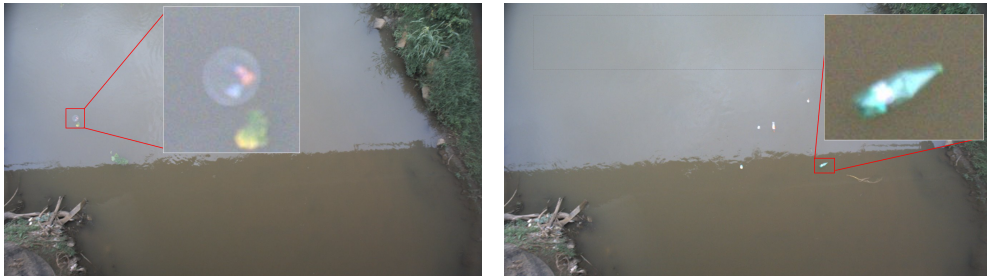


Figure 4.5: Examples of detected waste from the uMhlangane River dataset.

4.3.3 Weather-Waste Relationship

The overall rainfall during the deployment period was low. When there was no rainfall, the river remained relatively clean, although a significant amount of waste could be seen along the riverbanks. The litter boom, which was installed in collaboration with Green Corridors, was a good indicator of waste levels but often remained empty. However, a rainfall event on 27th November revealed a remarkable difference in the usually clean river, which now contained a continuous stream of waste, as shown in Figure 4.6. The rain continued until midnight, and by the next morning, when the sun was out again at 5 a.m., there was no trace of the sudden increase in waste and the river had returned to its calm and clean state.

This observation confirms that waste transportation in the river is closely linked to rainfall, as seen in other studies [56, 60]. However, it was remarkable how quickly the waste disperses afterward, likely sinking to the riverbed or being left along the sides

as the river narrows. There appears to be a threshold for rainfall intensity as events between 0.5 and 3 mm/h did not produce a noticeable increase in waste. These results suggest the need for further research to better understand the connection between rainfall intensity and the amount of waste washed into rivers, as well as the processes that influence its rapid dispersal.



Figure 4.6: uMhlangane river segment: clean during sunny day (top) and polluted during heavy rain (bottom)

4.4 Datasets Overview

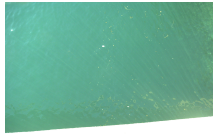
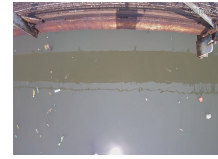
This section gives an overview of the datasets collected and used in this research. Table 4.2 summarizes the three datasets obtained during this thesis, alongside an additional dataset from van Lieshout et al. [30] collected in Jakarta and an example of each dataset is shown in Figure 4.7. The Zurich dataset (D_S) will later in Chapter 5 serve as the source domain, while the Durban (D_{T1}) and Jakarta (D_{T2}) datasets are considered target domains for evaluating the performance dip and the proposed domain adaptation strategies. The ARCHE dataset was primarily used for system validation and will not be included in the later experiments.

Table 4.2: Dataset Overview

Dataset	Location (River)	Days (#)	Altitude (m)	Labeled Images (#)	Instances (#)	Images (#)
ARCHE (Section 4.2.1)	Geneva (Test site)	1	3	0	0	155
D_S (Section 4.2.2)	Zurich (Limmat River)	1	10	504	533	14028
D_{T1} (Section 4.3)	Durban (uMhlangane River)	7	8	419	845	25238
$D_{T2}[30]$	Jakarta (Grogol River)	10	4.5	526	11,064	526



(a) Geneva

(b) Zurich (D_S)(c) Durban (D_{T1})(d) Jakarta (D_{T2})Figure 4.7: Example images from each dataset: Geneva (system validation), Zurich (D_S), Durban (D_{T1}), and Jakarta (D_{T2}).

Subsets of the Zurich and Durban datasets collected with ARCAM were labeled using a single waste class without a finer taxonomy. This decision was made for several reasons. (1) First, the resolutions at which data was collected in Zurich and Durban allow for the clear recognition of only the most obvious objects, such as plastic bottles. Many other objects are visibly waste, but their specific categories remain unclear due to limited image resolution, as illustrated in Figure 4.8. A finer taxonomy would have been more feasible with the resolution achieved at ARCHE (0.1 cm/px), but that would require low bridges (less than 3 meters) during data collection. (2) Second, the diversity of waste types flowing through the rivers makes it challenging to collect a sufficiently balanced dataset across multiple categories. (3) Third, given that the primary focus is on estimating waste flux rather than classifying specific waste types, a binary classification provides the most relevant information by capturing the amount and size of waste floating by rather than restricting detection to only predefined object categories. Lastly, the Jakarta dataset was also labeled using a single category. It was recorded from a bridge overlooking the Grogol River, resulting in a topview perspective that resembles the Zurich and Durban datasets, which makes it a suitable second target domain for evaluating generalization across different locations.

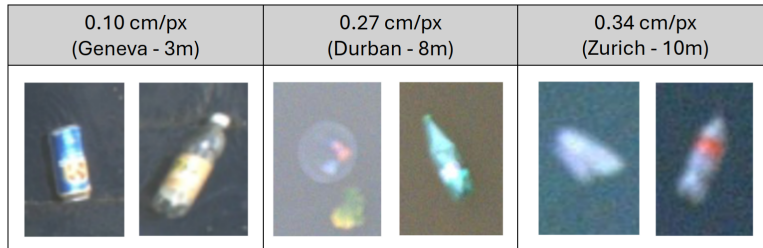


Figure 4.8: Image resolutions from ARCAM at different locations

As an initial analysis, the three datasets that will later be used in the experiments are compared based on pixel values and labeled instances. In terms of pixel values, the mean and standard deviation across all color channels highlight differences in color distributions between datasets, as visualized in Figure 4.9. D_S has greater variability in lighting conditions, whereas D_{T1} , D_{T2} have more consistent color distributions, which might be attributed to the murkier water in the uMhlangane and Grogol River and less variation in lighting throughout the day as they are closer to the equator.

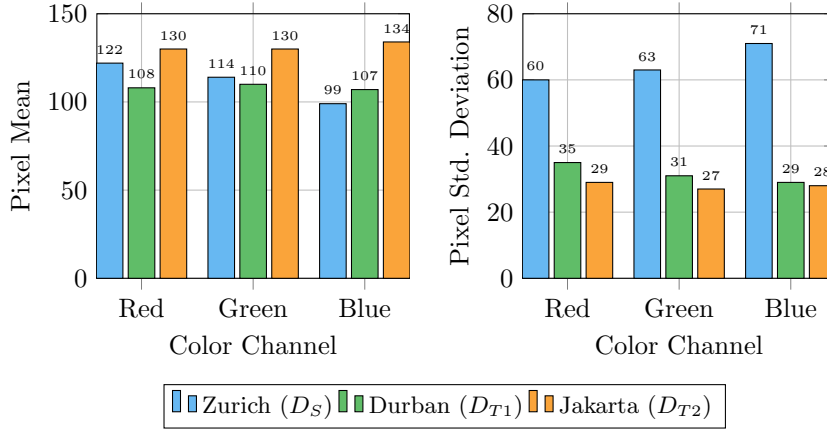





Figure 4.9: Pixel Statistics: (Left) Pixel Means, (Right) Pixel Standard Deviations

Beyond color distributions, the datasets also differ in terms of labeled instances and bounding box characteristics, as shown in Table 4.3. Bounding box sizes vary slightly depending on the camera setup and bridge height. Compared to Zurich (D_S), average bounding box sizes in Durban (D_{T1}) are about 13% larger, while in Jakarta (D_{T2}), they are approximately 27% smaller. These variations indicate that waste objects appear at slightly different scales across datasets, but within a range that is unlikely to be the primary cause of performance drops.

The number of bounding boxes per image is highest in Jakarta, whereas Zurich and Durban contain significantly fewer. Meaning that the Grogol River has much higher density of floating waste. In line with that Jakarta shows instance clustering, as indicated by an average intersection over union (IoU) between bounding boxes of 0.13, but Zurich and Durban show no bounding box overlaps. These dataset characteristics could impact model performance when applying a detector trained on the Zurich dataset to the other locations. This drop in performance and possible mitigations will be discussed in Chapter 5.

Table 4.3: Instance Statistics Across Datasets

Metric	Zurich (D_S)	Durban (D_{T1})	Jakarta (D_{T2})
Labeled Images	504	419	524
Avg. BBox Size (W × H)	54.64×52.14 	61.99×58.56 	39.57×44.56 
Avg. BBoxes/Image	1.06	2.03	21.03
Avg. BBox Overlaps	—	—	0.13

Chapter 5

Results & Discussion

This chapter presents the experimental results and analysis of the proposed methodologies for riverine waste detection. The performance of M_B (Section 5.1), M_F (Section 5.2) and M_P (Section 5.3) are evaluated in terms of detection performance and scalability. For this process the Street Parade dataset from the Limmat River in Zurich is the source dataset D_S , the one from uMhlangane River in Durban is target dataset one D_{T1} and the one collected by van Lieshout et al. [30] from the Grogol River in Jakarta is target dataset two D_{T2} . Lastly, open dictionary foundation models will be tested on all datasets as an alternative method. All of the approaches will be compared with each other in Section 5.4.

5.1 Baseline Detector

This section first establishes a well-performing detection model on the source domain, identifying the most valuable improvements, and then evaluates the performance drop on the target domains to set a reference point before applying adaptation techniques. Training the baseline model M_B was setup as follows:

- **Hardware:** Computer with an NVIDIA GeForce RTX 4090 GPU, Intel i7-12700k CPU, and 32GB RAM
- **Datasets:** Waste dataset from Limmat River with regular bounding boxes (D_S^{RBB}) and with improved oriented bounding boxes (D_S^{OBB}) from the pipeline discussed in Section 3.2.
- **Data Splitting:** The dataset was recorded over the course of one day and therefore contains lighting and weather variations (see Figure 5.1). To ensure fair evaluation, the train/test split follows an 80/20 ratio and is done per hour. To prevent the same waste item from appearing in both sets, a time difference threshold is applied. Additionally, 10% of images without waste objects are included to help the model recognize empty river conditions.
- **Model:** For real-time river waste monitoring, fast inference time is important. Therefore, only the smallest three YOLOv11 models (nano (n), small (s), and medium (m)) were tested to balance computational efficiency and detection performance. For regular bounding boxes (RBB), these models are pre-trained on the COCO dataset [61], while for oriented bounding boxes (OBB), are were pre-trained on DOTA v1 [62].
- **Hyperparameters:** A full overview of the selected hyperparameters can be found in Appendix D.1. Two key hyperparameters are early stopping

(*patience* = 10 epochs), used to prevent overfitting, and the input size, where both 640px and 1280px resolutions were tested due to the presence of very small-scale objects.

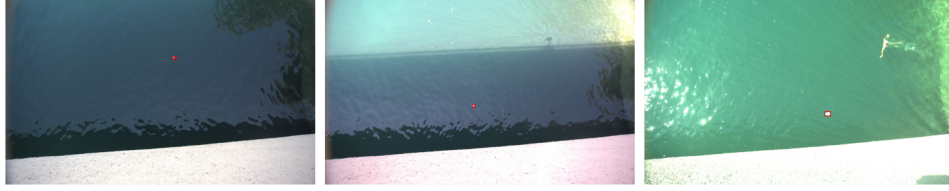


Figure 5.1: Images from Limmat River at Kornhausbrücke during the morning (left), mid-day (middle) and the afternoon (right)

Table 5.1 presents the performance results of the trained models. The reported results are evaluated at a fixed Intersection over Union (IoU) threshold of 0.5, meaning a detected object is considered correct if its bounding box overlaps with the ground truth by at least 50%. This threshold is sufficient for this application, as the primary goal is to count the number of waste items rather than achieving precise localization. The confidence threshold for computing Precision, Recall, and F1-score is selected to maximize the F1-score to get the best trade-off between precision and recall. The optimal threshold T_c is defined as:

$$T_c = \arg \max_T F1(T) \quad (5.1)$$

where $F1(T)$ represents the F1-score computed at confidence threshold T .

Lastly, as described in Section 3.5 the mean Average Precision (mAP) is also reported and will be used as the main comparison metric between models. The best-performing model in terms of mAP@50 is highlighted in blue. The smallest YOLOv11-nano variant achieves the highest performance for both the regular bounding box (RBB) configuration and the oriented bounding boxes (OBB). For all models, the larger input image size (1280x1280 pixels) results in a significant performance increase with higher confidence predictions due to the small scale of the instances that have to be detected. You also see that the model mostly struggles with actually recognizing the waste items from the image. Even at a very low confidence threshold, recall remains moderate while precision stays relatively high.

Bounding Box Improvement

To evaluate the impact of bounding box format, the models were trained with identical configurations, except for the bounding box type. Across all YOLOv11 variants, OBB consistently improves performance compared to RBB. This is likely because OBB improves alignment between predictions and ground truth at the standard 0.5 IoU threshold, leading to higher confidence scores and reducing the false positives and false negatives.

Beyond the numerical improvements, OBB provides several practical advantages for river waste detection. The bounding boxes themselves are more informative, as they better capture the shape and orientation of waste objects. This allows for a more accurate size estimation in the absence of a segmentation step. Additionally, it will handle cluttered waste more effectively. By reducing unnecessary overlaps, OBB minimizes ambiguities that could otherwise affect training when multiple waste items are close together. This is less relevant for the D_S dataset, which contains on average only one waste item per image, but will be useful for environments with high waste density, such as the D_{T2} dataset.

Model	Image Size	Precision	Recall	F1-Score	T_c	mAP@50
Regular Bounding Boxes (RBB)						
YOLO11n	640x640	0.57	0.52	0.54	0.01	0.58
YOLO11s	640x640	0.59	0.53	0.56	0.01	0.60
YOLO11m	640x640	0.78	0.40	0.53	0.47	0.41
YOLO11n	1280x1280	0.87	0.71	0.78	0.14	0.82
YOLO11s	1280x1280	0.87	0.58	0.70	0.22	0.72
YOLO11m	1280x1280	0.90	0.64	0.75	0.44	0.76
Oriented Bounding Boxes (OBB)						
YOLO11n-OBB	640x640	0.71	0.52	0.59	0.01	0.66
YOLO11s-OBB	640x640	0.79	0.48	0.60	0.03	0.66
YOLO11m-OBB	640x640	0.68	0.51	0.59	0.01	0.64
YOLO11n-OBB	1280x1280	0.85	0.81	0.83	0.06	0.87
YOLO11s-OBB	1280x1280	0.84	0.75	0.80	0.06	0.85
YOLO11m-OBB	1280x1280	0.90	0.76	0.81	0.14	0.86

Table 5.1: Performance metrics of YOLOv11 models with regular (RBB) and oriented bounding boxes (OBB). T_c represents the confidence threshold at which F1-score is maximized.

Domain Gap

To assess the generalization ability of M_B , it was tested on the two unseen datasets from Durban (D_{T1}) and Jakarta (D_{T2}) without any adaptation. As expected, a significant performance drop was observed for both RBB and OBB models, as shown in Figure 5.2. Several factors contribute to this domain shift, including differences in environmental conditions, waste types, camera setups, river surface appearances, and waste density.

The test sets used to evaluate the impact of domain shift are subsets of the datasets listed in Table 4.2. The details of these test sets are shown in Table 5.2. For Durban, waste imagery was collected over multiple days. Since images were sometimes captured at a rate of two frames per second, all labeled data from two days was selected for the test set to prevent overlap with the training set in later experiments. In Jakarta, images were captured every 15 minutes, eliminating the risk of overlap between train and test sets due to consecutive frames. The test sets remain constant across all experiments to ensure a fair comparison.

Dataset	Test Set Images	Test Set Instances
D_{T1}^{test} (Durban)	327	648
D_{T2}^{test} (Jakarta)	350	7933

Table 5.2: Test set details for target domains used in generalization and adaptation experiments.

To mitigate this performance drop, the next step is to investigate whether augmentations and synthetic data can improve generalization to unseen domains. Following this, Sections 5.2 and 5.3 analyze the impact of applying domain adaptation techniques and incorporating data from the target domains to bridge the performance gap.

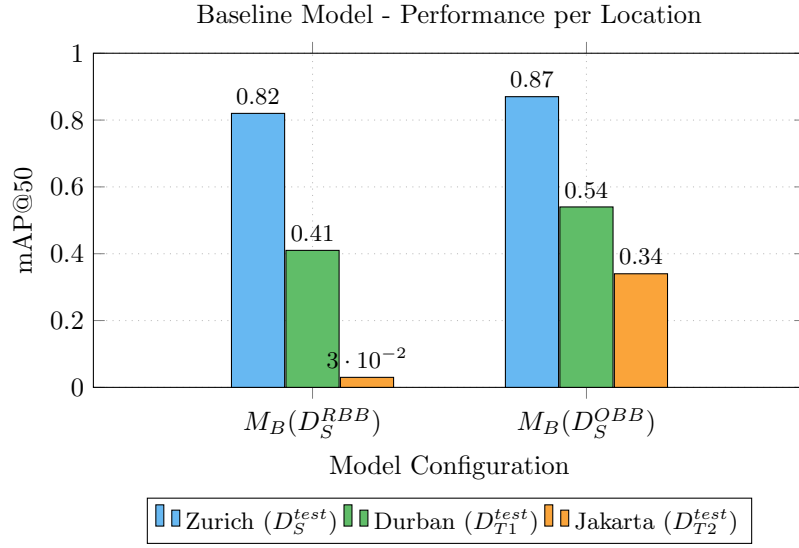


Figure 5.2: Performance drop due to domain shift to unseen datasets on the baseline model trained with regular and oriented bounding boxes.

Data Augmentation and Synthetic Data

To further improve performance and generalization, four augmentation techniques were applied: rotation, scaling, horizontal flipping, and mosaic augmentation. These were selected based on their effectiveness in previous studies [30, 31] and showed baseline improvements when applied one by one. Each augmentation introduces variations that help the model adapt to real-world changes in object orientation, size, and scene composition.

Besides increasing image variety, increasing the number of waste instances per image can also enhance performance since the Zurich dataset contains relatively few objects per image. Therefore, plastic objects from the TrashNet [28] dataset were extracted and randomly merged with Zurich images with varying positions and orientations. An example is shown in Figure 5.3.

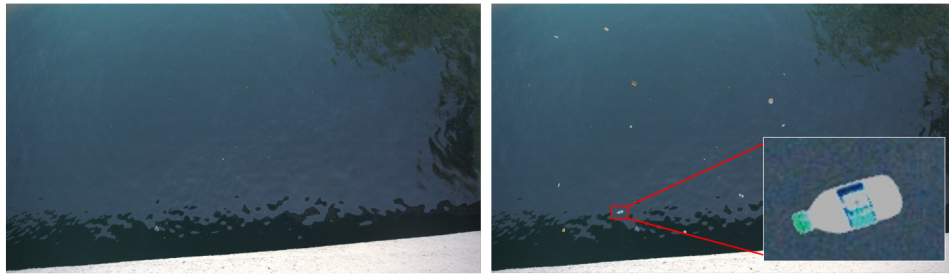


Figure 5.3: Example of synthetic data augmentation. The original image (left) is enhanced by adding synthetic waste objects (right) to simulate higher waste density.

Table 5.3 summarizes the impact of these techniques and their combinations. The applied augmentations include rotation, where small random rotations of 5 degrees are introduced. Scaling is applied at 50%, resizing images to help the model generalize to different object distances. Flipping is performed with a 50% probability, horizontally mirroring images to enhance viewpoint invariance. Additionally, mosaic augmentation is used to merge four images into one, increasing object density

and improving recall.

For the synthetic data, additional images were generated by taking one-third of the Zurich dataset and adding between 3 and 15 randomly placed waste objects per image. Each object is assigned a random rotation and color adjusted to merge with the background.

Method	Precision	Recall	F1-Score	T_c	mAP@50
Baseline	0.85	0.81	0.83	0.06	0.87
Augmented	0.90	0.82	0.86	0.35	0.89
Synthetic Data	0.85	0.76	0.80	0.12	0.83
Augmented + Synthetic Data	0.87	0.86	0.86	0.60	0.89

Table 5.3: Impact of augmentations and synthetic data on baseline detection performance for the best-performing OBB model from Table 5.1.

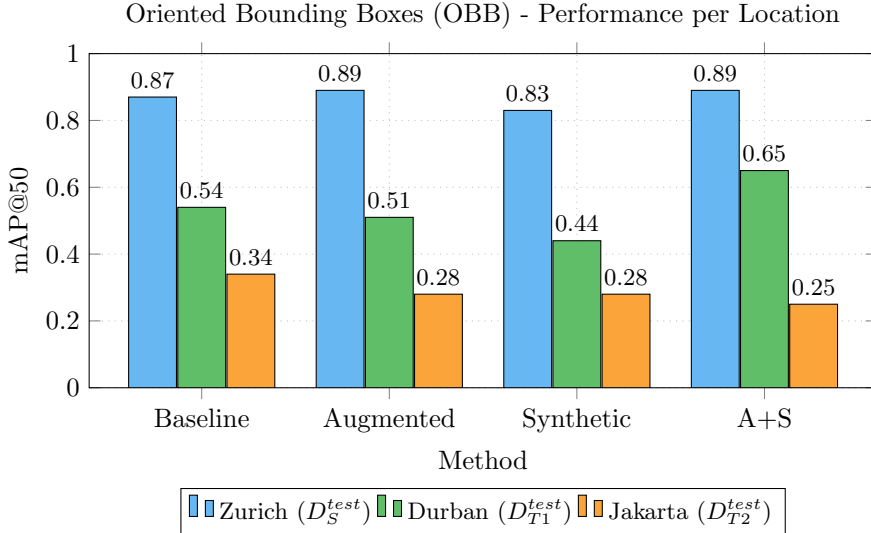


Figure 5.4: Impact of augmentations and synthetic data on domain generalization (OBB).

Final Baseline and Remaining Domain Gap

The best-performing configuration on the source dataset, using oriented bounding boxes (OBB) with both augmentations and synthetic data, achieves a final mAP@50 of 0.89. When compared to related works (see Appendix A), there is a wide variation in reported mAP scores (0.65 up to 0.98). However, direct comparison is difficult due to differences in test setups, including variations in viewpoint, dataset composition, and training strategies.

On the target datasets, OBB consistently performs better than regular bounding boxes, but the impact of augmentations and synthetic data varies. While they improve generalization to D_{T1} , they slightly reduce performance for D_{T2} .

The next sections explore domain adaptation techniques to further reduce performance loss in unseen environments by using data from the target domain. The best performing OBB model from Table 5.1, without any augmentation, will be used as the baseline for comparing the effectiveness of domain adaptation techniques in the following experiments.

5.2 Few-Shot Fine-Tuning

To counteract the performance drop due to domain shift, this section will investigate the minimal amount of labeled target domain data required to fine-tune the baseline model M_B , producing new models for each individual target domain, M_{F1} and M_{F2} , with the goal of improving performance. We are looking for the minimal subset size N_t^{opt} which is determined by either reaching threshold τ :

$$\text{mAP}_{50}(M_F(N_t), D_T^{\text{test}}) > \tau \quad (5.2)$$

or until the performance improvement rate satisfies a convergence condition, meaning no improvement has been observed in the last k iterations beyond a small threshold ϵ :

$$\frac{\text{mAP}_{50}(M_F(N_t), D_T^{\text{unseen}})}{\frac{1}{k} \sum_{i=1}^k \text{mAP}_{50}(M_F(N_t - i), D_T^{\text{unseen}})} - 1 < \epsilon \quad (5.3)$$

The fine-tuning process begins by adding one image ($N_t = 1$) and increases in increments of a chosen step size b . Since the aim is to bridge the domain shift, τ is set to the performance of the baseline model M_B on the source dataset: $\tau = 0.89$. Fine-tuning stops if no improvement is achieved in the last $k = 10$ iterations, where improvement is defined as an increase of at least $\epsilon = 0.01$.

The waste datasets vary in the number of instances per image, making it more informative to track fine-tuning progress based on the number of labeled instances rather than just the number of images. D_S and D_{T1} have significantly fewer waste instances per image compared to D_{T2} . Images are added in fixed increments, one by one ($b = 1$), but the number of new instances per step varies. This variability impacts domain adaptation, as images with more instances introduce greater object diversity, which can accelerate learning. To ensure a fair comparison of annotation effort across datasets, fine-tuning progress is plotted in terms of labeled instances rather than images.

Durban (D_{T1})

The fine-tuned baseline model M_{F1} on Durban is evaluated on D_{T1}^{test} and D_S^{test} , as shown in Figure D.2. Performance improves with labeled instances, reaching an optimal point at $N_t^{opt} = 16$. At this point mAP_{50} of D_{T1}^{test} has increased by 0.15 from the baseline performance (M_b).

Jakarta (D_{T2})

For Jakarta, the performance of M_{F2} on D_{T2}^{test} is plotted in Figure D.4. The model shows gradual improvement, with convergence after adding 25 images that contain $N_t^{opt} = 259$, resulting in a 0.22 increase in mAP_{50} from the baseline.

5.3 Pseudolabel Self Training

In this section, pseudolabel self-training is evaluated as a method for domain adaptation without manual annotation. The aim is to use the baseline model M_B to generate pseudolabels on the target datasets D_{T1} and D_{T2} and use these for self-training to improve detection performance while minimizing annotation effort. A new model will be created for each target domain, denoted as M_{P1} and M_{P2} , respectively.

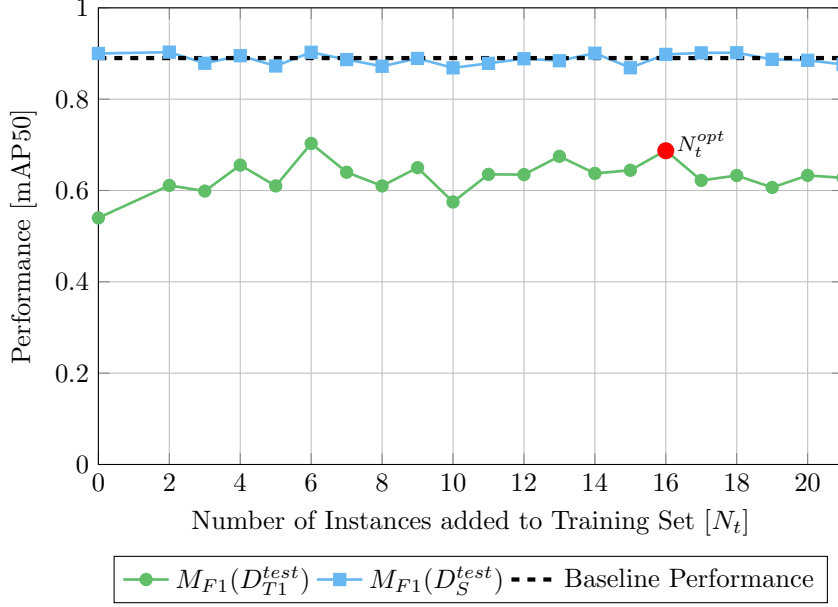


Figure 5.5: Model performance on D_S^{test} and D_{T1}^{test} test sets with N_t Durban instances added to the training set.

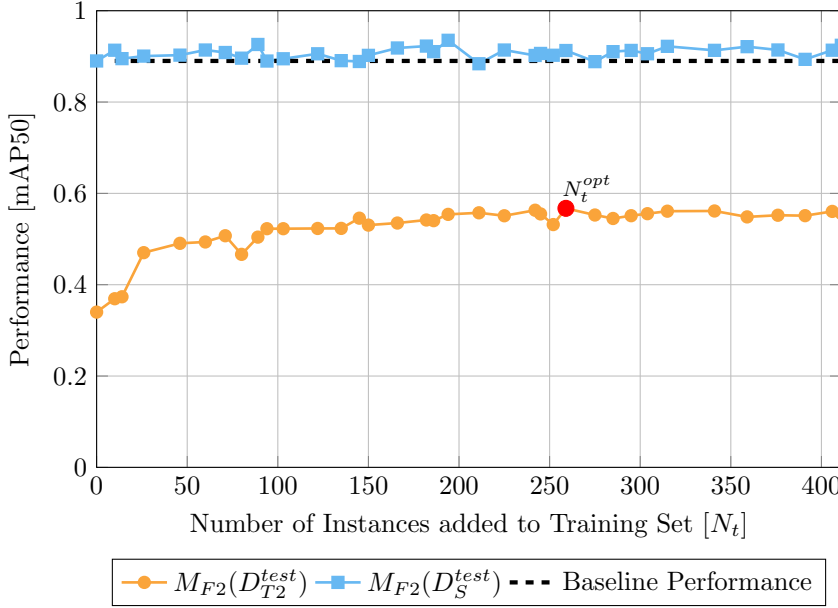


Figure 5.6: Model performance on D_S^{test} and D_{T2}^{test} test sets with N_t Jakarta instances added to the training set.

To establish a baseline, the performance of M_B on the target datasets was evaluated without pseudolabeling (see Section 5.1), which resulted in a significant performance drop. Next, a preliminary test is conducted where all generated pseudolabels are used, only filtering out the worst detections by setting a very low confidence threshold ($\theta = 0.001$). Then, high-confidence filtering and iterative filtering will be applied, followed by an alternative approach where low-confidence detections are filtered and augmented by adding synthetic data.

The test datasets are still the ones from Table 5.2. The set used for pseudolabeling is small for D_{T2} , as there are only a limited number of images of Jakarta (176 images in total). However, for the collected dataset in Durban (D_{T1}), the unlabeled images can also be used, so we will proceed with 300 images to generate pseudolabels and fine-tune the model.

Unfiltered Pseudolabeling

Figure 5.7 shows the confidence distributions of the generated pseudolabels for each dataset. For Jakarta with 176 images, 2,308 predicted labels were generated which is fewer than the actual number of ground truths present in those images. For Durban with 300 images, 396 pseudolabels were generated. These numbers again confirms the difficulty of the model in recalling all objects, even with a very low confidence threshold. Figure 5.8 shows two example images with pseudolabels generated for the target domains of Durban and Jakarta.

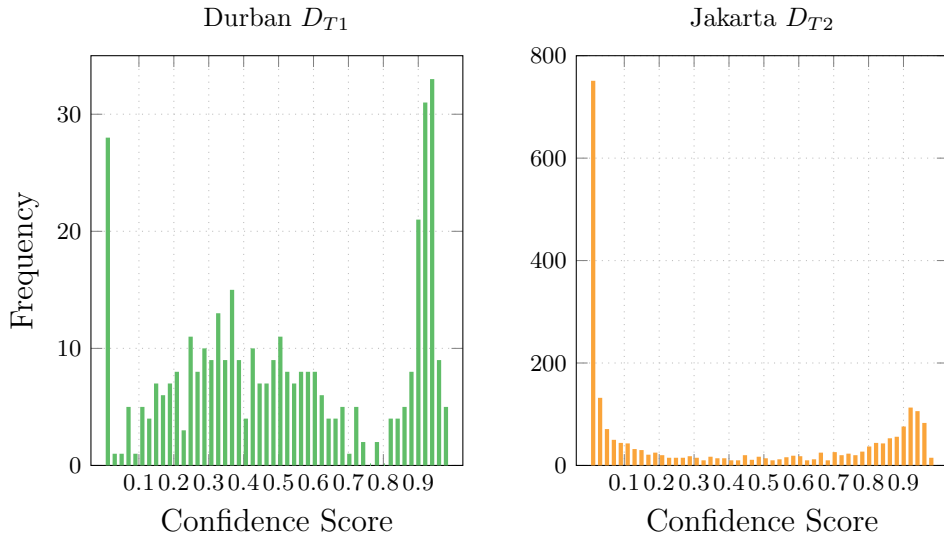


Figure 5.7: Pseudolabel confidence score distributions in Durban D_{T1} and Jakarta D_{T2} .

High-confidence Filtering

A strict confidence threshold is applied to filter out low-confidence pseudolabels. Figure 5.9 shows how performance varies with different confidence thresholds. As the threshold increases, fewer pseudolabels are retained, but the quality of the remaining labels improves. However, since few labels are detected at low confidence the optimal threshold is relatively low ($\theta^{opt} = 0.1$ for D_{T2} and $\theta^{opt} = 0.2$ for D_{T1}). With these optimal thresholds, the model is retrained several times, using the weights from the previous iteration as the starting point. This iterative process provides a significant improvement in performance, as shown in Figure 5.10.

The challenge with hard filtering is that potentially useful low-confidence labels are discarded. While these labels are uncertain, they may still be correct and contribute to model learning. Treating them as background can reinforce errors during training. Iterations might consistently keep improving and increasing the amount of labels above the threshold, which is explored next.



Figure 3 is a line graph showing the mAP50 (Y-axis, ranging from 0 to 1) versus the Confidence Threshold θ (X-axis, ranging from 0 to 1). Two series are plotted: D_{T1}^{test} (green line with circles) and D_{T2}^{test} (orange line with circles). The green line represents D_{T1}^{test} and the orange line represents D_{T2}^{test} . The green line is consistently higher than the orange line. The optimal confidence threshold for D_{T1}^{test} is marked as $\theta^{opt} = 0.7$, and for D_{T2}^{test} it is marked as $\theta^{opt} = 0.1$.

Confidence Threshold θ	D_{T1}^{test} mAP50	D_{T2}^{test} mAP50
0.05	0.60	0.35
0.10	0.59	0.38
0.15	0.58	0.30
0.20	0.58	0.32
0.25	0.58	0.30
0.30	0.58	0.32
0.35	0.58	0.30
0.40	0.58	0.32
0.45	0.58	0.30
0.50	0.58	0.32
0.55	0.58	0.30
0.60	0.58	0.32
0.65	0.58	0.30
0.70	0.62	0.30
0.75	0.58	0.32
0.80	0.58	0.30
0.85	0.58	0.32
0.90	0.58	0.30
0.95	0.58	0.32
1.00	0.58	0.35

A line graph showing the mAP50 performance of two models, D_{T1}^{test} and D_{T2}^{test} , over 5 iterations. The y-axis represents mAP50, ranging from 0 to 1.0. The x-axis represents Iterations, ranging from 1 to 5. D_{T1}^{test} (green line) maintains a constant mAP50 of approximately 0.6. D_{T2}^{test} (orange line) starts at approximately 0.38, increases slightly to 0.45 at iteration 3, and then decreases back to 0.4 at iterations 4 and 5.

Iterations	D_{T1}^{test} mAP50	D_{T2}^{test} mAP50
1	0.6	0.38
2	0.6	0.4
3	0.6	0.45
4	0.6	0.4
5	0.6	0.4

Low-confidence Filtering

For the Durban dataset, we first identify images without high-confidence detections, assuming that images with only very low-confidence labels ($\theta_{low} = 0.01$) contain no



Figure 5.11: Example D_{T1} without instances (left) and with synthetic waste (right)

instances. These images are then augmented with synthetic waste, and the model is fine-tuned using this augmented data. This approach will cannot be tested on D_{T2} due to its limited size and lack of object-free images, but for Durban, many images contain little to no waste. Figure 5.11 shows an example of synthetic waste on the uMhlangane.

Results

Table 5.12 provides an overview of the performance of each method applied to the target domains. The iterative filtering method showed the greatest impact, increasing performance by 7% and 10%. In Section 5.4, this method will be compared with the baseline M_B and the fewshot finetuned model M_F .

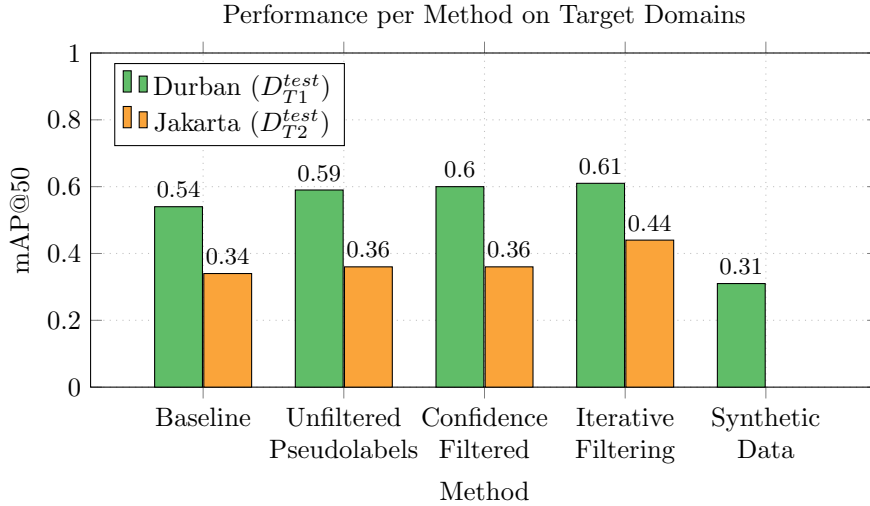


Figure 5.12: Variations of pseudolabeling tested on target domains

5.4 Comparison

Final comparison between the best performing results of the previous sections is shown in Figure 5.13. First of all, the baseline model is shown and how that performs on both the regular bounding boxes and oriented bounding boxes. This is immediately the biggest jump in performance, especially for D_{T2} . The augmentations and synthetic data improved the baseline model but not result in consistent improvements for the target domains and therefore the M_B^{OBB} without augmen-

tations was used as the baseline model in the next steps. Then, for the few-shot fine-tuning, reaching the desired goal was unlikely as some gap would always persist. Finally, for the pseudolabeling, M_F was most optimal when configured after several iterations, but since the model overall tends to underestimate the amount of waste a low threshold suffices.

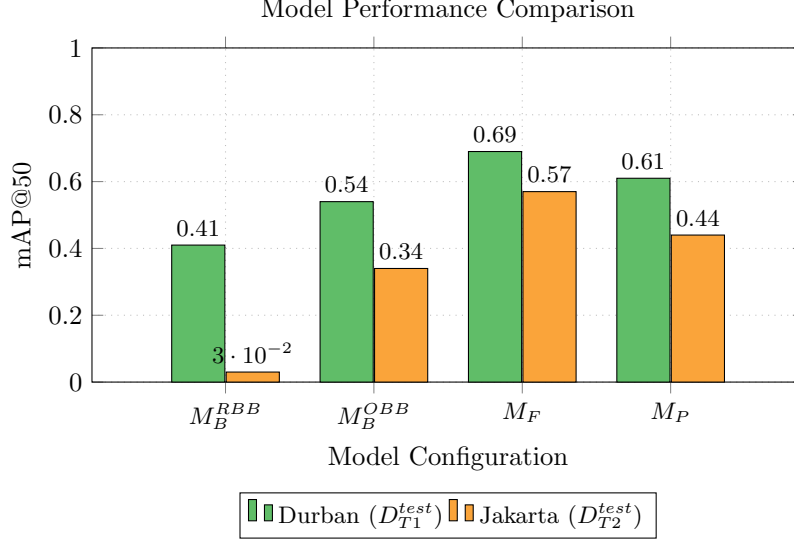


Figure 5.13: Model Performance Comparison

5.5 Multimodal Foundation Models

To compare multimodal foundation models, two have been selected for evaluation: Grounding DINO [46] and OWL-ViT [47]. Both models were tested on 15 images, 5 from each dataset, for a preliminary evaluation. When used off the shelf, these models struggled to detect waste items in full images. However, when cropped bounding boxes, generated by YOLO, were fed into the models, they successfully recognized several objects. This suggests that foundation models could be used to match cropped detected items to a database (e.g., identifying red plastic bottles as Coca-Cola bottles), potentially aiding policymakers with targeted queries about commonly detected waste items. Some objects, like see-through cups, remained difficult to identify (see Figure 5.14).



Figure 5.14: Example of waste detection using a multimodal foundation model.

Chapter 6

Conclusions & Outlook

This thesis explored camera-based monitoring techniques for detecting floating macroplastics in rivers and proposed methods to mitigate the drop in performance when moving from one location to another with minimal data and effort. The findings are structured around the three research questions outlined in the introduction. The first question, mainly discussed in Section 2, concerns the potential of camera-based monitoring systems to address the challenges of plastic waste detection in rivers. While mismanaged plastic waste is a widespread issue, there is a great deal of uncertainty surrounding effective mitigation strategies. Traditional methods, such as visual counting, are labor-intensive and prone to observer bias. Camera-based monitoring, with object detection algorithms, can overcome these drawbacks and have the potential to provide high spatio-temporal resolution data. However, current studies typically focus on specific locations, limiting the scalability policymakers often need. They need to understand plastic movement through water networks to develop, evaluate, and enforce mitigation strategies. Even simple one-class detection systems can help direct resources to areas in need. The challenge lies in the limited availability of real-world labeled data.

Scalability involves both hardware and software components, but this thesis primarily addresses the software aspect. To quantify the lack of scalability, we need to answer the second research question regarding the magnitude of domain shift. Therefore, two datasets were collected from two locations: the Limmat River in Zurich and the uMhlangane River in Durban, South Africa, as discussed in Section 4. During this time, it became clear how important it is to work with the weather. Existing research indicates that intense weather is linked to increased waste transport, but still, it was surprising to observe the sudden increase and rapid dispersal of the items in the river. For further experiments, Zurich served as the source domain and Durban as the target domain, with both datasets recorded using the same camera system. To further assess the scalability of the proposed methods, a third dataset from Jakarta was added as a second target domain.

A reasonable baseline performance was achieved with high image resolution and augmentations to represent the variations of river waste. However, when shifting to other rivers, the performance halved or more. An improvement was seen with the design of a SAM2-based semi-automated annotation pipeline, which helped mitigate inaccuracies in labeling, which are common because of the small scale of the objects. The use of oriented bounding boxes (OBB) led to a significant performance boost, particularly for the Jakarta dataset. An additional benefit is that the OBBs are more informative as they more accurately represent the size of the item.

Despite these improvements, a considerable gap remains, especially for Jakarta, where performance is still less than half. This brings us to the third question of how these shifts could be addressed. Few-shot fine-tuning was proposed as a low-

annotation-effort solution. In addition, human labeling can be less precise with the semi-automated pipeline, which speeds up the process. Initially, adding instances from the target domain improves the models, but it quickly stagnates. Jakarta showed a smoother improvement, whereas Durban was more volatile, likely due to differences in the number of instances per image. To imitate larger amounts of waste items, synthetic data was added to the images. This was expected to improve performance, but the results were inconsistent. Future research should explore more realistic synthetic objects that merge better with the background to see how it could improve performance and eliminate the need for hand labeling.

As an unsupervised method, pseudolabeling is proposed. The OBB baseline model struggled with recall, a challenge that persisted when applied to new domains. Even with a very low confidence threshold, fewer pseudolabels were generated than the number of ground truths present. In accordance with that, confidence filtering had a minimal impact, and the threshold could remain low. However, iterative training with the previous model's weights led to modest improvements.

Although the system still tends to underestimate waste amounts, integrating such systems into policy-making efforts sooner rather than later will allow for a better understanding of what is needed. Even when the quantities are not perfect, sudden increases or decreases in waste levels are also useful indicators. Beyond numerical data, the footage generated by the system will also be a resource for Green Corridors in education and awareness creation.

This research encountered several limitations that provide direction for future work. While multimodal foundation models show promise, they currently fall short of replacing YOLO-based pipelines in terms of performance and scalability. Additionally, upscaling the system and comparing more river waste datasets would allow for an analysis of which factors, such as camera systems or waste types, have the greatest impact. This would also allow for cross-method assessments, using different source domains to determine the consistency of the proposed methods' positive or negative impacts. In turn, that could help to tailor more effective adaptation strategies.

Several aspects were outside of the scope of this thesis but very important to achieve the overall mission. The current hardware setup is still a prototype and should be optimized to upscale monitoring efforts. Then there are societal challenges, such as combating corruption and advancing plastic waste beneficiation that are essential for achieving long-term success in waste management. Lastly, more research in the dispersal of plastics related to weather conditions would benefit preventive measures against plastic pollution.

You cannot improve what you do not measure!

Bibliography

- [1] R. Geyer, J. R. Jambeck, and K. L. Law, “Production, use, and fate of all plastics ever made,” *Science Advances*, vol. 3, 2017.
- [2] OECD, *Global Plastics Outlook*, 2022.
- [3] United Nations Environment , “Single-use plastics: A roadmap for sustainability,” <https://www.unep.org/ietc/resources/publication/single-use-plastics-roadmap-sustainability>, 2018, (Accessed on 03/27/2024).
- [4] J. R. Jambeck, R. Geyer, C. Wilcox, T. R. Siegler, M. Perryman, A. Andrady, R. Narayan, and K. L. Law, “Plastic waste inputs from land into the ocean,” *Science*, vol. 347, no. 6223, p. 768–771, Feb. 2015.
- [5] L. J. Meijer, T. van Emmerik, R. van der Ent, C. Schmidt, and L. Lebreton, “More than 1000 rivers account for 80% of global riverine plastic emissions into the ocean,” *Science Advances*, vol. 7, 2021.
- [6] A. Isobe and S. Iwasaki, “The fate of missing ocean plastics: Are they just a marine environmental problem?” *Science of The Total Environment*, vol. 825, p. 153935, Jun. 2022.
- [7] M. Kalina, J. Kwangulero, F. Ali, Y. G. Abera, and E. Tilley, ““where does it go?”: Perceptions and problems of riverine and marine litter amongst south africa and malawi’s urban poor,” *PLOS Water*, vol. 1, no. 3, p. e0000013, Mar. 2022.
- [8] S. C. Gall and R. C. Thompson, “The impact of debris on marine life,” *Marine Pollution Bulletin*, vol. 92, 2015.
- [9] Y. Mato, T. Isobe, H. Takada, H. Kanehiro, C. Ohtake, and T. Kaminuma, “Plastic resin pellets as a transport medium for toxic chemicals in the marine environment,” *Environmental Science and Technology*, vol. 35, 2001.
- [10] Z. Yuan, R. Nag, and E. Cummins, “Human health concerns regarding microplastics in the aquatic environment - from marine to food systems,” 2022.
- [11] S. G. Tetu, I. Sarker, V. Schrameyer, R. Pickford, L. D. Elbourne, L. R. Moore, and I. T. Paulsen, “Plastic leachates impair growth and oxygen production in prochlorococcus, the ocean’s most abundant photosynthetic bacteria,” *Communications Biology*, vol. 2, 2019.
- [12] Deloitte, “The price tag of plastic pollution: An economic assessment of river plastic,” <https://www2.deloitte.com/content/dam/Deloitte/nl/Documents/strategy-analytics-and-ma/deloitte-nl-strategy-analytics-and-ma-the-price-tag-of-plastic-pollution.pdf>, 2019, (Accessed on 03/28/2024).

- [13] Rijksoverheid Nederland, “Statiegeld op kleine plastic flesjes voor minder zwerfafval,” <https://www.rijksoverheid.nl/actueel/nieuws/2020/04/24/statiegeld-op-kleine-plastic-flesjes-voor-minder-zwerfafval>, 04 2020, (Accessed on 03/28/2024).
- [14] L. Lebreton and A. Andrady, “Future scenarios of global plastic waste generation and disposal,” *Palgrave Communications*, vol. 5, no. 1, Jan. 2019.
- [15] Plastics SA, “Extended producer responsibility: Creating a true circular economy for plastics packaging,” November 2024.
- [16] United Nations Environment Programme, “Unea resolution 5/14 entitled “end plastic pollution: Towards an international legally binding instrument”,” <https://digitallibrary.un.org/record/3999257>.
- [17] C. J. van Calcar and T. H. M. van Emmerik, “Abundance of plastic debris across european and asian rivers,” *Environmental Research Letters*, vol. 14, no. 12, p. 124051, 2019.
- [18] T. van Emmerik, T.-C. Kieu-Le, M. Loozen, K. van Oeveren, E. Strady, X.-T. Bui, M. Egger, J. Gasperi, L. Lebreton, P.-D. Nguyen, A. Schwarz, B. Slat, and B. Tassin, “A methodology to characterize riverine macroplastic emission into the ocean,” *Frontiers in Marine Science*, vol. 5, Oct. 2018.
- [19] R. Hurley, H. F. V. Braaten, L. Nizzetto, E. H. Steindal, Y. Lin, F. Clayer, T. van Emmerik, N. T. Buenaventura, D. P. Eidsvoll, A. Økelsrud, M. Norling, H. N. Adam, and M. Olsen, “Measuring riverine macroplastic: Methods, harmonisation, and quality control,” *Water Research*, vol. 235, p. 119902, 2023.
- [20] L. Schreyers, T. Van Emmerik, T. L. Nguyen, N. Phung, E. Castrop, T. Bui, E. Strady, S. Kosten, and L. Biermann, “A field guide for monitoring riverine macroplastic entrapment in water hyacinths,” *Frontiers in Environmental Science*, vol. 9, 2021.
- [21] K. Topouzelis, D. Papageorgiou, G. Suaria, and S. Aliani, “Floating marine litter detection algorithms and techniques using optical remote sensing data: A review,” *Marine Pollution Bulletin*, vol. 170, p. 112675, 2021.
- [22] European Space Agency, “S2 mission,” <https://sentiwiki.copernicus.eu/web/s2-mission>, (Accessed on 12/04/2024).
- [23] M. Geraeds, T. van Emmerik, R. de Vries, and M. S. bin Ab Razak, “Riverine plastic litter monitoring using unmanned aerial vehicles (uavs),” *Remote Sensing*, vol. 11, no. 17, 2019.
- [24] C. H. Chiang and J. G. Juang, “Application of uavs and image processing for riverbank inspection,” *Machines*, vol. 11, no. 9, 2023.
- [25] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, “Microsoft coco: Common objects in context,” 2014.
- [26] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, “Imagenet large scale visual recognition challenge,” *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

- [27] P. F. Proença and P. Simões, “Taco: Trash annotations in context for litter detection,” 2020.
- [28] G. Thung, “Trashnet: Dataset and neural network for trash classification,” <https://github.com/garythung/trashnet>, 2017, accessed: 2024-04-11.
- [29] Y. Cheng, J. Zhu, M. Jiang, J. Fu, C. Pang, P. Wang, K. Sankaran, O. Onabola, Y. Liu, D. Liu, and Y. Bengio, “Flow: A dataset and benchmark for floating waste detection in inland waters,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 10 2021, pp. 10 953–10 962.
- [30] C. van Lieshout, K. van Oeveren, T. van Emmerik, and E. Postma, “Automated river plastic monitoring using deep learning and cameras,” *Earth and Space Science*, vol. 7, no. 8, Aug. 2020.
- [31] T. Jia, A. J. Vallendar, R. de Vries, Z. Kapelan, and R. Taormina, “Advancing deep learning-based detection of floating litter using a novel open dataset,” *Frontiers in Water*, vol. 5, Dec. 2023.
- [32] R. Garelo, H.-P. Plag, A. Shapiro, S. Martinez, J. Pearlman, and L. Pendleton, “Technologies for observing and monitoring plastics in the oceans,” in *OCEANS 2019 - Marseille*. IEEE, Jun. 2019, p. 1–6.
- [33] S. Armitage, K. Awty-Carroll, D. Clewley, and V. Martinez-Vicente, “Detection and classification of floating plastic litter using a vessel-mounted video camera and deep learning,” *Remote Sensing*, vol. 14, no. 14, p. 3425, 07 2022.
- [34] Z. Pu, X. Geng, D. Sun, H. Feng, J. Chen, and J. Jiang, “Comparison and simulation of deep learning detection algorithms for floating objects on the water surface,” in *2023 4th International Conference on Computer Engineering and Application (ICCEA)*. IEEE, Apr. 2023, p. 814–820.
- [35] F. F. Putra and Y. D. Prabowo, “Low resource deep learning to detect waste intensity in the river flow,” *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 5, p. 2724–2732, Oct. 2021.
- [36] C. Nakkach and K. Moussi, “Deep learning and augmented reality based pollution detection system for sea surface applications,” in *2023 IEEE/ACIS 8th International Conference on Big Data, Cloud Computing, and Data Science (BCD)*. IEEE, Dec. 2023, p. 303–307.
- [37] P. F. Tasserou, L. Schreyers, J. Peller, L. Biermann, and T. van Emmerik, “Toward robust river plastic detection: Combining lab and field-based hyperspectral imagery,” *Earth and Space Science*, vol. 9, no. 11, Oct. 2022.
- [38] United Nations Environment Programme. (2024) Beat plastic pollution interactive. Accessed: 2024-07-04.
- [39] Y. Ganin and V. Lempitsky, “Unsupervised domain adaptation by backpropagation,” in *Proceedings of the 32nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, F. Bach and D. Blei, Eds., vol. 37. Lille, France: PMLR, 07 2015, pp. 1180–1189.
- [40] P. Oza, V. A. Sindagi, V. VS, and V. M. Patel, “Unsupervised domain adaptation of object detectors: A survey,” 2021.

- [41] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2017, p. 2242–2251.
- [42] A. RoyChowdhury, P. Chakrabarty, A. Singh, S. Jin, H. Jiang, L. Cao, and E. Learned-Miller, “Automatic adaptation of object detectors to new domains using self-training,” in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2019, p. 780–790.
- [43] M. Khodabandeh, A. Vahdat, M. Ranjbar, and W. G. Macready, “A robust learning approach to domain adaptive object detection,” 2019.
- [44] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. Van Gool, “Domain adaptive faster r-cnn for object detection in the wild,” 2018.
- [45] K. Saito, Y. Ushiku, T. Harada, and K. Saenko, “Strong-weak distribution alignment for adaptive object detection,” 2018.
- [46] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su, J. Zhu, and L. Zhang, “Grounding dino: Marrying dino with grounded pre-training for open-set object detection,” 2023.
- [47] M. Minderer, A. Gritsenko, A. Stone, M. Neumann, D. Weissenborn, A. Dosovitskiy, A. Mahendran, A. Arnab, M. Dehghani, Z. Shen, X. Wang, X. Zhai, T. Kipf, and N. Houlsby, “Simple open-vocabulary object detection with vision transformers,” 2022.
- [48] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer, “Sam 2: Segment anything in images and videos,” *Meta FAIR*, 2023, <https://ai.meta.com/research/publications/sam-2-segment-everything-in-images-and-videos/>.
- [49] Autonomous River Cleanup, “Autonomous River Cleanup,” <https://riverclean.ethz.ch/>, accessed: 2024-07-04.
- [50] S. Kalanathan, “Bridge-mounted detection of plastic waste quantification using cameras,” Bachelor Thesis, Computer Vision and Geometry Lab (CVG Lab), ETH Zurich, Zurich, Switzerland, Summer Semester 2021.
- [51] J. Habersatter, “The bottlespotter: Design of a river waste quantification system,” Master Thesis, ETH Zurich, 2022, spring Term.
- [52] M. Hauswirth, “Design of an efficient river-waste-quantification system using classic computer vision techniques implemented as an iot device,” 2022, semester Thesis, Autumn Term.
- [53] M. Hogenkamp, “Design of a hybrid detection pipeline for river waste quantification,” 2023, semester Thesis.
- [54] L. Meijer. (2021, April) In search of the rivers that carry plastic into the ocean. The Ocean Cleanup. Accessed: 2025-01-04.
- [55] Global Health Engineering, “Global Health Engineering.”
- [56] R. A. Bergen, “Evaluating the potential of extended producer responsibility returns for a small local waste collection company through a brand audit of riverine plastic waste in durban, south africa,” Master’s Thesis, ETH Zürich, 2023, eTH Supervisor: Dr. Prof. Elizabeth Tilley, Supervisor: Dr. Marc Kalina.

- [57] C. Meyer-Piening, “Examination of non-recycled marine plastic litter in order to identify recycling and beneficiation pathways in durban, south africa,” Master’s Thesis, Global Health Engineering Institute, ETH Zürich, 2022, supervisor: Marc Kalina.
- [58] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” 2017.
- [59] Esquivel Estay, Fidel, Kolvenbach, Hendrik, Strübin, Dario, Sun, Benjamin, Stolle, Jonas, Ensmenger, Adrian, Laporte, Lillian, Elbir, Emre, and Hutter, Marco, “Dataset and analysis of river waste pollution in the limmat river, ch, during a one-day city festival (zurich street parade, 2023),” Tech. Rep., 2024.
- [60] C. T. J. Roebroek, S. Harrigan, T. H. M. van Emmerik, C. Baugh, D. Eilander, C. Prudhomme, and F. Pappenberger, “Plastic in global rivers: are floods making it worse?” *Environmental Research Letters*, vol. 16, no. 2, p. 025003, Jan. 2021.
- [61] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” 2014, accessed: 2025-03-04.
- [62] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, W. Luo, M. Datcu, M. Pelillo, and L. Zhang, “Dota: A large-scale dataset for object detection in aerial images,” 2018, accessed: 2025-03-04.
- [63] M. Ahmed, N. Khan, P. R. Ovi, N. Roy, S. Purushotham, A. Gangopadhyay, and S. You, “Gadan: Generative adversarial domain adaptation network for debris detection using drone,” in *2022 18th International Conference on Distributed Computing in Sensor Systems (DCOSS)*. IEEE, May 2022, p. 277–282.
- [64] G. Aldric Sio, D. Guantero, and J. Villaverde, “Plastic waste detection on rivers using yolov5 algorithm,” in *2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. IEEE, Oct. 2022, p. 1–6.
- [65] L. Chen and J. Zhu, “Water surface garbage detection based on lightweight yolov5,” *Scientific Reports*, vol. 14, no. 1, Mar. 2024.
- [66] I. Cortesi, F. Mugnai, R. Angelini, and A. Masiero, “Mini uav-based litter detection on river banks,” vol. X-4/W1-2022. Copernicus GmbH, Jan. 2023, p. 117–122.
- [67] E. Gabor and P. Gabor, “Enhanced floating plastic waste detecting on offsets of river tisa, hungary,” *Chemical Engineering Transactions*, vol. 107, p. 73–78, Dec. 2023.
- [68] J. He, Y. Cheng, W. Wang, Y. Gu, Y. Wang, W. Zhang, A. Shankar, S. Selvarajan, and S. A. P. Kumar, “Ec-yolox: A deep-learning algorithm for floating objects detection in ground images of complex water environments,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 17, pp. 7359–7370, 2024.
- [69] X. Jiang, Z. Yang, J. Huang, G. Jin, G. Yu, X. Zhang, and Z. Qin, “Yolov5n++: An edge-based improved yolov5n model to detect river floating debris,” *Journal of Intelligent and Fuzzy Systems*, vol. 46, no. 1, p. 2507–2520, Jan. 2024.

- [70] B. O. Kelly, S. Chen, E. P. Zhou, and M. Elshakankiri, "Ai-enabled plastic pollution monitoring system for toronto waterways," in *2023 10th International Conference on Internet of Things: Systems, Management and Security (IOTSMS)*. IEEE, Oct. 2023, p. 53–58.
- [71] F. Lin, T. Hou, Q. Jin, and A. You, "Improved yolo based detection algorithm for floating debris in waterway," *Entropy*, vol. 23, no. 9, p. 1111, Aug. 2021.
- [72] M. Moshtaghi and E. Knaeps, "Combining spectral approaches and ai for marine litter detection and identification," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*. IEEE, 07 2021, p. 1130–1133.
- [73] J.-Y. Shen, C.-K. Lu, and L. G. Lim, "A novel deep convolutional neural network pooling algorithm for small floating objects detection," in *2023 International Conference on Consumer Electronics - Taiwan (ICCE-Taiwan)*. IEEE, 07 2023, p. 175–176.
- [74] J. P. Q. Tomas, M. N. D. Celis, T. K. B. Chan, and J. A. Flores, "Trash detection for computer vision using scaled-yolov4 on water surface," in *The 11th International Conference on Informatics, Environment, Energy and Applications*, ser. IEEA 2022. ACM, Mar. 2022, p. 1–8.
- [75] N. A. Zailan, M. M. Azizan, K. Hasikin, A. S. Mohd Khairuddin, and U. Khairuddin, "An automated solid waste detection using the optimized yolo model for riverine management," *Frontiers in Public Health*, vol. 10, Aug. 2022.
- [76] N. A. Zailan, A. S. Mohd Khairuddin, K. Hasikin, M. H. Junos, and U. Khairuddin, "An automatic garbage detection using optimized yolo model," *Signal, Image and Video Processing*, vol. 18, no. 1, p. 315–323, Sep. 2023.

Appendix A

Related Work Overview

Table A.1 shows an overview of recent studies on floating waste detection, focusing on the datasets that were used, the perspective they have, and the taxonomy type of the labeled instances. It also reports the type of detector employed along with the highest achieved results. Lastly, the location and time period of the datasets are noted, though these aspects are less relevant for generic waste datasets like TACO [27] and TrashNet [28].

Table A.1: Overview of Floating Waste Detection Studies

Reference	Waste Dataset (# Labeled Images)	Taxonomy	Perspective	Detector	Results	Location	Time Period	Pe-
Almed et al., 2022 [63]	Custom (500)	Binary	Top-view	YOLOv3	0.97 mAP	Baltimore, USA	3 months	
Sio et al., 2022 [64]	Custom (Unspecified)	Object-Based	Close-up	YOLOv5	0.79 precision, 0.58 recall	Unspecified	Unspecified	
Armitage et al., 2022 [33]	Custom (1920)	Object-Based	Ship-view	YOLOv5	0.68 mAP, 0.64 F1-score	Plymouth, UK	3 months	
Chen et al., 2024 [65]	Flow-Img (2000)	Object-Based	Ship-view	YOLOv5	0.85 mAP	Unspecified	3 months	
Chiang et al., 2023 [24]	TrashNet (2527), TACO (1500), HAIDA (3532), Custom (2595)	Binary	Close-up (TrashNet), Various (TACO, HAIDA), Top-view (Custom)	YOLOv5	0.75 mAP	Studio (TrashNet), Various (TACO, HAIDA), Taipei, Taiwan (Custom)	Not a time series	
Cortesi et al., 2023 [66]	UAVVaste (772)	Binary	Top-view	YOLOv4	0.86 precision, 0.80 recall, 0.83 F1-score	Various locations	Not a time series	
Gabor et al., 2023 [67]	Custom (195)	Binary	Ship-view	Faster R-CNN	Unspecified	Bodrogkisláhd, Hungary	13 months	
He et al., 2024 [68]	Lin2021	Object-Based	Ship-view	EC-YOLOX	0.82 mAP	Unspecified	Unspecified	
Jiang et al., 2024 [69]	Custom (Unspecified)	Object-Based	Top-view	YOLOv5	0.87 mAP	Unspecified	Unspecified	
Kelly et al., 2023 [70]	TACO (1500)	Binary	Various	Mask R-CNN	Unspecified	Various locations	Not a time series	
Lin et al., 2021 [71]	Custom (2400)	Object-Based	Ship-view	YOLOv5	0.79 mAP	Unspecified	Unspecified	
Moshaghni et al., 2021b [72]	Custom (Unspecified)	Binary	Top-view	Cascade-RCNN	0.75 mAP	Mol, Belgium	Unspecified	
Naklach et al., 2023 [36]	Custom (Unspecified)	Object-Based	Close-up	YOLOv8	0.96 precision, 0.95 recall, 0.95 F1-score	Unspecified	Not a time series	
Pu et al., 2023 [34]	Flow-Img (2000)	Object-Based	Ship-view	Faster R-CNN, Cascade-RCNN, YOLOv3, YOLOv5	0.92 mAP (YOLOv3)	Various locations	3 months	
Putra et al., 2021 [35]	Custom (370)	Binary	Ship-view	YOLOv3	0.65 mAP	Jakarta, Indonesia	Unspecified	
Shen et al., 2023 [73]	Flow-Img (2000)	Object-Based	Ship-view	YOLOv4	Unspecified	Various locations	3 months	
Tomas et al., 2022 [74]	Custom (Unspecified)	Material-Based	Close-up	YOLOv4	0.78 mAP	Manila, Philippines	Unspecified	
van Lieshout et al., 2020 [30]	Custom (528)	Binary	Top-view	Faster R-CNN	0.69 precision	Jakarta, Indonesia	10 days	
Zailan et al., 2022 [75]	AquaTrash (369)	Object-Based	Various	YOLOv4	0.89 mAP	Various locations	Not a time series	
Zailan et al., 2023 [76]	AquaTrash (369)	Object-Based	Various	YOLOv4	0.75 mAP	Various locations	Not a time series	

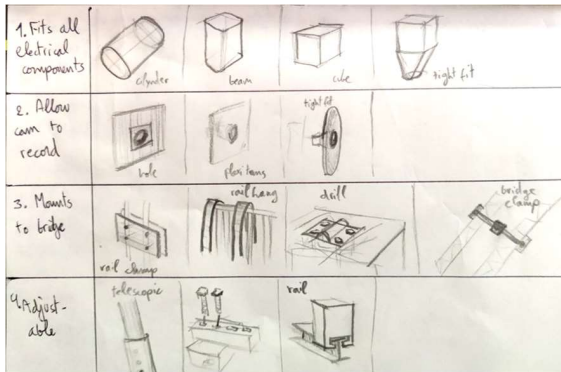
Appendix B

Design Evolution

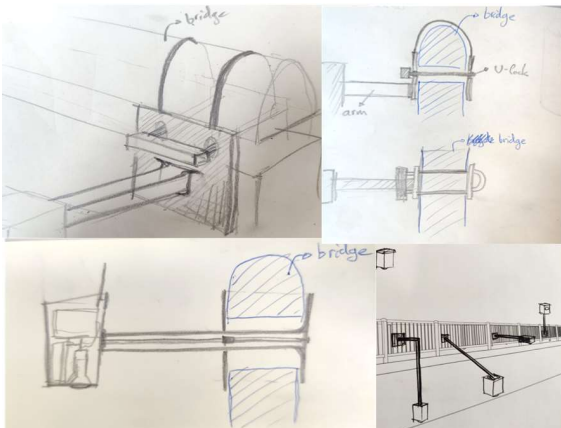
This appendix documents the development process of the ARCAM system. The evolution is divided into four phases: (1) Ideation, where initial concepts and requirements were defined; (2) Prototyping, which involved building and testing early versions of the system; (3) Field Adaptations, where modifications were made based on real-world deployment challenges; and (4) Deployment, where the final system was installed and used for river waste monitoring.

PHASE 1: Ideation

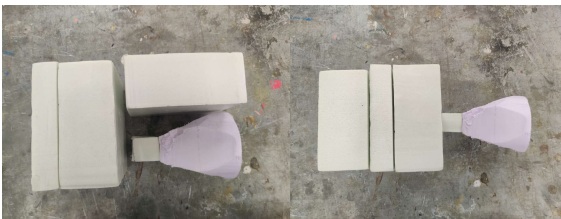
Morphological chart



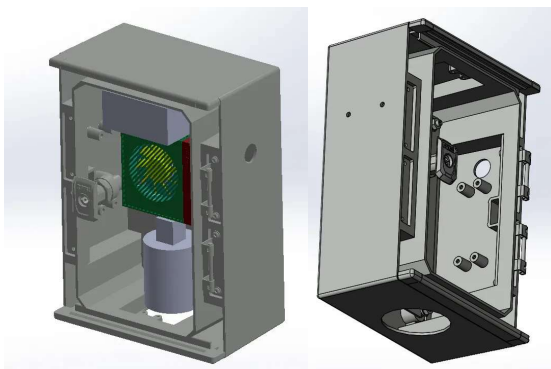
Mount system sketches



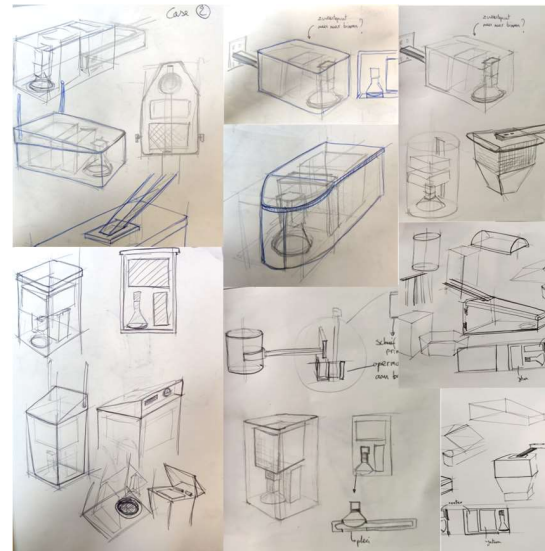
Foam models



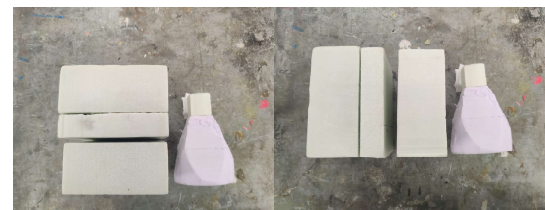
CAD Design



Detection box sketches



Foam models



PHASE 2: Prototyping

Initial prototype & clamp tests



Field-tests at ARCHE in Geneva



Implement learnings from field-tests



PHASE 3: Field Adaptations

Adapt bridge-mount and make cage



Deployment tests at Paradise Valley



PHASE 4: Deployment

Deployment at Riverhorse Valley



Appendix C

Stakeholder Interviews

C.1 Nick Swan - Green Spaces Programme Manager Green Corridors

Green Corridors Overview

Can you introduce yourself and describe your responsibilities at Green Corridors?

In general, environmental consulting with a focus on social ecology (nature-based solutions for communities). For Green Corridors:

- Operational responsibilities, rather than human resources.
- Program manager: transformative management of riverine (all operations related to riverine corridor).

Background information:

- City has 16 river catchments (uMgeni is one of them). They experience consequences of environmental changes caused by invasive alien plants in riverine corridors. Shallow-rooted alien invaded plants pile up against the bridges.
- These plants are 70-80 percent of solid waste that is washed up against the bridge (these numbers come from collection efforts).
- To mitigate this, maintenance of the landscape has to be funded. For every 1.8 rand spent on those efforts, you save 3.8 rand on breaking infrastructure.
- Name of city: historical Durban; name in local language: eThekweni metropolitan municipal authority.

Partnerships:

- Partnership between ETH and Green Corridors: another student now on absorbent hygiene products.
- eThekweni municipality: transformative riverine management program. City-level program for the last 10 years. Green Corridors is one of the organizations that has a formal agreement to do work in riverine management.
- Russel Stow is the project executive for the TRMP at the municipality.

Focus Areas: Studies, business cases, international funding, and analyzing the costs and benefits of maintaining river systems.

Project Description and Goals

Could you briefly describe what the project does and what the goals are?

Transformative riverine and open space management. Collecting and beneficiating waste for marketable products, such as processing Spanish reeds for useful products like shoe soles.

Challenges and Scale

What are the biggest challenges you face in these endeavours? Can you rank them?

1. Managing the malfunctioning state of governance.
2. Addressing waste management and environmental degradation.
3. Tackling rapid urbanization and its impacts, such as informal settlements not qualifying for normal services.
4. Amount of increasing solid waste and need for effective partnerships.

On what scale does Green Corridors operate when it comes to waste management?

- Focuses on the Umlanghani river, a tributary of uMgeni.
- Solid waste collection, litter booms, and public employment programs with 200 people working in the catchments to collect solid waste and control alien plant growth.
- Implements a beneficiation program to make marketable products from collected waste.

Waste Monitoring and Data Collection

Have you already done river waste monitoring? What data or knowledge came out of that?

Limited monitoring.

- Past projects include thesis students from ETH GHE (Raul and Chiara).
- Android smartphone-based app for photos before and after interventions, metrics of extracted waste, etc.
- Challenges in resources for optimal operation.
- Upcoming GIS data management improvements.

Major interest for GC: Improve waste data collection and characterize waste for actionable insights.

Policy and Collaboration

What does Green Corridors do on a policy-making level? What are they currently working on? Who enforces the EPR?

- Focus on policy implementation rather than making. EPR (Extended Producer Responsibility) is run by the industry.

- Goals: Informing actors, enhancing ongoing data collection, and facilitating compliance with EPR regulations.

Who do Green Corridors collaborate with for the management of disposed waste in rivers?

- Collaboration involves industry, government, and community.
- Partnerships with stakeholders like the Coca-Cola Foundation and others interested in compliance.

How is the relationship between Green Corridors and the municipality?
Close collaboration with systemic governance issues, though innovative and forward-thinking programs exist.

The Coca-Cola Foundation is mentioned as a partner/funder. What is the nature of the partnership?

- Strong interest from the industry, but challenges include governance issues and trust with the government.

C.2 Siphiwe Rakgabale - Coordinator Litter Boom Project

Litter Boom Project

Can you introduce yourself and describe your responsibilities in the litter boom project?

I am Siphwe Rakgabale, founder of Tri-Ecotos in partnership with GC. I oversee cleanup efforts and have been working in environmental conservation for 12 years. My responsibilities include coordinating the litter boom project. I monitor the performance of the installed litter booms by visiting each site to check and follow up on their operations. The project began in 2019 with GC, and currently, 21 litter booms are installed in the Umgeni area. Each area has assistants responsible for maintaining two litter booms.

Could you briefly describe what the project does and what the goals are?
The vision of the project is to prevent waste from reaching the sea or ocean while educating communities about waste management. The goal is to help communities understand the need to reduce waste and benefit from sustainable practices.

Short-term goals:

- Establish community waste stations for sorting and processing waste.
- Secure funding and partnerships, such as Sufferpool, which is the biggest funder of the project.

Targets:

- Install 50 litter booms by the end of 2025, expanding to other municipalities.
- Create two waste receiving stations (currently none) with a focus on sustainable waste processing, such as baling.

What are the biggest challenges you face in tracking and managing riverine waste? Can you rank them?

1. Lack of suitable storage facilities or stations.

2. High costs of purchasing necessary machinery.
3. Competitors stealing materials from the booms.
4. Difficulty estimating waste type and quantity due to variability (e.g., plastics, water hyacinths, trees, polystyrene).
5. Challenges from stormwater drain systems and illegal dumping.
6. Seasonal variations in waste composition and volume.

Common waste types: PET bottles, food residue, and household products.

Missing elements:

- Community education about waste management.
- Research on waste characteristics and litter boom effectiveness.

After collecting the waste, it is sorted. Could you describe the sorting process and categories you use?

The sorting process is divided into two types:

1. **Community project:** Focuses on plastics for recycling, particularly PP and hard plastics, and some PET. Cans and cardboard are also collected.
2. **Litter boom project:** Local assistants, trained for sorting, collect waste from booms and separate it into recyclable and landfill categories. They use a registration sheet to weigh and document the materials. Sorting occurs Monday to Friday, with collection every Wednesday.

Capacity for further sorting: Material-wise or even brand-specific sorting could be conducted temporarily to validate system performance.

What information about the waste is registered? Do you have an idea of the amounts you collect?

While waste is not counted item-wise, weights are recorded. For example, 13 bags of PET can fill one bulk bag depending on the material's compaction.

Is the waste wet? Waste is often contaminated. It is scooped, dried for a day, and then separated.

How frequently and for how long are the booms deployed? They are operational 24/7 and strategically positioned to facilitate waste collection.

What determines the locations of the litter booms? Locations are selected based on river flow and stormwater drain systems. Stakeholders collaborate to identify suitable sites.

Environmental Considerations

How does the period from October to December, with its rainfall, affect the amounts of waste and your operations? This period is the peak season for waste inflow, particularly due to increased rain pushing waste through stormwater drains.

Could you get me videos of the river that show waste floating by?

Videos can be provided. Pictures and videos via email.

Outlook/Automated Monitoring

If a bridge-mounted camera monitors floating plastics, what specific data would be valuable?

- General interest in research on plastic waste
- Waste composition and arrival speed.
- Water level and flow rate to optimize boom performance.

In what ways could you support the mounting and maintenance of AR-CAM? Rainy days, when waste volumes are high, provide ideal monitoring opportunities. All requirements and coordination for ARCAM installations go through Sips.

Appendix D

Experiments

D.1 Baseline Model Hyperparameters

Parameter	Baseline Model (M_B)	Explanation
Model	YOLO-n, YOLO-s, YOLO-m YOLO-n-obb, YOLO-s-obb, YOLO-m-obb	Starting with pre-trained YOLO models of varying sizes.
Epochs	100	Number of training epochs.
Patience	10	Number of epochs without improvement before early stopping to prevent overfitting.
Batch Size	16, 8	Number of images processed per batch.
Image Size	640, 1280	Resolution to which input images are resized before training. Larger sizes capture more detail but require more computation.
Initial Learning Rate	0.01	Influences how fast model weights are updated during training. Lower values improve stability.
Box Loss	7.5	Weight assigned to the bounding box regression loss. Higher values increase its impact on training.
Class Loss	0.5	Weight of the classification loss component, affecting how the model learns object classes.
Distribution Focal Loss	1.5	Weight of the focal loss, improving object detection for imbalanced datasets.
Data Augmentation Settings		
Rotation ($\pm 5^\circ$)	degrees = 5	Introduces small random rotations to improve detection of tilted objects.
Scaling (50%)	scale = 0.5	Resizes images by 50%, helping the model generalize to different object distances.
Flipping (50%)	fliplr = 0.5	Horizontally flips images with a 50% probability, enhancing viewpoint invariance.
Mosaic	mosaic = 1.0	Merges four images into one, increasing object density and improving recall.

Table D.1: Training parameters for the baseline model (M_B), including individual augmentation settings.

D.2 Fewshot Finetuning Results

The following graphs (Figures D.1, D.2, D.3, and D.4) present results from few-shot fine-tuning experiments using a small image resolution (640x640) for both the RBB and OBB datasets.

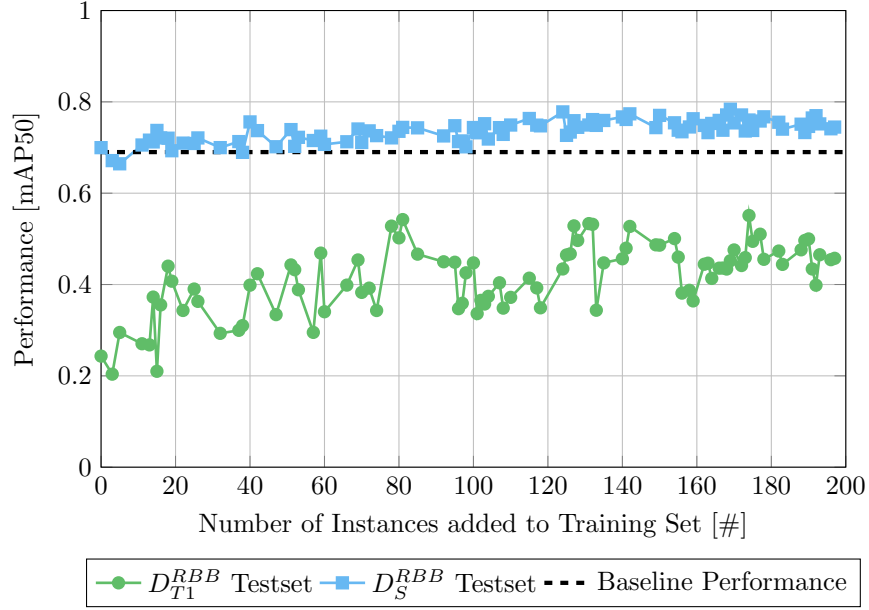


Figure D.1: Model performance on D_S^{RBB} and D_{T1}^{RBB} test sets with N_t new location instances added to the training set.

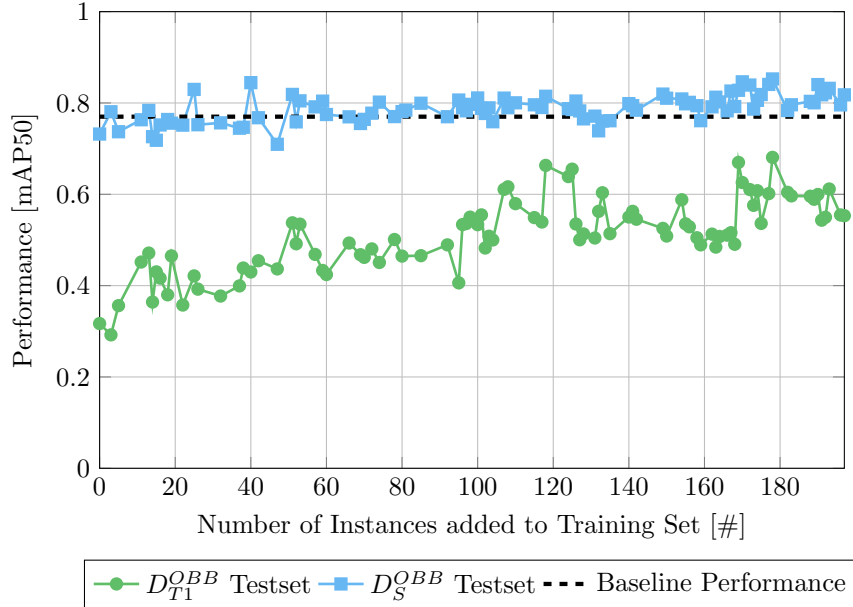


Figure D.2: Model performance on D_S^{OBB} and D_{T1}^{OBB} test sets with N_t new location instances added to the training set.

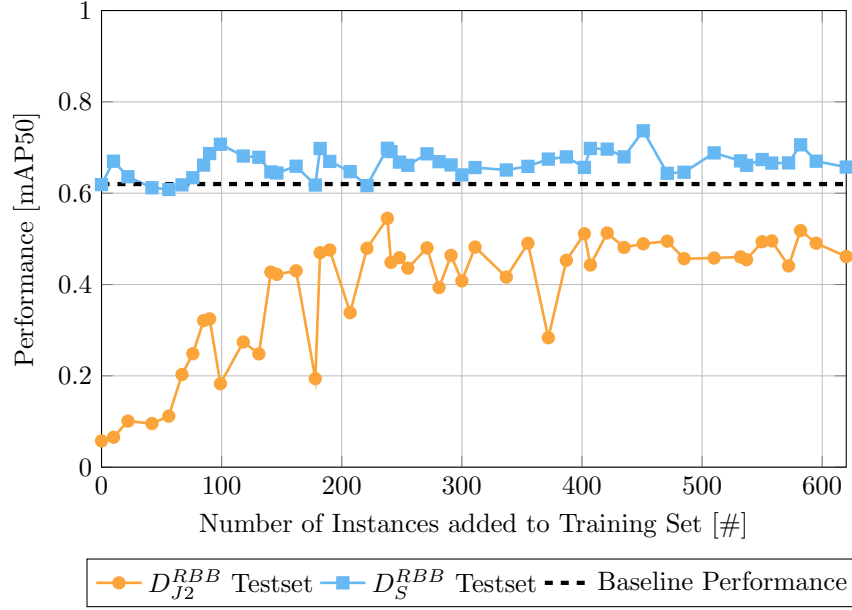


Figure D.3: Model performance on D_S^{RBB} and D_{T2}^{RBB} test sets with N_t new location instances added to the training set.

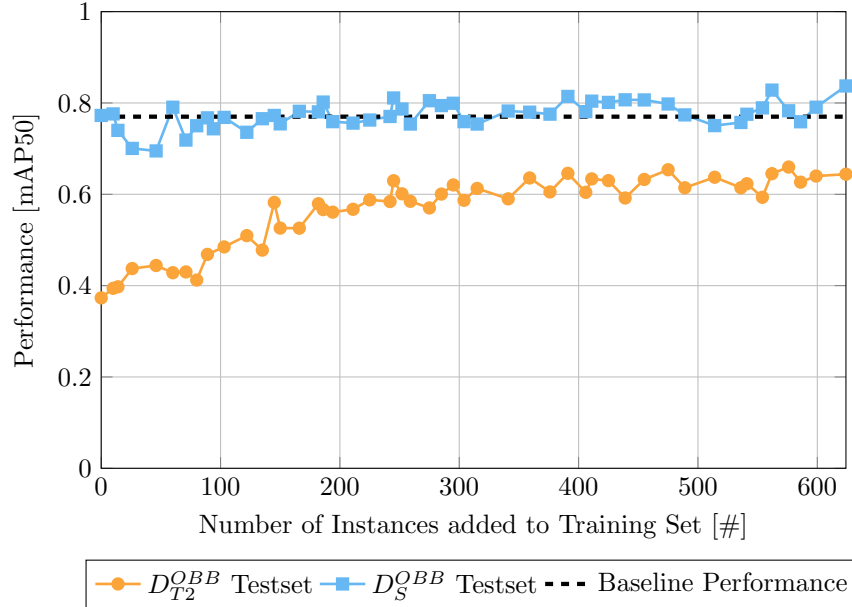


Figure D.4: Model M_{F1} performance on D_S^{OBB} and D_{T1}^{OBB} test sets with N_t new location instances added to the training set.