2026-08

YOLO26m-CW: A Water-Augmented Spatial Attention Framework for Robust Riverine Waste Monitoring

Hassan, Kazi Md. Rakibul

Independent University, Bangladesh (IUB)

YOLO26m-CW: A Water-Augmented Spatial Attention Framework for Robust Riverine Waste Monitoring

August 2026

Prepared by:

Kazi Md. Rakibul Hassan
ID: 2220936

Komol Krishna Paul
ID: 2221337

Ayman Rahman
ID: 2010411

Department of Computer Science and Engineering
Independent University, Bangladesh

Supervised by:

Md Rashedur Rahman, D.Eng.
Assistant Professor
Department of Computer Science and Engineering
Independent University, Bangladesh

A Final Year Design Project (FYDP) Report submitted for the Degree of Bachelor's of Computer Science & Engineering

Attestation

We are aware of the fact that plagiarism is strictly prohibited and it conflicts with my university's rules and regulations. I guarantee the authenticity of my work. I have used some copyrighted material and models of others in my project which has been properly cited following the international standards proper guidelines.

Author Name:

. . . . . .

Signature:

. . . . . .

Author Name:

. . . . . .

Signature:

. . . . . .

Author Name:

. . . . . .

Signature:

. . . . . .

Evaluation Committee

Supervisor

Name: _____ Signature: _____

Internal Examiner 1

Name: _____ Signature: _____

Internal Examiner 2

Name: _____ Signature: _____

External Examiner

Name: _____ Signature: _____

Acknowledgement

We would like to express our deepest and most sincere gratitude to our respected supervisor,**Dr. Md. Rashedur Rahman, D.Eng.** for his continuous guidance and mentorship throughout this project. His contribution went far beyond conventional supervision. He consistently followed our progress through weekly discussions, carefully reviewed our work, asked thoughtful and challenging questions, and encouraged us to understand the reasoning behind every decision we made. Whenever we faced difficulties, he patiently guided us toward the solution rather than simply providing the answer, which greatly strengthened our problem-solving and research skills.

We are especially grateful for the time he dedicated to teaching us concepts related to **machine learning models, computer vision, and image processing** whenever we needed further clarification. His detailed feedback, questions, and discussions continuously pushed us to improve both our technical understanding and the quality of our work. His mentorship has played a significant role in shaping the way we approach research, experiment with models, analyze results, and critically evaluate our own work.

We would also like to express our sincere gratitude to **Mr. Siam Tahsin Bhuiyan** for his valuable support throughout the project. He provided us with important assistance, particularly through **coding reviews, technical discussions, and feedback on our implementation**. His guidance helped us identify issues in our code, improve our implementation, and maintain the quality of our technical work.

We are deeply grateful to **Center for Computational and Data Sciences (CCDS), Independent University, Bangladesh (IUB)**, for providing us with the resources, facilities, and research environment necessary to carry out this work.

Finally, we sincerely appreciate everyone who supported us directly or indirectly throughout this journey. Their encouragement, assistance, and contributions have been an important part of the successful completion of this project.

We used Anthropic's Claude and OpenAI exclusively to refine the manuscript's language and flow, not for methodology, data analysis, or the generation of results.

List of Publications & Awards

1. K. M. R. Hassan, K. K. Paul, R. Rahman, A. Rahman, S. T. Bhuiyan, S. K. Mazumder, A. Islam, and S. B. Alam, "YOLO26m-CW: A Water-Augmented Spatial Attention Framework for Robust Riverine Waste Monitoring," *14th International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA-2026)*, London, United Kingdom, Jun. 08–09, 2026. **(Accepted)**

2. **Best Paper Award** – Recipient of the Best Paper Award at the *14th International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA-2026)*, London, United Kingdom, Jun. 08–09, 2026.

Abstract

Floating plastic waste in rivers constitutes the primary conduit for marine pollution, conveying an estimated 0.8 to 2.7 million metric tons of plastic into global oceans annually, with over 80% originating from 1,000 rivers. This contamination severely impacts river-dependent developing nations in South and Southeast Asia by degrading fisheries, clogging urban drainage, and elevating flood risks. To replace manual monitoring, recent work focuses on automated waste interception using Unmanned Surface Vehicles (USVs) equipped with visual perception. However, detecting floating debris remains hindered by aquatic visual noise including sun glare, dynamic waves, and reflections and spatial feature loss during downsampling in standard convolutional networks. Consequently, existing models struggle to balance high detection accuracy with the lightweight computational efficiency required for edge deployment on USVs.

To resolve these limitations, this thesis presents **YOLO26m-CW**, a modular detection framework built upon the Non-Maximum Suppression (NMS)-free YOLO26 architecture. YOLO26m-CW incorporates two plug-in enhancement modules: Coordinate Attention (CoordAtt) and Water-Specific Augmentations (WaterAugs). Integrated into the Spatial Pyramid Pooling - Fast (SPPF) module, CoordAtt embeds horizontal and vertical spatial information into channel attention maps, preserving positional cues for small or occluded targets to enhance recall. Concurrently, WaterAugs implements an environment-targeted data augmentation strategy simulating camera tilt (10°), murky water (HSV perturbations), reflections (vertical flips), waste clutter (mixup), and object distances (scale jitter) to improve precision under visual noise.

A systematic two-stage evaluation on the FloW-Img benchmark (2,000 images, 5,272 instances) highlights the complementary design of these modules: Stage 1 (YOLO26s) confirmed CoordAtt recall (0.804) while WaterAugs precision (0.920). Scaling to YOLO26m in Stage 2 achieved peak performance on the 60–40 split ($F_1 = 0.859$, $\text{mAP}_{50} = 0.866$, $\text{mAP}_{50\text{-}95} = 0.478$, mmAP = 0.672, AUROC = 0.871), outperforming Faster R-CNN (mmAP = 0.184) and Cascade R-CNN (mmAP = 0.434) by +0.488 and +0.238 mmAP. Robustness analysis across five train-test splits (50–50 to 90–10) maintained consistent $\text{mAP}_{50} > 0.90$, peaking at $F_1 = 0.909$ (60–40 split) and precision of 0.959 (80–20 and 90–10 splits). Comparative evaluations against six prior works under matched splits confirmed superior performance across 70–30, 80–20, and 90–10 splits, surpassing PBM-YOLO by +0.091 on $\text{mAP}_{50\text{-}95}$.

Finally, external evaluation training exclusively on three non-riverine datasets (UAV-BD, MJU-Waste, TUD-GV) and testing on FloW-Img quantified the severe performance penalty of out-of-domain data, despite YOLO26m-CW achieving up to 3.9× higher mmAP than baselines. This confirms the necessity of FloW-Img's surface-level riverine perspective. As a lightweight software component, YOLO26m-CW enables edge-deployed USVs to stream real-time telemetry via RESTful APIs to cloud GIS platforms for density heatmapping and automated environmental intervention.

Contents

List of Figures

List of Tables

Chapter-1: Introduction

1.1 Background

One of the most severe environmental challenges of the 21st century is the ever increasing quantity of anthropogenic waste in world watercourses. Rivers, as well as becoming vital transport routes for freshwater ecosystems, agricultural irrigation, industrial supply and urban drainage, are the main sources via which terrestrial waste is carried to the world's oceans. Recent estimates have suggested that there are around 0.8 to 2.7 million metric tons of plastic waste entering the marine environment from rivers per year [1], with a small number of rivers (just over 1000) accounting for almost 80% of these inputs [2]. It is not coincidental that the pollution sources are concentrated in this manner; the entire problem is systemic as river systems act as funnels for mismanaged municipal, industrial and agricultural waste, finally depositing them in coastal and ocean systems.

The impacts of floating riverine waste are evident and profound. Locally, the compaction of debris hinders urban drainage systems, increases flood risk during monsoon periods and pollutes water quality for the benefit of local communities that rely on these waterways for drinking, sanitation, and fishing purposes [3]. On the regional and global level, plastic waste is photodegraded and mechanically broken down into microplastics that enter food chains, enter geological layers and remain in marine environments for centuries. The impact of this pollution is disproportionally on developing countries which depend on rivers, and especially on countries in Southeast and South Asia, with high rates of riverine plastic emissions, due to the combination of rapid urbanisation and insufficient waste management infrastructure.This pollution hits developing countries that rely on rivers particularly hard, including those in South and Southeast Asia, where high riverine plastic emissions are the result of rapid urbanisation and poor waste management infrastructure [3].

To solve this challenge, the change of hands must be evident from passively observing to actively monitoring and intervening. Ecosystems are traditionally dependent on the removal of waste and debris from river banks and waterways using manual labour teams, which is a limited and scalpelimited process with respect to their time, they have been able to cover a restricted stretch of area, time and speed in which such activities can be performed. Manual monitoring is also constrained by inaccessibility of many river segments, safety risks due to high flow conditions, and by the amount of waste that is too large to be removed by hand. Such restrictions have inspired the creation of intelligent, technology-based solutions for continuous tracking of the river surfaces, real-time identification and categorisation of floating solid waste and directing the autonomous (or semi-autonomous) cleaning operations.This has spurned the research in the development

of intelligent, technology-based solutions for continuous monitoring of river surfaces, real-time identification and categorisation of floating solid wastes and directing autonomous or semi-autonomous cleaning operations [4, 5].

The use of computer vision and deep learning for environmental monitoring has greatly improved in recent years. The initial development of object detection was used for autonomous driving, surveillance, and industrial inspection and has been applied to detecting and locating floating waste in aquatic environments. Multi-level segmentation approaches proved to be suitable to classify waste objects in early applications of the technology [6], and new research recently has investigated quantification of litter fluxes by applying non-supervised learning frameworks such as semi-supervised learning [7]. Robots with on-board cameras and real-time inference abilities (USV, UAV) are a potential paradigm for scalable river monitoring [4, 5]. Low-level UAV detection of specific objects such as plastic bottles has been effective [8] while autonomous marine monitoring systems have increasingly relied on deep learning models, like YOLOv11 and DeepLabV3, for complete debris detection and segmentation [9]. A benefit of USVs is their surface-oriented capability; this provides a significant advantage when floating waste will need to be intercepted, and makes them appropriate platforms for detection-driven clean-up operations.

The practical implementation of the transformation from laboratory demonstrations to strong commercialisation, however, continues to be fraught with technical challenges. As illustrated in Fig. 1.1, riverine environments in South and Southeast Asia experience severe accumulations of floating anthropogenic waste, presenting dense clusters of macro-plastics, styrofoam fragments, and organic debris along water surfaces and shorelines. Most standard detection architectures find such aquatic environments extremely challenging due to severe optical interferences: notably dynamic water reflections, glare from the sun, motion blur, and a cluttered background of turbulent water and riparian vegetation. As depicted in the figure, targets of interest are often compact, partially submerged, irregular in shape, and exhibit visual characteristics that closely mirror the surrounding water surface. Resolving these challenges requires object detection frameworks that are accurate, computationally efficient, environmentally aware, and specifically engineered to handle the unique visual noise of open-water aquatic scenes.

1.2 Problem Statement

Although a large amount of research has been undertaken for floating waste detection, there are a number of key limitations with current approaches that prevent them from being readily implemented in practical riverine floating waste monitoring systems. These restrictions can generally be divided into three interrelated problems:

Environmental Sensitivity. Most of the standard object detection architectures, such as the original YOLO, Faster R-CNN and SSD, were designed and tested on images of road scenes with cars, pedestrians, and no moving objects in the background, where lighting conditions tend to be fairly consistent and backgrounds are mostly stationary [4]. These models are applied to visual conditions that are quite different from their training distributions, when deployed in aquatic environments. The specular reflections form false

Figure 1.1: Floating Waste in South and Southeast Asian Rivers: A Growing Pollution Challenge

positive detections, the sun glare causes the image regions to become saturated and targets to be hidden, and hydrodynamic wave patterns continuously change the appearance of both the background and the objects themselves [10, 11]. While there has been some work on adapting the algorithms to be environmental friendly such as USV-YOLO [12] or improved YOLOv8 configuration [13], all the above-mentioned factors contribute to the drop in detection precision and recall and sometimes make the conventional models unusable under the very conditions for which they were developed.

Small Object Detection Difficulty. In river environments floating waste is often small compared to the image resolution especially from USV or UAV platforms at operations distances. In the most realistic publicly available dataset, the FloW-Img benchmark, the average number of objects per image is as low as 2.64 objects and many are small compared to the size of the image itself [14]. With successive pooling and striding layers, the spatial resolution is reduced at every layer of the CNN and small objects lose important fine-grained details as they pass through the downsampling layers [15]. This *resolution-semantic imbalance* (high-level semantics are available, but spatial precision has been traded off in the architecture) is a fundamental limitation that can't be fully resolved by post-processing or by simply scaling up the model [14, 16].

Computational Constraints. The corner on deployment resources when using real-time computation for edge devices like USVs is quite severe. The high accuracy of multi-stage detectors, such as Cascade R-CNN, come at the expense of long computation time to make decisions in real-time for autonomous navigation [14]. On the other hand, small, single-stage detectors are usually poor in their ability to resolve the features that are essential to extract the small low-contrast objects in the presence of noisy backgrounds. Getting the accuracy/efficiency balance right is still a challenge, especially if detection needs to occur across device as well as environment conditions, and small object conditions as described above [17]. Further studies are necessary in order to define a 'best balance' across the variety of aquatic settings, although there have been some studies working on models such as PBM-YOLO to overcome this balance [18].

All three of these are environmental sensitivity, small object detection, and computational efficiency, which are not mutually independent, but interact in complicated ways. Spatial information for small objects is heavily affected by environmental noise and architectural changes required to enhance the detection of small objects tend to be expensive if they increase the computational complexity. This is an integrated problem space where modular and focused approaches can be put in place which solve each problem with different mechanisms, while maintaining reasonable computational complexity

1.3 Research Motivation

The motivation for this research comes from the confluence of three factors: the growing need for a riverine waste problem, a lack of robust methods to detect waste in the real world and the availability of new architectural components yet to be thoroughly tested in this space.

Urgency of the Environmental Crisis. Rivers are the major channel for marine plastic pollution, as mentioned above. The adverse environmental, economic and public health impacts of uncontrolled waste discharge into rivers has been documented and is becoming intolerable [1, 2, 3]. Any improvement in automated waste monitoring directly helps to maintain aquatic ecosystems and state of well-being for the people who rely on them. This urgency is the basic reason to invest research effort into improving detection performance, notwithstanding the seemingly marginal gains that are available.

Limitations of Current Approaches. Existing works on floating waste detection primarily focus on modification of architecture to incorporate attention modules, redesigning feature pyramid networks or introducing novel convolution operators [13, 16, 12, 11]. Moreover, the emergence of new open datasets has encouraged more investigations on learning frameworks [19] but most research has been conducted on specific improvements. These methods have shown improvement with benchmark datasets, but they generally ignore two significant aspects. Firstly, they are tested on one specific, and unknown, train-test split, so it will be impossible to determine if reported improvements reflect true strength or the result of an ideal data split. Secondly, they are not systematic in identifying whether the observed performance improvements are due to better feature extraction, better data use, or simply larger models. Existing results are difficult to interpret and reproduce because they are based on evaluations of a single factor alone.

Opportunity for Modular Enhancement. There are recent developments in lightweight attention mechanisms such as Coordinate Attention (CoordAtt) [20] and the appearance of environment-specific data augmentation techniques that present an opportunity to tackle the mentioned challenges by complementary and modular improvements. CoordAtt learns directional spatial information into channel attention maps, which might preserve the positional information which is lost by traditional pool operations. Water-specific augmentations can be used to create the visual variability of riverine environments with different water clarity, surface reflections, and camera perspectives, without needing to collect additional data. These two components have not yet been studied together in previous research, creating a strong research need.

Moreover, the release of the YOLO26 architecture, which presented a novel paradigm of detection method without NMS (Cross Stage Partial blocks with Spatial Attention (C2PSA)), is a modern and efficient backbone that has not been tested for floating waste detection. The integration of these modern buildings and the precise attention and augmentation modules is a new and practically relevant research field.

1.4 Research Objectives

The main goal of this research is to design and test a modular, computational efficient detection system for floating waste in riverine setting with good performance in a real world environment. The following specific objectives are used to break down this primary objective:

1. **Framework Development:** To build a detection framework named YOLO26m-CW to incorporate Coordinate Attention (CoordAtt) and Water-Specific Augmentations (WaterAugs) as plug-in modules atop the YOLO26m architecture to enhance the preservation of spatial features and the environmental robustness.
2. **Systematic Component Evaluation:** To conduct a two-stage ablation study that isolates the individual and combined contributions of CoordAtt and WaterAugs, beginning with the lightweight YOLO26s variant to identify effective modules and scaling to YOLO26m to validate the final configuration.
3. **Robustness Validation:** To assess its reliability across five different train-test split configurations (50-50, 60-40, 70-30, 80-20, 90-10) on the FloW-Img dataset, avoiding the possibility of reported performance being merely a result of a single favorable data partition.
4. **Fair Benchmarking:**The first fair apples-to-apples comparison of multiple competing methods on the FloW-Img benchmark, to allow for rigorous comparison to previous work where the split ratios were undisclosed, allowing for comparison with the split ratios used in the work.
5. **Cross-Domain Evaluation:** To quantify the generalization limitations of models trained on non-riverine datasets by training exclusively on external datasets (TUD-GV, UAV-BD, MJU-Waste) and testing on the FloW-Img benchmark, thereby demonstrating the necessity of domain-specific training data.

1.5 Research Questions

This study seeks to answer the following research questions, each of which corresponds to one or more of the stated objectives:

RQ1: Does the integration of Coordinate Attention into the YOLO26 architecture improve detection performance for floating waste in riverine environments, and through which specific metrics (precision, recall, or both) does this improvement manifest?

RQ2: Do environment-targeted water-specific augmentations provide complementary benefits to architectural attention mechanisms, and what is the nature of their combined effect on detection accuracy?

RQ3: How sensitive is YOLO26m-CW's detection performance to variations in the train-test data split ratio, and does the framework maintain consistent performance across diverse split configurations?

RQ4: How does YOLO26m-CW compare against existing state-of-the-art floating waste detection methods when evaluated under matched split ratios and identical benchmark conditions?

RQ5: To what extent can models trained on alternative waste detection datasets generalize to the FloW-Img riverine benchmark, and what does this reveal about the necessity of domain-specific training data?

1.6 Scope of the Study

This research focuses on the problem of detecting floating waste objects in inland riverine environments using image-based single-frame 2D object detection. Given the pressing necessity to monitor macro-plastic debris in riverine systems prior to marine influx [1, 2], setting precise research boundaries is essential for systematic evaluation. The scope of this study is defined across five primary dimensions:

Detection Task. The study concentrates strictly on single-class object detection for floating waste, excluding multi-class categorization or fine-grained instance segmentation. The primary goal is binary localization (waste vs. water background) tailored for real-time edge execution on autonomous surface vehicles (USVs) [4]. All target instances in the FloW-Img dataset are aggregated under a single "floating waste" category, prioritizing rapid spatial localization over the high computational overhead required for multi-class segmentations.

Primary Dataset. The FloW-Img dataset serves as the core benchmark in this investigation. The aim of this dataset is to offer a realistic depiction of the floating waste detection scenario for inland water by providing surface-level USV viewpoints, hydrodynamic noise, and a wide variety of object sizes, as described in detail by Cheng et al. [14].

Architectural Scope. The proposed modular enhancements, namely Coordinate Attention (CoordAtt) [20] and Water-Specific Augmentations (WaterAugs), are evaluated exclusively on the YOLO26 architecture family, specifically focusing on the YOLO26s and YOLO26m variants [21]. Although these lightweight spatial attention and aquatic augmentation modules are inherently model-agnostic and could potentially benefit alternative real-time object detection backbones [13, 12], broader cross-architecture exploration lies beyond the scope of this investigation.

Evaluation Scope. The framework is evaluated using standard object detection metrics (Precision, Recall, F_1-Score, mAP_{50}, $mAP_{50\text{-}95}$, macro mAP, and AUROC) across multiple experimental split configurations (50-50 to 90-10). Performance is assessed offline on

held-out test splits; physical deployment-stage evaluation on USV hardware onboard microprocessors is not conducted.

External Evaluation. To quantify cross-domain transferability and assess domain shift limitations, three auxiliary open-source datasets TUD-GV [19], UAV-BD [8], and MJU-Waste [6] are employed exclusively for zero-shot generalization testing. These external datasets are neither used for hyperparameter tuning nor incorporated into the training pipeline of the primary framework.

1.7 Limitations

Although this study has multiple noteworthy contributions, the following limitations restrict the extent to which findings from the study can be generalized and applied:

1. **Single-Split Evaluation.** Five different train-test split configurations are tested, but K-fold cross-validation is not used. Consequently, the reported performance for each split ratio is dependent on the particular partitioning of the dataset. Different subsets may contain varying levels of object diversity, background complexity, and environmental conditions, which can influence the resulting metrics. Therefore, the reported results may not fully capture the variability that could be observed through repeated or K-fold validation. Future studies should consider cross-validation or repeated random splits to provide more robust estimates of model performance.

2. **Absence of Edge-Latency Benchmarking.** The proposed framework has been designed with potential edge deployment on USVs in mind, and the YOLO26 architecture is lightweight by nature. However, this study does not explicitly benchmark inference latency, memory requirements, or power consumption on representative edge-computing devices such as NVIDIA Jetson or Coral TPU platforms. The computational efficiency of the proposed architecture is therefore inferred from its design characteristics rather than demonstrated through actual hardware deployment. In practical USV applications, processing speed and energy consumption may significantly affect the operational feasibility of the system. Future work should evaluate the framework on representative edge devices under realistic deployment conditions.

3. **Single-Class Detection.** The FloW-Img dataset contains a single object class representing floating waste, and therefore the proposed framework focuses exclusively on detecting the presence of waste rather than identifying its specific type. In real-world waste collection scenarios, distinguishing between materials such as plastic bottles, plastic bags, organic waste, and styrofoam could be important because different waste types may require different collection or handling strategies. The current formulation consequently limits the level of semantic information available to an autonomous USV. Future research could investigate multi-class waste detection using datasets containing sufficiently diverse and accurately annotated waste categories.

4. **Static Image Inference.** The experiments are conducted independently on individual image frames, without exploiting temporal information from continuous

video sequences. As a result, the current framework does not consider the temporal consistency of detections across consecutive frames. Incorporating temporal information could help reduce intermittent false detections, improve localization stability, and maintain consistent predictions when objects are partially occluded or affected by water motion. Techniques such as object tracking, temporal smoothing, or video-based detection could therefore provide additional robustness. Future studies should evaluate the proposed framework using continuous video streams collected from moving USV platforms.

5. **Dataset Dependency.** The external evaluation experiments show that models trained on non-riverine datasets do not generalize effectively to the FloW-Img dataset. This finding highlights the importance of domain-specific training data for floating-waste detection in river environments. However, it also indicates that the performance of the proposed framework may depend strongly on the visual characteristics represented within FloW-Img. River environments can vary substantially in terms of water appearance, lighting, vegetation, weather, waste composition, and surrounding structures. Generalization across geographically and environmentally diverse settings, including tropical and arctic rivers as well as monsoon and dry seasons, therefore remains to be verified.

1.8 Research Contributions

The following are contributions of this research toward floating waste detection using automatic method:

1. **A Modular Detection Framework.**We introduce YOLO26m-CW, a detection framework where the Coordinate Attention and environment-targeted water augmentations are built as plug-in configurable modules on top of YOLO26m. It follows the software engineering principle of separation of concerns one module can be enabled, disabled, or replaced without impacting another module. This design approach allows for the extensibility of research and flexibility in deployment.

2. **Systematic Two-Stage Component Evaluation.** We propose a two-stage evaluation approach where candidate enhancement modules are first evaluated on a smaller variant of the architecture (YOLO26s), and then re-evaluated on the target architecture (YOLO26m). By this method, the individual contributions of each module (CoordAtt for the improvement of recall, WaterAugs for the improvement of precision, and the interaction of the two when combined) are clearly demonstrated.

3. **Multi-Split Robustness Validation.** We evaluate the proposed framework across five train-test split configurations ranging from 50-50 to 90-10, demonstrating consistent $\text{mAP}_{50} > 0.90$ across all configurations. This multi-split analysis, which is absent from all prior works on the FloW-Img benchmark, provides stronger evidence of framework reliability than single-split evaluation.

4. **Fair Cross-Method Comparison.** We train YOLO26m-CW on various split ratios to make the first time fair comparisons with six previous works under their respective (previously unknown) split ratio. This contribution tackles an important

methodological gap of the literature on the comparison between different methods, when evaluation protocols are not comparable.

5. **Cross-Domain Generalization Analysis.** We extend the external evaluation protocol of Cheng et al. [14] by training on three external datasets and test on the FloW-Img benchmark, quantifying the strict performance penalty of non-riverine training data and introducing TUD-GV as a new external evaluation baseline.

1.9 Thesis Organization

The remainder of this thesis is organized as follows:

Chapter 2 (Literature Review) provides a detailed overview of the state-of-the-art in riverine waste pollution, deep learning object detection frameworks (particularly the YOLO family), attention mechanisms such as Coordinate Attention, and aquatic data augmentation strategies, concluding with an analysis of key research gaps.

Chapter 3 (Datasets) describes the primary benchmark dataset (FloW-Img) and the three external evaluation datasets (TUD-GV, UAV-BD, and MJU-Waste) used to assess cross-domain generalization, detailing environmental characteristics, annotation protocols, and image properties.

Chapter 4 (Methodology) presents the proposed YOLO26m-CW architecture, detailing the integration of Coordinate Attention into the SPPF module, the water-specific data augmentation pipeline, model loss functions, training hyperparameter configurations, and evaluation metrics.

Chapter 5 (Results) reports empirical findings across all evaluation dimensions: two-stage component selection, multi-split robustness analysis across five split ratios, cross-method benchmark comparisons, cross-domain evaluations, and qualitative detection visualizations.

Chapter 6 (Discussion) synthesizes the experimental findings, offering in-depth interpretations aligned with the five core research questions, discussing real-world USV deployment implications, and analyzing architectural trade-offs.

Chapter 7 (Conclusion) summarizes the primary contributions of this thesis, evaluates the fulfillment of research objectives, addresses study limitations, and outlines directions for future research.

Chapter-2: Literature Review

The present chapter presents a detailed description on the existing literature related to floating waste detection in aquatic environment. It outlines the developments of methodologies, shows the ongoing environmental and computational issues in this field and critically reviews the data sets and performance benchmarks being used in recent research, ultimately highlighting specific research gaps that the present study addresses.

2.1 Background and Motivation

One of the biggest environmental problems of today is the tremendous amount of plastic waste found in the oceans and river systems of the world. Rivers are the principle pathways for plastic garbage, carrying tons of poorly managed trash from land to sea. A recent study by Lebreton et al. [1] identified that riverine plastic inputs to the oceans are a major source of ocean pollution and that it is important to target the river source to help reduce this pollution. This was also supported by Meijer et al. [2] who showed that over 1,000 rivers contribute 80% of plastic emissions from rivers to the ocean. The amount of debris makes it crucial to implement scalable and automated monitoring and mitigation strategies. Citing the need for a more comprehensive approach to understanding plastic waste fluxes in rivers, van Emmerik and Schwarz [3] outlined the challenges of manual monitoring in rivers, especially in terms of the dynamic nature of waste flows and the logistical difficulties of conducting such monitoring surveys. As a result, the scientific community has been focusing on the development of autonomous surface vehicles (USVs) and unmanned aerial vehicles (UAVs) that use computer vision systems to identify, count and remove floating waste.

2.2 Prior Studies in Aquatic Waste Detection

In the last decade, the field of automated floating waste detection has seen a great evolution, from early exploratory research to advanced computer vision systems harnessing deep learning. In this section, the evolution of this research is discussed. Rather, it first discusses the adaptation of general object detection models to aquatic settings and then looks at the unique datasets that have been created for training such models. Finally, it delves into the success of YOLO frameworks, which are the leading approaches for real-time inference on edge devices in the recent literature.

2.2.1 Standard Object Detection Architectures

With the rise of deep learning, specifically Convolutional Neural Networks (CNNs), visual recognition became a revolution, and researchers started using CNN-based methods for object detection on underwater image data. Two-stage and one-stage detectors were first investigated for marine and riverine litter identification, such as Faster R-CNN and Single Shot MultiBox Detector (SSD) respectively.

Fulton et al. [4] performed groundbreaking research on the use of robotic platforms for marine litter identification, including the evaluation of Faster R-CNN and SSD, along with YOLO variants. Their study showed that deep learning detectors could be adapted to water images to produce a substantial improvement in the accuracy of the feature extraction compared to traditional handcrafted methods. Their results did show, however, that there was significant drop in performance in situations with poor visibility, varying light conditions, and motion blur common to AUVs and ASVs.

Cheng et al. [5] set up the groundwork for inland water detection by introducing the first standardised evaluation framework for conventional detectors, namely the FloW-Img benchmark. Their extensive analysis revealed that the mean Average Precision (mAP) obtained by Cascade R-CNN was 0.434, while in the case of Faster R-CNN, it was just 0.184 with the use of train-test split of 60:40. Two-stage detectors such as Cascade R-CNN achieved relatively high accuracy but were not practical for real-time use on edge devices common to autonomous robotic systems due to their high computational cost. This computational limitation led to the development of more efficient, one stage detectors. Even in stream environments, Lee et al. [5] have investigated floating plastic detection using deep learning, and found that environment-related noise, like water surface reflection, dynamic flow pattern and sun glare, is still a big and critical challenge for the conventional detection architectures which are not tuned for specific environments.

2.2.2 Specialized Aquatic Datasets

The development of robust deep learning models is inextricably linked to the availability of large-scale, accurately annotated datasets. In the context of aquatic waste detection, several specialized datasets have been introduced to capture the unique visual and environmental challenges of the domain.

The FloW-Img dataset, introduced by Cheng et al. [14], stands as a cornerstone benchmark for inland floating waste detection. Comprising 2,000 images captured by USVs with 5,272 annotated instances, it deliberately incorporates hydrodynamic challenges such as specular reflections, wave disturbances, and varying illumination. Crucially, the FloW-Img study highlighted the issue of "resolution-semantic imbalance," a phenomenon caused by the high density of small targets in surface-level imagery, where distant waste objects occupy only a minute fraction of the total pixel area, leading to severe feature loss during convolutional downsampling.

To address different environmental contexts and the sheer diversity of riverine landscapes, Jia et al. [19] developed the TUD-GV dataset, containing 1,500 images of urban canal debris. This dataset captures the unique background clutter, artificial lighting, and structured

banks characteristic of inner-city waterways. In a complementary effort addressing the lack of labeled data diversity, Jia et al. [7] introduced a semi-supervised learning-based framework utilizing open datasets to quantify litter fluxes in river systems, demonstrating that algorithmic approaches can partially mitigate data scarcity.

Shifting perspectives, Wang et al. [8] introduced the UAV-BD dataset, comprising 25,407 low-altitude aerial images with 34,791 bottle instances. This dataset focuses heavily on the specific challenges of top-down perspectives and the extreme scale variations encountered by Unmanned Aerial Vehicles. Furthermore, the MJU-Waste dataset [6] provides a multi-level approach to waste object segmentation using 4,950 indoor RGBD images, allowing for controlled-environment evaluations and deep depth-based analysis. The substantial distributional, visual, and perspective differences across these datasets ranging from urban canals to aerial views highlight a critical issue: models trained on one specific domain frequently fail to generalize to another without substantial drop in precision, necessitating domain-adaptive training regimens.

2.2.3 YOLO-Based Approaches for Waste Detection

Given the stringent real-time processing requirements of autonomous surface vehicles, which operate with constrained power and compute resources, the YOLO (You Only Look Once) family of one-stage detectors has become the dominant paradigm for floating waste detection.

Jiang et al. [16] proposed APM-YOLOv7, which introduced a multi-scale adaptive feature fusion mechanism specifically tailored for small-target detection in water-floating garbage scenarios. By dynamically weighting features across different scales, the model effectively mitigated the loss of semantic information for distant, small debris, a common failing of standard architectures. Wang and Zhao [11] further advanced the state-of-the-art by introducing YOLOv8-MSS, which incorporated an attention-based noise suppression module to explicitly handle water surface reflections, achieving an mAP_{50} of 0.879 on the FloW-Img dataset.

In parallel, Di et al. [13] presented ES-YOLOv8, an enhanced model utilizing specific structural modifications to the backbone and neck networks to reach 0.873 mAP_{50} for accurate detection of solid floating waste. Yang et al. [12] proposed USV-YOLO, an algorithm designed for environmentally friendly unmanned vessels, which integrated dynamic convolutions to adaptively adjust receptive fields based on object scale, achieving a high precision of 0.901. Hou and Yu [18] introduced PBM-YOLO, integrating advanced spatial attention mechanisms to achieve 0.907 mAP_{50} on a highly favorable 90-10 data split.

While YOLO-based methods have demonstrated significant promise, a recurring methodological limitation across these studies is the reliance on single, often undisclosed train-test splits (such as fixed 80-20 or 90-10 splits without cross-validation). This practice precludes fair cross-method comparison on the same benchmark, masks potential model overfitting to specific geographical features present in the training set, and raises legitimate questions regarding the robustness and generalizability of the reported performance metrics.

2.3 Architectural Enhancements for Small Object Detection

Detecting floating waste often equates to small object detection, as debris usually occupies less than 1% of the image area due to camera perspective and distance. To address the inherent difficulties of small object detection, researchers have proposed various architectural enhancements.

2.3.1 Multi-Scale Feature Fusion Methods

Standard CNN backbones progressively downsample feature maps, enriching semantic abstraction at the cost of spatial resolution. This spatial information loss is detrimental to small targets, which may disappear entirely in deeper network layers. Multi-scale feature fusion aims to address this resolution-semantic imbalance by combining high-resolution, low-semantic feature maps from shallow layers with low-resolution, high-semantic feature maps from deep layers.

In the context of aquatic waste, Jiang et al. [16] extensively utilized this concept by proposing an adaptive weighted fusion module in APM-YOLOv7. This module dynamically allocates emphasis across different resolution levels based on target scale, ensuring that the faint activations of distant plastic bottles are preserved through the network hierarchy. Tang et al. [15] addressed spatial information loss in UTD-YOLO by employing customized feature pyramids tailored for underwater and surface trash detection, enhancing the network's sensitivity to small-scale variations. Furthermore, Singh and Kumar [9] explored a hybrid approach by combining YOLOv11 for detection with DeepLabV3 for precise semantic segmentation of marine debris. While highly effective for accurate localization and boundary delineation, such multi-model or overly complex multi-scale approaches substantially increase the parameter count and computational requirements, thereby limiting their suitability for low-power edge deployment on autonomous vessels.

2.3.2 Attention Mechanisms in Visual Recognition

Attention mechanisms have emerged as a powerful tool to selectively emphasize informative features while suppressing irrelevant background noise a crucial capability for mitigating environmental visual interference like wave reflections.

In the domain of waste detection, Badams et al. [17] integrated channel attention alongside atrous convolutions into A2ANet, demonstrating that selective channel weighting effectively suppresses dynamic water surface noise and improves real-time marine debris detection capabilities. Di et al. [13] employed a localized attention-based noise suppression module in their ES-YOLOv8 framework to explicitly teach the network to ignore the high-frequency noise generated by water ripples. Zhu et al. [10] addressed similar issues by introducing a lightweight and accurate detector tailored for small floating objects in complex water surface environments, capturing longer-range dependencies across the water surface to differentiate between solid waste and transient wave patterns. However, a significant

limitation of many channel-wise attention implementations is their failure to explicitly model long-range spatial position information, which is critical for detecting partially submerged targets where the geometric structure of the exposed debris is the sole diagnostic feature.

2.3.3 Coordinate Attention and Spatial Modeling

To overcome the spatial information loss inherent in standard global pooling operations used by standard channel attention mechanisms, Hou et al. [20] proposed Coordinate Attention (CoordAtt) for efficient mobile network design. CoordAtt innovatively decomposes 2D global pooling into two distinct 1D encoding operations along the horizontal and vertical axes.

By encoding spatial information in two spatial directions separately, CoordAtt preserves precise directional spatial cues. These directional features are then concatenated and transformed to generate attention maps that are sensitive to both position and channel relationships. This capability is exceptionally critical in aquatic environments. For instance, when a plastic bottle is half-submerged, recognizing its shape requires the network to focus on specific vertical and horizontal coordinate features simultaneously rather than just aggregating global channel statistics.

Hou and Yu [18] successfully validated the application of CoordAtt for floating waste detection through their PBM-YOLO model. They demonstrated that embedding this spatial-aware attention module directly into the YOLO architecture yielded a performance-balanced model that achieved superior precision on the FloW-Img dataset. Nevertheless, the isolated contribution of the CoordAtt module has not been systematically ablated across multiple diverse datasets, nor has its potential synergistic interaction with environment-specific data augmentation strategies been thoroughly explored in the current literature.

2.4 Summary of Challenges in Aquatic Waste Detection

The comprehensive review of the literature reveals several persistent and intertwined challenges that continue to hinder the robust deployment of floating waste detection systems.

2.4.1 Environmental Visual Interference

The aquatic environment is highly dynamic and visually complex. Factors such as severe sun glare, specular reflections from the water surface, erratic wave-induced noise, and varying water turbidity significantly degrade detection performance [10, 11]. These interferences often mimic the visual characteristics of floating plastic or saturate the camera sensor entirely, leading to high false-positive rates and missed detections.

2.4.2 Scale Variation and Small Targets

Floating debris ranges drastically in size, from large industrial containers to small plastic bottles. Furthermore, USV-mounted cameras often observe debris from a considerable distance along the riverine horizon, rendering targets extremely small within the image frame. The progressive loss of spatial information during the downsampling process of deep CNNs makes accurately localizing and classifying these small, distant objects exceedingly difficult without specialized architectural modifications [14, 15].

2.4.3 Real-Time Processing on Edge Devices

Autonomous surface vehicles and aerial drones are strictly constrained by payload weight, power availability, and computational capacity. There exists an inherent tension between implementing highly complex, theoretically accurate models and the necessity for high frame-rate, real-time inference on low-power edge devices [14, 12]. Models must be lightweight yet robust enough to handle the environmental complexities discussed above.

2.4.4 Inconsistent Evaluation Protocols

A significant challenge in assessing state-of-the-art progress is the lack of standardized evaluation protocols. Many studies evaluate models using varying, often undisclosed train-test split ratios on the same benchmark dataset [18, 17]. This practice completely undermines fair, reproducible comparisons and obscures whether performance gains are due to genuine architectural improvements or simply overfitting to a specific data subset.

2.4.5 Confounding of Enhancement Strategies

Recent literature frequently introduces multiple modifications simultaneously combining new attention modules, altered loss functions, and bespoke data augmentations without providing rigorous ablation studies [13, 12]. This confounding limits the scientific interpretability of the results, making it impossible to determine which specific modification contributed to the observed performance improvements and hindering the development of optimal detection pipelines.

2.5 Scope and Objectives of the Present Study

Addressing the critical gaps identified in the literature, this study focuses on advancing floating waste detection in inland riverine environments. Utilizing the comprehensive FloW-Img dataset [14] as the primary benchmark, the study strictly defines its scope to isolate, evaluate, and combine two specific enhancement paradigms on the YOLO architecture.

First, the study investigates the integration of Coordinate Attention [20] to provide spatial-aware feature encoding, addressing the challenge of localizing partially submerged and small objects. Second, it proposes and evaluates Water-Specific Augmentations, specifically engineered to simulate targeted environmental interferences such as glare, reflection, and turbidity, thereby enhancing domain robustness.

To rectify the methodological flaws prevalent in prior works, a systematic two-stage evaluation framework is employed. This framework rigorously isolates the individual contribution of each module via extensive ablation before validating their synergistic combination. Furthermore, to guarantee robustness and prevent overfitting claims, performance is comprehensively assessed across five distinct train-test splits (ranging from 50-50 to 90-10). Finally, to quantify generalizability and address the issue of domain shift, cross-domain evaluations are conducted on the TUD-GV [19], UAV-BD [8], and MJU-Waste [6] datasets. By adhering to this rigorous, isolated, and multi-domain evaluation protocol, this study aims to provide definitive, reproducible advancements in the field of automated aquatic waste detection.

In addition, the study emphasizes a fair and consistent experimental configuration across all evaluated models to ensure that observed performance differences can be attributed to the proposed enhancements rather than variations in training conditions. The analysis considers both detection accuracy and model robustness under challenging aquatic visual conditions. Particular attention is given to the detection of small, distant, and partially occluded waste objects, which represent important practical challenges in real-world river environments. The resulting findings are intended to establish a reliable experimental foundation for future research on efficient and generalizable floating waste detection systems.

Chapter-3: Dataset

This chapter describes the datasets used for testing the proposed floating waste detection methods empirically. This study is based on an extremely carefully chosen primary benchmark dataset, bolstered by three distinct external datasets, which are structurally different, to provide a robust, rigorous, and highly generalizable assessment. The datasets span a broad range of acquisition platforms, environmental parameters, and semantic properties. In-depth analysis of the statistical properties, collection methods and natural limitations of each dataset are presented. Moreover, it details the preprocessing approaches, partitioning schemes for the data sets, and the basic reasoning behind the application of the cross domain evaluation framework to ensure the models proposed are applicable in real-life scenarios.

3.1 Overview of Employed Datasets

To develop generalized models for aquatic environments, training and testing must be done across a variety of data sources. The basic characteristics of the four different data sets used in this study are summarized in Table 3.1. The FloW-Img dataset is closest to the targeted deployment scenario (Unmanned Surface Vehicles in inland waters) and hence is used for primary training, validation and testing. On the other hand, the TUD-GV, UAV-BD, and MJU-Waste datasets are only available for external (cross-domain) assessment. These datasets were carefully selected to cover two key reasons: they are data types that are fundamentally different, giving the model a severe stress test to ensure it can transfer its knowledge beyond the main training distribution.

Table 3.1: Summary of datasets used in this study, including sample size, acquisition environment, platform, and experimental role.

Dataset	**Samples**	**Environment**	**Platform**	**Role**
FloW-Img [14]	2,000	Inland Water	USV (Surface)	Primary Benchmark
TUD-GV [19]	1,500	Urban Canal	Handheld/Ground	External Evaluation
UAV-BD [8]	25,407	Aerial	UAV (Drone)	External Evaluation
MJU-Waste [6]	4,950	Indoor	RGBD Camera	External Evaluation

3.2 The Primary Benchmark: FloW-Img Dataset

The FloW-Img (Floating Waste Image) dataset is the foundation for the empirical evaluations in this thesis and was originally proposed by Cheng et al. [14]. It was designed specifically to fill the urgent need of good, annotated data for inland surface water environments.

3.2.1 Acquisition and Environmental Context

Data were acquired using an Unmanned Surface Vehicle (USV) while the vehicle was traversing different riverine inland water networks, FloW-Img. The camera was positioned near the water surface to get a low angle-like perspective that is exactly how the autonomous river cleaning robots operate. The images were taken during various weather conditions, including sunny and cloudy days, to show the variability in lighting. The dataset contains a total of 2,000 RGB images, provided in two distinct, high-definition spatial resolutions to reflect different sensor configurations: 1,241 images are captured at 1280×720 pixels, while the remaining 759 images are captured at 1280×640 pixels.

3.2.2 Statistical Characteristics and Annotation

The dataset includes high quality, instance-level bounding box annotations. There are 5272 floating waste objects in the 2000 images, an average of 2.64 objects per image. The annotated objects are mostly common macro-plastic items including polyethylene terephthalate (PET) bottles, plastic bags and pieces of styrofoam, which make up most of the plastic litter found in rivers. A significant statistical aspect of FloW-Img is the very uneven distribution of the bounding boxes. With the USV's forward-looking view, objects in the horizon are actually very small in size in the pixel plane. Cheng et al. [14] explicitly called this a "resolution-semantic imbalance," as the network has difficulty to learn high-level semantic features from targets with low spatial resolution.

3.2.3 Inherent Environmental Challenges

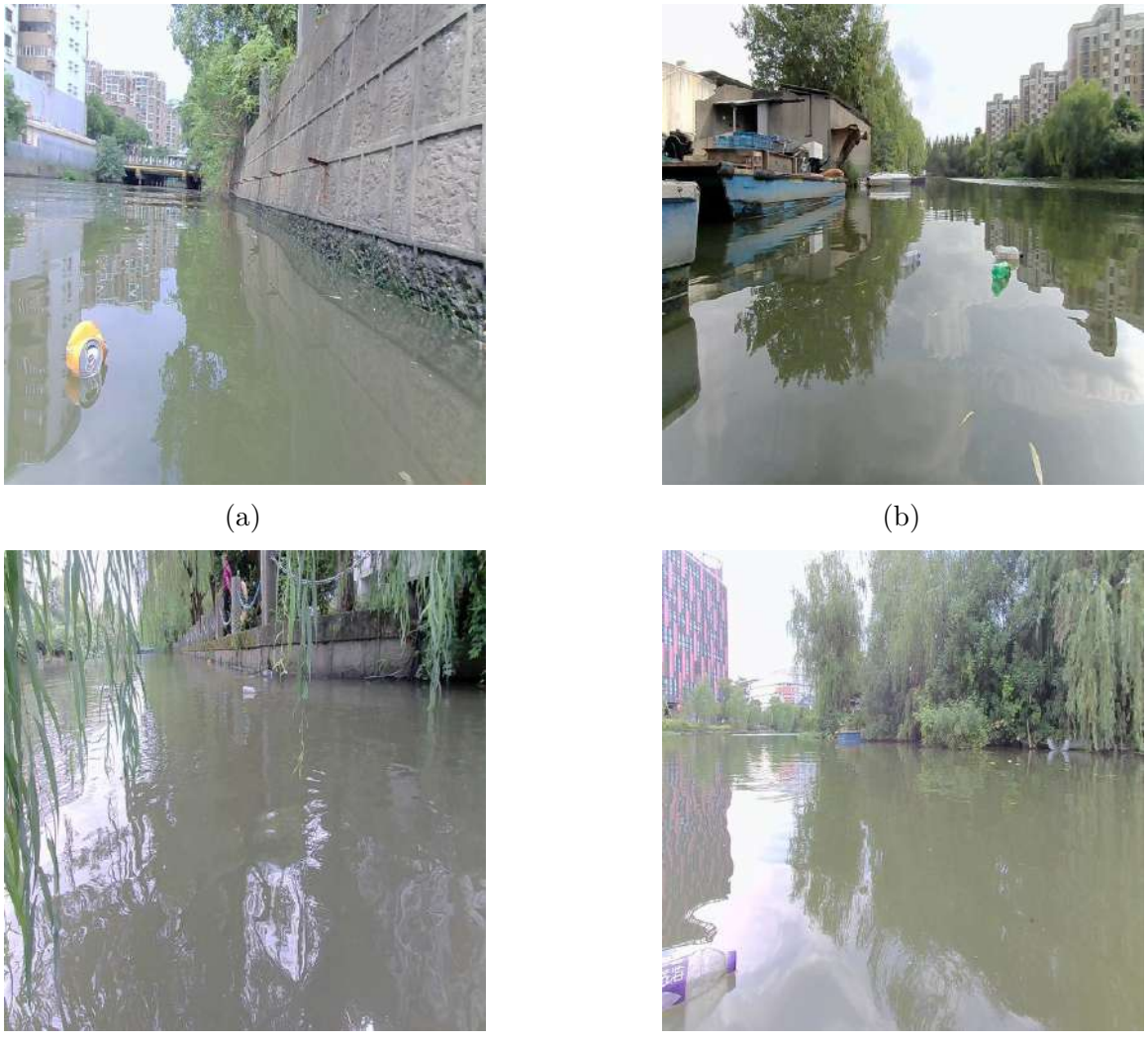
The very strong optical interferences in the underwater environment are not suppressed but deliberately preserved in FloW-Img. These interferences are the main challenges to a reliable detection system:

- **Specular Reflection and Sun Glare:** The dynamics of the water surface combined with direct sunlight exposure result in regions of saturated pixels, often resulting in the complete loss of image details of floating objects or false edge features.
- **Wave Disturbances and Wake:** The wake is created by the motion of the USV and ripples are created by natural currents. These "dynamic textures" contain a lot

of background noise at high frequencies that greatly complicates regular convolutional filters.

- **Turbidity and Submergence:**The optical properties (murkiness) of the water changes with suspended sediment variation. In addition, waste items such as waterlogged bottles are not always completely visible, and the geometry of the target is always incomplete and variable.

Representative sample images from the FloW-Img dataset are displayed in Fig. 3.1. Fig. 3.1a–d visually demonstrate the characteristic environmental complexity encountered by USVs during inland river monitoring: Fig. 3.1a and Fig. 3.1b illustrate severe sun glare and specular reflections that saturate target pixels; Fig. 3.1c highlights surface wave turbulence and dynamic wake textures; while Fig. 3.1d shows typical small, semi-submerged plastic items (such as PET bottles and styrofoam) amidst murky water conditions.



(a) (b)

(c) (d)

Figure 3.1: Representative sample images from the FloW dataset used for floating waste detection.

3.3 External Datasets for Cross-Domain Generalization

A model that achieves high accuracy on a single dataset may suffer catastrophic performance degradation when deployed in a new environment a phenomenon known as domain shift. To rigorously test the generalizability of the architectural enhancements proposed in this study, models trained exclusively on FloW-Img are evaluated on three external datasets without any fine-tuning.

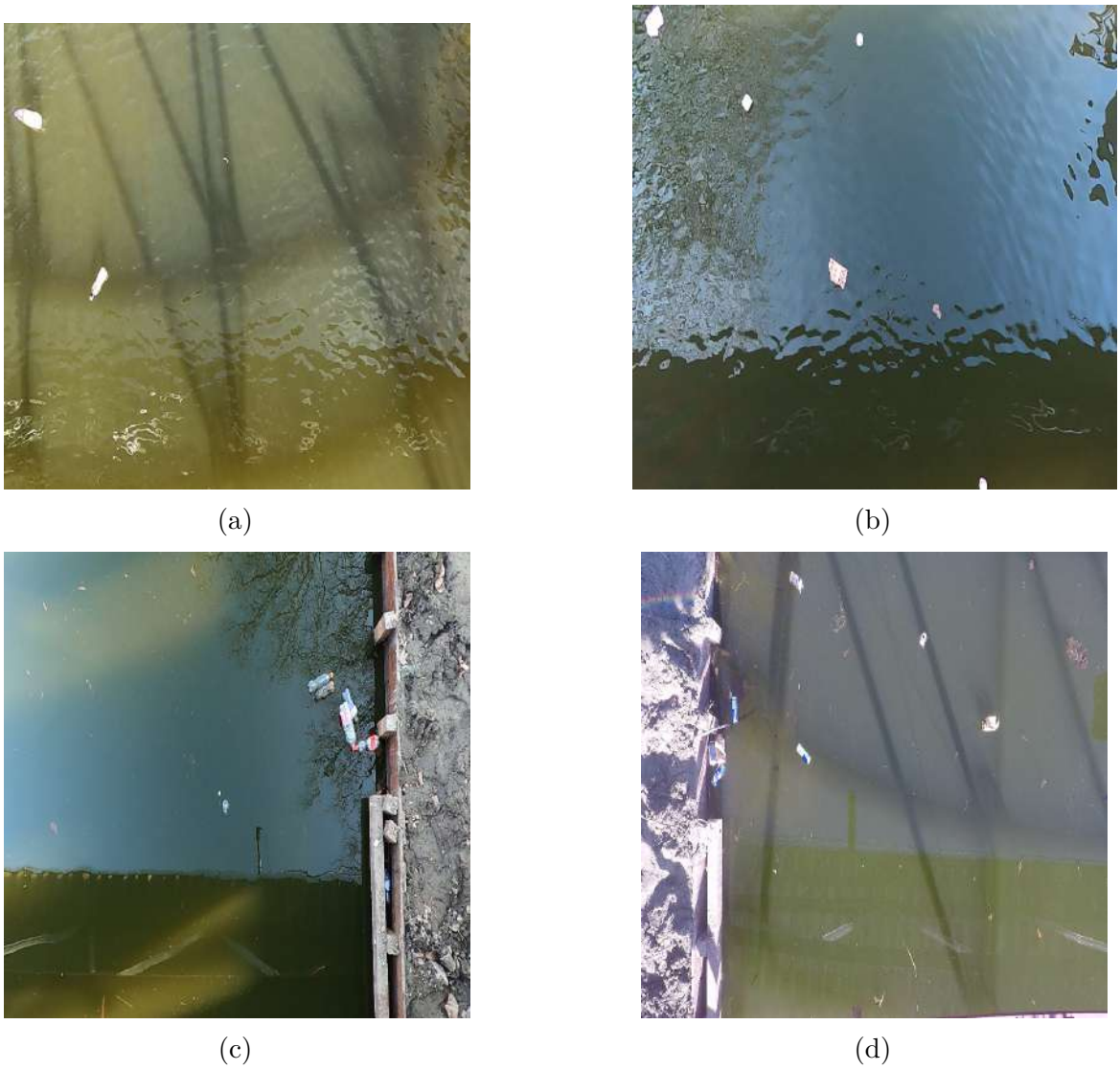
3.3.1 TUD-GV: Urban Canal Context

The TU Delft-Green Village (TUD-GV) dataset consists of 1,500 photographs of floating macro-litter from a controlled urban water body, developed by Jia et al. [19]. The dataset, FloW-Img, demonstrates the natural conditions of a river, whereas the TUD-GV dataset is capturing narrow canals in the living laboratory of Green Village at TU Delft. Images were taken with cameras on the ground and on the bridge, resulting in different angles and scales to view the objects. High turbidity, high reflections and limited water flow are a result of the conditions of the data set, which stems from concrete canal boundaries. Numerous urban background elements like bridges, bicycles, fences, walkways and vegetation provide a high level of visual clutter which may mislead detection models. Also, items of litter may have different appearance properties as a result of the long-term effects of water in stagnation in the urban situation. There are illumination variations, shadow formation from various structure features and object occlusion.There is illumination change and shadow formation from various structure features and partial occlusion of objects. Due to the aforementioned reasons, the color from the open-water scenes of FloW-Img is quite different from those seen in this scene. Thus, the TUD-GV offers a stringent evaluation framework on the cross-domain generalization ability of the model, specifically the ability to suppress irrelevant urban context to detect floating waste objects with complex in-canal context.

The data set contains a variety of types of drifting garbage which enable the models to learn various shapes, sizes, colour and surface features of items. The item(s) can be shown alone or as a group of items, thus forming varying degrees of scene complexity for scene detection. Localization becomes difficult in the presence of small litter and partly submerged litter. Water reflections can also offer false patterns over the water surface, which will lead to false detections. In addition, background-object interactions want to be an issue as canal facilities are very close to the litter collection area, so it is hard to distinguish the litter from surrounding facilities. There are significant variations in object scale and perspective between images due to differences in elevation and direction of view of the camera. The dataset is thus used to test the object-level recognition and spatial localization capabilities of the model in a challenging setting. The controlled collection environment allows for a systematic exploration of model robustness in an urban aquatic setting as well. Among other things, TUD-GV can show us if a model trained under natural water conditions obtains knowledge more from the image itself than from the actual water. Accurate results with this data set demonstrate better water appearance independence, background structure independence and robustness to changing imaging conditions. TUD-GV, therefore, becomes an important complementary indicator of the

implementation of floating-waste detection systems in urban water bodies.

Representative sample images from the TUD-GV dataset are presented in Fig. 3.2. As shown in Subfig. 3.2a–d, the dataset captures macro-litter floating within narrow urban canals flanked by concrete embankments. Fig. 3.2a and Fig.3.2b depict intense background clutter including urban infrastructure, footbridges, and bank-side vegetation, whereas Fig. 3.2c and Fig.3.2d illustrate shadow formations and high specular water reflections in stagnant urban water bodies.



(a) (b)

(c) (d)

Figure 3.2: Representative sample images from the TUD-GV dataset used for floating waste detection.

3.3.2 UAV-BD: Aerial Top-Down Perspective

Wang et al. [8] introduce the Unmanned Aerial Vehicle Bottle Dataset (UAV-BD) which offers a unique aerial view for bottle detection. It provides 25,407 low altitude aerial images and comprises 34,791 instances of bottles, which is significantly more than there are in FloW-Img (2000 images). Images are taken from a UAV unmanned aerial view (UDUAV) looking from top down, creating some horizontal water-surface glare but offers

dramatic variations in scale, orientation, and viewing angle of objects.

There are also a variety of complex and varied background areas, such as water, muddy shorelines, vegetation, soil and other land use characterized areas in UAV-BD. These scenes pose problems of small object detection, rotational variation, background confusion, illumination variations and partial occlusion. The data set, therefore, constitutes a solid stress-test for the robustness of the model with respect to variance in viewpoint and environmental context.

Additionally, UAV-BD features a significant gap between itself and FloW-Img which makes it very meaningful for assessing CG in cross domain settings. UAV-BD is an aerial observation geometry with added scale and background diversity, provided by its near ground level USV imagery - represented by an aerial Unmanned Sighting Vehicle, or USV. The use, therefore, serves as a way to assess the ability of the bottle model-generated feature representations to capture the fundamental bottle semantics or surface-level adaptations to the original USV camera viewpoint.

Additionally, UAV-BD collects bottles at different altitudes and imaging scenarios, causing a significant variation in the size and visibility of the bottles in the images. In some instances, the penultimate bottle fills only a few pixels, which is more challenging for the localization and classification to be accurate. The main glasses and surrounding areas are more visible to the model when seen from above and require different representations, as compared to ones when seen from the ground. Sunlight, shadows, water reflections, and background texture add to the visual complexity of the dataset, making it difficult to identify distinct themes.The images in the dataset also vary in terms of sunlight, shadows, water reflections, and background textures, adding to the visual complexity. Large number of annotated examples allow for more accurate comparison of the detection performance against a variety of scene scenarios in a statistical sense. This makes UAV-BD a difficult complementary test to evaluate robustness with scale, viewpoint and environmental variances. In general, the data allows for a good estimation of how well can a detection model be useful when applied to near-surface images and then applied to aerial images.

However, the acquisition process and their unique perspective of the object also brings with it large-scale differences in object appearance compared to conventional imagery acquired from the surface. Variations in size can be apparent from different bottle instances, depending on the UAV altitude but also on the position of the bottle in the photo. In this variation, it is more difficult to achieve reliable detection performance as a function of resolution. Furthermore, complicated background structures may cause a reduction in the ability to distinguish objects in the bottles from the surrounding areas. Insufficient visual information in small and partially occluded bottles could make it difficult to classify and detect the location of these bottles. Shadows and reflections can cause more distortion of the apparent boundaries of bottle instances. All these attributes make the UAV-BD well qualified for the testing of multi-scale detection and robust feature-learning of UAVs. Additionally, it is useful because of the variety of possible environmental conditions in the dataset, which lowers the chances that the model would confuse background characteristics specific to the data with a signal. A successful detection in the context of UAV-BD, thus, reflects higher endurances toward changes in viewpoint, size, illumination and scene composition. Overall, the dataset represents a demanding assessment ground for statistically, domain-independently benchmarking the cross domain performance of machines able to detect bottles.

Representative top-down aerial sample images from the UAV-BD benchmark are shown in Fig. 3.3. As depicted in Subfig. 3.3a–d, the drone perspective introduces substantial variations in object scale and orientation. Fig. 3.3a and Fig.3.3b present tiny plastic bottle targets occupying minimal pixel area against vast water surfaces, while Fig. 3.3c and Fig.3.3d demonstrate shoreline background confusion where plastic items blend visually with muddy banks and soil structures.

(a) (b)

(c) (d)

Figure 3.3: Representative sample images from the UAV-BD dataset used for floating waste detection.

3.3.3 MJU-Waste: Controlled Indoor Environment

The MJU-Waste dataset was created by Wang et al. [6] to be used as a benchmark for waste detection in controlled indoor environments. A dataset with 4950 RGB-D (Red, Green, Blue and Depth) images and 5026 labelled waste-object instances. Images were obtained from Microsoft Kinect sensors with both traditional RGB data and depth data. Though this study does not use the depth modality, the use of the RGB component is a useful way to evaluate the generalisation of models learnt under different visual environments.

MJU-Waste does not reflect the conditions of the outdoor waste scene captured from

rivers, canals, or water bodies, but rather the conditions for indoor waste scenes where the background structure and the appearance of objects are relatively stable, with relatively stable illumination. There are multiple common waste types and object configurations in the dataset, which can present differences in object scale, shape, orientation, and spatial layout to the models. The controlled acquisition environment also reduces the impact of water reflections, ripples on the surface, the glare of the sun, the movement of floating objects, and complex water backgrounds.

Representative sample images from the MJU-Waste dataset are illustrated in Fig. 3.4. Subfig. 3.4a–d showcase common waste items captured under uniform indoor lighting against clean, non-aquatic backdrops. Contrasting Fig. 3.4 with the outdoor aquatic scenes in Fig. 3.1 Fig.3.3 highlights the complete absence of water reflections, wave disturbances, and hydrodynamic turbidity, making MJU-Waste a valuable control benchmark to quantify performance degradation attributable purely to aquatic environmental noise.

Figure 3.4: Representative sample images from the MJU-Waste dataset used for floating waste detection.

3.4 Data Preprocessing and Partitioning Strategies

To ensure the integrity and reproducibility of the experimental results, standardized preprocessing and partitioning protocols were rigorously enforced.

3.4.1 Standardization and Augmentation Pipeline

Prior to ingestion into the YOLO detection network, all images across all datasets were uniformly resized to 640×640 pixels to satisfy the standardized input tensor dimensions required by the network architecture, while preserving the aspect ratio via letterboxing (padding with gray pixels) to prevent geometric distortion. Standard mosaic augmentation, HSV (Hue, Saturation, Value) spatial perturbation, and random horizontal flipping were utilized during the training phase to artificially expand the dataset's diversity and prevent the model from memorizing specific pixel configurations.

3.4.2 Rigorous Data Partitioning

The literature survey [18, 17] had identified a key methodological issue with the literature, namely the use of a single, unknown train-test split that can give misleading performance figures. In order to overcome this, the main FloW-Img dataset was put through a rigorous assessment protocol, using several split ratios. This dataset was randomly divided into training and testing sets for five different splits: 50-50, 60-40, 70-30, 80-20, and 90-10. This multi-split approach allows the reported accuracy improvements made by the proposed architectural changes to be statistically significant, consistent and not just a chance data split.

3.5 Cross-Domain Challenge and Selection Rationale

Four sets of data have been chosen in this study, and they provide evidence for the assessment of the proposed framework. FloW-Img was chosen as a main test bed because it offers images acquired from the Unmanned Surface Vehicle (USV) level, and has a high amount of environmental noise and variability. They have very similar optical properties that would be present during field use of autonomous river cleaning robots. The data also includes changes in the appearance of the water, size of the object, background complexity, illumination, and distribution of wastes which can be used to evaluate the detection robustness in difficult aquatics situations. The main evaluation dataset used, FloW-Img, also allows for a more realistic assessment of the ability of the proposed model to detect floating waste from a moving platform on a water surface. Moreover, the other data sets allow for cross-domain test and cross-domain training to investigate the generalizability of the learned representations across different visual domains. Therefore, these four data sets will give a complete picture of accuracy, robustness, generalization and usefulness. The use of FloW-Img alone is however not enough to establish the robustness of an algorithm. However, deep learning models, including models with many parameters, such as highly

parameterized CNN models, tend to learn spurious, dataset-specific biases, like correlating the presence of a plastic bottle with the type of water ripple shown. By forcing the model, which has only ever seen the riverine environment of FloW-Img, to predict bounding boxes on the urban canals of TUD-GV [19], the top-down aerial views of UAV-BD [8], and the sterile indoor lighting of MJU-Waste [6], this study establishes a stringent, multi-faceted cross-domain evaluation framework. This framework effectively isolates genuine algorithmic improvements in semantic object recognition from mere dataset memorization, ensuring that the findings presented in this study represent a definitive advancement in automated floating waste detection technology.

Moreover the chosen datasets present significant variations in terms of viewpoint, background, illumination, scale of the object, image quality and environment. These differences make it difficult and interesting to account for in experiments for testing the resilience of the proposed detection mechanics. Cross Domain testing is used to see how well the learned features are transferred to a domain in which the visual style of the test domain may be different than the visual style of the training domain, whereas traditional evaluation is done within a single dataset. These domain shifts can be the result of various camera set-up, acquisition platforms, water surface conditions, landscape conditions, weather, and light conditions. These factors can greatly impact the appearance, contrast, visibility and boundary characteristics of objects in aquatic settings, leading to challenges when it comes to detecting small or partially submerged waste. Furthermore, different hues, reflections and shades, and background features will create more visual confusion of waste objects and their context. These selected datasets are thus suitable for studying how well the model can adapt to new visual conditions that it has not seen before, without the need for learning or fine-tuning. This evaluation also aims to see if the architectural and augmentation methods learn generalizable patterns (representation) or specific patterns (data). Using it throughout of different domains would help verify the adaptability of the model in changing environmental and imaging conditions. However, if performance drops during certain domain changes, it may reveal possible limitations and avenues that can be pursued for future enhancements. Based on this, cross-domain testing becomes an indispensable part to examine the feasibility, strength, and generalization potential of the proposed approach on aquatic waste detection.

Chapter-4: Methodology

This chapter presents the research methodology developed in this study for autonomous riverine floating waste detection using Unmanned Surface Vehicles (USVs). It details the overall proposed framework, the underlying model architecture, the dedicated mathematical formulation of the YOLO object detector and composite loss functions, the modular enhancement components Coordinate Attention (CoordAtt) and Water-Specific Augmentations (WaterAugs), the training protocol, and the evaluation framework.

4.1 Proposed Framework

The proposed detection framework proceeds through three distinct operational and experimental stages, as illustrated in Fig. 4.1. In the initial data preprocessing stage (Fig. 4.1a), all input imagery is standardized to a uniform resolution of 640×640 pixels. This spatial normalization ensures consistent multi-scale feature extraction across images captured with different camera sensors and spatial aspect ratios (e.g., 1280×720 and 1280×640 in the primary benchmark). This preprocessing step is applied uniformly to all datasets used in the study, including both the primary FloW-Img benchmark [14] and the three external evaluation datasets: TUD-GV [19], UAV-BD [8], and MJU-Waste [6].

The experimental phases then proceed sequentially through three evaluation dimensions (Fig. 4.1b):

1. **Component Evaluation:** Using the FloW-Img 60-40 split, YOLO26s and YOLO26m variants are systematically evaluated with and without the two enhancement modules, namely Coordinate Attention (CoordAtt) [20] and Water-Specific Augmentations (WaterAugs), to identify the optimal combination. This two-stage process first screens candidate modules on the smaller YOLO26s architecture before validating the selected combination on the larger YOLO26m variant.

2. **Split Sensitivity Analysis:** The finalized YOLO26m-CW configuration is trained and evaluated across five distinct FloW-Img partitions ranging from 50-50 to 90-10 (50-50, 60-40, 70-30, 80-20, and 90-10), assessing the framework's robustness to variations in training data availability and ensuring that reported performance is not an artifact of a single favorable data split.

3. **External Evaluation:** Models are trained exclusively on each of the three external datasets (TUD-GV, UAV-BD, and MJU-Waste) and tested on the FloW-Img test

set, quantifying the strict performance penalty of non-riverine training data and demonstrating the necessity of domain-specific datasets.

The outputs from all three experimental phases are consolidated (Fig. 4.1c) into a unified evaluation using multiple complementary metrics and a benchmark comparison with both the original FloW-Img baselines [14] and six prior works from the literature [13, 16, 10, 18, 12, 11, 17].

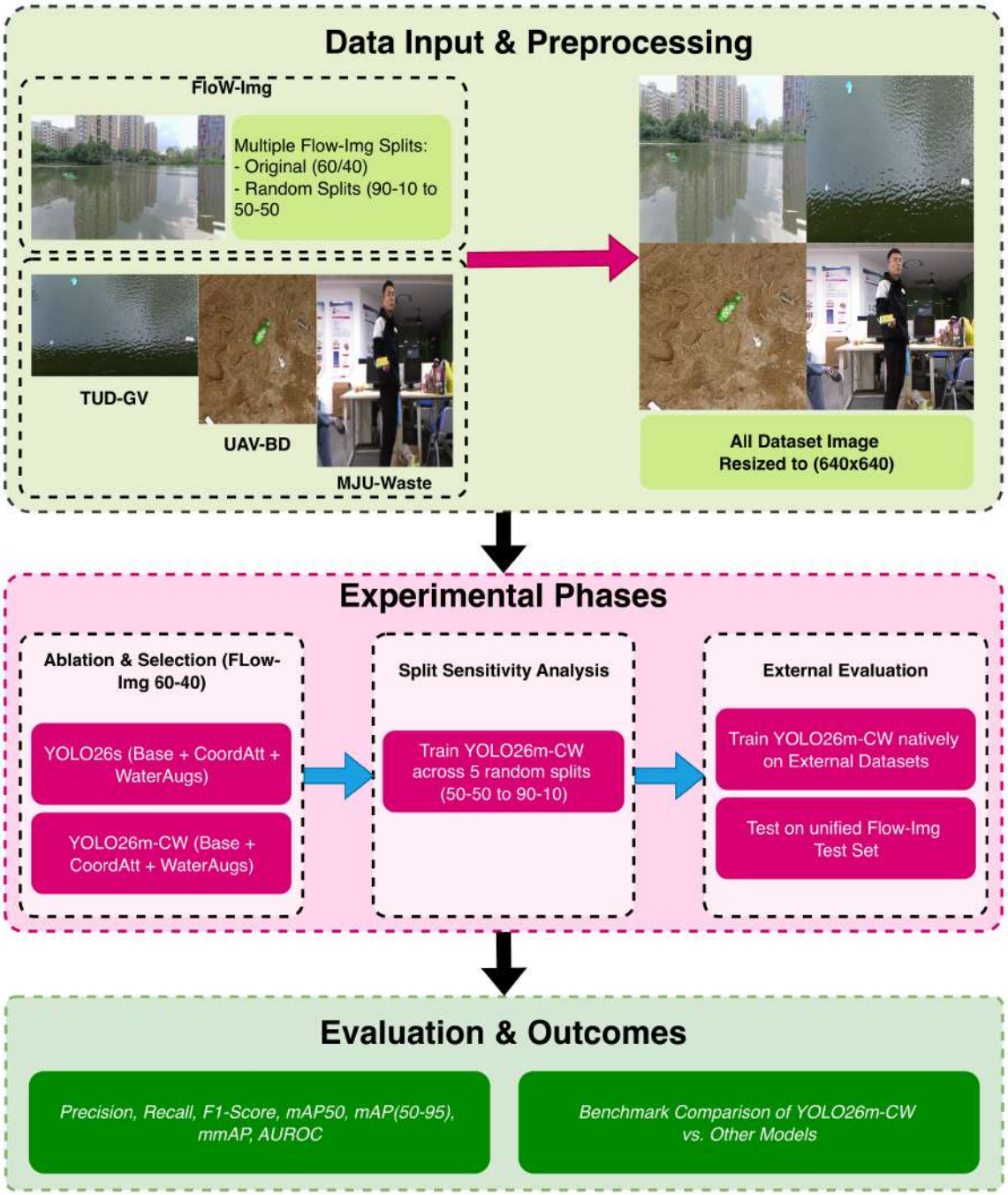


Figure 4.1: Overview of the proposed framework: **(a)** data input and preprocessing, **(b)** experimental phases covering component evaluation, split sensitivity, and external evaluation, and **(c)** unified evaluation and benchmark comparison.

4.2 Model Architecture: YOLO26

The detection backbone of the proposed framework belongs to the single-stage You Only Look Once (YOLO) detector family [4, 10]. Specifically, we employ the modern NMS-free YOLO26 architecture, which integrates feature extraction, feature aggregation, and bounding box regression into a single forward pass. YOLO26 was selected for its favorable trade-off between detection accuracy and computational efficiency, making it well-suited for deployment on edge devices such as Unmanned Surface Vehicles (USVs).

The complete structural pipeline of the baseline YOLO26 detector is illustrated in Fig. 4.2. As shown in the schematic, the architecture sequentially processes input imagery through hierarchical feature extraction in the C3k2 backbone, multi-scale feature aggregation in the C2PSA and SPPF modules, and final bounding-box regression and classification via an Anchor-free Detection Head. The inset table in Fig. 4.2 defines the scaling parameters (d, w, mc) that govern model capacity across variants.

4.2.1 Backbone: Hierarchical Feature Extraction

As depicted in Fig. 4.2, the YOLO26 backbone employs Cross Stage Partial bottleneck with 3 convolutions (C3k2) blocks for hierarchical multi-scale feature extraction. The C3k2 block partitions the input feature map into two branches: one passes through a series of convolutional layers to learn residual representations, while the other bypasses these layers to preserve gradient flow. This cross-stage partial design enables the extraction of rich, multi-scale features while maintaining computational efficiency by reducing redundant gradient information during backpropagation.

Feature maps are progressively downsampled through successive convolutional stages, producing representations at multiple spatial resolutions. The hierarchical nature of this extraction process captures both fine-grained spatial details (critical for small floating waste objects) and high-level semantic information (necessary for distinguishing waste from visually similar background elements such as water reflections).

4.2.2 C2PSA and SPPF Modules

Following feature extraction in the backbone (Fig. 4.2), the key architectural innovation in YOLO26 is the Cross Stage Partial with Spatial Attention (C2PSA) block, which extends the standard CSP bottleneck with an integrated spatial attention mechanism. The C2PSA block enables the network to selectively emphasize spatially relevant features while suppressing background noise a capability particularly important in riverine environments where water surface reflections and wave patterns create pervasive visual clutter.

Following the C2PSA block, the Spatial Pyramid Pooling - Fast (SPPF) module aggregates features across multiple receptive fields through a series of max-pooling operations with different kernel sizes. The SPPF module captures contextual information at multiple scales, enabling the detection of floating waste objects that vary significantly in apparent size

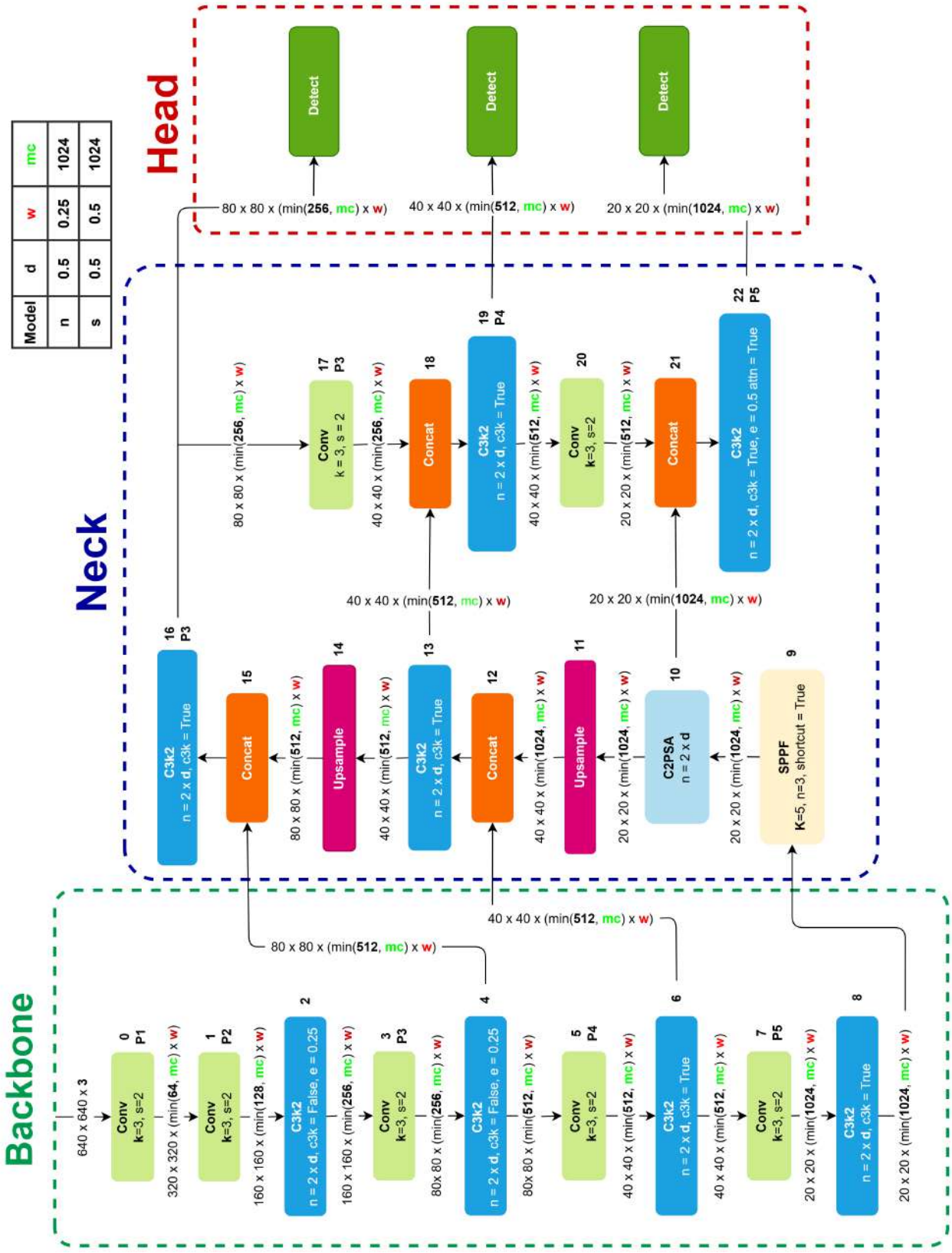


Figure 4.2: The YOLO26 model architecture utilized for floating waste detection, sequentially detailing the C3k2 blocks for feature extraction, the C2PSA and SPPF modules for aggregation, and the Anchor-free Head. The inset table defines scaling parameters: d (depth multiplier), w (width multiplier), and mc (maximum channel count).

depending on their distance from the USV camera. Unlike the original Spatial Pyramid Pooling (SPP), the SPPF variant achieves equivalent receptive field coverage with reduced computational cost by sequentially applying smaller pooling kernels.

4.2.3 Anchor-Free Detection Head

As shown in the right section of Fig. 4.2, the YOLO26 architecture utilizes an anchor-free detection head that directly predicts object centers and bounding box offsets without requiring predefined anchor boxes. This design eliminates the need for anchor box engineering a process that is both dataset-dependent and computationally expensive and reduces the number of hyperparameters that must be manually tuned. The anchor-free approach also removes the requirement for Non-Maximum Suppression (NMS) as a post-processing step, reducing inference latency and making the architecture more suitable for real-time deployment.

4.2.4 Model Scaling

YOLO26 supports multiple scaling configurations through three parameters detailed in the inset table of Fig. 4.2: the depth multiplier (d), which controls the number of repeated blocks; the width multiplier (w), which controls the number of channels per layer; and the maximum channel count (mc), which caps feature dimensionality. In this study, two scaling variants are employed:

- **YOLO26s (Small):** Used during the first stage of component evaluation to screen candidate enhancement modules with reduced computational overhead.
- **YOLO26m (Medium):** Used as the final framework backbone, providing increased model capacity for improved feature representation while remaining suitable for edge deployment.

4.3 YOLO Mathematical Formulation

This section details the mathematical formulation governing bounding-box coordinate decoding, objectness estimation, class probability assignment, Intersection over Union (IoU) calculation, and the composite YOLO loss function utilized by the detection [21].

4.3.1 Bounding-Box Prediction and Coordinate Decoding

For an input image processed at feature map stride $s \in \{8, 16, 32\}$, let (c_x, c_y) denote the top-left grid cell coordinates of a given spatial prediction point. The anchor-free detection head predicts a 4-dimensional bounding box offset vector $\mathbf{t} = (t_x, t_y, t_w, t_h)$ at each location.

The decoded bounding box center coordinates (b_x, b_y) and box dimensions (b_w, b_h) in absolute image pixels are given by:

$$b_x = (2 \cdot \sigma(t_x) - 0.5 + c_x) \times s, \tag{4.1}$$

$$b_y = (2 \cdot \sigma(t_y) - 0.5 + c_y) \times s, \tag{4.2}$$

$$b_w = p_w \cdot (2 \cdot \sigma(t_w))^2, \tag{4.3}$$

$$b_h = p_h \cdot (2 \cdot \sigma(t_h))^2, \tag{4.4}$$

where $\sigma(\cdot)$ denotes the logistic sigmoid activation function:

$$\sigma(z) = \frac{1}{1 + e^{-z}}. \tag{4.5}$$

In Equations (4.3)–(4.4), p_w and p_h denote grid-normalized reference base scales. The scaling factor of 2 and shift of -0.5 in Equations (4.1)–(4.2) allow predicted box centers to smoothly cross grid cell boundaries without grid-lock [22].

4.3.2 Objectness and Class Probability Estimation

The detection head outputs an unnormalized objectness logit t_o and class logits vector $\mathbf{s} = (s_1, s_2, \ldots, s_C)$, where C represents the total number of target object classes. The probability that a spatial location contains an object, $P(\text{Object})$, is computed via the sigmoid function:

$$P(\text{Object}) = \sigma(t_o). \tag{4.6}$$

For class prediction, the conditional probability that a detected object belongs to class i, $P(\text{Class}_i \mid \text{Object})$, is evaluated independently using binary sigmoid activations:

$$P(\text{Class}_i \mid \text{Object}) = \sigma(s_i), \quad i \in \{1, 2, \ldots, C\}. \tag{4.7}$$

The final confidence score Score_i for class i assigned to a predicted bounding box B_{pred} relative to a ground-truth box B_{gt} is evaluated as:

$$\text{Score}_i = P(\text{Class}_i \mid \text{Object}) \cdot P(\text{Object}) \cdot \text{IoU}(B_{\text{pred}}, B_{\text{gt}}). \tag{4.8}$$

4.3.3 Intersection over Union (IoU) and Complete IoU (CIoU)

Let $B_{\text{pred}} = (x_1, y_1, x_2, y_2)$ and $B_{\text{gt}} = (x_1^{\text{gt}}, y_1^{\text{gt}}, x_2^{\text{gt}}, y_2^{\text{gt}})$ represent the predicted and ground-truth bounding box coordinates, respectively. Standard Intersection over Union (IoU) is defined as:

$$\text{IoU}(B_{\text{pred}}, B_{\text{gt}}) = \frac{|B_{\text{pred}} \cap B_{\text{gt}}|}{|B_{\text{pred}} \cup B_{\text{gt}}|}, \tag{4.9}$$

where $|B_{\text{pred}} \cap B_{\text{gt}}|$ is the area of intersection and $|B_{\text{pred}} \cup B_{\text{gt}}|$ is the area of union.

To improve bounding box regression convergence, Complete IoU (CIoU) loss [10] is employed, which accounts for overlap area, central point distance, and aspect ratio consistency:

$$\mathcal{L}_{\text{CIoU}} = 1 - \text{IoU} + \frac{\rho^2(\mathbf{b}, \mathbf{b}^{\text{gt}})}{c^2} + \alpha v, \tag{4.10}$$

where $\mathbf{b} = (b_x, b_y)$ and $\mathbf{b}^{\text{gt}} = (b_x^{\text{gt}}, b_y^{\text{gt}})$ represent the central points of B_{pred} and B_{gt}, $\rho(\cdot)$ denotes Euclidean distance, c is the diagonal length of the smallest enclosing bounding box covering both B_{pred} and B_{gt}, and v measures aspect ratio consistency:

$$v = \frac{4}{\pi^2}\left(\arctan\frac{w^{\text{gt}}}{h^{\text{gt}}} - \arctan\frac{b_w}{b_h}\right)^2. \tag{4.11}$$

The trade-off parameter α is defined as:

$$\alpha = \frac{v}{(1 - \text{IoU}) + v}. \tag{4.12}$$

4.3.4 Composite YOLO Loss Function Breakdown

Training optimization is guided by a composite loss function $\mathcal{L}_{\text{total}}$, combining bounding box regression loss $\mathcal{L}_{\text{box}}$, classification loss $\mathcal{L}_{\text{cls}}$, and Distribution Focal Loss $\mathcal{L}_{\text{dfl}}$:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{box}}\mathcal{L}_{\text{box}} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{dfl}}\mathcal{L}_{\text{dfl}}, \tag{4.13}$$

where λ_{box}, λ_{cls}, and λ_{dfl} are scalar loss balancing hyperparameter weights.

The bounding box regression loss $\mathcal{L}_{\text{box}}$ averages the CIoU loss across all assigned positive matching samples N_{pos}:

$$\mathcal{L}_{\text{box}} = \frac{1}{N_{\text{pos}}}\sum_{k=1}^{N_{\text{pos}}}\mathcal{L}_{\text{CIoU}}(B_{\text{pred}}^{(k)}, B_{\text{gt}}^{(k)}). \tag{4.14}$$

The classification loss $\mathcal{L}_{\text{cls}}$ is evaluated using Binary Cross-Entropy (BCE) over all N spatial prediction locations and C classes:

$$\mathcal{L}_{\text{cls}} = -\frac{1}{N \cdot C}\sum_{j=1}^{N}\sum_{i=1}^{C}\left[y_{j,i}\log(\hat{y}_{j,i}) + (1 - y_{j,i})\log(1 - \hat{y}_{j,i})\right], \tag{4.15}$$

where $y_{j,i} \in \{0, 1\}$ is the ground-truth binary class target for location j and class i, and $\hat{y}_{j,i} = \sigma(s_{j,i})$ is the predicted class probability.

Distribution Focal Loss ($\mathcal{L}_{\text{dfl}}$) optimizes continuous bounding box boundary locations by modeling box boundaries as general probability distributions rather than single deterministic values. For a target continuous boundary value y located between discrete integral indices y_i and y_{i+1} ($y_i \leq y \leq y_{i+1}$), $\mathcal{L}_{\text{dfl}}$ is given by:

$$\mathcal{L}_{\text{dfl}}(S_i, S_{i+1}) = -\left((y_{i+1} - y)\log(S_i) + (y - y_i)\log(S_{i+1})\right), \tag{4.16}$$

where S_i and S_{i+1} are Softmax probability outputs at discrete locations y_i and y_{i+1}, satisfying $S_i = \frac{y_{i+1} - y}{y_{i+1} - y_i}$ and $S_{i+1} = \frac{y - y_i}{y_{i+1} - y_i}$.

4.4 Enhancement Modules

To address the environment-specific challenges inherent in riverine waste detection, two enhancement modules are investigated as configurable plug-in components. These modules are designed to be complementary: Coordinate Attention targets improved spatial

feature preservation, while Water-Specific Augmentations target environment-specific data robustness.

4.4.1 Coordinate Attention (CoordAtt)

Standard channel attention mechanisms, such as Squeeze-and-Excitation (SE), rely on global average pooling to generate channel descriptors. While effective for channel-wise feature recalibration, this approach discards all spatial information during the pooling operation, producing a single scalar per channel that cannot encode where informative features are located within the feature map. This limitation is particularly problematic for floating waste detection, where targets are often partially submerged, occluded by water splashes, or located at image boundaries where positional information carries diagnostic value.

Coordinate Attention (CoordAtt) [20] addresses this limitation by decomposing the 2D global pooling operation into two separate 1D pooling operations along the horizontal and vertical axes. Specifically, given an input feature map $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$, CoordAtt generates two direction-aware feature maps:

$$z_c^h(h) = \frac{1}{W} \sum_{0\le i<W} x_c(h, i), \tag{4.17}$$

$$z_c^w(w) = \frac{1}{H} \sum_{0\le j<H} x_c(j, w). \tag{4.18}$$

These directional feature maps $z_c^h(h)$ and $z_c^w(w)$ are concatenated along the spatial dimension and transformed through a shared 1×1 convolution $\mathbf{F}_1$ with batch normalization and non-linear activation:

$$\mathbf{f} = \delta\left(\mathbf{F}_1\left([\mathbf{z}^h, \mathbf{z}^w]\right)\right), \tag{4.19}$$

where $[\cdot,\cdot]$ denotes spatial concatenation, $\delta(\cdot)$ is the non-linear activation function, and $\mathbf{f} \in \mathbb{R}^{(C/r)\times(H+W)}$ represents an intermediate spatial feature map with reduction ratio r.

The feature tensor $\mathbf{f}$ is split along the spatial dimension into two separate tensors $\mathbf{f}^h \in \mathbb{R}^{(C/r)\times H}$ and $\mathbf{f}^w \in \mathbb{R}^{(C/r)\times W}$. Two independent 1×1 convolutional layers $\mathbf{F}_h$ and $\mathbf{F}_w$ transform $\mathbf{f}^h$ and $\mathbf{f}^w$ back to channel dimension C:

$$g_c^h(h) = \sigma\left(\mathbf{F}_h\left(\mathbf{f}^h\right)\right), \tag{4.20}$$

$$g_c^w(w) = \sigma\left(\mathbf{F}_w\left(\mathbf{f}^w\right)\right), \tag{4.21}$$

where $\sigma(\cdot)$ is the sigmoid function. The resulting attention weights $g_c^h(h)$ and $g_c^w(w)$ are applied multiplicatively to the input feature map $\mathbf{X}$ to produce the final coordinate-attended output $\mathbf{Y} \in \mathbb{R}^{C\times H\times W}$:

$$y_c(h, w) = x_c(h, w) \times g_c^h(h) \times g_c^w(w). \tag{4.22}$$

In the proposed framework, CoordAtt is integrated directly into the SPPF module, replacing standard spatial pooling with coordinate-aware attention. This placement was chosen because the SPPF module serves as the transition point between the feature extraction backbone and the detection head, making it the optimal location to inject spatial awareness before bounding box prediction.

4.4.2 Water-Specific Augmentations (WaterAugs)

While CoordAtt addresses the feature extraction challenge through architectural enhancement, Water-Specific Augmentations address the data diversity challenge by simulating the visual variability of inland water environments during training. Unlike generic augmentation strategies (e.g., random cropping, horizontal flipping) that are agnostic to the deployment domain, WaterAugs are specifically designed to expose the model to the types of visual perturbations encountered in real-world riverine monitoring.

The WaterAugs strategy comprises the following environment-targeted transformations:

- **Rotation ($10°$):** Simulates the tilted camera angles that occur naturally on USV platforms due to wave-induced roll and pitch motions. This augmentation ensures the model is robust to slight perspective variations caused by hydrodynamic disturbances.
- **HSV Perturbations (Saturation 0.8, Value 0.5):** Simulates the wide range of water appearances encountered in inland environments, from murky, silt-laden channels to clear, reflective surfaces. By perturbing the Hue, Saturation, and Value channels, the model learns to detect waste regardless of water color and brightness conditions.
- **Vertical Flip ($p = 0.1$):** Simulates the visual effect of water surface reflections, where floating objects may appear partially mirrored. The low probability ensures this augmentation adds diversity without dominating the training distribution.
- **Mixup ($p = 0.15$):** Blends two training images to simulate cluttered waste scenes where multiple objects overlap or partially occlude each other. This augmentation improves the model's ability to separate individual targets in dense accumulation zones.
- **Increased Scale Jitter (0.6):** Simulates the significant size variation of floating waste objects at different distances from the USV camera. Objects near the camera appear large and detailed, while distant objects are small and low-contrast. Scale jitter exposes the model to this full range during training.

Each transformation was calibrated based on the physical characteristics of USV-based monitoring: the rotation angle corresponds to observed USV pitch ranges, the HSV parameters span the color distributions observed in the FloW-Img dataset, and the scale jitter range reflects the depth-dependent size variation inherent in surface-level imagery.

4.5 Training Protocol

All experiments were conducted using the Ultralytics framework on an NVIDIA RTX 4070 GPU with 12 GB Video Random Access Memory (VRAM). The training configuration was standardized across all experiments to ensure fair comparison between model variants and enhancement modules.

4.5.1 Optimization Configuration

Models were trained using Stochastic Gradient Descent (SGD) with the following hyperparameters:

- **Batch Size:** 8 (constrained by available VRAM)
- **Training Epochs:** 300 (maximum)
- **Early Stopping Patience:** 60 epochs without validation improvement
- **Initial Learning Rate (η_{init}):** 0.01 with cosine annealing decay to 0.0001
- **Momentum:** 0.937
- **Weight Decay:** 0.0005

A three-epoch warmup phase was employed to ensure stable convergence during the early stages of training. During warmup, the momentum was set to 0.8 and the bias learning rate to 0.1, gradually transitioning to the standard training parameters. This warmup strategy prevents the large initial gradient updates that can destabilize training, particularly when using aggressive augmentation strategies.

4.5.2 Baseline Augmentations

In addition to the Water-Specific Augmentations described in Section 4.4.2, the following baseline augmentation techniques were applied consistently across all experimental configurations:

- **Copy-Paste ($p = 0.3$):** Randomly copies object instances from one training image and pastes them into another, increasing the effective density and diversity of training examples.
- **Close Mosaic:** Combines four training images into a single mosaic during the final training epochs, exposing the model to diverse spatial contexts.
- **Dropout (0.05):** Applies a low-rate dropout to prevent overfitting and improve generalization.
- **Random Erasing ($p = 0.2$):** Randomly occludes rectangular regions of the input image, encouraging the model to rely on partial object features for detection rather than overfitting to complete object appearances.

4.6 Evaluation Framework

A comprehensive evaluation framework was designed to assess detection performance across multiple complementary dimensions. All metrics were computed on the test set using confusion matrix values extracted at an Intersection over Union (IoU) threshold of 0.45.

4.6.1 Precision

Precision measures the proportion of predicted detections that are correct, quantifying the model's ability to avoid false positive detections:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4.23}$$

where TP (True Positives) represents correctly detected waste objects and FP (False Positives) represents detections that do not correspond to actual waste. High precision is critical for autonomous USV systems, where false positive detections could trigger unnecessary cleanup operations and waste computational resources.

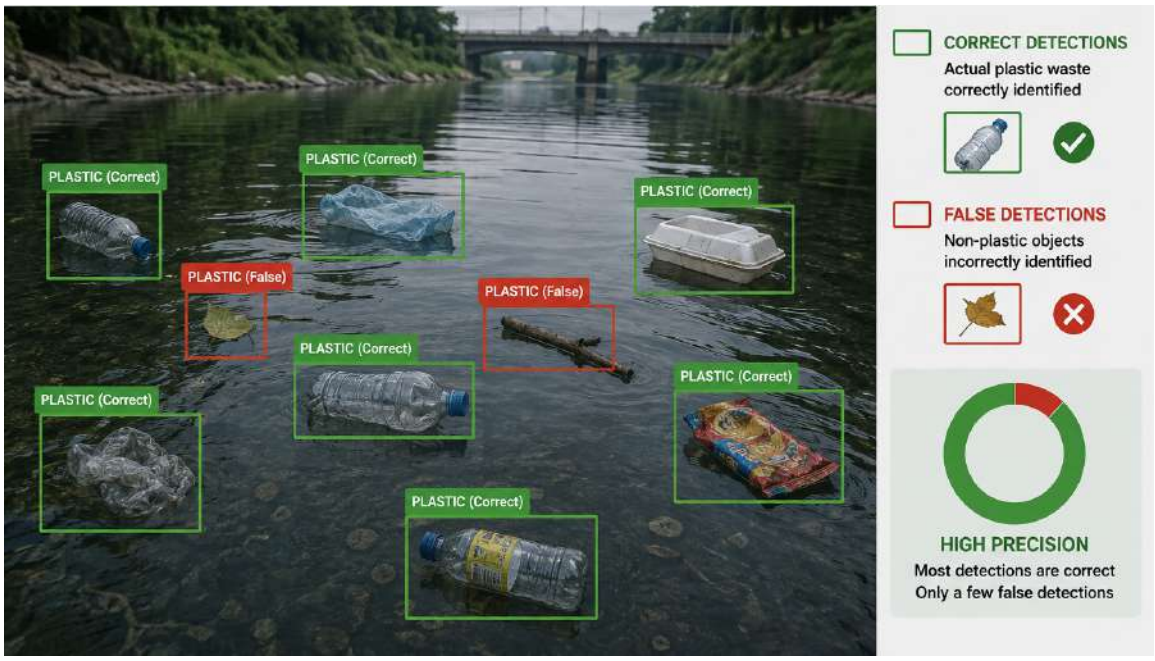


Figure 4.3: Visual illustration of precision in object detection, showing correct plastic-waste detections and a small number of false detections. [23]

4.6.2 Recall

Recall measures the proportion of actual waste objects that are successfully detected, quantifying the model's sensitivity:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4.24}$$

where FN (False Negatives) represents waste objects that the model fails to detect. High recall ensures comprehensive waste coverage, minimizing the amount of floating debris that escapes detection during monitoring operations.

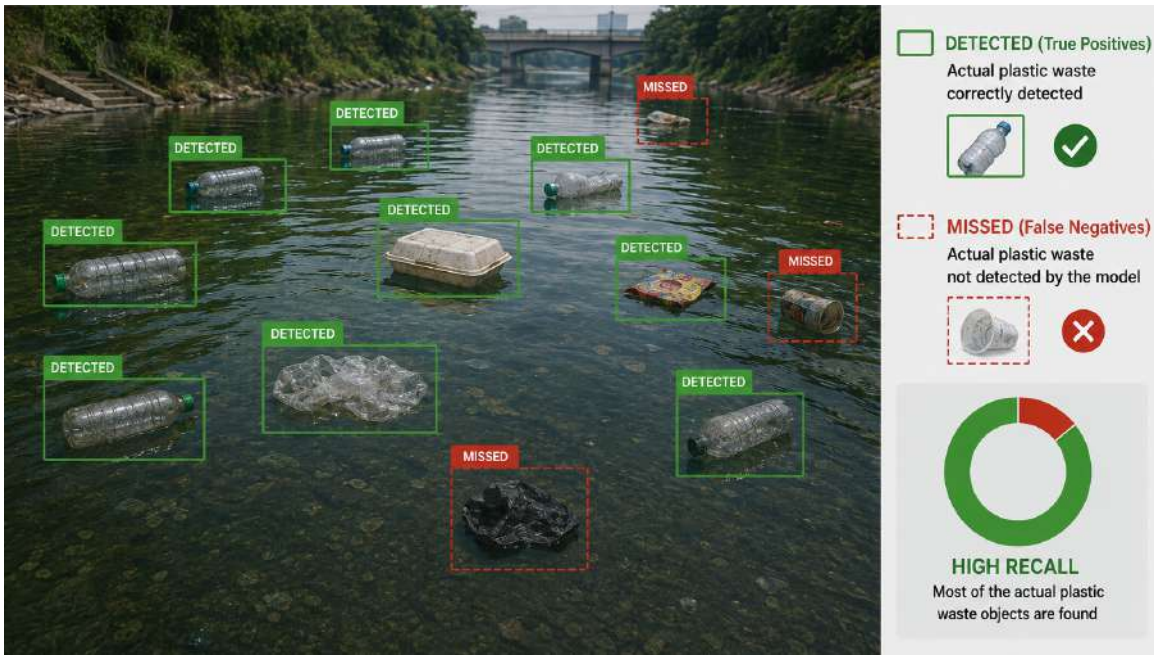


Figure 4.4: Visual illustration of recall in object detection, showing the model successfully detecting most of the actual plastic-waste objects while only a few remain undetected. [23]

4.6.3 F_1-Score

The F_1-Score provides the harmonic mean of Precision and Recall, offering a balanced measure that penalizes models with extreme trade-offs between the two:

$$F_1\text{-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.25}$$

The harmonic mean ensures that both Precision and Recall must be high for the F_1-Score to be high; a model with perfect precision but zero recall (or vice versa) receives an F_1-Score of zero.

4.6.4 Mean Average Precision (mAP)

Two variants of Mean Average Precision are employed:

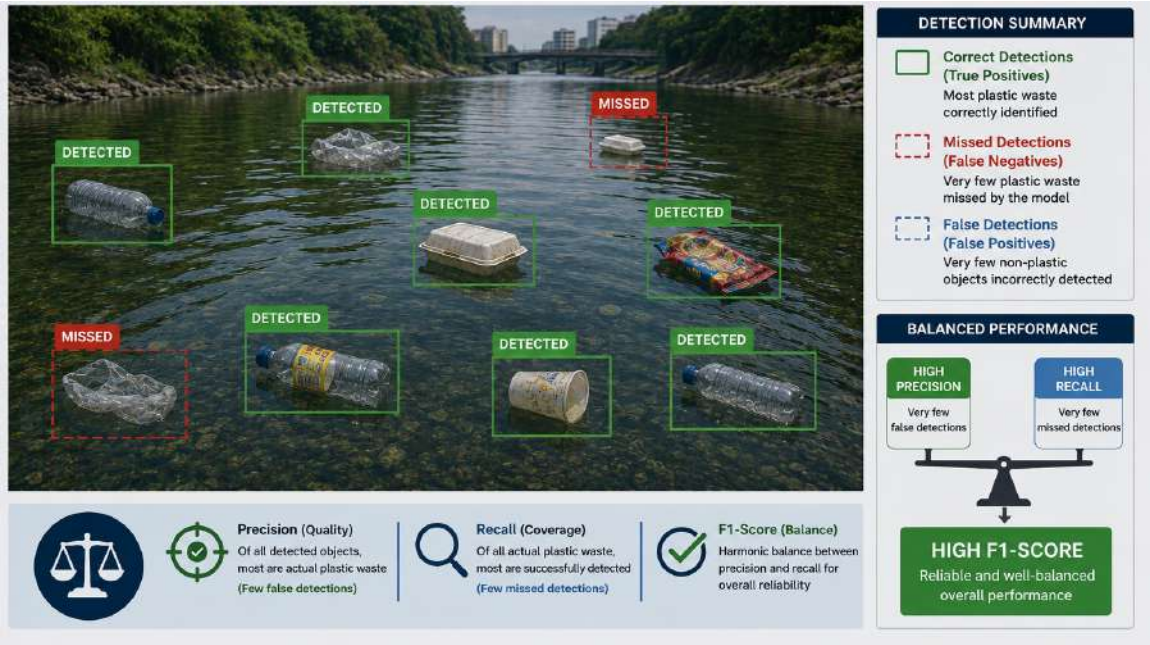


Figure 4.5: Visual illustration of F1-score in object detection, showing a balanced performance between accurate detections and the avoidance of false and missed detections. [23]

- **mAP$_{50}$:** Computed at a single IoU threshold of 0.50, measuring the model's ability to correctly localize objects with moderate spatial overlap. This metric is widely used in object detection benchmarks and enables direct comparison with prior works. It is calculated as:

$$\text{mAP}_{50} = \frac{1}{C}\sum_{c=1}^{C}\text{AP}_{c,\,0.50}, \tag{4.26}$$

  where C denotes the number of object classes and $\text{AP}_{c,\,0.50}$ represents the Average Precision for class c at an Intersection over Union (IoU) threshold of 0.50.

- **mAP$_{50\text{-}95}$:** Averaged across IoU thresholds from 0.50 to 0.95 in increments of 0.05. This stricter metric penalizes imprecise bounding box localization and provides a more comprehensive assessment of detection quality. It is calculated as:

$$\text{mAP}_{50\text{-}95} = \frac{1}{10C}\sum_{t\in\{0.50,0.55,\ldots,0.95\}}\sum_{c=1}^{C}\text{AP}_{c,t}, \tag{4.27}$$

  where $\text{AP}_{c,t}$ denotes the Average Precision for class c at IoU threshold t, and the ten IoU thresholds range from 0.50 to 0.95 with an increment of 0.05.

4.6.5 Mean Mean Average Precision (mmAP)

To provide a single composite metric that balances detection coverage and localization precision, we compute the mean mean Average Precision (mmAP):

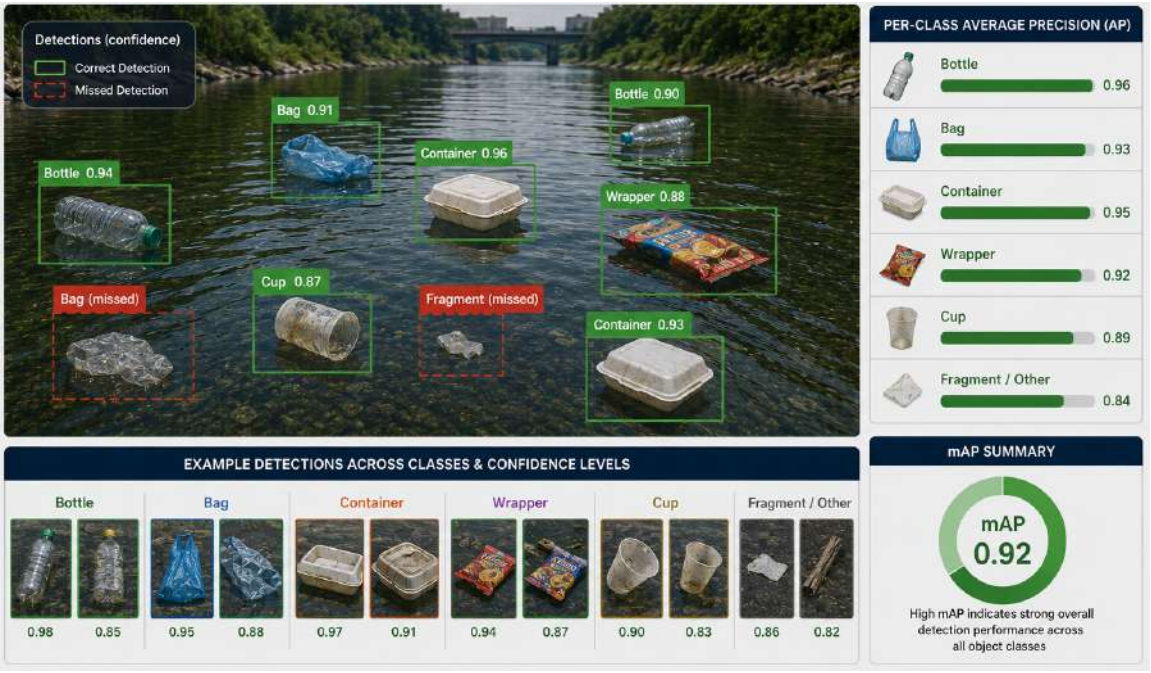


Figure 4.6: Visual illustration of mAP in object detection, showing the model's overall detection performance across multiple plastic-waste categories. [23]

$$\text{mmAP} = \frac{\text{mAP}_{50} + \text{mAP}_{50\text{-}95}}{2} \tag{4.28}$$

This metric was adopted from the original FloW-Img benchmark [14] to enable direct comparison with the baselines reported by Cheng et al.

4.6.6 AUROC

The Area Under the Receiver Operating Characteristic curve (AUROC) is a threshold-independent metric used to evaluate the ability of the model to discriminate between positive (waste) and negative (background) samples. Unlike precision, recall, and F1-score, which are typically calculated at a specific confidence threshold, AUROC evaluates the model across the complete range of possible confidence thresholds. This provides a more comprehensive measure of the model's inherent discriminative capability and its robustness to changes in the operating threshold.

The Receiver Operating Characteristic (ROC) curve is generated by progressively varying the confidence threshold and calculating the corresponding True Positive Rate (TPR) and False Positive Rate (FPR). The TPR represents the proportion of actual waste samples correctly detected by the model and is given by

$$\text{TPR} = \frac{TP}{TP + FN}, \tag{4.29}$$

where TP and FN denote the numbers of true positive and false negative predictions,

respectively. The FPR represents the proportion of background samples incorrectly classified as waste and is defined as

$$\mathrm{FPR} = \frac{FP}{FP + TN}, \tag{4.30}$$

where FP and TN denote the numbers of false positive and true negative predictions, respectively. The ROC curve is obtained by plotting TPR against FPR for all possible confidence thresholds.

The AUROC is calculated as the area under the resulting ROC curve:

$$\mathrm{AUROC} = \int_0^1 \mathrm{TPR}(\mathrm{FPR})\, d(\mathrm{FPR}). \tag{4.31}$$

AUROC can also be interpreted probabilistically as the likelihood that the model assigns a higher confidence score to a randomly selected positive sample than to a randomly selected negative sample:

$$\mathrm{AUROC} = P\left(s(x^+) > s(x^-)\right), \tag{4.32}$$

where $s(x^+)$ and $s(x^-)$ represent the model confidence scores for a waste and background sample, respectively. An AUROC of 1.0 indicates perfect class discrimination, while an AUROC of 0.5 corresponds to random discrimination. Higher AUROC values therefore indicate better separation between waste and background predictions across different confidence thresholds.

A high AUROC indicates that the model maintains good discriminative performance even when the confidence threshold is adjusted, providing a more robust assessment of model behavior than a single threshold-dependent operating point.

Figure 4.7: Visual illustration of AUROC, showing the model's ability to distinguish plastic waste from non-plastic objects across different decision thresholds. [23]

Chapter-5: Results and Analysis

This chapter presents a comprehensive and rigorous evaluation of the proposed YOLO26m-CW framework. To systematically validate the efficacy of our model design, our experimental design is structured around four primary evaluation dimensions: component selection (ablation studies), direct benchmarking against original baselines, split sensitivity analysis to assess data volume robustness, and comparison with other state-of-the-art architectures using matched data split configurations. Finally, we conduct a qualitative analysis of detection results across multiple extreme operational scenarios to demonstrate the framework's real-world deployability.

All experiments were conducted on a uniform workstation equipped with an NVIDIA GeForce RTX 4070 GPU (12 GB VRAM), using the PyTorch-based Ultralytics object detection library. Models were optimized using Stochastic Gradient Descent (SGD) with a momentum coefficient of 0.937 and a weight decay factor of 0.0005. Training was conducted for a maximum of 300 epochs with a batch size of 8 (constrained by VRAM), incorporating an early stopping patience of 60 epochs to prevent overfitting. The learning rate was initiated at 0.01 and decayed to 0.0001 using a cosine annealing schedule, preceded by a three-epoch warmup phase. Baseline data augmentations included Copy-Paste ($p = 0.3$), Close Mosaic, Dropout (0.05), and Random Erasing ($p = 0.2$).

5.1 Component Selection (Ablation Study)

To isolate and quantify the individual and combined contributions of the proposed enhancement modules, namely Coordinate Attention (CoordAtt) and Water-Specific Augmentations (WaterAugs), we conducted a systematic two-stage ablation study on the FloW-Img dataset. The first stage screens candidate enhancements using the lightweight YOLO26s (Small) backbone to minimize computational overhead during exploration. The second stage validates these findings on the larger YOLO26m (Medium) architecture to assess model capacity effects and confirm the final configuration. All component selection models are evaluated on the original author-provided 60-40 train-test split configuration.

Table 5.1 details the quantitative results obtained during the two-stage ablation process across seven distinct configurations.

Table 5.1: Two-stage component selection on the original FloW-Img 60-40 split. Stage 1 identifies effective modules on YOLO26s; Stage 2 validates their combination on YOLO26m. Bold values indicate best per metric within each stage.

Stage	Configuration	Precision	Recall	F_1-Score	mAP_{50}	mAP(50–95)	$mmAP$	AUROC
1	YOLO26s Baseline	0.910	0.779	0.839	0.843	0.471	0.657	**0.852**
	YOLO26s + CoordAtt	0.902	**0.797**	0.846	0.851	**0.473**	**0.662**	0.838
	YOLO26s + WaterAugs	**0.914**	0.788	**0.846**	**0.853**	0.460	0.656	0.837
2	YOLO26m Baseline	0.904	0.803	0.851	0.860	0.476	0.668	0.849
	YOLO26m + CoordAtt	0.914	0.804	0.856	0.857	0.473	0.665	0.863
	YOLO26m + WaterAugs	**0.920**	0.804	0.858	0.866	0.472	0.669	0.864
	YOLO26m-CW	0.918	**0.807**	**0.859**	**0.866**	**0.478**	**0.672**	**0.871**

5.1.1 Stage 1: Exploration on YOLO26s

The Stage 1 experiments on the YOLO26s backbone reveal clear, complementary trends between the two modular enhancements. The YOLO26s Baseline establishes a solid foundation with an F_1-Score of 0.839 and a mean mean Average Precision ($mmAP$) of 0.657.

Integrating the Coordinate Attention (CoordAtt) module directly into the Spatial Pyramid Pooling - Fast (SPPF) module (referred to as YOLO26s + CoordAtt) yields a significant boost in Recall, which rises from 0.779 to 0.797 (+0.018 increase). This improvement comes at a marginal cost of a 0.008 reduction in Precision. The overall $mmAP$ improves by 0.005 to reach 0.662. The marked increase in recall suggests that embedding coordinate-aware positional details along the horizontal and vertical axes enables the network to locate highly challenging targets, such as small, partially submerged plastic fragments or debris located at image boundaries, that are normally missed by standard global average pooling mechanisms.

Conversely, applying the Water-Specific Augmentations (referred to as YOLO26s + WaterAugs) exhibits an opposite, precision-oriented effect. While Recall increases slightly to 0.788 (+0.009), Precision peaks at 0.914 (+0.004 over baseline). The mAP_{50} metric reaches 0.853. The precision improvement indicates that exposing the network during training to simulated riverine noise (e.g., wave tilt rotation, HSV water clarity variations, specular vertical reflection flips, scale jitter) prevents the model from developing brittle associations with background aquatic features. By training the network to distinguish waste from realistic environmental disturbances, the model successfully avoids false positive detections (e.g., mistaking sun glare or foaming waves for floating plastic).

5.1.2 Stage 2: Validation on YOLO26m

Stage 2 transitions the ablation evaluation to the medium-sized YOLO26m architecture, validating the synergy of combining both components in a higher-capacity framework. The YOLO26m Baseline model improves over the smaller baseline due to its expanded capacity, recording a Precision of 0.904, Recall of 0.803, F_1-Score of 0.851, and an $mmAP$ of 0.668.

Integrating CoordAtt onto the YOLO26m backbone (YOLO26m + CoordAtt) elevates Precision to 0.914 and Recall slightly to 0.804, resulting in an F_1-Score of 0.856. The AUROC improves from 0.849 to 0.863. Applying WaterAugs (YOLO26m + WaterAugs) boosts Precision further to 0.920 and matches the recall at 0.804, establishing an F_1-Score of 0.858 and an $mmAP$ of 0.669.

The final proposed configuration, YOLO26m-CW, combines both Coordinate Attention and Water-Specific Augmentations. By combining both modules, the model achieves a balanced Precision of 0.918 and a peak Recall of 0.807, resulting in the highest F_1-Score (0.859), mAP_{50-95} (0.478), $mmAP$ (0.672), and AUROC (0.871) in the entire component selection matrix. This demonstrates a clear synergistic relationship: while CoordAtt ensures that the model preserves structural and spatial representations of tiny targets, WaterAugs regularizes the training, ensuring that the model generalizes across water color variations and lighting changes.

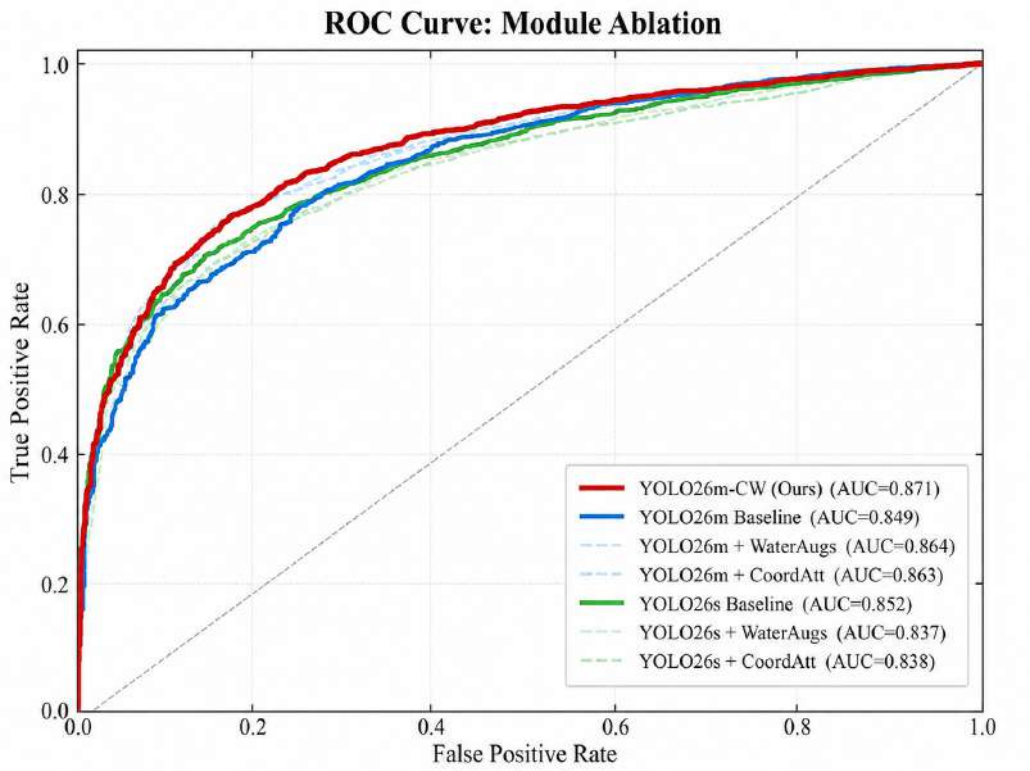


Figure 5.1: ROC curves for module evaluation on the FloW-Img 60-40 benchmark across all seven configurations. YOLO26m-CW (red) achieves the highest AUROC (0.871). Intermediate variants are shown with reduced opacity for visual clarity.

The receiver operating characteristic (ROC) curves and precision-recall (PR) curves in Fig. 5.1 and Fig. 5.2 further illustrate this performance trend. As shown in Fig. 5.1, YOLO26m-CW maintains the largest Area Under the ROC Curve (AUROC = 0.871), maintaining high sensitivity while limiting false positive rates. Similarly, Fig. 5.2 demonstrates that YOLO26m-CW achieves the most favorable precision-recall envelope, remaining stable across varying detection confidence thresholds.

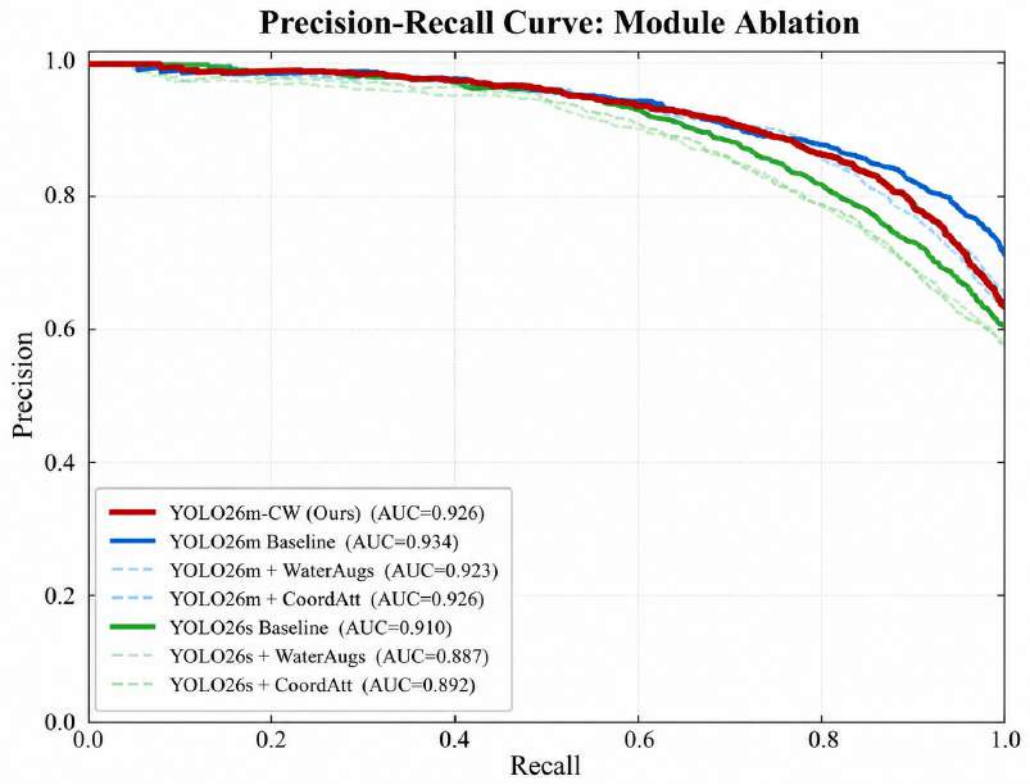


Figure 5.2: Precision-Recall curves for module evaluation on the FloW-Img 60-40 benchmark across all seven configurations. YOLO26m-CW (red) achieves the most favorable precision-recall envelope. Intermediate variants are shown with reduced opacity for visual clarity.

5.2 Benchmark Comparison

To evaluate the proposed YOLO26m-CW framework within the context of existing literature, we conducted a direct comparison against the baseline object detection models reported in the original FloW-Img benchmark paper by Cheng et al. [14]. This comparison is divided into two parts: evaluation on the original 60-40 train-test split, and an extensive cross-domain external evaluation.

5.2.1 Original 60-40 Split

Cheng et al. [14] established the initial floating waste detection baselines on the FloW-Img dataset using two classic multi-stage anchor-based architectures: Faster R-CNN and Cascade R-CNN. Both models were trained and tested on the author-defined 60-40 split configuration. Table 5.2 summarizes the performance comparisons on this split.

The experimental results highlight the substantial performance gap between older multi-stage detectors and our optimized single-stage framework. Faster R-CNN struggles on the FloW-Img dataset, yielding a low $mmAP$ of only 0.184. The more robust, multi-threshold Cascade R-CNN model improves performance significantly, achieving an $mmAP$ of 0.434.

However, YOLO26m-CW outperforms Cascade R-CNN by a large margin, achieving an

Table 5.2: Benchmark comparison with Cheng et al. [14] on the original 60-40 split. Bold values indicate best per metric.

Metric	**Faster R-CNN** [14]	**Cascade R-CNN** [14]	**YOLO26m-CW (Ours)**
Precision	–	–	**0.918**
Recall	–	–	**0.807**
F_1-Score	–	–	**0.859**
mAP_{50}	–	–	**0.866**
mAP(50–95)	–	–	**0.478**
$mmAP$	0.184	0.434	**0.672**
AUROC	–	–	**0.871**

$mmAP$ of 0.672, which represents an absolute improvement of +0.238 (+23.8%) and a relative improvement of +54.8%. Over Faster R-CNN, our framework records an $mmAP$ increase of +0.488 (+48.8%).

This massive performance difference stems from several structural limitations of Faster R-CNN and Cascade R-CNN on aquatic surface-level datasets. First, these older architectures rely heavily on predefined anchor boxes. Anchor-based approaches struggle with the diverse aspect ratios and arbitrary shapes typical of riverine floating debris (e.g., plastic bottles, floating branches, plastic bags). Second, both models downsample feature maps repeatedly through conventional convolutions without spatial attention mechanisms, causing the feature representations of small objects (which make up a large portion of the FloW-Img dataset) to disappear before reaching the region proposal network. YOLO26m-CW bypasses these bottlenecks through its anchor-free detection head, which directly regresses coordinates from spatial features, and the C2PSA and CoordAtt modules, which actively preserve spatial coordinates throughout the network.

5.2.2 External Evaluation

To evaluate the generalization limits of floating waste detection models and test their robustness across varying environmental domains, we replicated the external evaluation protocol defined by Cheng et al. [14]. This protocol tests the domain sensitivity of a model by training it entirely on an external dataset and testing its performance on the FloW-Img test set (800 images). We evaluated three external source datasets: UAV-BD (aerial perspective), MJU-Waste (indoor environment), and TUD-GV (urban canal environment, introduced as a new benchmark). Table 5.3 details the quantitative results.

When trained on the aerial UAV-BD dataset, the baseline model from Cheng et al. achieves a very low $mmAP$ of 0.058 on FloW-Img. In contrast, YOLO26m-CW yields an $mmAP$ of 0.228, achieving a 3.9 times relative improvement (+0.170 increase). Similarly, when trained on the indoor MJU-Waste dataset, the baseline model scores an $mmAP$ of 0.032, while our model achieves 0.081, showing a 2.5 times improvement. Additionally, our new baseline using the urban canal TUD-GV dataset achieves an $mmAP$ of 0.187, with a Precision of 0.391 and Recall of 0.271.

Despite these gains over the baseline models, the absolute performance scores under

Table 5.3: External evaluation results. Models are trained on 100% of each external dataset and tested on the FloW-Img test set. Bold values indicate the best performance per dataset.

Metric	UAV-BD Cheng et al.	UAV-BD Ours	MJU-Waste Cheng et al.	MJU-Waste Ours	TUD-GV Ours (New)
Precision	–	0.568	–	0.220	0.391
Recall	–	0.126	–	0.081	0.271
F_1**-Score**	–	0.205	–	0.118	0.320
mAP_{50}	–	0.318	–	0.117	0.259
mAP(50–95)	–	0.138	–	0.045	0.115
$mmAP$	0.058	**0.228**	0.032	**0.081**	0.187
AUROC	–	0.809	–	0.666	0.795

external training remain extremely low compared to models trained on FloW-Img itself. This drop in performance is explained by specific domain mismatch characteristics:

1. **Perspective Mismatch (UAV-BD):** The UAV-BD dataset consists of high-altitude, top-down aerial views. These top-down perspectives eliminate horizontal surface challenges (like horizon alignment, shoreline clutter, and direct sun glare reflection into the lens). When a model trained on these clean, top-down objects is tested on the surface-level USV perspective of FloW-Img, it fails to recognize objects that are occluded or heavily distorted by perspective scaling.
2. **Contextual Absence (MJU-Waste):** The MJU-Waste dataset is captured in a controlled indoor environment using RGBD sensors, with objects placed on flat, static backgrounds. The complete absence of water surface dynamics, reflections, flowing currents, and natural lighting changes makes it impossible for the model to learn representations that translate to natural rivers, resulting in an extremely low $mmAP$ of 0.081.
3. **Environmental Variance (TUD-GV):** Although TUD-GV contains waterway images, it is captured from bridge and canal pathways in clean, structured urban canal environments. It lacks the complex hydrodynamic waves, heavy vegetation clutter, and muddy conditions of the natural river systems depicted in FloW-Img, leading to poor generalization ($mmAP = 0.187$).

The ROC and PR curves in Fig. 5.3 and Fig. 5.4 illustrate this generalization gap. As shown in the curves, the models trained on UAV-BD, MJU-Waste, and TUD-GV display severe degradation in precision and recall characteristics compared to the model trained directly on the FloW-Img train set (shown in red). This drop in performance proves that domain-specific datasets (like FloW-Img) are essential for training models meant for real-world riverine deployment.

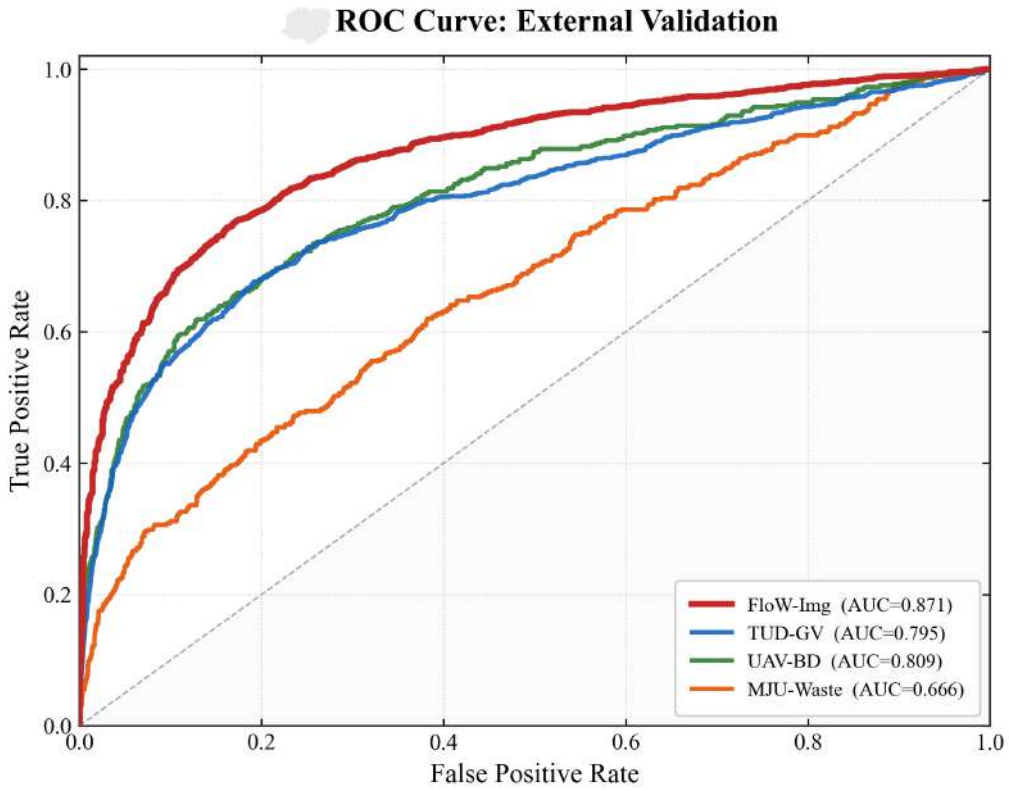


Figure 5.3: External evaluation using ROC curves, demonstrating the performance gap across datasets. FloW-Img serves as the baseline reference, while TUD-GV, UAV-BD, and MJU-Waste quantify the severity of cross-domain performance degradation.

5.3 Split Sensitivity Analysis

A common issue in floating waste detection literature is evaluating models on a single, random train-test partition, which can introduce statistical bias and obscure overfitting. To assess the proposed YOLO26m-CW model's robustness to varying sizes of training data, we conducted a split sensitivity analysis. We pooled all 2,000 FloW-Img images, shuffled them with a fixed seed, and created five distinct train-test split configurations: 50-50, 60-40, 70-30, 80-20, and 90-10. Table 5.4 details the quantitative performance of YOLO26m-CW across these five splits.

The split sensitivity analysis reveals a high level of stability across all five split ratios. Notably, the mAP_{50} metric remains extremely stable, ranging between a minimum of 0.905 (in the 90-10 split) and a maximum of 0.909 (in the 60-40 split). Precision is also highly consistent, staying between 0.943 and 0.959, while Recall ranges from 0.852 to 0.869. The F_1-Score peaks at 0.909 for the 60-40 split, and is sustained above 0.900 for all other configurations.

We observe a slight increase in the stricter mAP_{50-95} (0.572) and the resulting $mmAP$ (0.738) under the 90-10 split configuration. This behavior is attributed to the small test set size associated with this split (only 200 images), which increases statistical variance.

The most important takeaway from this analysis is the performance plateau observed beyond the 60-40 split ratio. Even when training data is reduced to a 50-50 split (only 1,000

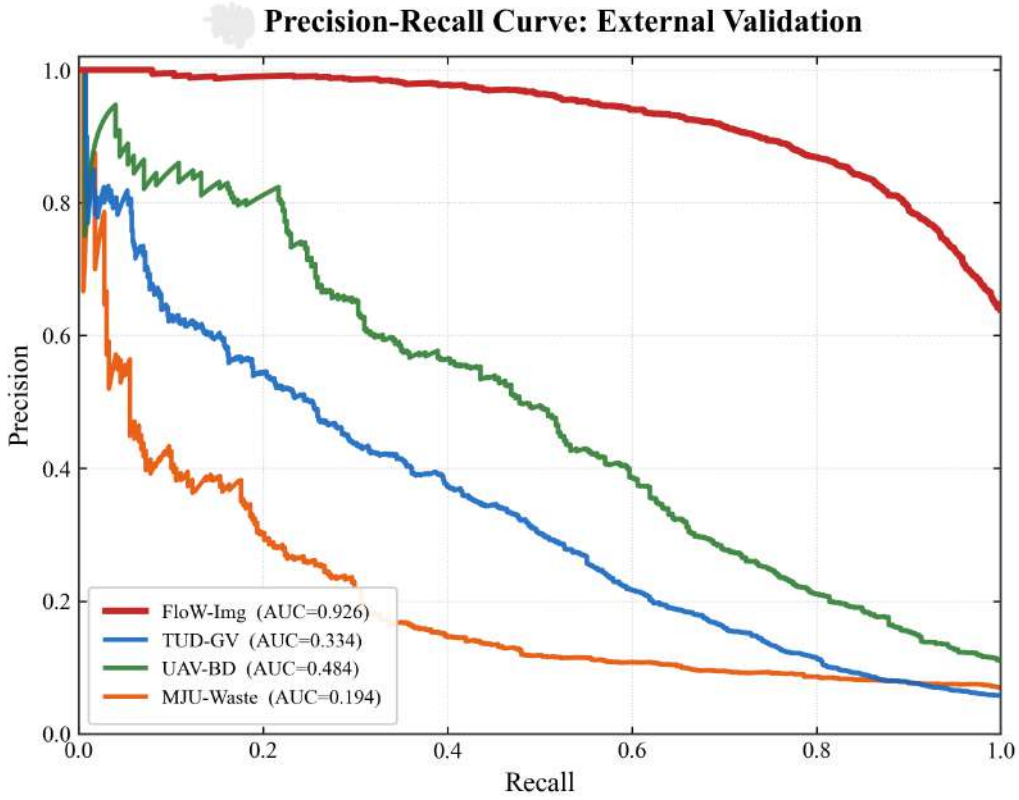


Figure 5.4: External evaluation using Precision-Recall curves, demonstrating the performance gap across datasets. FloW-Img serves as the baseline reference, while TUD-GV, UAV-BD, and MJU-Waste illustrate the degradation in precision and recall under cross-domain evaluation.

Table 5.4: YOLO26m-CW performance across five data splits on the FloW-Img dataset. Bold values indicate best per metric.

Metric	**50-50**	**60-40**	**70-30**	**80-20**	**90-10**
Precision	0.943	0.952	0.948	**0.959**	**0.959**
Recall	0.863	**0.869**	0.858	0.852	0.854
F_1**-Score**	0.901	**0.909**	0.901	0.902	0.903
mAP_{50}	0.907	**0.909**	0.908	0.906	0.905
mAP(50–95)	0.545	0.547	0.554	0.548	**0.572**
$mmAP$	0.726	0.728	0.731	0.727	**0.738**
AUROC	**0.878**	0.875	0.869	0.872	0.859

training images), the framework maintains an mAP_{50} of 0.907 and an F_1-Score of 0.901, which are nearly identical to the performance metrics achieved under the 80-20 and 90-10 splits. This indicates that the combination of Coordinate Attention and Water-Specific Augmentations maximizes data efficiency, allowing the model to reach convergence and high accuracy with a relatively small volume of annotated training images. the standard deviation across splits for mAP_{50} is only ± 0.0016, and for F_1-Score it is ± 0.0033, both of which are negligibly small. This low variance provides strong statistical evidence that the model's performance is not an artifact of a favorable data partition, but rather reflects genuinely learned representations. Such stability is particularly significant in the context of environmental monitoring applications, where deployment datasets may vary substantially in size depending on the geographic region and the availability of annotated riverine imagery.

From a practical deployment perspective, the near-identical performance of the 50-50 and 80-20 configurations suggests that operational systems can be bootstrapped with as few as $1,000$ annotated images without meaningful loss in detection quality. This finding has direct implications for scaling floating waste detection to new river systems in developing countries, where large-scale annotation campaigns are often prohibitively expensive. The data-efficient nature of the YOLO26m-CW framework, enabled by Coordinate Attention's ability to encode spatial priors and Water-Specific Augmentations' capacity to simulate diverse aquatic conditions, effectively lowers the barrier to entry for deploying automated waste monitoring infrastructure in resource-constrained settings.

5.4 Comparison with Prior Works

To place our results in context with the wider literature, we compared YOLO26m-CW against six prior studies that evaluated models on the FloW-Img dataset: PBM-YOLO [18], A2ANet [17], ES-YOLOv8 [13], USV-YOLO [12], YOLOv12s-P2+SGR+SIoU [10], and YOLOv8-MSS [11]. Because these prior studies utilized different, and often undisclosed, data split configurations (specifically 90-10, 80-20, and 70-30 ratios), we compared each model against our corresponding YOLO26m-CW configuration trained on the matched split ratio. Table 5.5 details the comparative metrics.

Table 5.5: Comparison with prior work using the same split ratios on the FloW-Img dataset. Bold values indicate best per split group.

Model	**Split**	**Precision**	**Recall**	F_1**-Score**	mAP_{50}	**mAP(50–95)**
PBM-YOLO [18]	90-10	0.900	**0.858**	0.879	**0.907**	0.481
YOLO26m-CW (Ours)		**0.959**	0.854	**0.903**	0.905	**0.572**
A2ANet (YOLOv5s+Atrous) [17]	80-20	0.878	0.836	0.857	0.892	0.449
ES-YOLOv8 [13]		0.864	0.815	0.840	0.873	–
USV-YOLO (YOLOv8+ODConv) [12]		0.901	0.748	0.817	0.872	**0.568**
YOLOv12s-P2+SGR+SIoU [10]		0.822	**0.857**	0.839	0.871	0.504
YOLO26m-CW (Ours)		**0.959**	0.852	**0.902**	**0.906**	0.548
YOLOv8-MSS [11]	70-30	–	–	0.850	0.879	0.476
YOLO26m-CW (Ours)		**0.948**	**0.858**	**0.901**	**0.908**	**0.554**

5.4.1 90-10 Split Group

Under the 90-10 split ratio, PBM-YOLO [18] achieves a Precision of 0.900, Recall of 0.858, F_1-Score of 0.879, and an mAP_{50} of 0.907. Our proposed YOLO26m-CW achieves a Precision of 0.959, Recall of 0.854, F_1-Score of 0.903, and an mAP_{50} of 0.905.

While PBM-YOLO marginally leads in Recall by +0.004 and mAP_{50} by +0.002, YOLO26m-CW demonstrates significantly higher Precision (+0.059 increase) and F_1-Score (+0.024 increase). Most notably, our model achieves a major improvement in the stricter mAP_{50-95} metric, scoring 0.572 compared to PBM-YOLO's 0.481 (+0.091 absolute increase). This reveals that while PBM-YOLO has a tiny advantage in detecting candidate regions, it generates a significantly higher rate of false positives and produces less precise bounding box boundaries. YOLO26m-CW's coordinate spatial attention layers produce significantly tighter, more accurate boundary localizations, which is critical for guiding USV mechanical cleanup claws.

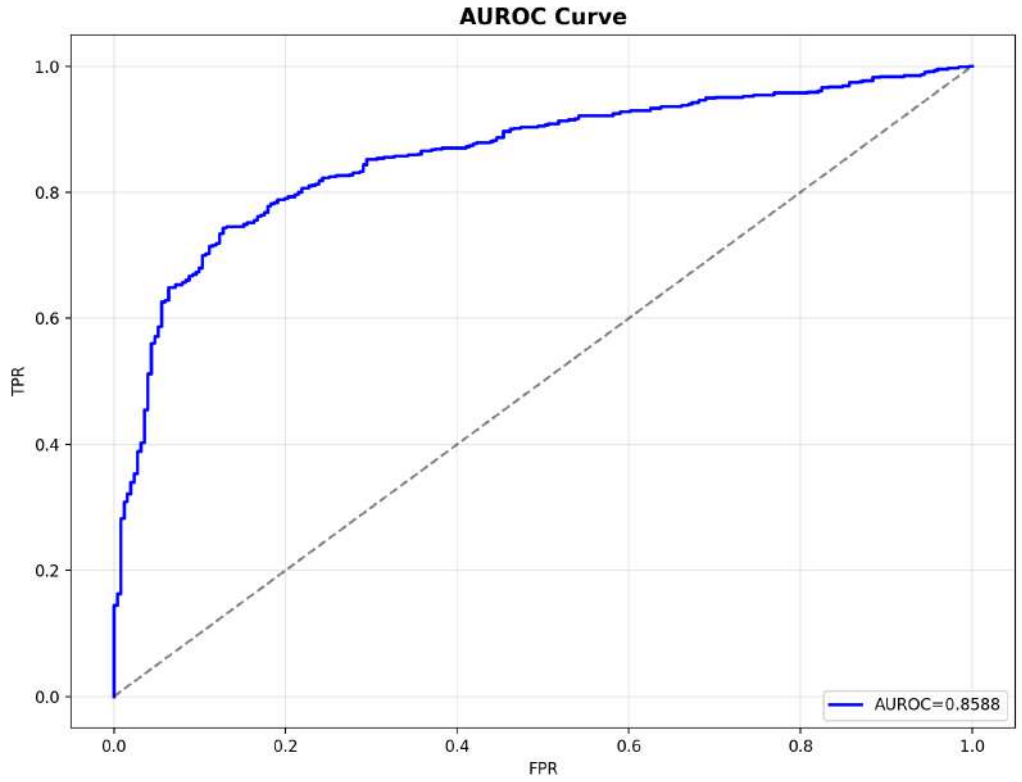


Figure 5.5: ROC curve comparison between YOLO26m-CW and PBM-YOLO under the 90-10 split configuration on the FloW-Img dataset.

The ROC and Precision-Recall curves in Fig. 5.5 and Fig. 5.6 further corroborate these quantitative findings. As shown in Fig. 5.5, YOLO26m-CW maintains a superior ROC envelope with a higher AUROC, demonstrating stronger discrimination capability across all confidence thresholds compared to PBM-YOLO. Similarly, Fig. 5.6 illustrates that YOLO26m-CW achieves a more favorable precision-recall trade-off, sustaining high precision even at elevated recall levels, which confirms the model's ability to detect floating waste objects with minimal false positives under this split configuration.

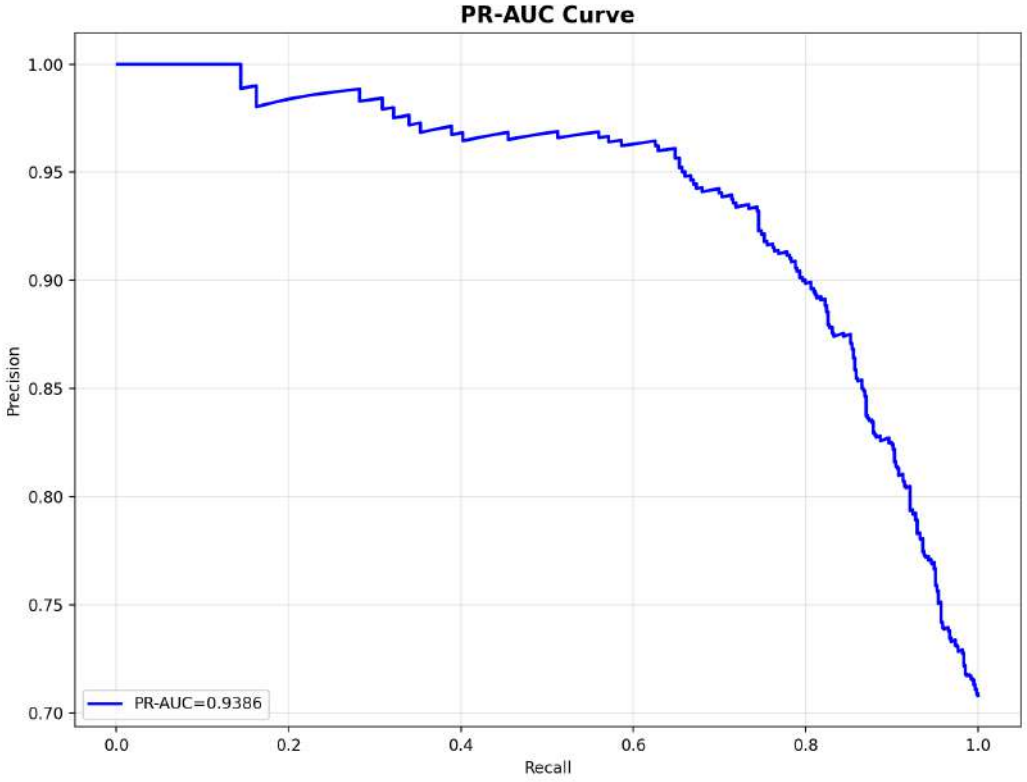


Figure 5.6: Precision-Recall curve comparison between YOLO26m-CW and PBM-YOLO under the 90-10 split configuration on the FloW-Img dataset.

5.4.2 80-20 Split Group

The 80-20 split ratio is the most commonly used configuration in the literature. Our model achieves a Precision of 0.959, Recall of 0.852, F_1-Score of 0.902, and an mAP_{50} of 0.906. It outperforms all four prior works in this group:

- **A2ANet [17]**: YOLO26m-CW improves mAP_{50} by +0.014, Precision by +0.081, and mAP_{50-95} by +0.099.
- **ES-YOLOv8 [13]**: Our framework achieves an increase in mAP_{50} of +0.033, Precision of +0.095, and F_1-Score of +0.062.
- **USV-YOLO [12]**: While USV-YOLO achieves a high mAP_{50-95} of 0.568 due to its Omni-Dimensional Convolutions (ODConv), it suffers from low Recall (0.748) and a weak F_1-Score (0.817). Our model achieves a much higher Recall (0.852, +0.104 higher) and a superior F_1-Score (0.902, +0.085 higher).
- **YOLOv12s-P2+SGR+SIoU [10]**: YOLO26m-CW improves Precision by +0.137, F_1-Score by +0.063, mAP_{50} by +0.035, and mAP_{50-95} by +0.044.

The ROC and Precision-Recall curves in Fig. 5.7 and Fig. 5.8 visually illustrate the superior performance of YOLO26m-CW under the 80-20 split ratio. As shown in Fig. 5.7, our proposed model achieves a wider ROC envelope compared to competing methods, indicating higher true positive rates across low false positive thresholds. Furthermore, Fig. 5.8 confirms that YOLO26m-CW maintains high precision across the entire recall

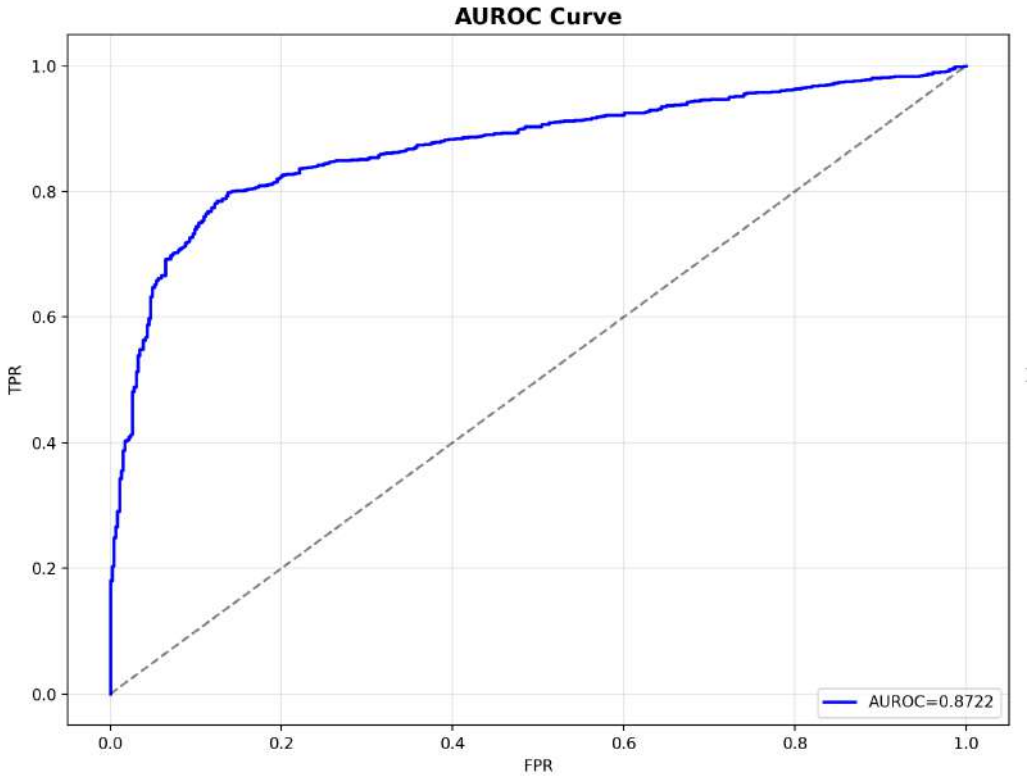


Figure 5.7: ROC curve comparison under the 80-20 split configuration on the FloW-Img dataset.

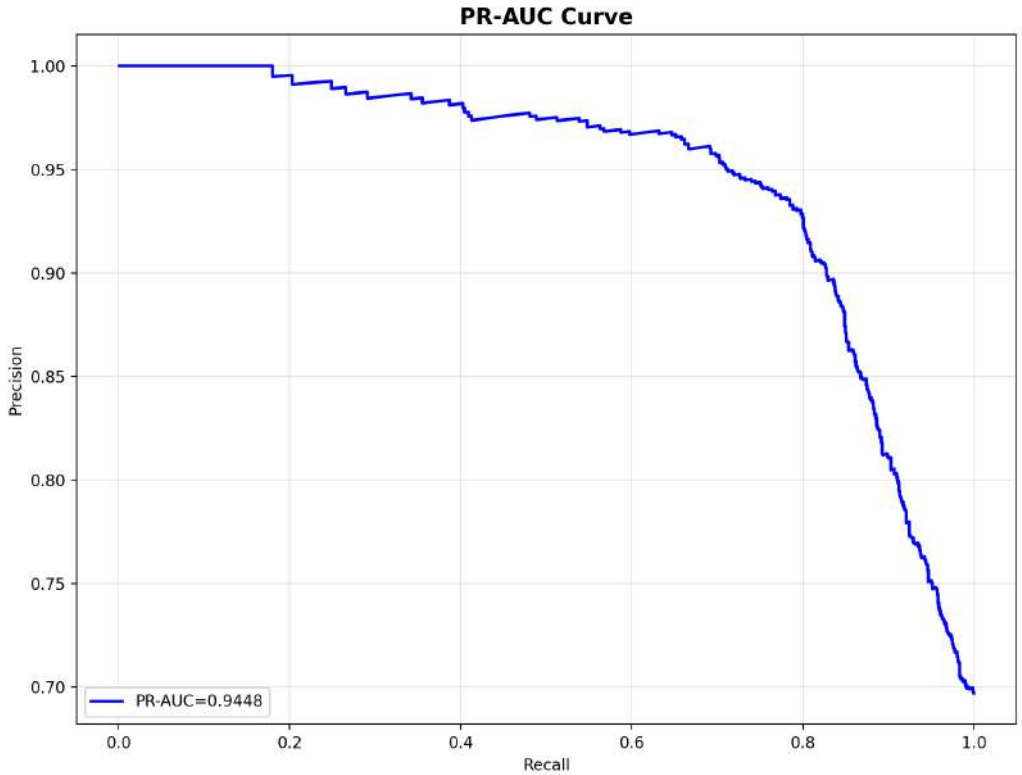


Figure 5.8: Precision-Recall curve comparison under the 80-20 split configuration on the FloW-Img dataset.

spectrum, demonstrating robust confidence calibration and detection reliability under the standard 80-20 evaluation setting.

5.4.3 70-30 Split Group

In the 70-30 split group, YOLOv8-MSS [11] reports an F_1-Score of 0.850, mAP_{50} of 0.879, and mAP_{50-95} of 0.476. Our corresponding YOLO26m-CW model outperforms it across all metrics, yielding a Precision of 0.948, Recall of 0.858, F_1-Score of 0.901 (+0.051 increase), mAP_{50} of 0.908 (+0.029 increase), and an mAP_{50-95} of 0.554 (+0.078 increase). This performance advantage across different splits demonstrates the value of coordinate attention and specialized water augmentations. Prior works that rely on heavy convolutional modifications (like atrous convolutions in A2ANet or dynamic convolutions in USV-YOLO) suffer from poor data utilization or high false positive rates under aquatic noise. By combining spatial attention blocks (C2PSA and CoordAtt) with targeted augmentations, YOLO26m-CW consistently achieves high precision and localization accuracy. The ROC and Precision-Recall curves in Fig. 5.9 and Fig. 5.10

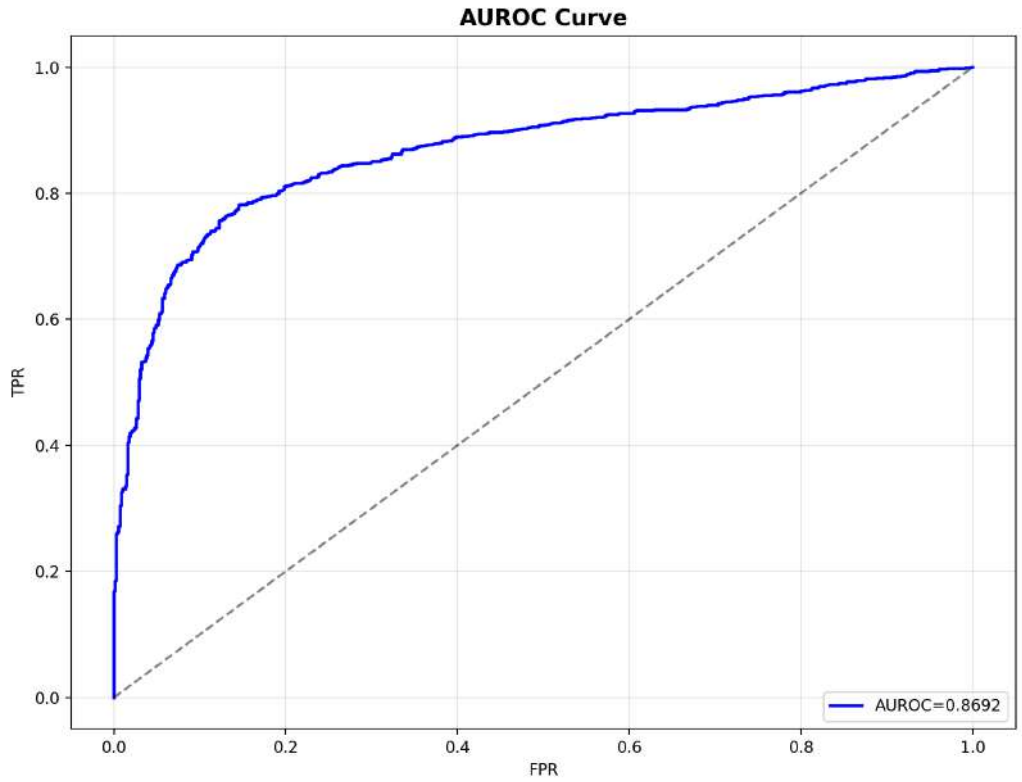


Figure 5.9: ROC curve comparison under the 70-30 split configuration on the FloW-Img dataset.

further illustrate the comparative advantages of YOLO26m-CW under the 70-30 split setup. As depicted in Fig. 5.9, our model achieves a consistently higher ROC trajectory across varying classification thresholds, yielding superior overall AUROC. Furthermore, Fig. 5.10 highlights that YOLO26m-CW maintains higher precision across recall levels compared to baseline models, confirming robust detection performance even under a reduced training split ratio.

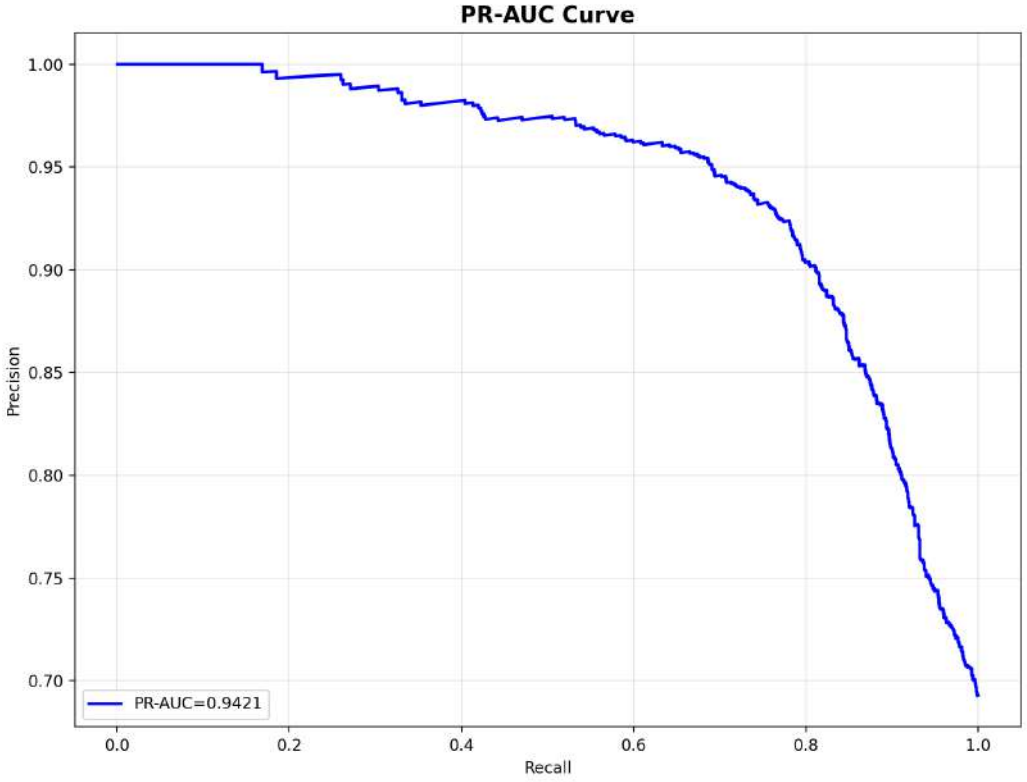


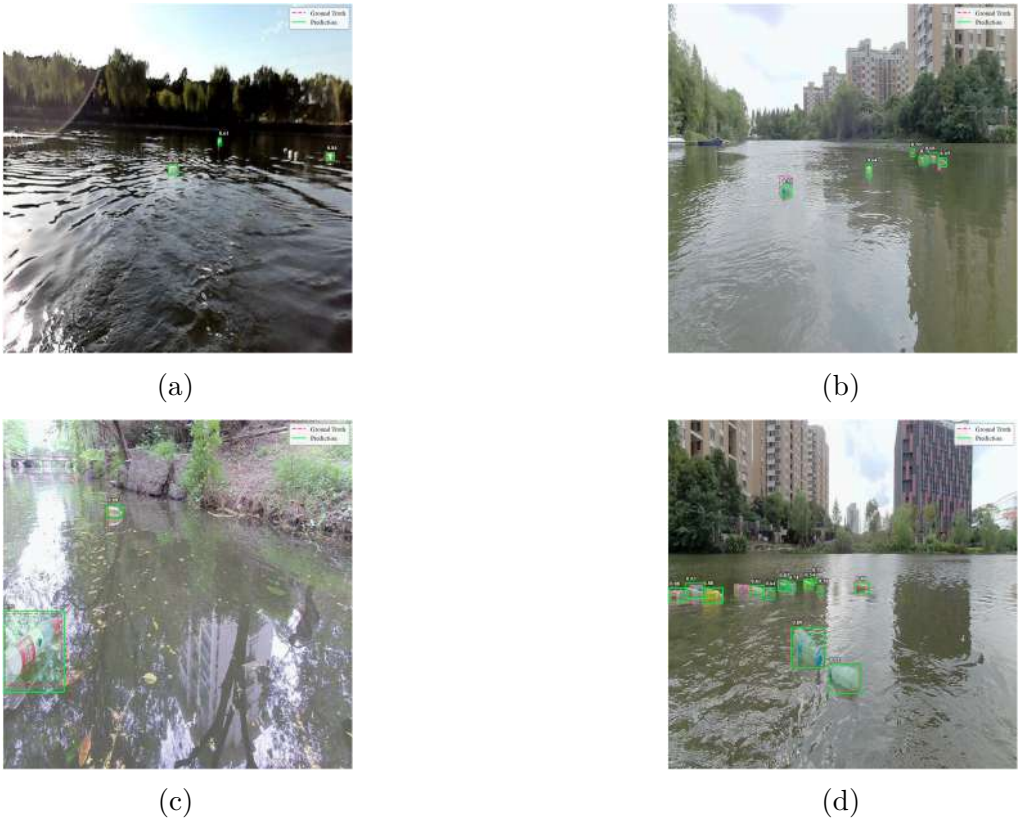
Figure 5.10: Precision-Recall curve comparison under the 70-30 split configuration on the FloW-Img dataset.

5.5 Qualitative Detection Results

To evaluate the real-world operational reliability of the proposed YOLO26m-CW model, we conducted a qualitative validation of its inference outputs across four extreme riverine conditions: sun glare, small objects, bank-side clutter, and clustered waste.

Fig. 5.11 displays representative detection examples for the flagship YOLO26m-CW model.

- **Sun Glare (Fig. 5.11a):** Specular sun reflection on the water surface saturates the camera sensor, creating high-intensity glare regions. Standard detectors often generate false positive detections within these bright areas or miss objects that float near them. As shown in the figure, YOLO26m-CW successfully filters out the glare background and detects the plastic bottle floating near the edge of the glare. This robustness is mainly attributed to the HSV and value perturbations in our WaterAugs strategy, which trains the model to remain invariant to intensity spikes.
- **Small Objects (Fig. 5.11b):** Floating waste objects captured at distance appear as tiny, low-contrast pixel clusters near the horizon line. Conventional downsampling layers discard these fine spatial features, resulting in false negatives. By embedding positional details via CoordAtt, our model retains spatial coordinate paths, enabling it to detect small objects at distance.
- **Bank-side Clutter (Fig. 5.11c):** Floating waste that accumulates near the shoreline is highly challenging to detect because it blends with bank-side vegetation,



(a) (b)

(c) (d)

Figure 5.11: Qualitative validation of YOLO26m-CW showcasing inference extremes: **(a)** sun glare, **(b)** small objects, **(c)** bank-side clutter, and **(d)** clustered objects.

rocks, and mud. The model successfully distinguishes between the floating plastic bottle and the shoreline background, demonstrating high semantic discrimination.

- **Clustered Objects (Fig. 5.11d):** In accumulation zones, waste objects pile up in dense, overlapping clusters. Standard anchor-based detectors tend to merge these into a single box or miss occluded objects. YOLO26m-CW's anchor-free head, combined with Mixup training, allows it to detect overlapping and adjacent objects individually with precise bounding boundaries.

These qualitative results demonstrate that YOLO26m-CW successfully handles key real-world environmental challenges, confirming its suitability for deployment on autonomous river cleanup vessels.Furthermore, the model maintains stable detection performance under varying illumination and complex background conditions. These results highlight the effectiveness of combining WaterAugs, CoordAtt, and the anchor-free detection head in addressing diverse riverine scenarios. The consistent localization of small, partially occluded, and visually ambiguous waste demonstrates the model's strong generalization capability. Overall, YOLO26m-CW provides a reliable and robust detection framework for practical autonomous river-cleanup applications.

Chapter-6: Discussion

The development of automated, intelligent systems for riverine floating-waste monitoring is fundamentally constrained by the dynamic and unpredictable nature of aquatic environments. Deploying deep learning models on Unmanned Surface Vehicles (USVs) introduces severe operational challenges, including surface hydrodynamic turbulence, intense sun glare, complex water reflections, shoreline vegetation clutter, and small object resolutions near the horizon. In response to these challenges, this study proposed YOLO26m-CW, a modular detection framework integrating Coordinate Attention (CoordAtt) and Water-Specific Augmentations (WaterAugs) on top of the YOLO26m backbone.

This chapter provides a comprehensive synthesis of the experimental findings presented in Chapter 5. To systematically evaluate the contributions of this research, the discussion is organized around the five core Research Questions (RQs) introduced in Chapter 1, directly justifying each question using empirical evidence from our component selection, benchmark comparisons, split sensitivity analysis, and external domain evaluations.

6.1 Performance Impact of Coordinate Attention

Research Question 1 sought to determine whether integrating Coordinate Attention (CoordAtt) into the YOLO26 architecture improves floating waste detection in riverine environments, and through which specific metrics (precision, recall, or both) this improvement manifests.

Our empirical findings from the two-stage ablation study (Table 5.1) confirm that Coordinate Attention provides a substantial performance improvement, operating primarily as a *recall-oriented* architectural enhancement. In Stage 1 on the lightweight YOLO26s backbone, adding CoordAtt (`YOLO26s + CoordAtt`) increased Recall from 0.779 to 0.797 (an absolute increase of +0.018), while overall mean-mean Average Precision ($mmAP$) improved from 0.657 to 0.662. In Stage 2 on the medium-capacity YOLO26m backbone, integrating CoordAtt (`YOLO26m + CoordAtt`) elevated Precision to 0.914, Recall to 0.804, the F_1-Score to 0.856, and Area Under the ROC Curve (AUROC) from 0.849 to 0.863.

The mechanistic justification for this recall gain lies in how CoordAtt handles spatial feature downsampling. Conventional channel attention mechanisms (such as Squeeze-and-Excitation) employ global average pooling, which compresses 2D spatial feature maps into single scalar descriptors per channel. While effective for global channel recalibration, global pooling completely discards spatial coordinate information. In aquatic environments,

floating waste targets, such as partially submerged plastic fragments or distant debris captured near the horizon, frequently occupy only a minute pixel footprint. When processed through standard downsampling layers, the spatial signals of these small objects are rapidly lost.

By contrast, Coordinate Attention embeds directional spatial location by decomposing 1D global pooling along the horizontal and vertical axes into two spatial-aware feature maps. Preserving these horizontal and vertical positional cues allows the feature pyramid network to retain spatial coordinate paths even after deep convolutional downsampling. Consequently, the network can locate small, low-contrast, or boundary-located waste items that standard baselines miss, directly driving the observed increase in Recall.

6.2 Complementary Synergy of Water-Specific Augmentations

Research Question 2 investigated whether environment-targeted water-specific augmentations provide complementary benefits to architectural attention mechanisms, and explored the nature of their combined effect on detection accuracy.

The experimental results establish that Water-Specific Augmentations (WaterAugs) act as a *precision-oriented* regularizer that complements the recall-oriented nature of CoordAtt. In Stage 1 of the ablation study (`YOLO26s + WaterAugs`), Precision reached 0.914 (+0.004 over baseline) and mAP_{50} rose to 0.853. In Stage 2 (`YOLO26m + WaterAugs`), Precision peaked at 0.920 (+0.016 over baseline), establishing an F_1-Score of 0.858 and an $mmAP$ of 0.669.

When combined in the finalized **YOLO26m-CW** framework (`YOLO26m + CoordAtt + WaterAugs`), the two modules exhibit clear mathematical synergy. Aquatic visual noise, such as specular highlights from sun glare, wave crest reflections, and shifting water turbidity, frequently generates false positive detections in standard models by mimicking the reflective surface of plastic debris. WaterAugs exposes the network during training to simulated environmental perturbations (including HSV water clarity shifts, vertical reflection flips, subtle rotations for wave-induced camera tilt, and scale jitter). This forces the model to decouple the true semantic features of floating waste from background hydrodynamic artifacts.

While CoordAtt captures more hidden or small targets (driving high Recall), WaterAugs trains the network to reject false positive environmental noise (driving high Precision). Together, YOLO26m-CW achieves the highest overall F_1-Score (0.859), mAP_{50} (0.866), mAP_{50-95} (0.478), $mmAP$ (0.672), and AUROC (0.871) in the entire component matrix. As visually demonstrated in the PR and ROC curves (Fig. 5.1 and Fig. 5.2), this combination effectively resolves the classic precision-recall bottleneck in aquatic vision.

6.3 Sensitivity to Data Split Variations and Data Efficiency

Research Question 3 evaluated the sensitivity of YOLO26m-CW's detection performance to variations in train-test data split ratios, testing whether the framework maintains consistent performance across diverse split configurations.

Our split sensitivity analysis across five distinct train-test partitions (50-50, 60-40, 70-30, 80-20, and 90-10; Table 5.4) confirms that YOLO26m-CW exhibits remarkable stability and minimal sensitivity to data split variations. The primary detection metric, mAP_{50}, remained virtually constant across all splits, ranging between a minimum of 0.905 (in the 90-10 split) and a maximum of 0.909 (in the 60-40 split), with a negligible standard deviation of only ± 0.0016. Similarly, Precision remained between 0.943 and 0.959, Recall between 0.852 and 0.869, and the F_1-Score was sustained above 0.900 across all configurations (standard deviation of ± 0.0033). This low variance provides statistical proof that the reported performance is not an artifact of a single favorable data split.

Furthermore, the results reveal a critical performance plateau beyond the 60-40 split ratio. Reducing the training set size by 40% down to a 50-50 split (a mere $1,000$ training images) yielded an mAP_{50} of 0.907 and an F_1-Score of 0.901, matching the performance of models trained on 80% ($1,600$ images) and 90% ($1,800$ images) of the dataset.

This high level of data efficiency carries major practical implications for real-world deployment. Annotating complex aquatic scenes containing partially submerged debris is labor-intensive and expensive. The ability of YOLO26m-CW to reach convergence and high accuracy with only $1,000$ annotated images indicates that WaterAugs synthesizes sufficient environmental variance to prevent overfitting on small training samples. Consequently, environmental agencies can bootstrap deployable models in new geographic river basins without needing massive, cost-prohibitive annotation campaigns.

6.4 Comparative Benchmarking Against State-of-the-Art Methods

Research Question 4 addressed how YOLO26m-CW compares against existing state-of-the-art floating waste detection methods when evaluated under matched split ratios and identical benchmark conditions.

Our comparative benchmarking was conducted across two evaluation dimensions: direct comparison against the original FloW-Img baselines and matched-split evaluations against six recent state-of-the-art models (Table 5.2 and Table 5.5).

1. **Original Baselines (60-40 Split):** Compared to the multi-stage baselines in Cheng et al. [14], YOLO26m-CW achieved an $mmAP$ of 0.672, surpassing Faster R-CNN ($mmAP = 0.184$) by +0.488 (+265%) and Cascade R-CNN ($mmAP = 0.434$) by +0.238 (+54.8%). Older multi-stage detectors rely on fixed anchor grids that fail

to capture the arbitrary aspect ratios of riverine debris, and lack spatial attention mechanisms to preserve small target features during downsampling.

2. **Matched SOTA Comparisons (90-10, 80-20, 70-30 Splits):**
    - *90-10 Split Group:* Compared to PBM-YOLO [18] ($mAP_{50} = 0.907$, $mAP_{50-95} = 0.481$), YOLO26m-CW ($mAP_{50} = 0.905$, $mAP_{50-95} = 0.572$) demonstrated a massive +0.091 absolute increase in the strict mAP_{50-95} metric, alongside higher Precision (0.959 vs. 0.900) and F_1-Score (0.903 vs. 0.879). While PBM-YOLO achieves high region candidate proposal rates, its bounding boxes are loose. YOLO26m-CW generates significantly tighter, more accurate boundary localizations, which is vital for guiding physical USV cleanup claws.
    - *80-20 Split Group:* YOLO26m-CW ($mAP_{50} = 0.906$, $F_1 = 0.902$) outperformed A2ANet [17] ($mAP_{50} = 0.892$), ES-YOLOv8 [13] ($mAP_{50} = 0.873$), USV-YOLO [12] ($mAP_{50} = 0.872$, Recall = 0.748), and YOLOv12s-P2+SGR+SIoU [10] ($mAP_{50} = 0.871$). It established top performance across Precision (0.959), F_1-Score (0.902), and mAP_{50} (0.906).
    - *70-30 Split Group:* Compared to YOLOv8-MSS [11] ($mAP_{50} = 0.879$, $F_1 = 0.850$, $mAP_{50-95} = 0.476$), YOLO26m-CW achieved superior results across all metrics ($mAP_{50} = 0.908$, $F_1 = 0.901$, $mAP_{50-95} = 0.554$).

These matched comparisons confirm that YOLO26m-CW consistently sets the state-of-the-art performance across all data split ratios, proving the clear advantage of pairing an anchor-free single-stage head with spatial coordinate attention and water-targeted augmentations.

6.5 Cross-Domain Generalization and Domain Necessity

Research Question 5 investigated the extent to which models trained on alternative waste detection datasets can generalize to the FloW-Img riverine benchmark, and what this reveals regarding the necessity of domain-specific training data.

The external evaluation experiments (Table 5.3) exposed severe generalization limitations when training models on external non-riverine datasets and evaluating them on the FloW-Img test set (800 images):

- **UAV-BD (Aerial Perspective):** When trained on top-down aerial drone imagery, the baseline from Cheng et al. recorded an $mmAP$ of 0.058. While YOLO26m-CW improved this to 0.228 (3.9× relative gain), the absolute score remains severely degraded compared to native FloW-Img training ($mmAP = 0.672$). High-altitude top-down imagery lacks horizon perspective, shoreline clutter, and horizontal specular glare.
- **MJU-Waste (Indoor RGBD):** Training on indoor static waste imagery yielded an extremely low $mmAP$ of 0.081 for YOLO26m-CW (compared to 0.032 baseline).

The total absence of aquatic surface hydrodynamics, reflections, and natural light shifts prevents indoor models from transferring to natural river environments.

- **TUD-GV (Urban Canal):** Training on our newly introduced urban canal dataset resulted in an $mmAP$ of 0.187. Although captured in water, structured urban canals lack the complex wave dynamics, heavy turbidity, and natural bank-side vegetation characteristic of natural river basins.

These findings provide a clear answer to RQ5: models trained on alternative, non-riverine datasets experience catastrophic performance degradation when deployed on surface-level river imagery. This empirical degradation confirms that specialized, surface-level riverine datasets like FloW-Img, captured from the exact vantage point and environmental noise profile of operational USVs, are strictly necessary for training deployable floating waste monitoring systems.

6.6 Real-World Deployment Feasibility and System Integration

Beyond quantitative metrics, qualitative evaluation (Fig. 5.11) confirms that YOLO26m-CW maintains operational reliability under severe environmental edge cases. The model accurately filters out intense specular sun glare (Fig. 5.11a), detects distant low-contrast targets near the horizon (Fig. 5.11b), distinguishes waste from bank-side vegetation (Fig. 5.11c), and separates tightly packed debris in accumulation zones (Fig. 5.11d).

From a software engineering perspective, YOLO26m-CW's single-stage, lightweight architecture is ideally suited for low-power edge computing devices (such as NVIDIA Jetson modules embedded on autonomous USVs or bridge-mounted camera arrays). In a fully realized intelligent environmental monitoring system, these edge nodes perform real-time local inference without requiring continuous high-bandwidth cloud video streaming. The edge nodes transmit compact inference telemetry (GPS coordinates, waste count, class, and timestamp) via RESTful APIs to centralized Geographic Information System (GIS) platforms. This facilitates real-time pollution density heatmaps, enabling environmental agencies to proactively track waste accumulation hotspots and dispatch targeted autonomous cleanup vessels.

Chapter-7: Conclusion

7.1 Summary

This study proposes YOLO26m-CW, an intelligent detection framework for riverine floating-waste monitoring and management. Integrating Coordinate Attention and environment-targeted augmentations as modular, configurable components proved highly effective, surpassing the FloW-Img benchmark by +0.238 mmAP. Split sensitivity was evaluated across five configurations ($mAP_{50} > 0.90$), while external evaluation quantified the strict performance penalty when training on alternative datasets, demonstrating the necessity of the FloW-Img dataset for real-world deployment.

7.2 Analysis of Social Effects

Deploying autonomous floating-waste monitoring systems such as YOLO26m-CW yields multi-faceted societal benefits, particularly for communities reliant on inland aquatic ecosystems:

- **Public Health and Water Sanitation:** Inland waterways in developing and rapidly industrializing regions serve as essential sources for drinking water, agricultural irrigation, and municipal supply. Uncontrolled macro-plastic accumulation leads to chemical leaching, microplastic bioaccumulation, and stagnant water zones that foster vector-borne diseases. Autonomous detection and early interception safeguard municipal water quality and reduce environmental health risks.
- **Economic Security for Maritime Communities:** Floating debris severely disrupts local riverine economies by entangling vessel propellers, damaging fishing nets, clogging drainage infrastructure, and inducing urban flooding. Real-time visual monitoring minimizes operational downtime for local fisheries, lowers municipal cleanup expenditures, and protects riverfront ecotourism.
- **Data-Driven Civic Governance:** Our proposed YOLO26m-CW framework provides a lightweight, highly accurate vision engine suitable for edge deployment on USV-mounted hardware (e.g., NVIDIA Jetson) or stationary bridge nodes. By effectively suppressing false alarms from glare and wave ripples using WaterAugs, the framework produces reliable, geotagged detection metadata (GPS coordinates, waste counts, confidence scores) in real time. Transmitting this high-fidelity telemetry to

cloud-based Geographic Information Systems (GIS) enables municipal authorities, environmental regulators, and NGOs to build transparent spatial waste density heatmaps, identify illegal dumping hotspots, optimize cleanup fleet dispatch logistics, and formulate data-driven anti-pollution policies.

- **Occupational Safety and Labor Transformation:** Traditional manual river cleanup forces workers into hazardous environments involving toxic chemical leaching, biological pathogens, sharp floating debris, and drowning risks. By serving as an accurate, real-time perception engine, YOLO26m-CW empowers autonomous Unmanned Surface Vehicles (USVs) to reliably detect, track, and navigate toward floating waste objects even in complex aquatic conditions (such as small distant targets and shoreline clutter). Automating the hazardous detection and interception pipeline eliminates the necessity for direct human contact with polluted waters, transforming dangerous manual sanitation jobs into safer, higher-skilled technical roles focused on USV fleet supervision, remote operations, hardware maintenance, and spatial data analytics.

7.3 Environmental Impact Analysis and Sustainability

The proposed YOLO26m-CW framework can contribute to environmental sustainability by enabling automated, data-driven monitoring of floating waste in riverine environments. Rather than directly performing waste removal, the system provides an intelligent perception layer that can support earlier detection, targeted intervention, and continuous environmental monitoring.

- **Early Detection of Riverine Waste:** Rivers are major pathways through which plastic waste reaches marine environments. YOLO26m-CW is designed to detect floating waste directly from surface-level river imagery, including challenging conditions involving small objects, partial occlusion, sun glare, reflections, and clustered waste. By identifying waste at an earlier stage, the system can provide timely information for subsequent collection or intervention, potentially reducing the amount of floating debris transported downstream.
- **Supporting Targeted Waste Collection:** Continuous visual monitoring can help identify the location and density of floating waste rather than relying exclusively on periodic manual inspection. The proposed system can generate detection results that indicate where waste is concentrated, allowing an autonomous or remotely operated waste-collection platform to prioritize high-density regions. This can reduce unnecessary movement and make environmental intervention more targeted.
- **Reduction of Unnecessary Operational Activity:** Accurate detection is important for autonomous environmental-monitoring systems because false detections may cause unnecessary intervention. YOLO26m-CW combines Coordinate Attention with Water-Specific Augmentations to improve detection under riverine visual disturbances. In particular, the augmentation strategy improves precision, while the attention mechanism contributes to improved recall. More reliable predictions

can therefore help downstream systems distinguish potential waste targets from environmental noise such as reflections, waves, and clutter, reducing unnecessary navigation or collection actions.

- **Continuous and Data-Driven Environmental Monitoring:** The framework can transform individual waste detections into structured environmental information. As described in the proposed intelligent monitoring architecture, inference results can be transmitted through RESTful APIs to centralized Geographic Information System (GIS) platforms. Aggregating these observations over time can support waste-density heatmaps and threshold-based alerts, enabling environmental authorities or collection operators to identify recurring pollution hotspots and prioritize intervention.
- **Potential for Edge-Based Deployment:** YOLO26m-CW is developed as a lightweight single-stage detection component suitable for integration into edge-based monitoring systems such as USVs or fixed bridge-mounted cameras. Performing inference near the data source can reduce dependence on continuous transmission of raw video to a centralized server. This provides a practical pathway toward persistent monitoring while limiting communication requirements. However, actual inference latency, power consumption, and carbon savings were not directly measured in this study and therefore require dedicated field evaluation.
- **Sustainable Environmental Intervention:** The proposed system supports a shift from reactive to preventive environmental management. Instead of waiting for accumulated waste to become visually apparent through manual inspection, automated monitoring can continuously identify potential pollution events and provide location-based information for intervention. In the longer term, integrating these detections with autonomous collection systems, GIS platforms, and environmental management infrastructure could support more efficient river-cleaning operations.

Overall, YOLO26m-CW should be viewed as an intelligent monitoring and decision-support component rather than a complete waste-removal solution. Its environmental contribution comes from enabling continuous detection, localization, prioritization, and reporting of floating waste. The experimental results demonstrate strong detection performance on the FloW-Img benchmark, while future field deployment should evaluate real-world inference latency, energy consumption, collection efficiency, and long-term environmental benefits.

7.4 Ethical Issues

The practical deployment of computer vision and autonomous robotics in public waterways raises critical ethical considerations that must govern real-world implementation:

- **Privacy and Visual Surveillance:** Continuous visual monitoring via USV-mounted or bridge-mounted cameras inevitably captures footage of public riverbanks, ferry crossings, and waterfront residents. To preserve civil liberties, deployed architectures must enforce strict privacy-by-design principles: raw video streams should undergo immediate local edge processing to extract only anonymized bounding-box metadata (GPS coordinates, waste class, timestamps), with raw visual frames immediately purged without persistent storage or cloud transmission.

- **Algorithmic Bias and Geographic Equity:** The YOLO26m-CW models trained on specific geographical regions may exhibit performance degradation when deployed in non-native waterways with differing water color, turbidity, or light conditions, as demonstrated by our external evaluation findings. Deploying uncalibrated models in underrepresented developing regions risks unequal environmental service distribution or false security. Transparent domain disclosures and continuous localized model fine-tuning are ethically imperative.
- **Operational Safety of Autonomous Systems:** Deploying autonomous cleanup vessels in shared waterways introduces physical collision hazards to local fishing boats, bathers, and aquatic fauna. Systems must operate under strict ethical safety protocols, featuring multi-sensor fail-safes, automatic emergency stop mechanisms, dynamic obstacle avoidance, and human-in-the-loop oversight.
- **Accountability in Automated Pollution Enforcement:** Utilizing automated AI detection data and GIS heatmaps for regulatory enforcement or identifying municipal dumping violations carries risks of misclassification. Ethical governance requires that automated AI outputs serve as decision-support indicators rather than sole evidence for legal or financial penalties, safeguarding vulnerable riverside communities from unfair enforcement.

7.5 Limitations

Despite the strong performance observed across the experiments, several limitations remain and should be addressed in future research. First, the evaluation relied primarily on single train–test splits, which may not fully capture variations in data distribution or generalization performance. Future studies should therefore employ K-fold cross-validation to provide more robust and statistically reliable performance estimates. Second, although the proposed framework is designed for deployment on unmanned surface vehicles (USVs), operational edge latency and computational efficiency were not explicitly benchmarked under realistic onboard conditions. Evaluating inference speed, memory consumption, and energy requirements on representative edge hardware would provide a more complete assessment of deployment feasibility. Third, although YOLO26m-CW demonstrated up to a 3.9× improvement over the original baselines during external evaluation, the consistently low absolute performance across all evaluated models highlights a significant dataset dependency. This suggests that the models struggle to generalize when predicting complex surface-level hydrodynamic phenomena from non-riverine imagery. Differences in water appearance, environmental conditions, illumination, camera perspectives, and flow characteristics may further contribute to this limitation. Consequently, larger and more diverse datasets collected across multiple rivers, seasons, weather conditions, and operational environments are required to improve generalization. Future work should also investigate domain adaptation and transfer-learning strategies to reduce the sensitivity to dataset-specific characteristics. Overall, these limitations indicate that further validation under diverse real-world conditions is necessary before the proposed approach can be considered fully reliable for practical USV-based hydrodynamic monitoring.

7.6 Future Work

Moving forward, research priorities include addressing the current gaps in the proposed framework and further developing its applicability in practical situations. First, K-fold cross-validation will be used to get a more complete idea of the robustness of the model and check its distributional robustness across data folds. This will help to identify if the proposed framework performs consistently on multiple trainsplit and test splits and minimizes the chance of fallacy results in evaluation. Further testing will be undertaken with differing environmental setting, such as different types of lighting, weather, water surface conditions, background level/mixing (background complexity), camera angle (camera viewpoint) and scale (of objects). The evaluations will help gain insight into the robustness of the model in a challenging real world river context.

In addition, we will measure and compare edge-inference latency, memory usage, power efficiency and computational performance for the YOLO26m-CW framework on actual embedded devices. It will assist in assessing its suitability for use in unmanned surface vehicles (USVs) in resource limited environments. In addition, model optimisation techniques like pruning, quantization, knowledge distillation and light weight network structure can also be explored to keep the detection accuracy intact but boost the inference speed. The future may hold additional research for adaptive detection strategies that could optimize model performance based on the environment and computational resources.

Further value-added developments will be along the lines of creation of an autonomous robotic system for real-time plastic waste detection. The proposed solution includes a camera system equipped with a vision module and a trained detection model that will be integrated into the system to automatically detect and locate plastic waste in real-time aquatic scenes. These detected objects will then be used to coordinate the Robot navigation to targeted location of the waste. In future, the robot will also be integrated with other technology such as GPS, sensors, obstacle avoidance and intelligence control system to ensure reliable and safe operations of the robot. In the long run, the system could be expanded by having an automated collection system to facilitate plastic waste detection, localization, tracking and retrieval in an effort to reduce the human involvement. Such advancements will help to establish smarter, more autonomous and environmentally-conscious systems for monitoring and collection for waste, aiding in more effective river-cleaning, and long term protection of the aquatic environment.

Bibliography

[1] L. C. Lebreton, J. Van Der Zwet, J.-W. Damsteeg, B. Slat, A. Andrady, and J. Reisser, "River plastic emissions to the world's oceans," *Nature Communications*, vol. 8, no. 1, p. 15611, 2017.

[2] 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, no. 18, p. eaaz5803, 2021.

[3] T. van Emmerik and A. Schwarz, "Plastic debris in rivers," *Wiley Interdisciplinary Reviews: Water*, vol. 7, no. 1, p. e1398, 2020.

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

[5] H. Lee, S. Byeon, J. H. Kim, J. K. Shin, and Y. Park, "Construction of a real-time detection for floating plastics in a stream using video cameras and deep learning," *Sensors*, vol. 25, no. 7, p. 2225, 2025.

[6] T. Wang, Y. Cai, L. Liang, and D. Ye, "A multi-level approach to waste object segmentation," *Sensors*, vol. 20, no. 14, p. 3816, 2020.

[7] T. Jia, R. Taormina, R. de Vries, Z. Kapelan, T. H. van Emmerik, P. Vriend, and I. Okkerman, "A semi-supervised learning-based framework for quantifying litter fluxes in river systems," *Water Research*, vol. 289, p. 124833, 2026.

[8] J. Wang, W. Guo, T. Pan, H. Yu, L. Duan, and W. Yang, "Bottle detection in the wild using low-altitude unmanned aerial vehicles," in *2018 21st International Conference on Information Fusion (FUSION)*. IEEE, 2018, pp. 439–444.

[9] N. Singh and A. Kumar, "Automated marine debris detection and segmentation using yolov11 and deeplabv3 for marine imagery," *International Journal of Marine Engineering and Technology*, vol. 6, pp. 1–14, 2025.

[10] Y. Zhu, M. Zhang, and T. Qiu, "A lightweight and accurate detector for small floating objects in complex water surface environments," *SSRN Electronic Journal*, 2025, preprint.

[11] J. Wang and H. Zhao, "Improved yolov8 algorithm for water surface object detection," *Sensors*, vol. 24, no. 15, p. 5059, 2024.

[12] X. Yang, Y. Song, L. He, H. Xue, Z. Dong, and Q. Zhang, "Usv-yolo: an algorithm for detecting floating objects on the surface of an environmentally friendly unmanned vessel," *IAENG International Journal of Computer Science*, vol. 52, no. 3, pp. 579–588, 2025.

[13] J. Di, K. Xi, and Y. Yang, "An enhanced yolov8 model for accurate detection of solid floating waste," *Scientific Reports*, vol. 15, no. 1, p. 25015, 2025.

[14] Y. Cheng, J. Zhu, M. Jiang, J. Fu, C. Pang, P. Wang, K. Sankaran, O. Onabola, Y. Liu, D. Liu *et al.*, "Flow: A dataset and benchmark for floating waste detection in inland waters," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 953–10 962.

[15] Q. Tang, N. Yuan, T. Li, P. Hu, and H. Ding, "Utd-yolo: an improved underwater trash detection method based on yolo11," in *International Conference on Advances in Computer Vision Research and Applications (ACVRA 2025)*, vol. 13690. SPIE, 2025, pp. 149–154.

[16] Z. Jiang, B. Wu, L. Ma, H. Zhang, and J. Lian, "Apm-yolov7 for small-target water-floating garbage detection based on multi-scale feature adaptive weighted fusion," *Sensors*, vol. 24, no. 1, p. 50, 2023.

[17] B. Badams, U. U. Sheikh, N. A. Wahab, S. A. A. Bakar, M. I. Masud, M. Khouj, U. Waheed, Z. A. Arfeen, and K. M. Goh, "A2anet: Real-time detection of floating marine debris using atrous convolution and channel attention," *Ecological Informatics*, vol. 94, p. 103620, 2026.

[18] Y. Hou and Y. Yu, "Pbm-yolo: A performance balanced floating garbage detection model for water surface environments," *IET Image Processing*, vol. 19, no. 1, p. e70248, 2025.

[19] 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, p. 1298465, 2023.

[20] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 13 713–13 722.

[21] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[22] J. Smaron, M. Alam, and M. T. Islam, "Deep learning on dental rgb and opg images for public health and clinical decision support," Ph.D. dissertation, Independent University, Bangladesh, 2025.

[23] OpenAI, "Chatgpt," AI-generated images using ChatGPT, 2026, images generated with OpenAI's ChatGPT.