Journal of Applied Engineering and Technological Science
Vol 7(2) 2026: 1015-1044

YOLOV11-LCA: YOLOV11 ENHANCED WITH THE LOW-COMPLEXITY ATTENTION MECHANISMS FOR A ROBUST WATERWAY-FLOATING TRASH DETECTION

Muhammad Rafly Arjasubrata[1], Mahmud Dwi Sulistiyo[2*]
School of Computing, Telkom University, Bandung 40257, Indonesia[1,2]
muhammadrafly@student.telkomuniversity.ac.id[1],
mahmuddwis@telkomuniversity.ac.id[2*]

Received: 11 June 2025, Revised: 30 November 2025, Accepted: 10 January 2026
**Corresponding Author*

ABSTRACT
Accurate trash detection in aquatic environments remains a significant challenge for detection models, which exhibit persistent limitations in identifying small and partially submerged objects. Additionally, a notable gap exists in methodologies for fine-tuning the detection model to optimize performance for a specific waterways. To address these limitations, the first objective is to develop a detection model designed to enhance performance on small and partially submerged trash, and the second is to establish a framework for efficiently adapting the model to achieve high accuracy within local waterways. First, the YOLOv11 architecture is enhanced by integrating LCAM and LCBHAM attention mechanisms and pre-trained on various combinations of public datasets to establish a robust, baseline model. For the second objective, this baseline model is adapted using a data-efficient framework. This study process introduces the BojongTrash dataset, captured from a specific waterway, and involves systematically fine-tuning the model on incremental subsets of this data to determine the minimum quantity of images and training epochs required to achieve high accuracy in the target environment. The proposed YOLOv11s-LCA architecture demonstrated a statistically validated improvement over its baseline, increasing the mAP_{50} score from 0.779 to 0.836 on the FloW-Img dataset with only a 0.1% parameter increase. Furthermore, the research establishes a highly efficient fine-tuning framework, demonstrating peak mAP_{50} performance of 0.908 that achieved by fine-tuning on 1,000 images for only 3-5 epochs. Therefore, this research validates lightweight attention mechanisms as an efficient strategy for enhancing detection in complex environments and provides a practical framework that enables the rapid deployment of tailored, high-accuracy monitoring systems.
Keywords : *Floating Trash Detection, YOLOv11, Attention Mechanism, Deep Learning, Waterway Monitoring.*

1. Introduction

The pollution of aquatic environments by floating trash is an urgent global environmental crisis, posing significant risks to marine ecosystems and human health worldwide (Napper & Thompson, 2019). Internationally, rivers and waterways serve as critical pathways that transport vast quantities of land-based waste into the ocean, with over 1,000 rivers globally accounting for approximately 80% of annual riverine plastic emissions (Meijer et al., 2021). Asia contributes the highest continental estimate, responsible for 80.99% of global plastic emissions to the ocean (R.-S. Yu et al., 2023). Transitioning to the national scale, Indonesia is recognized as a major contributor to the aquatic waste contamination, generating approximately 7.8 million tons of plastic waste annually, with 4.9 million tons often mismanaged (Sakti et al., 2023), and was previously estimated to be the fifth-largest global source of marine plastic debris (Meijer et al., 2021). This issue is especially severe within the country's urban centres, such as Jakarta, where 13 local rivers transport a continuous flow of mismanaged plastic pollution into Jakarta Bay (Sari et al., 2022). Furthermore, critical waterways such as the Mahakam River face waste contamination from industry and agriculture (Sukmono et al., 2024), and a localized study, such as one on the Deli River in Medan, show that microplastic comprised 34.40% of the total macro litter collected, exacerbated by significantly increased plastic consumption (Hasibuan et al., 2022). The widespread contamination observed in these waterways is a visible consequence of longstanding failures in waste management, highlighting the critical need for data-driven interventions and advanced monitoring systems to effectively target cleanup efforts.

Traditional methods for monitoring waterway pollution, particularly floating trash, which rely on direct human effort and physical sampling, are labor-intensive, costly, and lack the scalability required to address the problem effectively (van Lieshout et al., 2020). Intermediate attempts at automation, such as the deployment of fixed camera monitoring systems (L. Zhang et al., 2021) or passive collection devices like Seabins (Kelly et al., 2023), are limited by restricted geographical coverage or the necessity of laborious manual review of image data. Consequently, there is an urgent transition toward robust, automated detection technologies utilizing Artificial Intelligence (AI) and computer vision (Bhuvaneswary et al., 2025). These modern systems deploy object detection models, notably the YOLO family of algorithms, on platforms such as Unmanned Aerial Vehicle (UAV) and Unmanned Surface Vehicle (USV) (Nguyen & Tran, 2022; Niu et al., 2019; Peng et al., 2024). While this technological shift enables large-scale monitoring, it also underscores the central challenge of developing detection models that strike a crucial balance between computational efficiency and the robustness required for reliable performance in dynamic waterway conditions.

The evolution of AI for floating trash detection began with accurate two-stage algorithms like R-CNN and Faster R-CNN, but their heavy computational demands and slow inference speeds made it impractical for real-time deployment (Devi et al., 2024; van Lieshout et al., 2020). To overcome this deficiency, the field shifted toward using single-stage You Only Look Once (YOLO) detectors, as their ability to detect floating trash in a single forward pass delivers the high speeds and balanced accuracy necessary for real-time applications (J. Wang & Zhao, 2024; G. Wu et al., 2023; Zhao et al., 2024). Despite establishing real-time feasibility, YOLO based models performance often degraded in complex environments, particularly in reliably detecting small or partially submerged trash due to visual noise from water reflections and background clutter (Devi et al., 2024; Fulton et al., 2019). Even recent, powerful architectures like YOLOv5 and YOLOv8, while strong general-purpose detectors, still show limitations in this specific domain and often require heavy modifications that re-introduce computational overhead. The current state-of-the-art addresses these fine-grained challenges by employing modified YOLO architecture integrated with specialized attention mechanisms, such as implementing Coordinate Attention to the YOLOv7 (K. Li et al., 2023). These mechanisms were designed to enhance feature representation, suppress background clutter, and improve detection robustness for challenging targets (Z. Jiang et al., 2023; J. Yu et al., 2023). However, a significant limitation associated with implementing these complex attention modules was the considerable computational overhead it introduced (Peng et al., 2024; J. Yu et al., 2023), which often threatened to negate the real-time efficiency gains achieved by the underlying YOLO architecture. This creates a clear research opportunity to investigate if the latest, highly efficient baselines, like the YOLOv11 family, can be paired with emerging Low Complexity Attention (LCA) mechanisms to solve this trade-off.

Despite the demonstrated capabilities of enhanced architecture like YOLOv7-CA in tackling visual difficulties such as small object and background clutter (K. Li et al., 2023), a critical challenge remains in reliably deploying these models for operational monitoring because of the lack of a standardized and efficient framework for adapting baseline models to unique, individual waterway conditions. Current studies frequently utilize models pre-trained on general datasets such as ImageNet or aggregated public floating trash datasets (Jia et al., 2024), but these generalized models often falter, showing substantial performance degradation when applied to varied locations or sensor settings due to insufficient transferability of high-level features (Maharjan et al., 2022). This pervasive issue of poor generalization is primarily rooted in the scarcity of publicly available, location specific annotated datasets which are essential for targeted fine-tuning (Maharjan et al., 2022; van Lieshout et al., 2020). Therefore, this lack of effective location-specific fine-tuning significantly impedes the transition to reliable, real-world monitoring systems, underscoring the critical need to develop a systematic framework for efficiently adapting these advanced detection models to the unique conditions of individual waterways.

Based on the outlined challenges, this research identifies two critical gaps in the current state-of-the-art floating trash detection. Firstly, while specialized attention mechanisms have shown promise in enhancing YOLO models' ability to handle complex visual noise and detect

challenging targets like small or partially submerged trash (Z. Jiang et al., 2023; K. Li et al., 2023; J. Yu et al., 2023), they often introduce significant computational overhead, undermining the real-time efficiency crucial for practical development (Peng et al., 2024; J. Yu et al., 2023). This highlights a need for attention mechanisms that improve robustness without substantial computational cost. Secondly, despite the availability of models pre-trained on diverse datasets, there is a lack of standardized and efficient framework for effectively adapting these generalized models to the unique conditions of specific individual waterways. This deficiency, stemming from poor generalization due to insufficient feature transferability (Maharjan et al., 2022) and the scarcity of location-specific annotated data (Maharjan et al., 2022; van Lieshout et al., 2020), severely hinders reliable operational monitoring. Therefore, a systematic methodology for data-efficient fine-tuning is critically needed to bridge this gap. The novelty of this research lies in its dual-pronged approach. It presents the first study, to our knowledge, that specifically designs and validates the integration of specialized Low Complexity Attention mechanisms for the task of floating trash detection. This contribution is paired with another crucial first, as the study also simultaneously develops a systematic, data-efficient framework for a model's practical adaptation to new, real-world waterways.

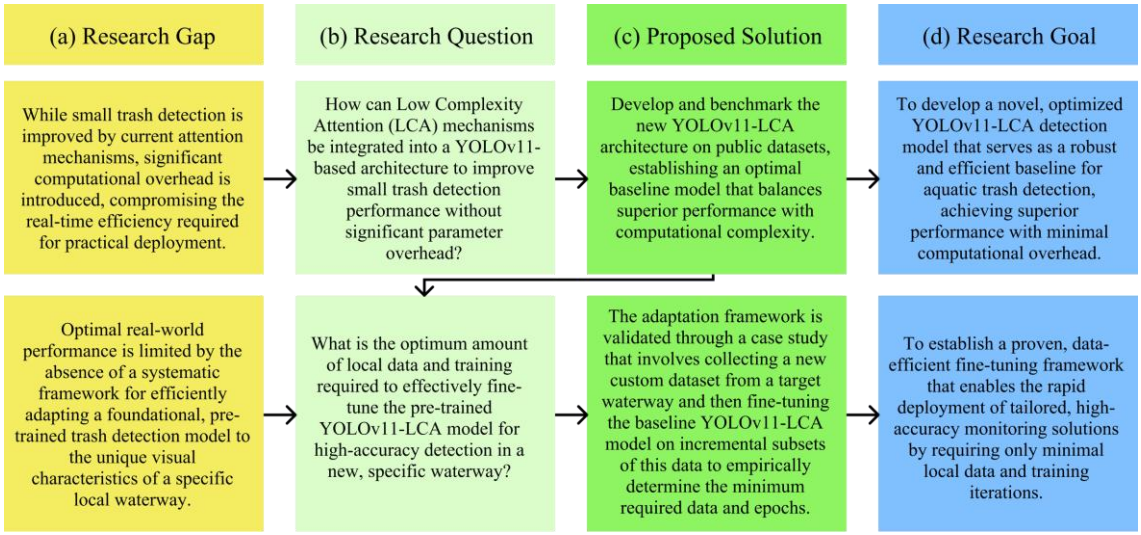


Fig. 1. Visual summary of the research framework, outlining the progression from identified gap to key contributions.

To address the first challenge outlined in Fig. 1(a) of enhancing small trash detection performance without compromising computational efficiency, this study proposes the development of a novel YOLOv11-LCA architecture. Following the strategic approach described in Fig. 1(b), this method integrates low complexity attention mechanisms into a YOLOv11 architecture. The core objective is to improve the model's ability to focus on small and partially submerged trash features, while actively suppressing background noise like water reflections and surface clutter. By using computationally lightweight attention modules, this solution is designed to achieve a superior balance between high detection accuracy and the minimal overhead required for real-time practical deployment. The proposed solution, as outlined in Fig. 1(c) is to develop and benchmark the architecture on public datasets, establishing a robust baseline model that demonstrates significantly improved performance with only a marginal increase in computational efficiency.

To bridge the second gap identified in Fig. 1(a), which is the lack of an efficient adaptation framework for specific waterway, this research establishes and validates a data-efficient adaptation framework. This framework provides a systematic methodology for fine-tuning the baseline YOLOv11-LCA model, aligning with the strategic approach described in Fig. 1(b). This process, as port of the proposed solution described in Fig. 1(c) involves collecting a new, custom dataset from a target location and then systematically fine-tune the baseline model on incremental subsets of the local data. By empirically analyzing the performance at each stage, the framework aims to determine the minimum quantity of local images and training iterations required to

achieve high-accuracy. This proposed solution delivers a proven, efficient fine-tuning process that enables the rapid deployment of tailored monitoring solutions to new environments with minimal data and computational resources.

This research provides several key contributions to the field of automated aquatic trash monitoring presented in Fig. 1. The primary outcomes deliver both a novel, enhanced detection architecture and a practical framework for real-world deployment, supported by a new specific dataset. The key contributions of this study are as follows:

- A novel, optimized YOLOv11-LCA detection model that serves as a robust baseline for aquatic trash detection, demonstrating significantly improved performance with minimal computational overhead.
- A proven, data-efficient fine-tuning framework that enables the rapid deployment of tailored, high-accuracy monitoring solutions to new environments by requiring only minimal local data and training iterations.
- The introduction of the "BojongTrash" dataset, a new, custom dataset collected from a specific waterway, which is used to validate the adaptation framework and serves as a resource for location-specific fine-tuning.

2. Literature Review

To situate this study, this section provides a critical review of recent deep learning advancements in automated waste management, addressing two primary research gaps. We first analyze the foundational task of waste classification to differentiate it from the more complex challenge of detection in real world scenarios. Following this, we evaluate the evolution of trash detection in waterway scenarios, focusing on the YOLO family and the integration of attention mechanisms, critically assessing the trade-off between model accuracy and computational cost. Concurrently, this review investigates the practical challenge of model deployment, highlighting the common absence of systematic frameworks for efficiently fine-tuning pre-trained models to new, specific local waterways. This two-pronged analysis will identify the persistent gaps in both architectural efficiency and practical adaptation, thereby establishing the dual necessities for the methodology proposed in this research.

Effective trash management and recycling efforts initially focused on accurate trash classification, which involves categorizing waste based on material type and recyclability. The TrashNet dataset (Aral et al., 2018), a standard benchmark in this domain, includes categories such as cardboard, glass, metal, paper, plastic, and general rubbish. The baseline classification is essential for the subsequent sorting process that influences the efficiency of recycling systems. Deep learning, particularly convolutional neural networks, has been widely explored for this task. Studies have shown that models like DenseNet121 and InceptionResNetV2, when fine-tuned with data augmentation techniques to compensate for limited dataset sizes, can achieve notable accuracy with the example of a DenseNet121 model achieve up to 95% accuracy on the test dataset. Further optimization, such as the integration of Genetic Algorithm (GA) to improve DenseNet12, has pushed classification accuracy to an impressive 99.6% (Mao et al., 2021). However, these results are achieved only on highly controlled scenarios featuring isolated items on simple backgrounds. This scenario does not address the far more complex challenge of detecting waste in real-world environments. While classification is optimized for sorting, robustly detecting trash in cluttered, reflective waterways remains a significant, unresolved issue. This study therefore moves beyond classification to focus on the more difficult problem of real-time detection, the necessary first step for automated monitoring.

While classification was a well-defined problem, real-world detection remained a significant challenge, prompting researchers to adapt using object detection models. Initial applications were dominated by two-stage object detectors such as R-CNN and its successor, Faster R-CNN, which became a foundational method for floating trash detection task (Devi et al., 2024; N. Li et al., 2022; van Lieshout et al., 2020). However, the practical application of these models for monitoring dynamic waterways is severely limited. Their two-stage process, which first proposes regions and then classifies them, results in heavy computational demands (van Lieshout et al., 2020). General speed benchmarks on the COCO datasets quantify this trade-off,

showing that two-stage Faster R-CNN models are generally slower, often only achieving sub-10 Frames Per Second (FPS), whereas single-stage detectors like SSD and YOLO can be significantly faster, running as fast as 25 – 33 FPS (Huang et al., 2017). This critical speed bottleneck, which makes real-time monitoring of floating trash impractical, is compounded by a severe performance gap in the aquatic domain. Recent benchmarking on the FloW-Img dataset, for example, demonstrates that a standard Faster R-CNN model achieves a low mean Average Precision (mAP) of only 0.180 (Devi et al., 2024). This dual bottleneck in both efficiency and accuracy led the field to rapidly shift toward single-stage detectors, most notably the YOLO family. By framing detection as a single regression problem, YOLO models are capable of processing an entire image in one pass, delivering high speeds essential for real-time applications (Redmon et al., 2016).

Moving beyond the slow two-stage models, the YOLO family of algorithms was rapidly adopted for real-time trash detection due to its superior balance of speed and accuracy (Carolis et al., 2020; Wahyutama & Hwang, 2022). This real-time capability led to its widespread adaptation in diverse, general-world scenarios, from smart-city based surveillance (Carolis et al., 2020) to lightweight systems on embedded hardware (Y. Liu et al., 2018). However, these successes in general urban scenarios fail to address the unique challenges of aquatic environments. This critical gap is demonstrated by quantitative benchmarks, which show that while a baseline model like YOLOv7 achieves real-time speed of 32.57 FPS (Z. Jiang et al., 2023), the model still has inherent difficulties with small scale trash and complex backgrounds (Z. Jiang et al., 2023; PENG et al., 2024). This results in false and missed detections (PENG et al., 2024), proving that while the YOLO architecture provides a necessary real-time foundation, it is quantitatively insufficient for this specific task. This established a clear need for architectural enhancements, leading researchers to integrate specialized attention mechanisms.

To address the documented failures of baseline YOLO models in floating trash detection task, further research commonly integrates specialized attention mechanisms to enhance feature representation. This strategy proved highly effective for improving model performance. For instance, a study integrating coordinate attention into a YOLOv3 model reported a 7.7 percentage improvement in AP score on a challenging dataset (Tharani et al., 2020). Another model, YOLOW, enhanced a YOLOv5s model by adding novel attention and fusion modules to improve small trash detection performance in water-crossing environments (Xu et al., 2023). This modification successfully increased the mAP_{50} from 0.787 to 0.821. However, this accuracy gain came at a high quantifiable cost, creating a new trade-off that compromised real-time efficiency. The YOLOW model parameter count increased nearly 190%, from 7.2M to 20.9M. Similarly, the YOLOv8MMS model computational load increased by over 50%, from 8.1 to 12.3 GFLOPs (J. Wang & Zhao, 2024). This quantifiable trade-off, where significant accuracy gains are achieved at the cost of major increases in model parameters and computational load, highlights a critical, unresolved gap. It directly reinforces the research question of whether it is possible to gain the benefits of feature enhancement without the computational overhead, establishing the clear need for truly low-complexity attention mechanisms.

To understand how a low-complexity module can be achieved, the foundational theory of attention mechanisms must be explained. The exploration of channel-wise or "what" attention was a foundational starting point, notably defined by the Squeeze and Excitation (SE) block, which introduced an effective and lightweight method for model inter-channel dependencies (Hu et al., 2018). This method operates by first "squeezing" global spatial information into a channel descriptor using global average pooling. This is followed by an excitation step, where fully connected layers, including a dimensionality-reducing bottleneck, learn a set of weights to recalibrate each channel. While widely adopted for its efficiency, the SE network's primary drawbacks are its complete disregard for spatial information and the potential for information loss within its bottleneck, which motivated efforts to improve channel attention. To address this limitation and re-introduce the "where" component, foundational work began to explore spatial attention. The Convolutional Block Attention Module (CBAM) was a seminal model that combined both concepts, sequentially inferring a channel attention map followed by a spatial attention map (Woo et al., 2018). Although CBAM effectively demonstrated the power of combining both dimensions, its sequential processing and reliance on a channel bottleneck were

identified as limitations. This sequential approach also assumes channel and spatial attention are separable, preventing the model from capturing more complex, cross-dimensional interactions. These limitations opened the door for the development of more efficient and hybrid strategies.

Subsequent research sought to overcome the limitations of foundational modules, primarily the information bottleneck in the SE block and the neglect of positional information. A direct response to the SE bottleneck was the Efficient Channel Attention (ECA) module, which empirically demonstrated that avoiding dimensionality reduction is critical for learning effective channel attention (Q. Wang et al., 2020). ECA-Net discards the two fully connected layers of SE and instead employs a single, fast 1D convolution to capture local cross-channel interactions efficiently, adding only a negligible number of parameters (Q. Wang et al., 2020). While ECA improved the efficiency of channel attention, it did not solve the problem of neglected positional information, which is critical for detection tasks. This led to the development of hybrid, cross-dimensional strategies, most prominently Coordinate Attention (CA), which explicitly embeds positional information into the channel attention mechanism (Hou et al., 2021). Unlike the 2D global pooling in SE that collapses all spatial information, CA factorizes channel attention into two parallel 1D feature encoding processes, one aggregating features along the horizontal axis and the other along the vertical axis (Hou et al., 2021). This method allows the network to capture long-range dependencies while preserving precise positional information, making it highly effective for localization and establishing it as a new standard for improving mAP in challenging, small-object detection scenarios.

To address the critical, unresolved gap for an attention mechanism that provides accuracy without major computational overhead, recent study have focused on developing lightweight modules. A state-of-the-art example is a family of Low Complexity Attention mechanisms, which includes the Low-Complexity Channel Attention Module (LCAM), Lightweight Detail Spatial Attention Module (LD-SAM), and a convolution replacement Low Complexity Bottleneck Hybrid Attention Module (LCBHAM). These are explicitly designed as lightweight, plug-and-play units for YOLO frameworks (Y. Zhang et al., 2024). Their design motivation is to enhance features for small object while maintaining minimal overhead, which was successfully validated in the domain of industrial defect detection. The LCAM module recalibrates "what" features to focus on by using efficient 2D convolutions instead of heavy fully-connected layers, thus minimizing parameters. The full LCBHAM sequentially combines this with a LD-SAM which refines "where" to focus by using a small 3×3 kernel that is more sensitive to fine-grained details and computationally cheaper than the larger kernels used in other modules. The effectiveness of this low-complexity design is quantifiable. When applied to a YOLOv8 baseline for industrial defect detection, the LCBHAM-enhanced model called DsP-YOLO significantly improved the performance on small, multi-scale defects. The enhanced model increased the overall mAP by 3.6% on the NEU-DET dataset, 2.1% on the PCB-DET dataset, and 3.9% on the GC10-DET dataset. Critically, this accuracy was achieved with a negligible parameter increase and almost zero increase in computational load, with the GFLOPs remaining identical. This proven ability to deliver significant mAP gains without computational overhead makes these LCA modules the ideal candidates for answering this study's first research question.

Having established a promising low-complexity architecture, the second major gap identified in literature is not the model design but in practical deployment. A robust baseline model, even one pre-trained on diverse public datasets, often experiences significant performance degradation when deployed in a new, specific waterway (Jia et al., 2024; Maharjan et al., 2022). This domain gap occurs because the high-level features learned from general datasets do not effectively transfer to the unique visual characteristics of a new aquatic environment (Maharjan et al., 2022). While fine-tuning is the logical solution, its application is hindered by the scarcity of annotated, local-specific datasets (Maharjan et al., 2022; van Lieshout et al., 2020). Creating a new, comprehensive dataset for every target river is operationally impractical, being both costly and labor-intensive. This problem necessitates a strategy of data-efficient fine-tuning, which has been validated in other data-scarce domains such as medical imaging (Alinsaif & Lang, 2020) and gesture recognition, where training on as little as one-third of data achieved over 90% accuracy (Escobedo-Gordillo et al., 2024). Despite this cross-domain success, a similar, systematic framework for the specific domain of floating trash detection is notably absent from

the literature (Maharjan et al., 2022). This absence confirms the second critical gap, which is the lack of a proven, data-efficient framework for adapting models to new waterways, which significantly impedes the transition to reliable, real-world monitoring. Validating such a framework, however, is dependent on having access to a specialized dataset designed specifically for this adaptation task.

The need for a specialized adaptation dataset reveals another critical resource gap, as publicly available datasets like FloW-Img (Cheng et al., 2021) and WaterTrash (Tharani et al., 2020) are designed for general baseline training, not for local adaptation. These general datasets cannot capture the specific water conditions, unique reflection patterns, and distinct debris types of every target river. (Jia et al., 2023; Liao & Juang, 2023). This limitation is why many researchers have had to create their own custom datasets to address specific environmental needs (Jia et al., 2023; Liao & Juang, 2023; van Lieshout et al., 2020). Therefore, to develop and validate a robust fine-tuning framework, it is necessary to first collect a new custom dataset from a specific target waterway. This new dataset will serve as the testbed for the adaptation study. These two related gaps, one in architecture and one in practical methodology, directly inform the contribution of this paper.

In summary, the critical review of the literature has identified two distinct, yet related, gaps. First, the field requires a detection architecture that successfully integrates attention to solve the persistent problem of small-object detection in aquatic environments without incurring the prohibitive computational overhead and high parameter counts seen in previous heavy attention-based solutions (L. Jiang et al., 2024; Xu et al., 2023). Second, the literature lacks a practical, data-efficient framework for adapting a robust foundational model to new, specific waterways, which is a critical barrier to real-world deployment (Maharjan et al., 2022). This study will address both gaps. It first proposes and benchmarks the novel YOLOv11-LCA architecture, integrating the theoretically efficient Low Complexity Attention mechanisms (Y. Zhang et al., 2024) to prove a balance of high accuracy and low overhead is achievable. It then develops and validates a systematic fine-tuning framework using the newly collected BojongTrash dataset to empirically determine the minimum amount of data and training epochs required for effective local adaptation. The following methodology section details the experiment design used to test these two contributions.

3. Research Methods

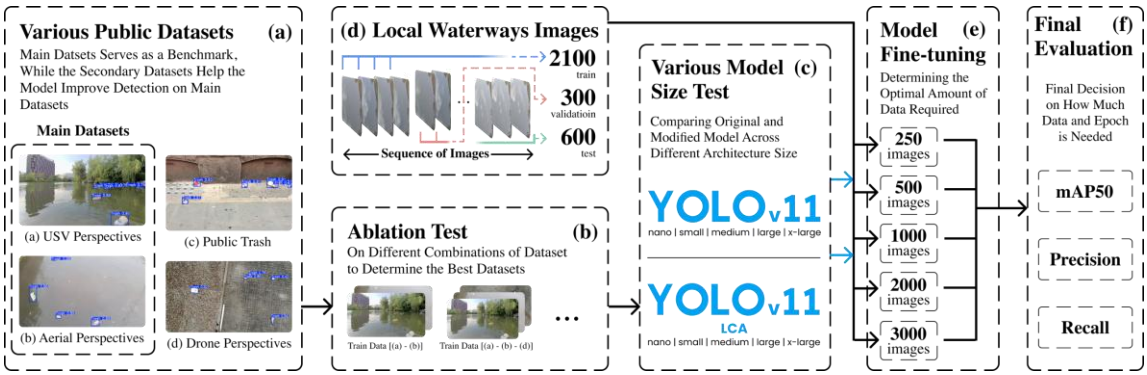


Fig. 2. Overview of the systematic experimental process. (**a**) Initial public dataset from diverse perspectives. (**b**) Dataset ablation testing for optimal combination. (**c**) YOLOv11 model variant and attention mechanism evaluation. (**d**) Local waterway image dataset collection and splitting. (**e**) Fine-tuning with varying local data amount. (**f**) Final model evaluation using key metrics.

This study employs a systematic, multi-stage experimental process, visually outlined in Fig. 2, to develop and validate an enhanced YOLOv11 model for detecting floating trash in waterways. This process is designed to directly address the first research gap on the need for a model that balances high accuracy for small objects with the low computational overhead required for real-time deployment. The initial phase, depicted in Fig. 2(a) begins by utilizing several public datasets captured from diverse viewpoints. An ablation test shown in Fig. 2(b) is then conducted using combinations of these public datasets to identify the optimal data combination. This optimal

combination subsequently serves as the basis for evaluating different YOLOv11 model sizes and, critically, for testing various configurations of the proposed attention mechanisms presented in Fig. 2(c). The objective of this baseline stage is to empirically prove that YOLOv11-LCA can quantifiably outperform the baseline YOLOv11 in detecting challenging small objects while maintaining minimal computational overhead, thereby providing a robust solution to the accuracy and efficiency trade-off.

Following the establishment of the optimal YOLOv11-LCA architecture, the research transitions to addressing the second research gap on the lack of a practical, data-efficient framework for fine-tuning a general model to a new specific waterway. The optimal YOLOv11-LCA weights derived from training on the public datasets serve as the initial weights for fine-tuning. This fine-tuning stage utilizes this established YOLOv11-LCA model and the "BojongTrash" temporal datasets, employing a specific data splitting strategy detailed in Fig. 2(d). As illustrated in Fig. 2(e), different quantities of this BojongTrash data are tested during the fine-tuning process, along with varied training iterations. This systematic exploration aims to determine the optimal amount of local data and training duration needed to effectively adapt the baseline YOLOv11-LCA to the unique conditions of the target river scenario. The performance of each trained model throughout all stages is rigorously assessed using the evaluation metrics presented in Fig. 2(f). In essence, this entire stage empirically defines a practical, data-efficient framework for adapting the optimized YOLOv11-LCA to new target environments.

3.1. Public Datasets

To establish a robust baseline for the proposed enhanced YOLOv11 model and enable comparative analysis with existing research that addresses the first research goal, this study leverages several publicly available datasets. The primary datasets selected for direct model evaluation and comparison are WaterTrash (Tharani et al., 2020), shown in Fig. 3(a), and FloW-Img (Cheng et al., 2021) with a sample depicted in Fig. 3(b). These were chosen as they are established benchmarks representing the two most common and distinct monitoring scenarios. In the context of this study, WaterTrash is utilized for its representation of trash in water channels from an aerial-like perspective, which often features clustered debris or objects obscured by shadows. In contrast FloW-Img provides a dataset of floating waste captured from the viewpoint of a USV. These two primary datasets are crucial as they represent distinct viewpoints relevant to waterway trash monitoring. Incorporating both datasets guarantees a thorough evaluation of the baseline model against the two main visual difficulties inherent in aquatic monitoring. To ensure a robust evaluation against the two main visual challenges in aquatic monitoring, utilizing both datasets is essential, with FloW-Img covering the difficulty of low-angle surface-level detection and WaterTrash addressing the specific hurdles of high-angle aerial detection.

(a) (b) (c) (d)

Fig. 3. Sample images from public dataset used in this study. (**a**) WaterTrash. (**b**) FloW-Img. (**c**) UAVVaste. (**d**) TACO.

To improve model generalization and prevent overfitting to the specific visual context of waterways, two secondary datasets, UAVVaste (Kraft et al., 2021), illustrated in Fig. 3(c), and Trash Annotation in Context (TACO) (Proença & Simões, 2020) with an example image depicted in Fig. 3(d) are incorporated into the experimental design. UAVVaste consist of images from public litter captured from a low-altitude drone's perspective, primarily in urban and natural land environments. TACO is a diverse, crowd-sourced dataset of general litter viewed from a pedestrian level, encompassing a wide variety of trash types and backgrounds. The rationale for this combination is to force the model to learn the intrinsic features of trash itself, such as shape and texture, rather than incorrectly associating trash with its background, like water reflections. This highly varied land based contexts in TACO build a robust general understanding of litter,

while UAVVaste strengthens this by providing non-water aerial perspectives. This approach is hypothesized to create a more resilient foundational model. A summary of all the public datasets, including their number of images and instances, is provided in Table 1.

Table 1 – Instance distribution across Train, Validation, and Test dataset for each of the public dataset.

Dataset Name	Instance		
	Train	*Validation*	*Test*
WaterTrash	28,449	7,122	10,193
FloW-IMG	2,598	650	2,023
UAVVaste	2,415	638	-
TACO	3,824	961	-
Total	33,453	9,371	12,216

3.2. Own Captured Dataset

While public datasets are used to build the foundational model, validating the second research goal of establishing a data efficient fine tuning framework required a new custom dataset. This study introduces a proprietary dataset named BojongTrash, specifically curated to reflect the unique conditions of local waterways near Telkom University. Data collection was introduced across three different waterways using a fixed position high definition camera with the sample shown in Fig. 4. The dataset introduces 9,000 images derived from 3,000 images captured from each waterway with the capture rate of one frame per second, predominantly around noon. This period was intentionally selected because it introduces the most difficult visual noise for a detection model, including harsh glare, intense water surface reflections, and deep shadows, which are known to cause false positives, thereby providing a rigorous test for model robustness.

(a) (b) (c)

Fig. 4. Sample images from BojongTrash dataset, captured from three distinct local waterway near Telkom University.

The BojongTrash dataset comprises data collected from three distinct rivers. Each river contributed 3,000 collected images that are first divided into three sequential temporal partitions, with each partition containing 1,000 images. Subsequently, each of these partitions undergoes a further split into training, validation, and testing subsets. Specifically, as illustrated by the breakdown in Fig. 5. The first 70% of the images are allocated for training, the next 10% for validation, and the final 20% are for testing. This hierarchical and sequential splitting strategy is critically important as it ensures temporal segregation. This method prevents data leakage by training the model on one block of time and testing it only on unseen footage from a later period, simulating a realistic deployment scenario. This partitioning structure forms the basis for the fine-tuning ablation study, which empirically determines the minimum data required for effective adaptation.

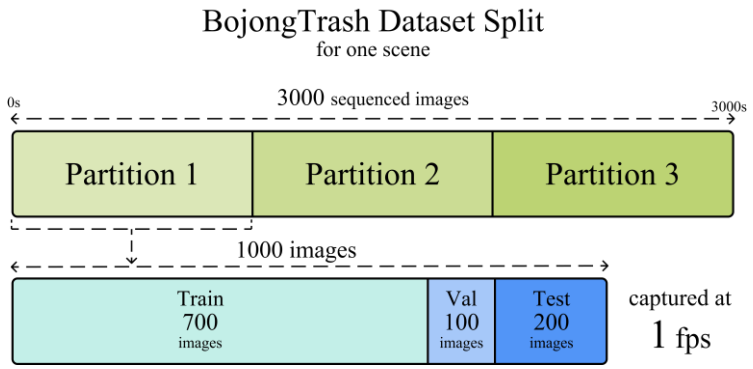


Fig. 5. Visualization of Hierarchical splitting strategy for each scene within the BojongTrash dataset.

The BojongTrash dataset captures a diversity of real-world aquatic conditions by sourcing images from three distinct waterways, each presenting unique visual challenges. The first waterway depicted in Fig. 4(a) is characterized by medium water flow speed and moderate levels of sunlight reflection. A notable characteristic of this location is the frequent presence of birds and their corresponding reflections on the water surface, introducing dynamic, non-trash objects that can test the model specificity. In contrast, the second waterway shown in Fig. 4(b) exhibits a high-water flow speed accompanied by very high and intense sunlight reflection, creating an overall very bright visual scene that can significantly impact feature extraction and object visibility. Finally, the third waterway presented in Fig. 4(c) presents a different set of difficulties with very slow water flow. While also bright, it features medium levels of reflection, but it distinguished by a high prevalence of small trash items, specifically challenging the model ability to detect diminutive and potentially cluttered waste. This diversity ensures the BojongTrash dataset provides a comprehensive testbed for rigorously evaluating the model's robustness and adaptability to the varied, challenging scenarios found in real world deployments.

3.3 YOLOv11 Architecture

YOLOv11 represents the latest advancement in the You Only Look Once based architecture (Redmon et al., 2016) that represents balance between high detection accuracy and real time inference speed. Developed by Ultralytics (Jocher & Qiu, 2024), YOLOv11 incorporates several architectural innovations designed to push the boundaries of real-time object detection across a variety of computer vision tasks. This balance is critical for the research goal of developing a practical, deployable system, and its efficient design serves as an ideal foundation for integrating new attention mechanisms without prohibitive computational cost. The core of the architecture retains the fundamental components of a Backbone for feature extraction, a Neck for feature fusion from different feature map scales, and a Head for generating the bounding box and class predictions. Fig. 6 depicts this overall architecture flow, illustrating how the Backbone extracts multi scale features, the Neck fuses these features, and the Head generates the final detections.

A significant architectural enhancement in YOLOv11 is the implementation of the C3k2 block, which plays a crucial role in both the backbone and neck structures. The C3k2 module is a highly computationally efficient implementation of the Cross Stage Partial (CSP) concept, its designed, which uses a smaller kernel size and a feature splitting and merging strategy, directly reduces the computational load and parameter count (Xiao et al., 2025). This block is designed to efficiently extract and fuse multi scale features, which reduces computational bottlenecks. This efficiency is precisely why it serves as a strong foundation and provide high performance to which our proposed LCA mechanisms can be added with only a marginal increase in computational cost. The detailed structure of the C3k2 block and its sub-blocks is shown in Fig. 7. The diagram also illustrates other critical components, such as the C2PSA block which handles cross stage feature fusion, and the SPPF block which effectively pools spatial features to help detect objects at different sizes. These efficient building blocks make the module critical for YOLOv11's strong performance in real-time detection tasks.

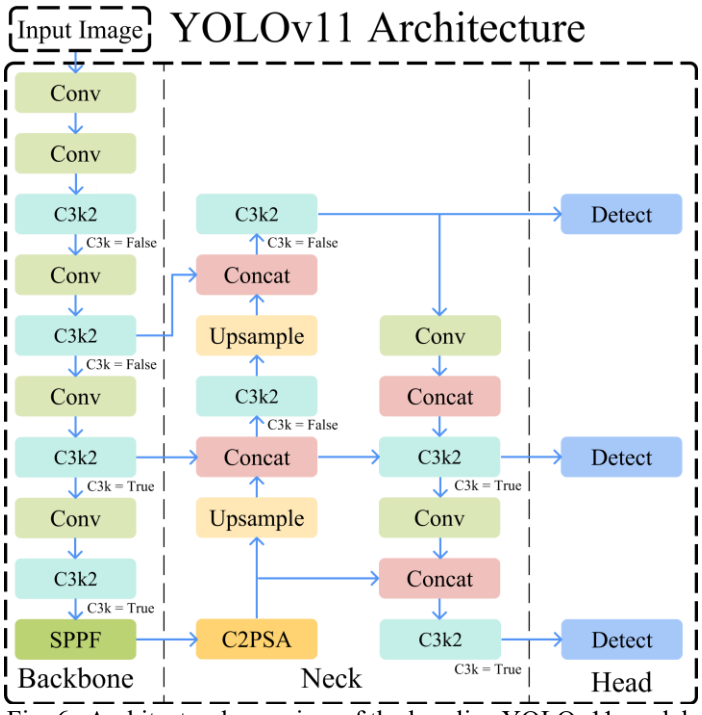


Fig. 6. Architectural overview of the baseline YOLOv11 model.

The YOLOv11 series, offers a range of model sizes, from the compact YOLOv11n (nano) to the larger YOLOv11x (extra-large). This study systematically evaluates these different models variants to establish a strong baseline and identify the optimal starting point for real-time performance. Table 2 quantifies the critical trade-off for this selection, detailing each model's parameter counts, FLOPs, benchmark inference speeds, and detection accuracy. This comparative analysis is a crucial methodological step. While larger models offer higher accuracy, their slow inference speeds and high computational cost make them unsuitable for the research objective of a real time system. Conversely, the smallest models are fast but may lack the feature extraction capacity to reliably detect small objects. This evaluation is therefore performed to select the base model that provides the best compromise between efficiency and accuracy before any attention mechanisms are added.

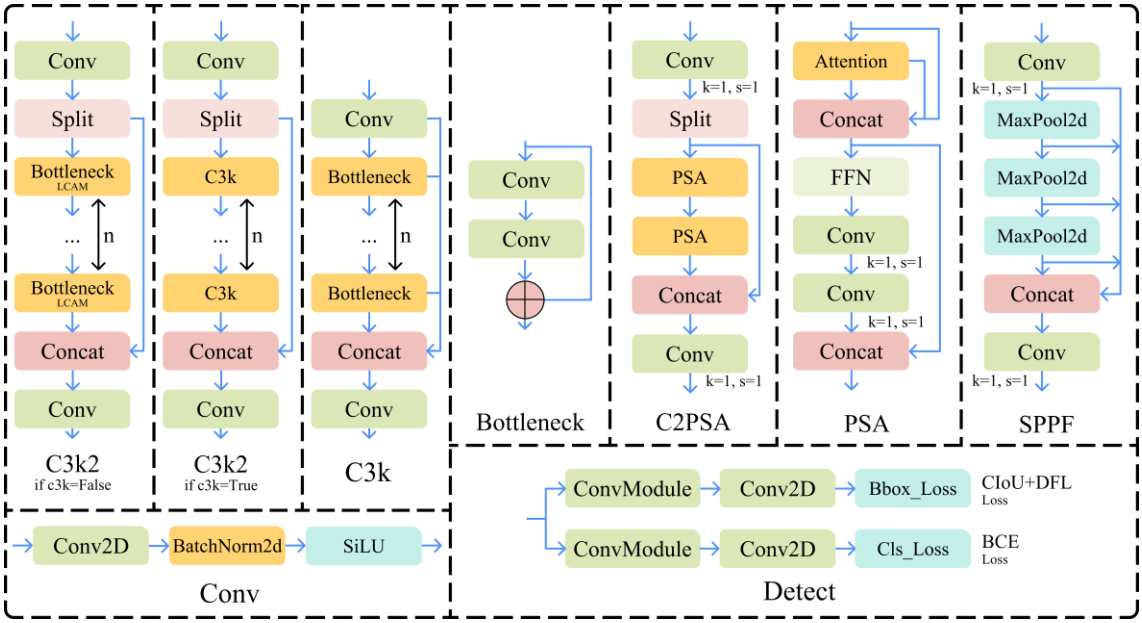


Fig. 7. Detailed architectural diagrams of key building blocks within the YOLOv11 framework.

Table 2 – Performance on MS COCO test set and complexity comparison of different baseline YOLOv11 model variants.

Model	mAP_{val} 50-95	$Speed_{CPU}$ (ms)	Params (M)	FLOPs (B)
YOLOv11n	39.5	56.1	2.6	6.5
YOLOv11s	47.0	90.0	9.4	21.5
YOLOv11m	51.5	183.2	20.1	68.0
YOLOv11l	53.4	238.6	25.3	86.9
YOLOv11x	54.7	462.8	56.9	194.9

3.4 Low Complexity Attention Mechanism

To address the first research gap, which is the significant computational overhead introduced by conventional attention mechanisms, this study selected the Low Complexity Attention modules (Y. Zhang et al., 2024). These modules were chosen because they are explicitly designed as efficient, plug and play units to enhance feature representation, especially for the challenging task of detecting small objects. This selection directly aligns with the research goal of improving small object detection accuracy while maintaining the real-time efficiency. (Y. Zhang et al., 2024). The LCBHAM architecture begins by processing an input feature map through an initial 2D convolution, followed by Batch Normalization and hard swish activation function. The core of LCBHAM attention capability then unfolds through two subsequent, specialized sub-modules: the Low Complexity Attention Module, which handles channel-wise feature recalibration, and the Lightweight Detail Spatial Attention Module (LD-SAM) which refines spatial features. This sequential application, where LCAM first identifies "WHAT" channels are important and LD-SAM then pinpoints "WHERE" the critical spatial details lie, allows LCBHAM to effectively focus on the minute yet discriminative characteristics often present in small objects. Such details might otherwise be diluted or lost in deeper network layers or when using less targeted attention mechanisms.

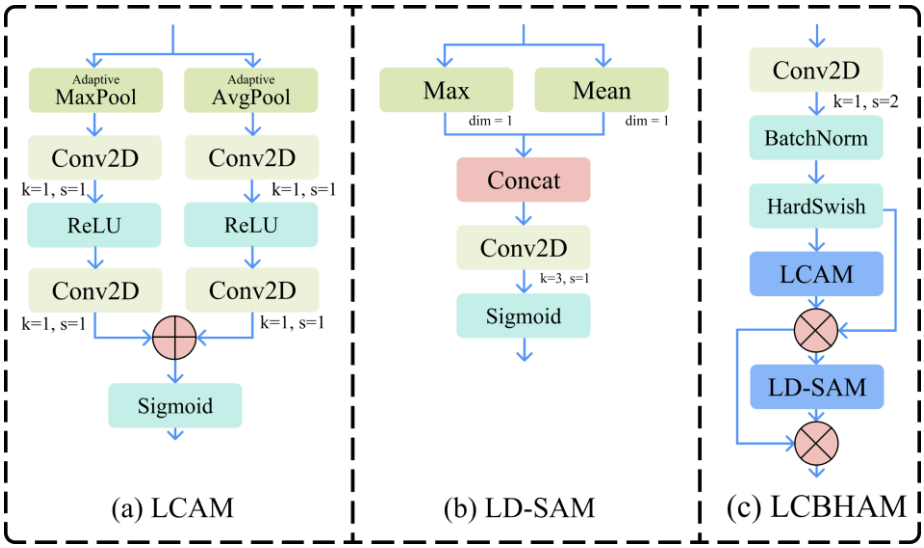


Fig. 8. Architecture of the proposed attention block. (a) The Low-complexity Channel Attention Module (LCAM). (b) the Lightweight Detail Spatial Attention Module (LD-SAM). (c) LCBHAM module leveraging LCAM and LD-SAM attention mechanisms.

The first attention component is the LCAM, responsible for channel-wise feature recalibration, as detailed in Fig. 8(a). Given an input feature map X, LCAM first generates channel descriptors by applying both Global Average Pooling (*AvgPool(X)*) and Global Max Pooling (*MaxPool(X)*). These two distinct descriptors are then independently processed through separate lightweight 1×1 convolutional layers ($Conv2d_{k=1,s=1}(X)$), each followed by ReLU activation. The resulting feature vectors from these two paths are then elementwise added together. Finally, a Sigmoid activation function (σ) is applied to this combined vector to produce the channel weights M_c. This dual-pooling strategy ensures that LCAM captures both global contextual information

across channel-specific features, which is particularly beneficial for distinguishing subtle characteristics that might define small objects. The formulation of LCAM channel attention map M_c is:

$$F_{avg}(X) = ReLU(Conv2d_{k=1,s=1}(AvgPool(X))$$
$$F_{max}(X) = ReLU(Conv2d_{k=1,s=1}(MaxPool(X))$$
$$M_c(X) = \sigma(F_{avg}(X) + F_{max}(X))$$

Following the channel refinement by LCAM, the LD-SAM is used to identify and emphasize salient spatial regions, with its architecture illustrated in Fig. 8(b). LD-SAM takes the channel-attended feature map X which is output of $X \otimes M_c$ as its input. It computes two 2D spatial descriptor maps by applying Average Pooling and Max Pooling with the dimension of 1 along the channel dimension. These two descriptor maps, which highlight average and peak responses across channels for each spatial location, are then concatenated along their channel dimension. This concatenated map is subsequently passed through a single 2D convolution with a kernel of 3 ($Conv2d_{k=3}(X)$), followed by Sigmoid activation function (σ) to generate the final spatial attention map M_S. The use of 3×3 kernel in LD-SAM is a deliberate design choice for efficiency. This kernel is small enough to be computationally cheap, yet large enough to capture the local spatial context needed to highlight small objects. It avoids the significant parameter increase of larger kernels used in other modules, thereby preserving the fine grained details crucial for small object detection without compromising the low complexity requirement. The spatial attention M_s is thus formulated as:

$$M_s(X) = \sigma(Conv2d_{k=3}([AvgPool_{k=1}(X);\ MaxPool_{k=1}(X)]))$$

In summary, the selection of these Low Complexity Attention modules is a deliberate methodological choice that directly addresses the first research gap. While standard attention mechanisms often introduce significant computational overhead, the research objective required a solution that could enhance small object detection without compromising real time efficiency. The LD-SAM and LCAM modules were chosen as the basic block of the attention because they are specifically designed as lightweight, plug and play units for YOLO architectures. Their design, which avoids parameter heavy layers in favour of efficient convolutions, has been proven in other complex detection tasks to provide significant mAP gains on small objects with almost zero increase in computational load. Therefore, these modules represent the ideal candidates to test if detection robustness and computational efficiency can be simultaneously achieved.

3.5. Low Complexity Attention Mechanism Implementation on YOLOv11

To enhance the YOLOv11 architectures capability, particularly in detecting small objects, this work proposes two primary architectural modifications. The first one is the integration of the LCAM module within the model's core C3k2 blocks Fig. 9 provides the detailed building blocks for this integration, showing how the LCAM is used to create a new BottleneckLCAM component, which in turn forms the final C3k2LCAM block. Specifically, LCAM is inserted at the final sequence of the bottleneck component. This specific placement is a deliberate design choice. By applying channel attention after the bottleneck's main feature extraction but before the final concatenation, the module can recalibrate and select the most informative channels. This ensures that only the most relevant features are passed on, aiming to improve its feature extraction and fusion capabilities without significantly increasing its computational overhead, leveraging LCAM's efficient design.

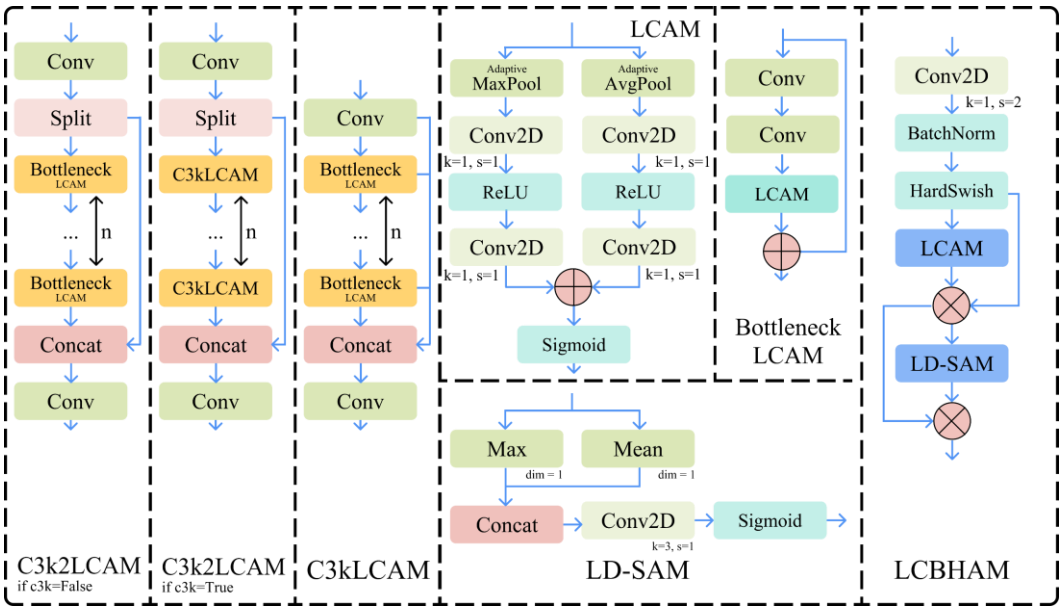


Fig. 9. Detailed architectural diagrams of the proposed attention-enhanced modules and their integration.

The second modification strategically replaces standard convolutional layers in specific, impactful locations within the YOLOv11 neck with the LCBHAM, as illustrated in Fig. 10. The reason for this replacement is that the neck is where multi-scale feature are fused. LCBHAM which itself leverages LCAM for channel attention and LD-SAM for spatial attention, is positioned to refine features at critical junctures in the Feature Pyramid Network (FPN) and Path Aggregation Network (PAN) structures of the neck. For instance, LCBHAM is applied to the feature maps before concatenation operations that merge features from different scales. This placement allows LCBHAM to enhance the salient features from each scale and suppress noise or less relevant information before they are fused, which is particularly beneficial for preserving the distinct characteristics of small objects that might otherwise be overwhelmed during multi-scale fusion. The sequential channel and spatial attention within LCBHAM provide a more comprehensive feature refinement compared to using LCAM alone, making it suitable for these key feature interaction points. Fig. 10 illustrates the exact placement of both modifications, showing the new C3k2LCAM blocks integrated throughout the model and the LCBHAM module replacing a key convolutional layer in the neck's up sampling path.

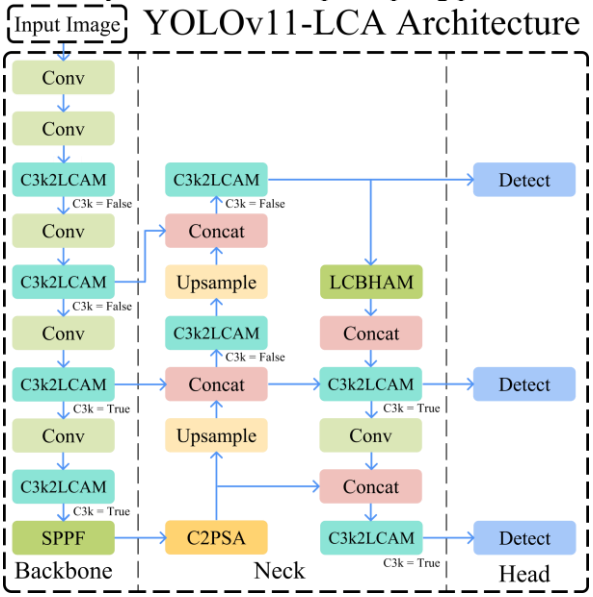


Fig. 10. Architecture of the enhanced YOLOv11 model incorporating Low Complexity Attention mechanisms.

To empirically validate these integration strategies, a comprehensive ablation study is designed. This study aims to understand the individual and synergistic effects of applying these

attention mechanisms at different stages and granularities within the architecture. Table 3 details the configurations for this experiment, which is designed to methodically test what each modification contributes. The baseline YOLOv11s, representing the vanilla model without any attention enhancements, serves as the reference. The YOLOv11s-LCBHAM model configuration isolates the impact of the comprehensive LCBHAM module by replacing standard convolution in the neck, allowing an evaluation of its standalone benefit at critical feature fusion points. The YOLOv11s-Backbone configuration implements LCAM to the C3k2 blocks solely in the backbone, focusing on early-stage feature recalibration, while YOLOv11s-Neck applies the LCAM on the C3k2 only in the neck, targeting later-stage refinement during feature fusion, and YOLOv11s-Full integrates LCAM to all the C3k2 inside the YOLOv11 architecture for widespread LCAM deployment. Further configurations like YOLOv11s-Backbone-LCBHAM explore the combined effects by pairing these C3k2LCAM integration with the LCBHAM in the neck. This structured approach, with the model configurations, detailed in Table 3, allows for a clear attribution of performance gains or losses to these specific architectural modifications, ultimately guiding the design of an optimally enhanced YOLOv11 architecture.

Table 3 – Configurations for the ablation study on attention mechanism integration within the YOLOv11s architecture.

Model Code	LCBHAM	C3k2LCAM Backbone	C3k2LCAM Neck
YOLOv11s	–	–	–
YOLOv11s-LCBHAM	✓	–	–
YOLOv11s-Backbone	–	✓	–
YOLOv11ls-Neck	–	–	✓
YOLOv11s-Full	–	✓	✓
YOLOv11s-Backbone-LCBHAM	✓	✓	–
YOLOv11s-Neck-LCBHAM	✓	–	✓
YOLOv11s-Full-LCBHAM	✓	✓	✓

Code	Explanation
LCBHAM	Applies the LCBHAM to replace conv block at neck
Backbone	Integrate LCAM into C3k2 blocks in the backbone
Neck	Integrate LCAM into C3k2 blocks in the neck
Full	Integrate LCAM into C3k2 blocks in the entire block

3.6. Building the Proposed Model

The construction of the proposed model followed a systematic, multi-stage experimental process designed to empirically validate each methodological choice. This process consisted of three main stages. First, a dataset ablation study was conducted to determine the optimal combination of public datasets described in Table 1, aiming to create the most robust foundational model. Second, an attention implementation ablation study was performed using the configuration shown in Table 3 to identify the most effective and efficient integration of the LCA to YOLOv11. Third, a final comparative analysis evaluated this optimal enhanced architecture against the baseline across all model sizes to identify the configuration that best balances detection accuracy with computational efficiency. All experiments were conducted using a consistent hardware and software environment, detailed in Table 4, to ensure fair and reproducible comparisons.

Table 4 – Experimental hardware and software setup.

Component	Specification
CPU	Intel Core i7-12700f
RAM	64GB
GPU	Nvidia RTX 3080
Python	3.11
Pytorch	2.5.0

The first stage was a dataset ablation study to empirically test the model generalization to different angle and environments. The YOLOv11m architecture was selected as the baseline model for this specific test because its medium size provides a stable balance of feature extraction capacity and training time, making it an ideal testbed for evaluating data composition without the

excessive cost of larger models. Table 5 details the exact dataset combinations that were tested. Each combination was used to train a separate YOLOv11m model, and its performance was independently evaluated on both the FloW-Img and WaterTrash test datasets. The combination yielding the highest and most consistent mAP_{50} between both test sets was selected as the optimal data composition and was used for all subsequent training stages.

With the optimal dataset established from the first stage, the second stage proceeded to evaluate the impact of integrating the proposed attention mechanisms. For this architectural ablation study, the baseline was switched to the YOLOv11s architecture. This was a deliberate choice because a smaller model has less redundant capacity, meaning any performance gains are more clearly attributable to the architectural change itself, rather than being masked by the high capacity of a larger model. Each of the enhanced model configurations from Table 3, such as YOLOv11s-LCBHAM, was trained on the optimal dataset. Their performance was then benchmarked against the baseline YOLOv11s on the FloW-Img and WaterTrash test datasets to identify the most effective attention strategy.

Table 5 – Dataset combinations used in the ablation study to determine the optimal training data composition.

WaterTrash Primary	FloW-Img Primary	UAVVaste Secondary	TACO Secondary
✓	–	–	–
–	✓	–	–
✓	✓	–	–
✓	✓	✓	–
✓	✓	–	✓
✓	✓	✓	✓

Finally, the third stage was conducted to ascertain the most effective overall model configuration. This comparison involved evaluating both the baseline YOLOv11 variants and the variants enhanced with the best-performing attention mechanism configuration identified in the second stage. The purpose of this final analysis was to determine the sweet spot model that provides the best trade-off between detection accuracy and computational cost, directly addressing the first research objective. All models in this phase were trained using the optimal dataset combination from the first stage and evaluated on the FloW-Img and WaterTrash test datasets. Throughout all three experimental stages on the public dataset, a consistent training hyperparameters, detailed in Table 6, were maintained to ensure fair and comparable evaluations, guiding the selection of the final proposed model architecture.

Table 6 – Training hyperparameters utilized for model training on public datasets.

Parameter	Value
Epoch	300
Input Size	640×640
Batch Size	16
Initial Learning Rate	0.01
Momentum	0.937
Workers	2
Deterministic	True
Optimizer	Adam

3.7. Fine-Tuning the Proposed Model on Local Waterway

Following the identification of the baseline YOLOv11-LCA on public datasets, the research transitioned addressing the second research goal of establishing a proven, data efficient framework for adapting this model to a new, specific waterway. To achieve this, an experiment was designed to answer the critical question of the minimum quantity of local data required for effective adaptation. The fine-tuning process commenced by utilizing the weights from the best performing model as a starting point. Specifically, the proposed model and its corresponding baseline counterpart for comparison were fine-tuned on incrementally larger subsets of the BojongTrash dataset, including 250, 500, 1,000, 2,000, and 3,000 images. This comparison is

methodologically crucial to determine if the architectural enhancements also improve data efficiency, meaning the enhanced model learns faster or achieves higher accuracy with the same limited amount of local data. The performance of each fine-tuned instance was evaluated on a held-out portion of the BojongTrash test set to identify the data size that offers the best trade-off between performance improvement and data collection effort. Table 7 details the incremental data subsets used in this ablation study, which was chosen to empirically identify the point of diminishing returns. The image sequences refer to the temporal partitions defined in Section 3.2, ensuring the test set remained temporally separate at every step.

Table 7 – Image sequence used in each ablation study.

Amount of Image	Dataset Split		
	Train	*Validation*	*Test*
250	0 – 174	175 – 199	200 – 249
500	0 – 349	350 – 400	400 – 499
1,000	0 – 699	700 – 799	800 – 999
2,000	0 – 699; 1,000 – 1,699	700 – 799; 1,700 – 1,799	800 – 999; 1,800 – 1,999
3,000	0 – 699; 1,000 – 1,699; 2,000 – 2,699	700 – 799; 1,700 – 1,799; 2,700 – 2,799	800 – 999; 1,800 – 1,999 2,800 – 2,999

Once the most effective data quantity from the BojongTrash dataset was established, the next phase focused on the second component of the minimum number of fine-tuning epochs required. This is a critical step for a data efficient framework, as training for many epochs is computationally expensive and impractical for rapid deployment. Using the optimal data subset, the proposed model was further trained on a limited number of additional epochs. Performance was monitored after each epoch to identify a suitable point where the model sufficiently adapted to the local data characteristics without beginning to overfit. This step is crucial for identifying the minimum training time needed to achieve peak performance. Table 8 details the consistent hyperparameters used for all fine-tuning experiments. Notably, the initial learning rate was set to 0.0005, a significant reduction from the 0.01 used in initial training. This is a deliberate and standard practice for fine-tuning. A smaller learning rate allows the model to make small, careful adjustments to its powerful pre-trained weights to adapt to the new data, rather than destroying those learned features with large, disruptive updates.

Table 8 – Fine-tuning hyperparameters for adapting to BojongTrash dataset.

Parameter	Value
Epoch	25
Input Size	640×640
Batch Size	16
Initial Learning Rate	0.0005
Momentum	0.937
Workers	2
Deterministic	True
Optimizer	Adam

3.8 Model Evaluation

To quantitatively assess the performance of the developed trash detection models, and to validate the research objective of achieving high accuracy, a set of standard object detection metrics were employed: Precision, Recall, and mean Average Precision at Intersection of Union (IoU) threshold of 0.50 (mAP_{50}). The choice of an IoU threshold of 0.50 is deliberate for this application where the primary objective is to detect the presence of trash items, and a highly precise bounding box localization is secondary to successfully identifying an object as trash. Therefore, a detection is considered correct if the IoU between the predicted bounding box and the ground truth bounding box is 0.50 or greater.

Precision measures the accuracy of the positive prediction made by the model. It answers the question “Of all the objects the model predicted as trash, what proportion was actually trash?”. Recall measures the model’s ability to find all relevant instances of trash within the dataset. It answers the question “Of all the actual trash items present in the images, what proportion did the model successfully detect?”.

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

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

Description:

1. TP (True Positive): The number of trash items correctly identified as trash by the model (IoU ≥ 0.50 with the ground trash item).
2. FP (False Positive): The number of predictions made by the model that do not correspond to an actual trash item or correspond to a trash item but with an IoU < 0.50.
3. FN (False Negative): The number of actual trash items present in the image that the model failed to detect.

Tracking both Precision and Recall is critical for this application. High Recall ensures the model finds most of the trash, which is vital for effective monitoring. High Precision ensures the model is reliable and does not generate false alarms by misidentifying reflections or leaves as trash, which is crucial for an automated system.

The mAP_{50} metric is the primary metric used in this study to compare model performance. It is calculated as the area under the Precision-Recall curve and provides a single, comprehensive score summarizing performance across all confidence thresholds. This metric is widely used as it offers a balanced measure of both precision and recall. This makes it the ideal metric for the ablation studies, as it allows for a robust and direct comparison of the overall effectiveness of different model architectures and training configurations.

To rigorously validate that the observed performance improvements are statistically significant and not artifacts of specific test variations, a non-parametric bootstrap analysis was conducted. This process involved generating 1,000 independent bootstrap samples by resampling the test images from both FloW-Img and WaterTrash datasets with replacement. For each iteration, the mAP_{50} metric was recalculated, creating a distribution of performance scores. From this distributions, the 95% confidence intervals were derived, establishing a solid statistical foundation for comparing the generalization capabilities of the baseline architecture against the proposed attention-enhanced model.

Together, these three metrics, reinforced by the statistical validation of the bootstrap analysis, provide a comprehensive understanding of the model's performance. mAP_{50} serves as the primary indicator of overall accuracy for comparing models. Precision and Recall offer crucial diagnostic insights into specific failure modes, balancing reliability against completeness. Evaluating all three is essential to robustly validate the model's effectiveness for the practical task of trash detection.

4. Results and Discussions

This section presents the empirical validation for the two primary research goals. The results are structured to first address the development of a novel, robust, and computationally efficient detection model, which answers the first research question. This involves a detailed analysis of the dataset and attention ablation studies, a comprehensive comparison of the YOLOv11-LCA model against its baseline and the relevant literature, and a qualitative assessment of its performance. Following this, the discussion transitions to the second research question, presenting the results from the fine-tuning experiments to establish and validate a practical, data-efficient framework for adapting the model to new, specific waterway.

Table 9 – Comparative performance of YOLOv11m on FloW-Img and WaterTrash test sets based on ablation test of diverse training data configurations from public datasets.

Training Data				FloW-Img Test Set		WaterTrash Test Set	
WaterTrash	*FloW-Img*	*UAVVaste*	*TACO*	*mAP_{50}*	*F1-Score*	*mAP_{50}*	*F1-Score*
✓	–	–	–	0.214	0.323	0.925	0.888
–	✓	–	–	0.780	0.770	0.238	0.234
✓	✓	–	–	0.817	0.794	0.925	0.888
✓	✓	✓	–	0.820	0.796	0.925	**0.890**

✓	✓	–	✓	**0.834**	**0.807**	0.924	0.887
✓	✓	✓	✓	0.827	0.804	**0.926**	0.887

The initial phase of experimental evaluation focused on determining the optimal combination of publicly available datasets for training a robust baseline trash detection model. The YOLOv11m was selected for this ablation study because it represents a stable balance of feature extraction capacity and training time. The results summarized in Table 9, show the importance of data composition. Among the tested combinations, the configuration that used the primary WaterTrash and FloW-Img datasets with the addition of the TACO dataset yielded the most significant improvement in detection performance, achieving an mAP50 of 0.834 on the FloW-Img test set. The TACO dataset comprising diverse crowd-sourced images of litter from many different land based context, forced the model to learn the intrinsic features of the trash itself, rather than incorrectly associating trash with water. This finding is a key contribution to model design, as it implies that a robust waterway detection model should not be trained exclusively on aquatic data. Consequently, the combination of WaterTrash, FloW-Img and TACO datasets was identified as the optimal training data composition and was utilized for all subsequent experiments.

Table 10 – Ablation study on YOLOv11s with different attention mechanism integration strategies on FloW-Img and WaterTrash test sets.

Model	FloW-Img Test Set			WaterTrash Test Set			Params (M)	GFLOPs
	P	*R*	*mAP50*	*P*	*R*	*mAP50*		
YOLOv11s	0.816	0.737	0.779	0.900	0.852	0.918	9.413	21.301
YOLOv11s+LCBHAM	0.836	0.778	0.835	0.915	0.860	0.924	9.415	21.307
YOLOv11s+Backbone	0.820	0.773	0.825	0.909	0.868	0.924	9.418	21.314
YOLOv11s+Neck	0.836	0.783	0.832	0.911	0.864	0.923	9.421	21.318
YOLOv11s+Full	0.800	0.748	0.800	0.908	0.845	0.918	9.423	21.319
YOLOv11s+Backbone+LCBHAM	0.834	0.765	0.822	0.906	0.861	0.923	9.421	21.320
YOLOv11s+Neck+LCBHAM	**0.838**	**0.792**	**0.836**	**0.916**	**0.869**	**0.925**	9.424	21.323
YOLOv11s+Full+LCBHAM	0.832	0.782	0.828	0.912	0.868	0.923	9.429	21.336

The second stage of the investigation evaluated the efficacy of the proposed attention mechanisms. The YOLOv11s model was selected for this ablation study, as its smaller architecture allows for a clearer observation of performance gains attributable to specific architectural modifications. As presented in Table 10, the baseline YOLOv11s model achieved an mAP50 of 0.779 on the FloW-Img test dataset and 0.918 on the WaterTrash test datasets. These baseline figures serve as the reference against which all attention-enhanced configurations were compared. The results distinctly highlight the benefits of integrating the LCA mechanisms. Among the various configurations tested, the YOLOv11s-Neck-LCBHAM emerged as the most effective configuration. This configuration achieved the highest performance, yielding a 0.057-point improvement in mAP50 on the FloW-Img test set to 0.836. This underscores the synergistic effect of applying attention at the critical feature fusion stage, both before fusion with LCBHAM and within the neck C3k2 blocks. While many studies report gains from general purpose modules like CBAM or SE, this result demonstrates that a targeted, specialized, and combined low complexity integration is a more effective and efficient for this specific detection problem.

Table 11 – Performance comparison of baseline YOLOv11 with the modified YOLOv11 across different model sizes on FloW-Img and WaterTrash test sets.

Model	FloW-Img Test Set			WaterTrash Test Set			Params
	P	*R*	*mAP50*	*P*	*R*	*mAP50*	(M)
YOLOv11n	0.816	0.696	0.768	0.880	0.816	0.894	2.582
YOLOv11n-LCA	0.824	0.739	0.805	0.894	0.848	0.911	2.585
FRL-YOLO (Xian et al., 2024)	0.839	0.721	0.793	–	–	–	~6.400
YOLOv11s	0.816	0.737	0.779	0.900	0.852	0.918	9.413
YOLOv11s-LCA	**0.838**	**0.792**	**0.836**	**0.913**	**0.862**	**0.924**	**9.424**
USD-YOLO	–	0.813	0.862	–	–	–	12.350
YOLOv11m	0.834	0.757	0.825	0.907	0.859	0.923	20.030
YOLOv11m-LCA	0.848	0.787	0.845	0.914	0.871	0.928	20.052
YOLOW (Xu et al., 2023)	0.870	0.751	0.821	–	–	–	~20.900
RTDETR-MARD (Sun et al., 2025)	0.853	0.832	0.866	–	–	–	~22.500
STE-YOLO (J. Yu et al., 2023)	–	–	0.832	–	–	–	32.740

YOLOv11l	0.841	0.782	0.841	0.913	0.861	0.927	25.280
YOLOv11l-LCA	0.846	0.786	0.856	0.916	0.872	0.926	25.315
YOLOv7-CA (K. Li et al., 2023)	–	–	0.811	–	–	–	~51.500
YOLOv11x	0.824	0.784	0.832	0.908	0.862	0.923	56.828
YOLOv11x-LCA	0.823	0.802	0.857	0.914	0.877	0.928	56.906

Following the identification of the optimal attention strategy from the previous stage, a comprehensive evaluation was conducted across the full spectrum of YOLOv11 model sizes. This stage was designed to identify the optimal balance between performance and model complexity. The comprehensive results shown in Table 11 demonstrate a consistent performance uplift from the LCA integration across all model scales, suggesting that the attention mechanism effectively captures fundamental features of floating trash regardless of the base feature extractor's capacity. This consistent, minimal overhead underscores the design efficiency of the attention modules. For instance, the proposed YOLOv11s-LCA, achieve a 0.057-point improvement in mAP_{50} on the FloW-Img test set, with only a 0.011M parameter increase. This represent a negligible growth in model size of roughly 0.1%, ensuring that the computational budget remains virtually unchanged while detection capability significantly improves. While other recent models like YOLOW (Xu et al., 2023) and STE-YOLO (J. Yu et al., 2023) achieve a comparable mAP_{50} scores of 0.821 and 0.832 respectively, they require more than double the parameters at 20.9M and 32.7M. Similarly, RTDETR-MARD (Sun et al., 2025), which reported a 0.866 mAP_{50}, is a much heavier model at 22.5M parameters. Even the YOLOv7-CA (K. Li et al., 2023), which achieved a lower 0.811 mAP, is a massive 51.5M parameter model. Therefore, the proposed YOLOv11s-LCA is novel because it proves that a strategic, low complexity attention integration can achieve superior or highly competitive accuracy with a fraction of the computational cost, establishing a new state-of-the-art in efficiency for this task. Such a drastic reduction in computational cost is critical for enabling the deployment of autonomous monitoring system on resource-constrained edge devices.

Table 12 – Statistical validation metrics comparing the baseline and modified YOLO11s models across 1,000 bootstrap samples.

Dataset	Model Variant	Average mAP_{50}	Standard Deviation	95% Confidence Interval	Improvement Probability
FloW-Img	YOLOv11s	0.776	0.0129	[0.750, 0.801]	–
	YOLOv11s-LCA	0.834	0.0139	[0.806, 0.860]	99.90%
WaterTrash	YOLOv11s	0.916	0.0030	[0.910, 0.923]	–
	YOLOv11s-LCA	0.925	0.0029	[0.920, 0.931]	98.50%

To validate the superiority of the YOLOv11s-LCA, which was identified as the optimal configuration in the comparative analysis, a targeted statistical evaluation was conducted specifically against the baseline YOLOv11s. This focused analysis ensures that the gains achieved by the proposed best model are statistically significant and not merely incidental to specific data splits. As summarized in Table 12, the results from 1,000 bootstrap samples confirm that the YOLOv11s-LCA consistently outperforms the baseline across both testing environments. On the challenging FloW-Img dataset, the modified model increased the average mAP_{50} from 0.811 to 0.834, demonstrating a remarkable improvement probability of 99.90%. The performance gain is also robust on the WaterTrash dataset, where the model achieved an improvement probability of 98.50%, raising the average mAP_{50} to 0.925 compared to the baseline score of 0.916.

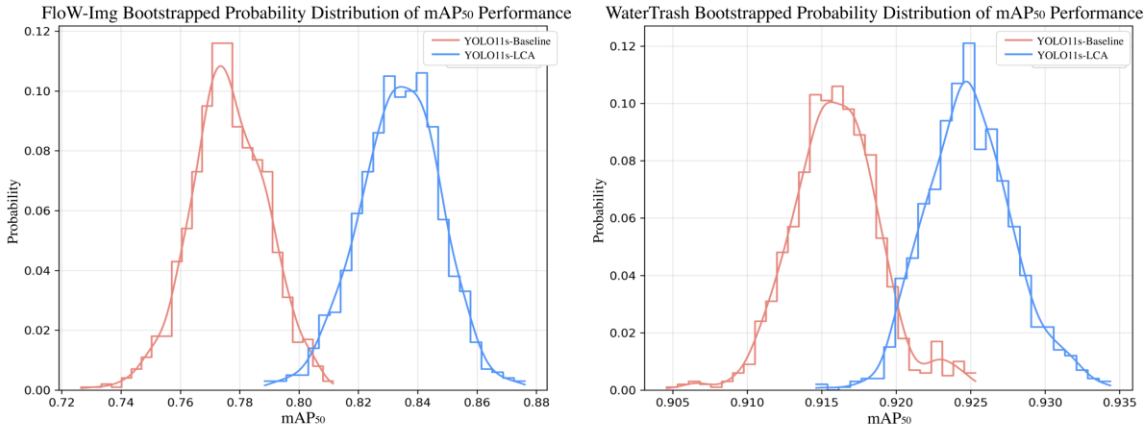


Fig. 11. Bootstrapped probability density distributions of mAP50 scores. (a) Distribution on the FloW-Img test set. (b) Performance distribution on the WaterTrash test set.

This statistical distinction is visually corroborated by the probability density distributions plotted in Fig. 11. In the results for both the FloW-Img and WaterTrash datasets, presented in Fig. 11(a) and Fig. 11(b) respectively, the distribution for the YOLOv11s-LCA represented by the blue line exhibits a clear rightward shift compared to the baseline YOLOv11s shown by the red line. This systematic displacement toward higher accuracy values illustrates that the architectural improvement are fundamental and effective regardless of the specific image composition in the test subset. This shift indicates that the enhanced model does not merely achieve a higher maximum score but consistently yields higher mAP_{50} scores across the vast majority of resampled test sets. The minimal overlap between the curves, which is particularly pronounced in the FloW-Img dataset, reinforces the conclusion that the integration of the low-complexity attention mechanism provides a stable and reliable performance enhancement rather than a variance-dependent outlier. Furthermore, this distinct separation confirms that the performance gap is statistically significant, providing a high degree of confidence that reported gains would be reproducible in diverse real-world deployment scenarios.

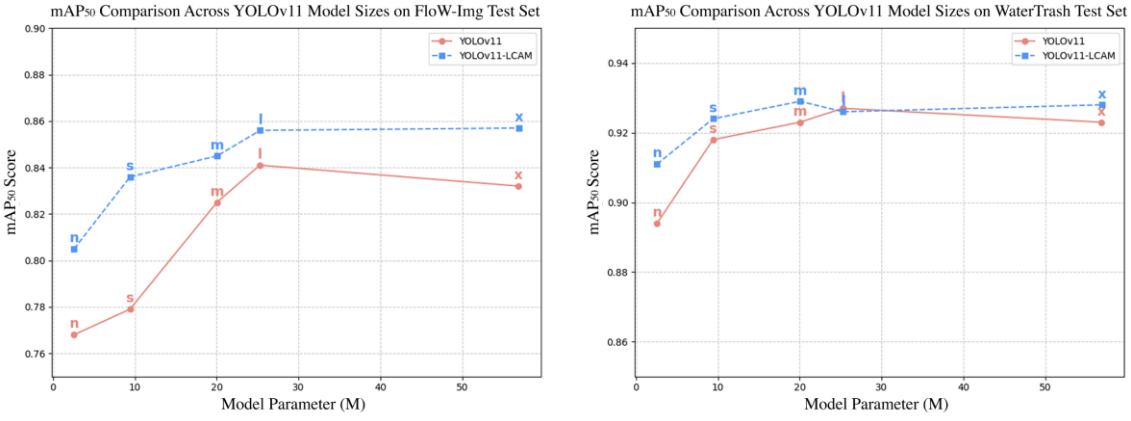


Fig. 12. Comparison of mAP50 scores between baseline and modified YOLOv11 models across different model sizes. (a) Performance on FloW-Img test set. (b) Performance on WaterTrash test set.

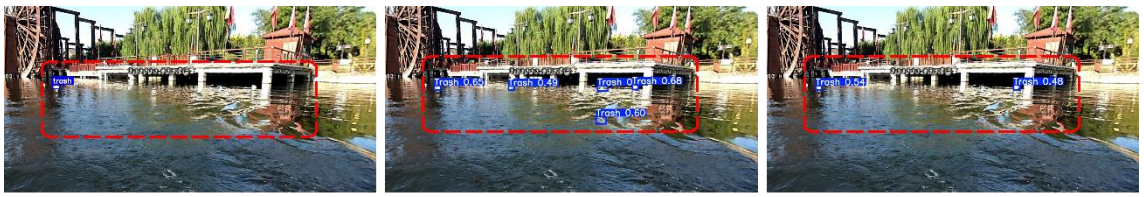
With the model's performance gains statistically validated, the practical implication of this superior accuracy-efficiency trade-off is visualized in Fig. 12 which plots mAP_{50} against the number of model parameters. In both Fig. 12(a) and Fig. 12(b), the blue line representing the enhanced YOLOv11-LCA, consistently lies above the baseline models. This visual evidence reinforces that the attention-enhanced models achieve superior accuracy for any given parameter count.. Critically, the graph for the FloW-Img test set in Fig. 12(a) shows that the YOLOv11s-LCA model at 9.4M parameters achieves a higher mAP_{50} than the baseline YOLOv11m at 20.0M parameters. This visually proves that the proposed attention strategy is a far more efficient method for improving performance than simply scaling up the model size. This result solidifies the

YOLOv11s-LCA as the optimal balanced model, as it achieves its high accuracy without the high computational cost of larger variants, efficiently realizing the research objective.

(a) (b) (c)

Fig. 13. Visual comparison of detection capabilities on a challenging FloW-Img test image. (a) Ground truth showing multiple trash items, including three small objects. (b) Baseline YOLOv11s predictions, unable to detect the trash. (c) Modified YOLOv11s-LCA predictions, demonstrating improved detection of small trash.

To further illustrate the practical impact of the attention mechanism, Fig. 13 presents a qualitative comparison of detection results from the challenging FloW-Img test set. This provides a qualitative explanation for the quantitative mAP jump. Fig. 13(a) shows the ground truth, highlighting several trash items, including a cluster of three small objects within the red dashed box. The result from the baseline YOLOv11s model shown in Fig. 13b fails to detect these smaller, clustered items. In contrast, the modified YOLOv11s-LCA in Fig. 13(c) successfully detects all three of these small trash objects. This visual evidence underscores the enhanced feature discrimination capabilities from the low complexity attention mechanism. Specifically, the LD-SAM component likely empowers the YOLOv11s-LCA to retain the fine-grained spatial detail needed to discern and localize these small trash items, highlighting the practical benefits of the proposed architectural enhancements.



(a) (b) (c)

Fig. 14. Example of false positive suppression by the modified model in a reflective BojongTrash environment. (a) Original image with ground truth label. (b) Baseline YOLOv11s output with numerous false positives. (c) Modified YOLOv11s-LCA output, showing reduction of false positive detection.

The model's enhanced robustness is further highlighted in Fig. 14. The ground truth image shown in Fig. 14(a) indicates a single trash item in a scene with numerous sharp reflections. The baseline YOLOv11s model in Fig. 14(b), generates multiple false positive detections, misinterpreting reflections and water ripples as additional trash items. This susceptibility to false positives significantly reduces the model reliability. In stark contrast, the modified YOLOv11s-LCA, with its prediction result depicted in Fig. 14(c) demonstrates a markedly improved ability to differentiate true trash from environmental noise. It correctly detects the target trash item while effectively suppressing most false positives, erroneously identifying only a single reflection. This showcases the attention mechanism's crucial role in enhancing the model discriminative power and its practical utility for robust monitoring settings.

Having established the YOLOv11s-LCA as a robust and efficient baseline model, the research discussion now transitions to its practical application and adaptability. The next phase of the study addresses the challenge of efficiently deploying this model in a new, specific waterway. The following results, derived from the experiments on the BojongTrash dataset, are presented as the empirical evidence used to construct and validate a practical fine-tuning framework. This framework aims to define the optimal, minimal amount of local data and training required for effective, real-world adaptation.

Table 13 – Impact of training data quantity on fine-tuning performance for baseline YOLOv11s and modified YOLOv11s on the three BojongTrash dataset scene.

Images	Model	BojongTrash Scene 1			BojongTrash Scene 2			BojongTrash Scene 3		
		P	*R*	*mAP50*	*P*	*R*	*mAP50*	*P*	*R*	*mAP50*
250	YOLOv11s	0.757	0.601	0.698	0.532	0.495	0.438	0.806	0.711	0.795
	YOLOv11s-LCA	0.828	0.665	0.742	0.579	0.556	0.488	0.824	0.755	0.832
500	YOLOv11s	0.798	0.638	0.746	0.631	0.538	0.562	0.891	0.710	0.834
	YOLOv11s-LCA	0.801	0.674	0.818	0.598	0.579	0.587	0.842	0.770	0.856
1,000	YOLOv11s	0.898	0.777	0.873	0.769	0.565	0.681	0.890	0.754	0.855
	YOLOv11s-LCA	0.911	0.816	0.908	0.760	0.594	0.697	0.863	0.785	0.875
2,000	YOLOv11s	0.934	0.828	0.915	0.809	0.576	0.716	0.871	0.775	0.887
	YOLOv11s-LCA	0.919	0.858	0.925	0.791	0.601	0.728	0.872	0.805	0.897
3,000	YOLOv11s	0.934	0.828	0.915	0.764	0.582	0.739	0.872	0.790	0.904
	YOLOv11s-LCA	0.919	0.858	0.925	0.771	0.607	0.755	0.881	0.795	0.913

The results detailed in Table 13 consistently show that the YOLOv11s-LCA model outperforms the baseline across all data quantities and all three scenes. This suggests the architectural enhancements not only improve baseline accuracy but also improve data efficiency during adaptation. For example, with only 1,000 images on the first scene, the YOLOv11s-LCA model achieved an mAP_{50} score of 0.908, significantly higher than the baseline model's. This trend holds across all scenes, but Table 13 also reveals the critical point of diminishing returns. While performance for Scene 1 jumps to 0.908 with 1,000 images, it only incrementally increases to 0.925 when using 3,000 images. This finding is the core of the data efficient framework, suggesting approximately 1,000 images is the optimal choice. Notably, scene 2 presented the most significant challenge for both models, yielding comparatively lower mAP50 scores, as is visually evident in Fig. 15(b). This reduced performance is directly attributable to the unique characteristics of scene 2, which, as previously described in Section 3.2 is characterized by intense sunlight glare that occludes objects. This demonstrates that data quantity alone cannot easily solve extreme visual noise. However, it is crucial to note that the modified YOLOv11s model still maintained a clear performance advantage over the baseline model in scene 2, proving its refined attention capabilities provide superior resilience even in highly challenging environments.

Fig. 15. Visual comparison of mAP50 scores achieved by baseline and modified YOLOv11s as the number of fine-tuning images from BojongTrash dataset increases. (a) Performance on Scene 1. (b) Performance on Scene 2. (c) Performance on Scene 3.

Fig. 15 visually confirms this point of the impact of the diminishing returns on the addition of the training data, while also reinforcing that the modified model remains superior at all data increments. In all three scenes, the performance curves for both the baseline and modified models climb steeply until the 1,000 image mark, after which they begin to flatten significantly. The gains achieved by adding data beyond 1,000 images are marginal compared to the large initial jump, suggesting the model has effectively learned the core features of the new environment by that point. This graph provides the clear visual evidence to support the framework's recommendation of approximately 1,000 images as the optimal, data efficient target for adaptation, balancing high performance with minimal data collection effort.

Table 14 – Epoch-wise mAP50 and F1-Score progression during fine-tuning with the modified YOLOv11s-LCA with 1,000 images, evaluated on respective test sets.

Epoch	Scene 1		Scene 2		Scene 3	
	mAP50	*F1-Score*	*mAP50*	*F1-Score*	*mAP50*	*F1-Score*
1	0.781	0.785	0.596	0.579	0.825	0.791
2	0.830	0.810	0.605	0.602	0.878	0.813
3	0.856	0.827	0.616	0.603	0.890	0.824
4	0.843	0.832	0.611	0.600	0.892	0.822
5	0.842	0.817	0.608	0.601	0.892	0.828
6	0.842	0.820	0.608	0.599	0.892	0.826
7	0.844	0.820	0.609	0.600	0.891	0.828
8	0.838	0.814	0.608	0.600	0.891	0.828
9	0.840	0.816	0.609	0.599	0.892	0.829
10	0.839	0.813	0.608	0.598	0.892	0.830

To further investigate the second component of the framework, training efficiency, with the identified optimal 1,000 training images, Table 14 presents a detailed breakdown of the mAP50 and F1-Score progression for the modified YOLOv11s-LCA model over 10 epochs for each of the three BojongTrash scenes. The results in Table 14 indicate that significant performance gains are typically achieved within the first few epochs of fine-tuning. For all three scenes, both mAP50 and F1-Score show a clear trend of improvement that rapidly peaks and then plateaus. The epoch-wise fine-tuning progression detailed in Table 14 is visually supported by Fig. 16, which plots the mAP50 and F1-Score for the modified YOLOv11s-LCA model against the number of epochs for each of the BojongTrash scenes. In these plots, both mAP50 and F1-Score curves demonstrate a rapid that flattens considerably after epoch 5, indicating that further epochs do not lead to substantial improvements in either metric. This strongly suggests that the fine-tuning process converges within only 3-5 epochs, making extended training inefficient. These results, combined with the findings from the data quantity ablation, establish a complete, practical, and data-efficient framework for real-world deployment. This framework demonstrates that a user can achieve near-peak performance by collecting only approximately 1,000 images.

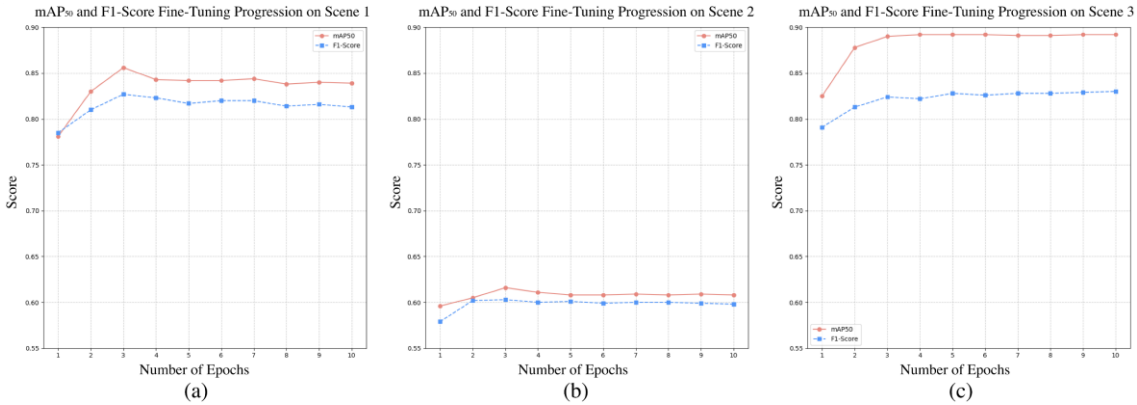


Fig. 16. Visual Representation of mAP50 and F1-Score fine-tuning progression for the modified YOLOv11s-LCA model over 10 epochs using 1,000 training images from the BojongTrash dataset. (a) Progression on Scene 1. (b) Progression on Scene 2. (c) Progression on Scene 3.

Despite the overall strong performance of the established framework, the BojongTrash dataset also highlights several inherent challenges and areas for future refinement, as shown in Fig. 17. These instances reveal scenarios where the model, even with attention mechanisms, can be confounded by complex environmental factors or objects with visual characteristics similar to trash. Fig. 17(a) illustrates instances where flying birds, likely due to their size or shape are misclassified as trash. Similarly, Fig. 17(b) demonstrates that complex water surface phenomena, such as prominent ripples or small waves, are misinterpreted as trash, and Fig. 17(c) shows strong reflections also causing false positives. These occurrences underscore the complexity of reliably distinguishing genuine trash from dynamic non trash elements. These specific misclassifications

highlight a clear path for future work such as incorporating more diverse negative sampling or specialized training data, which could teach the model to better discriminate these specific problem cases.

(a) (b) (c)

Fig. 17. Common false positive detections encountered in the BojongTrash dataset (a) Birds misidentified by floating debris. (b) Water ripples misclassified for trash. (c) Reflection on the water surface leading to incorrect trash detection.

(a) (b) (c)

Fig. 18. Example of detection inconsistency across sequential frames from the BojongTrash dataset. (a) A small trash item is detected. (b) The same trash item is missed in the subsequent frame. (c) The trash item is detected again in the third frame.

Another observed challenge, illustrated in Fig. 18, pertains to detection inconsistency across sequential frames. Fig. 18(a) shows a small trash item, indicated within the red box, is correctly detected. However, in the subsequent frame shown in Fig. 18(b), the same trash item is missed by the model, only to be detected again in the third frame as shown in Fig. 18(c). This intermittent detection can be problematic for applications like accurately quantification trash over time. These inconsistencies likely arise from subtle changes in the object appearance or reflection between frames. This is a known issue that is typically solved with post processing. Future work should integrate the powerful detection output of the YOLOv11s-LCA model with an object tracking algorithm like SORT or DeepSORT. These methods would maintain object identities across frames, smooth out intermittent detections, and provide more robust trajectory tracking, thereby improving the overall reliability and utility of the system for dynamic trash monitoring.

5. Conclusion

This research successfully addressed the dual challenge of creating an accurate yet computationally efficient model for floating trash detection. The primary contribution is the novel YOLOv11-LCA architecture, which strategically integrates Low Complexity Attention modules, LCAM and LCBHAM, into the YOLOv11 neck. This approach proved highly effective, increasing the mAP50 on the challenging FloW-Img dataset from 0.779 to 0.836 with a negligible 0.1% parameter increase and a statistically confirmed improvement probability of 99.90%, establishing a superior, state-of-the-art balance of performance and efficiency. This finding contributes to lightweight object detection theory by demonstrating that targeted, Low Complexity Attention can be more effective than just scaling model parameters. A second major contribution is the validation of a practical, data-efficient fine-tuning framework, which empirically determined that this robust model can be rapidly adapted to new, specific waterways using only approximately 1,000 local images and just 3 to 5 training epochs. This framework has significant practical implications for environmental automation, lowering the barrier for real-world deployment and enabling scalable, cost-effective, and sustainable water quality monitoring. Future work should focus on enhancing this system's robustness by using targeted negative sampling to reduce false positives from environmental elements like water ripples and by integrating object tracking algorithms to ensure temporal consistency in dynamic monitoring scenarios

Acknowledgement

The authors extend our sincere gratitude to the Directorate of Research and Community Service, Telkom University for the financial support under the Grant numbers KWR4.068/LIT06/PPM-LIT/2024.

References

Abdu, H., & Noor, M. H. M. (2022). Domestic Trash Classification with Transfer Learning Using VGG16. *2022 IEEE 12th International Conference on Control System, Computing and Engineering (ICCSCE)*, 137–141. https://doi.org/10.1109/ICCSCE54767.2022.9935653

Alinsaif, S., & Lang, J. (2020). Histological Image Classification using Deep Features and Transfer Learning. *2020 17th Conference on Computer and Robot Vision (CRV)*, 101–108. https://doi.org/10.1109/CRV50864.2020.00022

Aral, R. A., Keskin, S. R., Kaya, M., & Haciomeroglu, M. (2018). Classification of TrashNet Dataset Based on Deep Learning Models. *2018 IEEE International Conference on Big Data (Big Data)*, 2058–2062. https://doi.org/10.1109/BigData.2018.8622212

Bhuvaneswary, N., Fedrick, A. A., Alltrin, K. S., Jaspher, C. J. J., & Sounder, K. (2025). AI-Based Trash Collector Boat for Autonomous Waterway Pollution Management. *2025 International Conference on Sustainable Energy Technologies and Computational Intelligence (SETCOM)*, 1–5. https://doi.org/10.1109/SETCOM64758.2025.10932499

Bianco, S., Gaviraghi, E., & Schettini, R. (2024). Efficient Deep Learning Models for Litter Detection in the Wild. *2024 IEEE 8th Forum on Research and Technologies for Society and Industry Innovation (RTSI)*, 601–606. https://doi.org/10.1109/RTSI61910.2024.10761805

Carolis, B. De, Ladogana, F., & Macchiarulo, N. (2020). YOLO TrashNet: Garbage Detection in Video Streams. *2020 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS)*, 1–7. https://doi.org/10.1109/EAIS48028.2020.9122693

Cheng, Y., Zhu, J., Jiang, M., Fu, J., Pang, C., Wang, P., Sankaran, K., Onabola, O., Liu, Y., Liu, D., & Bengio, Y. (2021). FloW: A Dataset and Benchmark for Floating Waste Detection in Inland Waters. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 10933–10942. https://doi.org/10.1109/ICCV48922.2021.01077

Devi, B. S., Nagaraja, K. V, & Singh, R. P. (2024). Optimization Enhancements for Faster R-CNN in Floating Bottle Detection. *2024 IEEE 11th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, 1–7. https://doi.org/10.1109/UPCON62832.2024.10982812

Escobedo-Gordillo, A., Brieva, J., Moya-Albor, E., Ponce, H., Franco-Gaona, E., & Cruz-Aceves, I. (2024). Optimal Dataset Size for Fine-Tuning sEMG-Based Hand Gesture Recognition in Rehabilitation Prosthesis. *2024 20th International Symposium on Medical Information Processing and Analysis (SIPAIM)*, 1–5. https://doi.org/10.1109/SIPAIM62974.2024.10783516

Fulton, M., Hong, J., Islam, M. J., & Sattar, J. (2019). Robotic Detection of Marine Litter Using Deep Visual Detection Models. *2019 International Conference on Robotics and Automation (ICRA)*, 5752–5758. https://doi.org/10.1109/ICRA.2019.8793975

Ganvir, K. D., Nerkar, P. R., Ghate, L. W., & Bhagat, H. H. (2019). The impact of water pollution and preliminary study on river trash collecting mechanism. *International Journal of Technical Research and Applications*, *7*(1), 85–87.

Hasibuan, N. H., Salsabila, R., Perdana, Z., Khair, H., Husin, A., Suryati, I., Nurfahasdi, M., & Patumona, S. (2022). Assessment of macro litter in Deli River Medan during pandemic COVID-19. *IOP Conference Series: Earth and Environmental Science*, *977*(1), 012106. https://doi.org/10.1088/1755-1315/977/1/012106

Hou, Q., Zhou, D., & Feng, J. (2021). Coordinate Attention for Efficient Mobile Network Design. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13708–13717. https://doi.org/10.1109/CVPR46437.2021.01350

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-Excitation Networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7132–7141. https://doi.org/10.1109/CVPR.2018.00745

Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., Fischer, I., Wojna, Z., Song, Y., Guadarrama, S., & Murphy, K. (2017). Speed/Accuracy Trade-Offs for Modern Convolutional Object Detectors. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3296–3297. https://doi.org/10.1109/CVPR.2017.351

Jambeck, J. R., Geyer, R., Wilcox, C., Siegler, T. R., Perryman, M., Andrady, A., Narayan, R., & Law, K. L. (2015). Plastic waste inputs from land into the ocean. *Science*, *347*(6223), 768–771. https://doi.org/10.1126/science.1260352

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

Jia, T., Vallendar, A. J., de Vries, R., Kapelan, Z., & Taormina, R. (2023). Advancing deep learning-based detection of floating litter using a novel open dataset. *Frontiers in Water*, *5*. https://doi.org/10.3389/frwa.2023.1298465

Jiang, L., Liu, F., Lv, J., Liu, B., & Wang, C. (2024). GST-YOLO: a lightweight visual detection algorithm for underwater garbage detection. *Journal of Real-Time Image Processing*, *21*(4), 114. https://doi.org/10.1007/s11554-024-01494-w

Jiang, Z., Wu, B., Ma, L., Zhang, H., & Lian, J. (2023). APM-YOLOv7 for Small-Target Water-Floating Garbage Detection Based on Multi-Scale Feature Adaptive Weighted Fusion. *Sensors*, *24*(1), 50. https://doi.org/10.3390/s24010050

Jocher, G., & Qiu, J. (2024). *Ultralytics YOLO11*. https://github.com/ultralytics/ultralytics

Kelly, B. O., Chen, S., Zhou, E. P., & Elshakankiri, M. (2023). AI-Enabled Plastic Pollution Monitoring System for Toronto Waterways. *2023 10th International Conference on Internet of Things: Systems, Management and Security (IOTSMS)*, 53–58. https://doi.org/10.1109/IOTSMS59855.2023.10325803

Kraft, M., Piechocki, M., Ptak, B., & Walas, K. (2021). Autonomous, Onboard Vision-Based Trash and Litter Detection in Low Altitude Aerial Images Collected by an Unmanned Aerial Vehicle. *Remote Sensing*, *13*(5), 965. https://doi.org/10.3390/rs13050965

Li, K., Wang, Y., & Hu, Z. (2023). Improved YOLOv7 for Small Object Detection Algorithm Based on Attention and Dynamic Convolution. *Applied Sciences*, *13*(16), 9316. https://doi.org/10.3390/app13169316

Li, N., Huang, H., Wang, X., Yuan, B., Liu, Y., & Xu, S. (2022). Detection of Floating Garbage on Water Surface Based on PC-Net. *Sustainability*, *14*(18), 11729. https://doi.org/10.3390/su141811729

Liao, Y.-H., & Juang, J.-G. (2022). Real-Time UAV Trash Monitoring System. *Applied Sciences*, *12*(4), 1838. https://doi.org/10.3390/app12041838

Liao, Y.-H., & Juang, J.-G. (2023). Automatic Marine Debris Inspection. *Aerospace*, *10*(1), 84. https://doi.org/10.3390/aerospace10010084

Liu, C., Xie, N., Yang, X., Chen, R., Chang, X., Zhong, R. Y., Peng, S., & Liu, X. (2022). A Domestic Trash Detection Model Based on Improved YOLOX. *Sensors*, *22*(18). https://doi.org/10.3390/s22186974

Liu, T., Luo, R., Xu, L., Feng, D., Cao, L., Liu, S., & Guo, J. (2022). Spatial Channel Attention for Deep Convolutional Neural Networks. *Mathematics*, *10*(10), 1750. https://doi.org/10.3390/math10101750

Liu, Y., Ge, Z., Lv, G., & Wang, S. (2018). Research on Automatic Garbage Detection System Based on Deep Learning and Narrowband Internet of Things. *Journal of Physics: Conference Series*, *1069*, 012032. https://doi.org/10.1088/1742-6596/1069/1/012032

Maharjan, N., Miyazaki, H., Pati, B. M., Dailey, M. N., Shrestha, S., & Nakamura, T. (2022). Detection of River Plastic Using UAV Sensor Data and Deep Learning. *Remote Sensing*, *14*(13), 3049. https://doi.org/10.3390/rs14133049

Mao, W.-L., Chen, W.-C., Wang, C.-T., & Lin, Y.-H. (2021). Recycling waste classification using optimized convolutional neural network. *Resources, Conservation and Recycling*, *164*, 105132. https://doi.org/10.1016/j.resconrec.2020.105132

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

Misra, D., Nalamada, T., Arasanipalai, A. U., & Hou, Q. (2021). Rotate to Attend: Convolutional Triplet Attention Module. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 3138–3147. https://doi.org/10.1109/WACV48630.2021.00318

Mittal, G., Yagnik, K. B., Garg, M., & Krishnan, N. C. (2016). SpotGarbage. *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 940–945. https://doi.org/10.1145/2971648.2971731

Mo, R., Lai, S., Yan, Y., Chai, Z., & Wei, X. (2022). Dimension-aware attention for efficient mobile networks. *Pattern Recognition*, *131*, 108899. https://doi.org/10.1016/j.patcog.2022.108899

Napper, I. E., & Thompson, R. C. (2019). Environmental Deterioration of Biodegradable, Oxo-biodegradable, Compostable, and Conventional Plastic Carrier Bags in the Sea, Soil, and Open-Air Over a 3-Year Period. *Environmental Science & Technology*, *53*(9), 4775–4783. https://doi.org/10.1021/acs.est.8b06984

Nguyen, T.-T., & Tran, H.-L. (2022). An Efficient Model for Floating Trash Detection based on YOLOv5s. *2022 9th NAFOSTED Conference on Information and Computer Science (NICS)*, 230–234. https://doi.org/10.1109/NICS56915.2022.10013413

Niu, G., Li, J., Guo, S., Pun, M.-O., Hou, L., & Yang, L. (2019). SuperDock: A Deep Learning-Based Automated Floating Trash Monitoring System. *2019 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, 1035–1040. https://doi.org/10.1109/ROBIO49542.2019.8961509

PENG, C., HE, B., XI, W., & LIN, G. (2024). Improved YOLOv7 Algorithm for Floating Waste Detection Based on GFPN and Long-Range Attention Mechanism. *Wuhan University Journal of Natural Sciences*, *29*(4), 338–348. https://doi.org/10.1051/wujns/2024294338

Politikos, D. V., Fakiris, E., Davvetas, A., Klampanos, I. A., & Papatheodorou, G. (2021). Automatic detection of seafloor marine litter using towed camera images and deep learning. *Marine Pollution Bulletin*, *164*, 111974. https://doi.org/10.1016/j.marpolbul.2021.111974

Proença, P. F., & Simões, P. (2020). *TACO: Trash Annotations in Context for Litter Detection*.

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

Sakti, A. D., Sembiring, E., Rohayani, P., Fauzan, K. N., Anggraini, T. S., Santoso, C., Patricia, V. A., Ihsan, K. T. N., Ramadan, A. H., Arjasakusuma, S., & Candra, D. S. (2023). Identification of illegally dumped plastic waste in a highly polluted river in Indonesia using Sentinel-2 satellite imagery. *Scientific Reports*, *13*(1), 5039. https://doi.org/10.1038/s41598-023-32087-5

Salman, N. (2021). ANALYSIS AND MONITORING OF RIVER WATER QUALITY IN TASIKMALAYA CITY. *Journal of Community Based Environmental Engineering and Management*, *5*(1), 33–40. https://doi.org/10.23969/jcbeem.v5i1.3786

Sari, M. M., Andarani, P., Notodarmojo, S., Harryes, R. K., Nguyen, M. N., Yokota, K., & Inoue, T. (2022). Plastic pollution in the surface water in Jakarta, Indonesia. *Marine Pollution Bulletin*, *182*, 114023. https://doi.org/10.1016/j.marpolbul.2022.114023

Shi, C., Xia, R., & Wang, L. (2020). A Novel Multi-Branch Channel Expansion Network for Garbage Image Classification. *IEEE Access*, *8*, 154436–154452. https://doi.org/10.1109/ACCESS.2020.3016116

Sukmono, Y., Hadibarata, T., Kristanti, R. A., Singh, A., Al Farraj, D. A., & Elshikh, M. S. (2024). Occurrence and visual characterization of microplastics from Mahakam River at Tenggarong City, Indonesia. *Journal of Contaminant Hydrology*, *267*, 104440. https://doi.org/10.1016/j.jconhyd.2024.104440

Sun, B., Tang, H., Gao, L., Bi, K., & Wen, J. (2025). RTDETR-MARD: A Multi-Scale Adaptive Real-Time Framework for Floating Waste Detection in Aquatic Environments. *Journal of Marine Science and Engineering*, *13*(5), 996. https://doi.org/10.3390/jmse13050996

Tamin, O., Moung, E. G., Dargham, J. A., Yahya, F., Farzamnia, A., Sia, F., Naim, N. F. M., & Angeline, L. (2023). On-Shore Plastic Waste Detection with YOLOv5 and RGB-Near-Infrared Fusion: A State-of-the-Art Solution for Accurate and Efficient Environmental Monitoring. *Big Data and Cognitive Computing*, *7*(2). https://doi.org/10.3390/bdcc7020103

Tharani, M., Amin, A. W., Maaz, M., & Taj, M. (2020). *Attention Neural Network for Trash Detection on Water Channels*.

van Lieshout, C., van Oeveren, K., van Emmerik, T., & Postma, E. (2020). Automated River Plastic Monitoring Using Deep Learning and Cameras. *Earth and Space Science*, *7*(8). https://doi.org/10.1029/2019EA000960

Wahyutama, A. B., & Hwang, M. (2022). YOLO-Based Object Detection for Separate Collection of Recyclables and Capacity Monitoring of Trash Bins. *Electronics*, *11*(9), 1323. https://doi.org/10.3390/electronics11091323

Wang, J., & Zhao, H. (2024). Improved YOLOv8 Algorithm for Water Surface Object Detection. *Sensors*, *24*(15), 5059. https://doi.org/10.3390/s24155059

Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., & Hu, Q. (2020). ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11531–11539. https://doi.org/10.1109/CVPR42600.2020.01155

Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018). *CBAM: Convolutional Block Attention Module* (pp. 3–19). https://doi.org/10.1007/978-3-030-01234-2_1

World Bank. (2020). *Stemming the Plastic Tide in Indonesia: Policy, Investments, and Research*. https://www.worldbank.org/en/news/feature/2020/10/06/stemming-the-plastics-tide-in-indonesia

Wu, C. M., Sun, Y. Q., Wang, T. J., & Liu, Y. L. (2022). Underwater trash detection algorithm based on improved YOLOv5s. *Journal of Real-Time Image Processing*, *19*(5), 911–920. https://doi.org/10.1007/s11554-022-01232-0

Wu, G., Ge, Y., & Yang, Q. (2023). UTD-YOLO: underwater trash detection model based on improved YOLOv5. *Journal of Electronic Imaging*, *32*(06). https://doi.org/10.1117/1.JEI.32.6.063034

Xia, Z., Zhou, H., Yu, H., Hu, H., Zhang, G., Hu, J., & He, T. (2024). YOLO-MTG: a lightweight YOLO model for multi-target garbage detection. *Signal, Image and Video Processing*, *18*(6–7), 5121–5136. https://doi.org/10.1007/s11760-024-03220-2

Xian, R., Tang, L., & Liu, S. (2024). Development of a Lightweight Floating Object Detection Algorithm. *Water*, *16*(11), 1633. https://doi.org/10.3390/w16111633

Xiao, R., Wang, H., Wang, L., & Yuan, H. (2025). C3Ghost and C3k2: performance study of feature extraction module for small target detection in YOLOv11 remote sensing images. In S. S. Agaian (Ed.), *Second International Conference on Big Data, Computational Intelligence, and Applications (BDCIA 2024)* (p. 139). SPIE. https://doi.org/10.1117/12.3059792

Xu, S., Tang, H., Li, J., Wang, L., Zhang, X., & Gao, H. (2023). A YOLOW Algorithm of Water-Crossing Object Detection. *Applied Sciences*, *13*(15), 8890. https://doi.org/10.3390/app13158890

Yu, J., Zheng, H., Xie, L., Zhang, L., Yu, M., & Han, J. (2023). Enhanced YOLOv7 integrated with small target enhancement for rapid detection of objects on water surfaces. *Frontiers in Neurorobotics*, *17*. https://doi.org/10.3389/fnbot.2023.1315251

Yu, R.-S., Yang, Y.-F., & Singh, S. (2023). Global analysis of marine plastics and implications of control measure strategies. *Frontiers in Marine Science*, *10*. https://doi.org/10.3389/fmars.2023.1305091

Zhang, L., Wei, Y., Wang, H., Shao, Y., & Shen, J. (2021). Real-Time Detection of River Surface Floating Object Based on Improved RefineDet. *IEEE Access*, *9*, 81147–81160. https://doi.org/10.1109/ACCESS.2021.3085348

Zhang, Y., Wang, X., Shakeel, M. S., Wan, H., & Kang, W. (2022). Learning upper patch attention using dual-branch training strategy for masked face recognition. *Pattern Recognition*, *126*, 108522. https://doi.org/10.1016/j.patcog.2022.108522

Zhang, Y., Zhang, H., Huang, Q., Han, Y., & Zhao, M. (2024). DsP-YOLO: An anchor-free network with DsPAN for small object detection of multiscale defects. *Expert Systems with Applications*, *241*, 122669. https://doi.org/10.1016/j.eswa.2023.122669

Zhao, P., Guo, Y., Yang, Z., Wang, Z., Wang, H., & He, Y. (2024). *YOLOv8_CB: An Improved YOLOv8 Model with CBAM and BiFPN for Pipeline Girth Weld Defect Detection* (pp. 372–383). https://doi.org/10.1007/978-981-96-0313-8_28