NCE Journal of Science and Engineering (NJSE), Vol. 2, June 2025

Unveiling Waste and its Ripple Effect: Analyzing the Impact of Waste of Water Resources Quality

Bipun Man Pati[1,*], Bishnu Khadka[1], Ukesh Thapa[1], Sujay Kumar Pal[1], Dhiraj Pyakurel[1]
[1]AI Research Center, Advanced College of Engineering and Management, Tribhuvan University Kathmandu, Nepal
[*]Correspondence: bipunmanpati, bishnu.khadka, ukesh.thapa, sujay.pal, dhiraj}@acem.edu.np

Manuscript received October 22, 2024; accepted January 12, 2025

Abstract- **Object detection and image segmentation have become powerful tools for performing most of computer vision tasks, including the detection of waste in aquatic environments. However, these approaches have yet to be implemented for waste detection in Nepali rivers. To fill this gap, this research assesses the effectiveness of the current State-of-The-Art (SOTA) object detection models, namely YOLOv5 and YOLOv7, and image segmentation techniques such as Fully Convolutional Network (FCN) and DeepLabv3+. We evaluate the method at two different sites: Dhobi Khola (D1) and Mahadev Khola (D2). YOLOv5 and YOLOv7 are evaluated in terms of mean Average Precision (mAP) at an Intersection over Union (IoU) threshold of 0.5 and F1 score. For segmentation models, we use the Dice score, mean intersection over Union (mIoU), and the F1 score as evaluation criteria. The metrics reveal that DeepLabv3+ is the best-performing model for segmentation tasks, with 0.811 and 0.832 mIoU in D1 and D2, respectively. For object detection, YOLOv7x is the better performing model on dataset D1 with an F1 score of 0.866 and an mAP of 0.862, while YOlOv5m outperformed other models with an mAP score of 0.915 and an F1 score of 0.854 on D2. The findings of this study highlight the effectiveness of the proposed deep learning models in detecting waste in riverine environments. In addition, the superior performance of the selected models underscores their potential as promising approaches for large-scale environmental monitoring and waste management applications.**

Keywords - Waste Detection, Object Detection, Semantic Segmentation, Yolov5 · Yolov7 · DeepLabv3+ · FCN

I. INTRODUCTION

Rivers are a major contributor to marine plastic pollution. More than 1000 rivers account for approximately 80 percent of global riverine plastic emissions into the ocean [1]. In South Asian countries such as Nepal, river pollution is severe and critical near urban areas due to the huge amount of pollution caused by urban activities. The Bagmati River in Kathmandu Valley is currently experiencing severe pollution with a biochemical oxygen demand (BOD) in the range of 20-30 mg/liter, and the total coliform is 104-105 MPN/100 ml [2]. The Bagmati River is heavily polluted due to uncontrolled and increasing urbanization, solid waste disposal, and uncontrolled discharge of domestic and industrial wastewater [3]. This improper disposal of solid waste in and around river systems degrades not only water quality but also aquatic ecosystems and human health; for instance, research by Devkota et al. found that haphazard disposal of solid waste on the banks of the Bishnumati River has severely deteriorated surface and subsurface water quality [4]. To avoid such situations, it is necessary to study not only the major rivers but also their tributaries, as they serve as critical pathways for the transport of pollutants [5]. The research on river waste detection and quantification in river systems is essential for understanding pollution levels, designing clean-up measures, and planning waste management policies. Traditionally, sampling and visual observations have been the primary methods for detecting and assessing waste in rivers. This approach, while straightforward, is labor-intensive, time-consuming, and often inconsistent due to human subjectivity [6]. However, automation aided by Machine Learning (ML) helps mitigate these limitations. Deep Learning (DL), a subset of ML, offers more accurate and scalable solutions compared to traditional methods [7]. The recent developments in Convolutional Neural Networks (CNN) have created a thriving environment for learning features from images, resulting in image classification, segmentation, and detection. Building on these advancements, a considerable amount of literature has been published on the detection of waste in water bodies using DL, specifically for macroplastic litter (plastic items > 5 mm), with most studies relying on camera images due to their availability and dataset size [8]. The effectiveness of such models depends not only on the algorithms but also on the quality and accessibility of image data. Using high-resolution aerial imagery from satellites is expensive. Unmanned Aerial Vehicles (UAVs) are considered a flexible and more economical alternative for aerial imagery without compromising on the quality of images. This is because UAVs can provide images with a high resolution and high image acquisition frequency [9]. However, manual interpretation of UAV images is slow, error-prone, and resource-intensive. To address this challenge, modern DL models, particularly CNN-based approaches, are increasingly used for automated waste detection using UAV-captured imagery [7].

The detection of waste in rivers is typically addressed using object detection and image segmentation [8]. Object detection focuses on identifying and localizing waste items in an image, drawing bounding boxes around the object. This is particularly useful for the quantification of waste. Segmentation, on the other hand, goes a step further by precisely delineating the boundaries of each waste region, pixel by pixel. This helps in the quantification of the area covered by the waste, which can be a better indicator of pollution severity than just the number of items. Our study employs object detection and segmentation to take advantage of the strengths of each approach and improve waste detection in riverine environments.

Research addressing waste detection in rivers for an object detection task has predominantly relied on models from the You Only Look Once (YOLO) [10] family of models [11, 12, 13, 14, 15], or adaptations derived from it [16, 17, 18]. Numerous studies chose YOLOv5 as their model of choice [11, 14, 13, 12], while Yang et al. used an adaptation derived from YOLOv5 [17]. In contrast, Tharani et al. [19] and van Lieshout et al. [20] used MLDet (VGG) [21] and Faster-RCNN [22] with Inceptionv2 [23] for river litter detection, respectively. A study by Maharjan et al. on river plastic detection using UAVs and YOLO-based DL techniques compared a number of YOLO models, from YOLOv1 to YOLOv5, with YOLOv5s achieving the highest mAP of 0.81 with low computational costs [11]. However, a study by Cordova et al. recorded an improved AP@0.5 of 84.9% for YOLOv5x over 79.9% for YOLOv5s on the PlastOPol dataset [24]. As such, the YOLOv5 family (both YOLOv5s and YOLOv5x) was selected, with different performance and efficiency trade-offs suitable for different deployment scenarios in river waste detection. In addition to that, a study by Roshni et al. employed YOLOv5 and YOLOv7 for trash detection in aquatic environments, including underwater trash detection [25]. Based on the study by Yusof et al., YOLOv7 performed better than YOLOv5 and YOLOv6 in road defect detection with the highest mAP of 0.79 [26]. This was also supported by the research of Gasparovic et al., proving that YOLOv7 scored a high 0.963 mAP, which exceeded YOLOv6's 0.953 mAP [27]. Therefore, the YOLOv7 application was considered to be useful for this study, in alignment with its proved superiority over YOLOv5 and YOLOv6 in different detection tasks.

Several studies have explored image segmentation techniques for litter detection in various water bodies, like seashores [28] [29], marine surfaces [30], marine underwater [31], lakes [32], and rivers [19, 32, 33, 34]. U-Net [35] and DeepLabv3+ [36] are popular CNNs that have found extensive applications in diverse segmentation tasks due to their flexibility in use across various domains [37]. Given that the U-Net model is inspired by the Fully Convolutional Network (FCN) [38], using FCN as a baseline for our study is beneficial, as it provides a foundation for comparison with more advanced segmentation architectures like DeepLabv3+. Additionally, a study by Shi et al. [39] showed that DeepLabv3+ performs with a high accuracy of 91.7% in waste area segmentation tasks for real-time operations like application in sweeping robots. Furthermore, adaptations of DeepLabv3+ have shown to achieve a 0.82 mIoU score on manually labeled multispectral imagery data of floating plastics [33]. Given its strong performance in waste detection and segmentation, we incorporate DeepLabv3+ in addition to FCN for our study.

The primary contributions of this article are summarized as follows:

We assess the effectiveness of YOLOv5 and YOLOv7, two widely used object detection models, for the detection of waste in riverine environments. We analyze the performance of DeepLabv3+ and FCN in segmenting waste regions within river systems. The remainder of the paper is organized as follows: Section 2 provides the materials and methods of the study, which include the study area, material used, methodology, and the performance indicators of the models. Section 3 then describes the results. Discussion of the results is in the Section 4 and finally the conclusion is in Section 5.

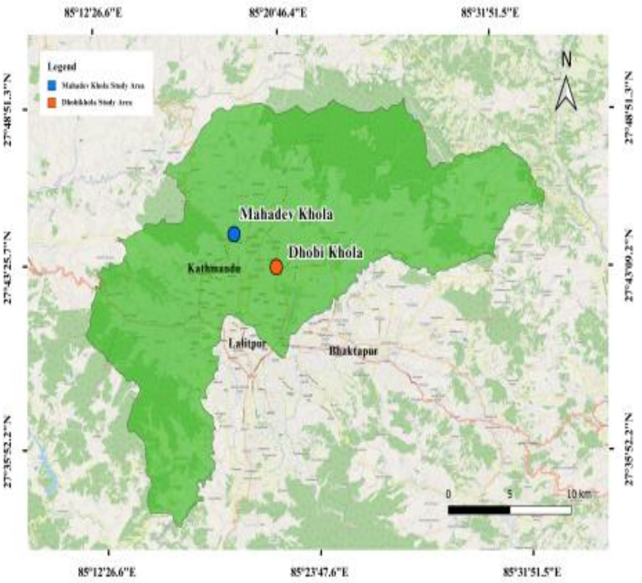

Fig. 1: Location of the study site.

II. Materials and Methods

A. Study Area

Data collection was conducted in the Dhobi Khola (D1) and Mahadev Khola (D2) located in Kathmandu, Nepal, as shown in Figure 1. D1 is a tributary of the larger Bagmati River based in Kathmandu, the capital city. On the other hand, D2 is a tributary of the Bishnumati River. The Bishnumati River flows into the Bagmati River, which is also located in Kathmandu. The Bagmati River is part of the Koshi River basin, which ultimately drains into the Ganges basin. We chose the study sites based on considerations of data collection accessibility and their role as major sources of waste flowing into downstream areas.

B. Materials

For this study, aerial surveys were performed using a DJI Mavic 3 Enterprise drone equipped with a 20-megapixel imaging sensor. This drone was selected for its high-resolution imaging capabilities and stable flight performance, making it well-suited for environmental monitoring. Photos were taken at a 5280 x 3956 pixel resolution with the camera's ISO set to 100. To maintain consistent image quality, flight paths were planned to ensure uniform coverage and minimize variations in lighting conditions. The shutter speed and aperture were configured for automatic adjustment. UAV flights were conducted at an altitude of 45 meters above ground level. It yielded a Ground Sample Distance (GSD) of 0.6125 cm or less across the study area, ensuring a high level of spatial detail in the collected data.

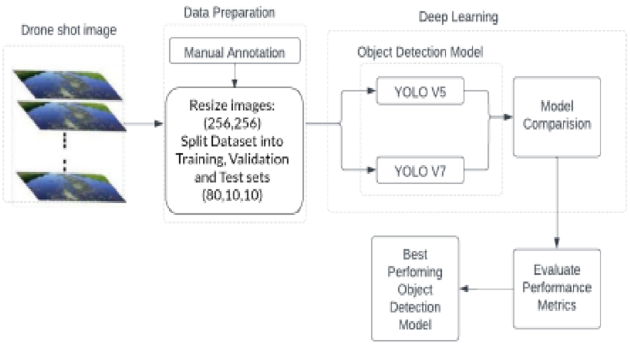
Drone shot image
Data Preparation
Manual Annotation
Resize images: (256,256) Split Dataset into Training, Validation and Test sets (80,10,10)
Deep Learning
Object Detection Model
YOLO V5
YOLO V7
Model Comparision
Evaluate Performance Metrics
Best Perfoming Object Detection Model

Fig. 2. Object detection methodology.

C. Methodology

This section provides an overview of the various DL model architectures employed in the study, along with the proposed methodologies for plastic waste detection in rivers. The primary objective is to assess model performance in detecting and classifying waste in riverine environments using aerial imagery and advanced DL techniques. To achieve this, we first collected image data from the D1 and D2. The acquired images were preprocessed by cropping them to a standardized resolution of 256 × 256. Subsequently, these images were manually annotated to create ground truth labels for both object detection and semantic segmentation tasks, as illustrated in Figures 2 and 3 respectively. Annotation was conducted to ensure high-quality training data, enabling accurate model predictions. Both datasets D1 and D2, annotated for the object detection and segmentation tasks, were split into an 80:10:10 ratio for training, validation, and testing, respectively. In the semantic segmentation task, we trained FCN and DeepLabv3+ architectures on the train set of the dataset. The best-performing model was selected based on evaluation metrics, including mean Intersection 3 over Union (mIoU), Dice coefficient, Precision, and Recall scores, as depicted in Figure 3. Models are evaluated on the test set of the corresponding dataset. These metrics provide a comprehensive assessment of model effectiveness in segmenting plastic waste from river imagery. Similarly, for the object detection task, we trained multiple variations of the YOLOv5 and YOLOv7 architectures, selecting the optimal model based on key performance metrics, as shown in Figure 2. These models were evaluated using standard object detection metrics such as mean Average Precision (mAP), Precision, Recall, and F1-score. By comparing the evaluation metrics across different models, we identify the best-performing model for the dataset.

YOLOv5: The YOLOv5 model, part of the "You Only Look Once" (YOLO) family of architectures, is a SOTA object detection framework known for its speed and accuracy. YOLOv5, like other versions from the YOLO family of architectures, specializes in detecting and localizing multiple objects within an image by simultaneously predicting bounding boxes and class probabilities. This makes it particularly effective for tasks requiring the detection of diverse objects in complex environments, such as waste detection in rivers from aerial imagery. The network is divided into the backbone, neck, and detection head. It uses CSPDarkNet53 as its backbone. In the backbone, the Cross-Stage Partial (CSP) network maintains the original accuracy while improving the inference time. It also solves the problem of network learning duplicate gradients for different layers of the network, caused by multiple skip connections. It solves the problem by 4 dividing the input into two parts, one of which goes through the usual in-between layers and another gets concatenated after that block. Due to this advantage of the CSP network, the DarkNet53 backbone from the YOLOv3 [40] is modified to utilize the CSP network, resulting in the CSPDarkNet53 backbone for feature extraction. To further enhance this feature extraction process, the neck part of the model is used. YOLOv5 achieves this enhanced feature extraction by using Spatial Pyramid Pooling Fast (SPPF), a faster version of Spatial Pyramid Pooling (SPP), and CSP_PANet, an adaptation of PANet that uses CSP between some of the PANet layers. It concatenates the feature map instead of adding it, as performed in PANet. This improves the processing speed of the model, making the feature enhancement process faster. Lastly, anchors are used for the detection head of the network, like in the YOLOv3 model. For this study, the YOLOv5 model was used without any modifications.

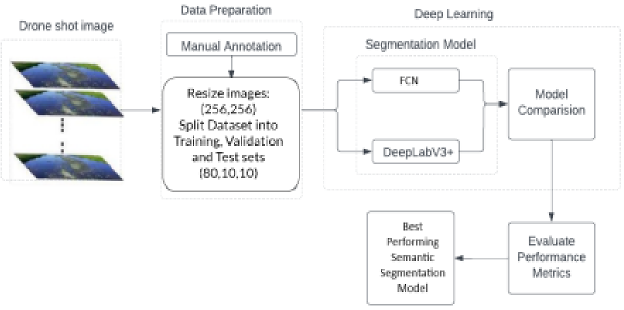
Drone shot image
Data Preparation
Manual Annotation
Resize images: (256,256) Split Dataset into Training, Validation and Test sets (80,10,10)
Deep Learning
Segmentation Model
FCN
DeepLabV3+
Model Comparision
Evaluate Performance Metrics
Best Performing Semantic Segmentation Model

Fig. 3. Semantic Segmentation Methodology

YOLOv7: With a significant improvement in speed and improvement in accuracy, YOLOv7 marked a substantial improvement from its predecessors. The YOLOv7 model introduces novel modifications such as Extended Efficient Layer Aggregation Network (E-ELAN), model scaling, planned re-parameterized convolution, and penalty for lead loss. Similar to the YOLOv5 model, the YOLOv7 network is also divided into three parts: the backbone, the neck, and the detection head. The use of E-ELAN enhanced the management of the gradient path, hence improving its efficiency. The ELAN layer consists of multiple CBS structures. Here, CBS structure means convolutional layer, batch normalization layer, followed by SiLU activation. Additionally, the implementation of model scaling allowed for scalable models, i.e., the ability to create models of varying sizes. This helps in maintaining the optimal model structure while mitigating hardware resource consumption. The CUC layer is the basic unit of feature map combination, including convolution, up-sampling, and combining feature maps. The REP layer is a novel concept that uses the skill of structural reparameterization to adjust the structure in inference to improve the performance of the model. There are multiple heads in the YOLOv7 architecture. The auxiliary head contributes to the training process, while the lead head produces the final output. YOLOv7 assigns coarse labels for the auxiliary head, whereas fine labels are assigned for the lead head. Therefore, two types of losses are employed. (auxiliary loss and main loss). With the help of an assistant loss, the weights of the auxiliary heads are updated. Additionally, a technique is used after training called reparameterization to improve the model. It does not increase the training time but improves the inference results. For this study, we used two variants of the YOLOv7 model: the regular YOLOv7 model and a larger YOLOv7x model.

Fully Convolutional Network (FCN): FCN is a segmentation model that only uses only convolutional layers. It uses a CNN to extract image features, then transforms the number of channels into the number of classes via a 1×1 convolutional layer, and finally transforms the height and width of the feature maps to those of the input image via the transposed convolution. In FCN, fully connected layers are omitted, making the network capable of producing dense, pixel-wise predictions directly while also having capabilities to support input of multiple sizes. ResNet101 is used as the backbone. The FCN architecture used for this study is provided in the torchvision [41] package by PyTorch. This architecture was chosen to ensure reproducibility, leaving behind community-tested implementations. For training the model, we used an output stride of 16 and a learning rate of 0.04 using a poly-learning rate scheduler.

DeepLabv3+: DeepLabv3+ is a SOTA semantic segmentation model, specializing in detailed mask generation. The use of atrous (dilated) convolution instead of the normal convolution with varying rates in the backbone enables better feature extraction without losing resolution. In DeepLabv3+, output from the backbone is passed to the Atrous Spatial Pyramid Pooling (ASPP) module to capture multiscale context using atrous convolutions with varying rates. This enables better feature extraction without losing resolution quality. ResNet101 is utilized as the backbone for this study. Additionally, DeepLabv3+ uses an encoder-decoder architecture. The use of the decoder helps recover spatial details lost during down sampling. Furthermore, in both the ASPP and the decoder, the use of separable convolutions reduces computational complexity and improves speed; however, it does not reduce accuracy. For training the DeepLabv3+ model, similar to the FCN simulation, we used an output stride of 16 and a poly learning rate scheduler with a starting learning rate of 0.04.

Performance Indicators: For object detection, the classification of True Positives (TP), False Positives (FP), and False Negatives (FN) occurs at the object level. This is based on an Intersection over Union (IoU) threshold, typically set at 0.5. IoU is used to access the degree of overlap between the ground truth and prediction and is calculated using Equation 1.

$$IoU_{class} = \frac{\text{Area of Overlap}}{\text{Area of Union}} = \frac{|A \cap B|}{|A \cup B|}, \qquad (1)$$

where, |A ∩ B| is the area of overlap (intersection) of the predicted and ground truth mask or bounding box, |A ∪ B| is the area of union of the predicted and ground truth mask or bounding box.

A detected bounding box is considered a TP if its IoU with the corresponding ground truth bounding box exceeds this threshold; if not, it is labeled as an FP. Undetected objects are categorized as FNs. On the other hand, image segmentation evaluates these metrics at the pixel level, where each pixel is assigned to either the waste or non-waste class. Unlike object detection, segmentation does not rely on a fixed IoU threshold. Instead, a pixel is deemed a TP if it is correctly classified as part of the target class. Misclassified pixels contribute to the FP and FN counts. This approach allows for a more detailed and fine-grained evaluation of segmentation performance.

Precision: Precision measures how accurate our waste detection is. It tells us what proportion of the objects we identified as waste were actually waste. It is calculated as the number of TP divided by the sum of TP and FP, as is given in Equation 2 below.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}. \qquad (2)$$

Recall: Recall measures the ability of the model to find all the actual positive instances, i.e., waste instances. It is calculated as the number of TP divided by the sum of TP and FN and is given in Equation 3.

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}. \qquad (3)$$

where FN is when a model fails to identify waste material.

mIoU: The mIoU for binary semantic segmentation is given in Equation 4.

$$\text{mIoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}. \qquad (4)$$

Dice-Score: The Dice Score is one of the important measurement metrics to determine how a model performs in semantic segmentation, measuring the overlap between the predicted and ground truth segmentation masks. The Dice Score can be defined as:

$$\text{Dice Score} = \frac{2 \cdot |A \cap B|}{|A| + |B|}, \qquad (5)$$

where, |A ∩ B| is the area of overlap (intersection) of the predicted and ground truth mask or bounding box, |A| is the total area of the predicted mask or bounding box, and |B| is the total area of the ground truth mask or bounding box. **mAP@50:** The mAP is a comprehensive performance metric for multi-class object detection. The Average Precision (AP) scores for each class are averaged to determine it. Since there are only two classes (positive and negative) in binary classification, this reduces to the AP of the single positive class.

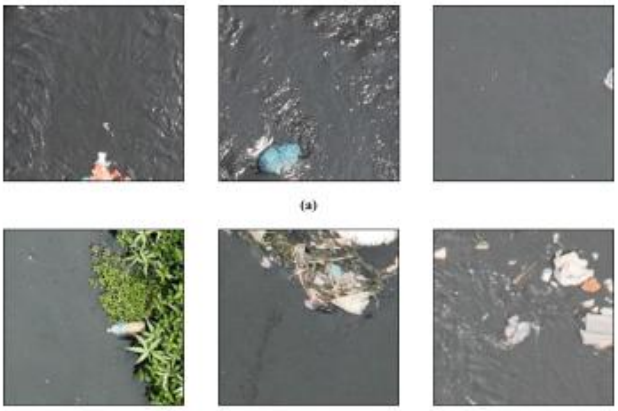

Fig. 4. Sample images from datasets used for training DL models for waste detection in rivers. (a) D1 (b) D2.

For a particular class, the AP itself measures the trade-off between recall and precision. The area under the Precision-Recall (P-R) curve is its formal definition. By adding up the accuracy values at distinct recall thresholds where the precision varies, this region is computationally estimated. mAP@50 is a 6 popular variation of mAP, in which a detection is deemed TP if its IoU with a ground truth bounding box for the same class is larger than or equal to 0.5. Accordingly, mAP (or mAP@50) in our case denotes the AP of the positive (waste) class, as determined by the given IoU threshold.

F1-score: The F1 score is a measure of a model's accuracy on a dataset at a specific confidence level and IoU threshold. It is the harmonic mean of precision and recall of the model. It is given in Equation 6.

$$\text{F1-score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (6)$$

III. Results and Discussion

Dataset Preparation: The image dataset consists of images cropped to a size of 256 × 256 pixels. We annotated the images using Supervisely [42] to record the bounding box and segments for each identifiable piece of waste in each image. Manual labeling of waste in the image is a work-intensive task. Though some labeling errors are inevitable due to difficulty detecting the substance, labelers have made every effort to identify just waste. After annotations, images are randomly associated with training, validation, and test sets with a split ratio of 80:10:10, respectively. The dataset represents only floating or above-water waste. Sample images from the D1 and D2 datasets are shown in Figure 4.

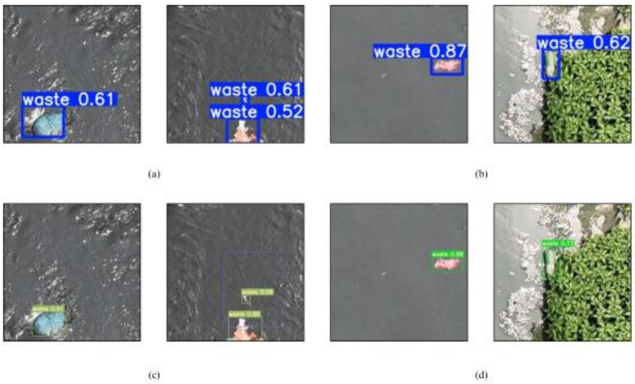

Fig. 5. Object Detection Result: (a) YOLOv5l results on dataset D1 (b) YOLOv5m results on dataset D2 (c) YOLOv7x results on dataset D1 (d) YOLOv7 results on dataset D2.

TABLE 1: OBJECT DETECTION RESULTS

Dataset	Model	Precision	Recall	F1 Score	mAP@50
D1	YOLOv5n	0.917	0.733	0.815	0.808
	YOLOv5s	0.865	0.858	0.861	0.868
	YOLOv5m	0.833	0.733	0.780	0.782
	YOLOv5l	**0.959**	0.730	0.828	**0.875**
	YOLOv5x	0.867	0.860	0.863	0.862
	YOLOv7	0.824	0.633	0.716	0.715
	YOLOv7x	0.867	**0.866**	**0.866**	0.862
	YOLOv7w6	0.869	0.733	0.793	0.764
	YOLOv7e6	0.893	0.831	0.835	0.839
	YOLOv7d6	0.846	0.761	0.801	0.761
	YOLOv7e6e	0.888	0.800	0.842	0.844
D2	YOLOv5n	0.872	0.798	0.833	0.875
	YOLOv5s	0.881	0.750	0.810	0.869
	YOLOv5m	0.940	0.782	0.854	**0.915**
	YOLOv5l	0.888	**0.883**	**0.885**	0.883
	YOLOv5x	0.860	0.800	0.829	0.884
	YOLOv7	0.941	0.800	0.865	0.895
	YOLOv7x	0.864	0.849	0.856	0.893
	YOLOv7w6	0.891	0.817	0.852	0.877
	YOLOv7e6	**0.956**	0.733	0.830	0.874
	YOLOv7d6	0.923	0.800	0.857	0.880
	YOLOv7e6e	0.860	0.816	0.837	0.865

Experimental Results for Object Detection: For the D1 dataset, YOLOv5l achieved the highest precision of 0.959. However, in terms of recall and F1 score, YOLOv7x outperformed all models, achieving a score of 0.866 for both metrics. While YOLOv5l recorded the highest mAP@50 score of 0.875, YOLOv7x attained a comparable mAP@50 score of 0.862. Given its superior F1 score and competitive mAP@50 performance, YOLOv7x is identified as the best-performing model for D1. For the D2 dataset, YOLOv7-e6 demonstrated the highest precision at 0.956, whereas YOLOv5l exhibited superior recall and F1-score, both at 0.883 and 0.885, respectively. Nonetheless, YOLOv5m achieved the highest mAP@50 score of 0.915, emerging as the best model for D2, as mAP@50 serves as a critical benchmark for overall model performance. Among the YOLOv5 variants, YOLOv5l was the best-performing model for the D1 dataset, achieving the highest mAP@50 score of 0.875. Similarly, among the YOLOv7 variants, YOLOv7x demonstrated the best performance with a mAP@50 score of 0.862. While the F1 scores for YOLOv5l (0.828) and YOLOv7x (0.866) indicate strong precision-recall balance, our primary metric remains mAP@50, as it provides a more comprehensive evaluation of detection performance. For the D2 dataset, YOLOv5m outperformed other YOLOv5 variants, attaining a mAP@50score of 0.915, while among the YOLOv7 models, YOLOv7 achieved the highest mAP@50 score of 0.895. Although YOLOv5m and YOLOv7 also exhibited high F1 scores (0.854 and 0.865, respectively), we prioritize mAP@50 as our key performance indicator for a more reliable assessment of model effectiveness.

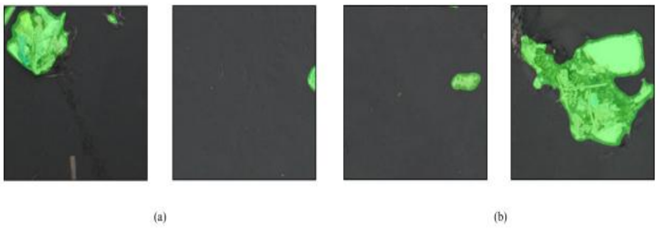

Fig. 6. Segmentation Result dataset (a) DeepLabsV3+ results on dataset D1 (b) DeepLabsV3+ results on dataset D2.

TABLE 2: SEMANTIC SEGMENTATION RESULTS

Dataset	Model	mIoU	Dice	Precision	Recall
D1	FCN	0.782	0.866	0.874	0.858
	DeepLabv3+	**0.811**	**0.886**	**0.900**	**0.873**
D2	FCN	0.792	0.872	0.872	0.871
	DeepLabv3+	**0.832**	**0.901**	**0.903**	**0.899**

Experimental Results for Semantic Segmentation: DeepLabv3+ outperformed the FCN model across all evaluation metrics on Dataset D1. Specifically, DeepLabv3+ achieved a higher mIoU score of 0.811 compared to 0.782 for FCN, as well as a superior Dice score of 0.886 against 0.866. Additionally, it demonstrated higher precision (0.900 vs. 0.874) and better recall (0.873 vs. 0.858). Similarly, in dataset D2, DeepLabv3+ again surpasses the FCN model in all the different metrics under consideration. Instead of FCN's 0.792, DeepLabv3+ produced a mIoU of 0.832. As for Dice coefficient, DeepLabv3+ achieved 0.901, with FCN registering 0.872. On Precision, DeepLabv3+ got 0.903 while FCN had 0.872; and on Recall, DeepLabv3+ had an edge of 0.899 over FCN's 0.871. The segmentation result is visualized in Figure 6. Therefore, across both datasets, DeepLabv3+ consistently demonstrated superior performance, making it the better-performing model for waste segmentation.

Minimum Detectable Waste Object Sizes: Detection of the tiniest waste item observed in datasets D1 and D2 is accomplished by applying both object detection and semantic segmentation methods. The study is carried out with the top models specific to each task on the respective datasets. Specifically, for object detection, YOLOv7x is applied to D1, whereas YOLOv5m is applied to D2. In the case of semantic segmentation, DeepLabV3+ is applied to both datasets. In D1, the minimum size of an object detected is 207.09 cm2 for object detection and 69.27 cm2 for segmentation. In D2, the minimum size of an object detected is 76.16 cm2 for object detection and 165.44 cm2 for segmentation. They provide us with some information regarding the model's capability to detect fine-scale waste objects in the datasets. The smallest detectable waste is illustrated in Figure 7.

For object detection, YOLOv7x was best for D1, with high recall (0.866) and F1-score (0.866), and YOLOv5m dominated D2 with better mAP@50 (0.915). YOLOv7x utilizes expanded E-ELAN to maximize feature reuse and dynamic label assignment for the sake of improving occluded object detection [43]. Such advances account for its improved recall on D1, which potentially holds overlapping waste instances. YOLOv5m's anchor-based approach favors computational efficiency and accuracy [44] and this proves effective on D2, where object appearances can be more homogeneous. Its moderate depth and width probably avoid overfitting to intricate backgrounds. Notably, YOLOv5l achieved the highest precision on D1 (0.959), but its fixed anchor boxes struggled with occluded objects, resulting in lower recall (0.730 vs. YOLOv7x's 0.866). Conversely, YOLOv7-e6's lower precision on D2 (0.956 vs. YOLOv5m's 0.940) suggests overfitting due to its larger parameter count—a limitation in smaller datasets. Dataset characteristics further modulated performance: D2's higher mAP@50 across models (max 0.915 vs. D1's 0.875) indicate fewer annotation inconsistencies or less background clutter.

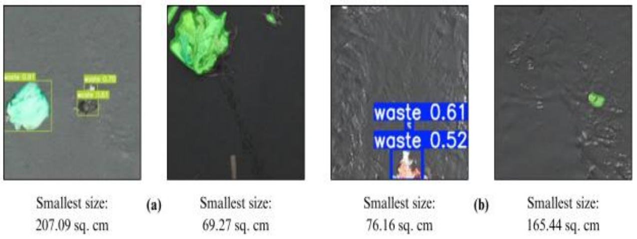
waste 0.61
waste 0.52
Smallest size: 207.09 sq. cm
(a)
Smallest size: 69.27 sq. cm
Smallest size: 76.16 sq. cm
(b)
Smallest size: 165.44 sq. cm

Fig.7. Visualization of the smallest detected waste objects for datasets (a) D1 and (b) D2 using the best-performing models for object detection (YOLOv7x for D1, YOLOv5m for D2) and semantic segmentation (DeepLabV3+ for both datasets).

DeepLabv3+ consistently outperformed FCN across both datasets, achieving superior mIoU, Dice, precision, and recall (Table 2). This dominance is attributable to its ASPP module, which employs multi-scale dilated convolutions to capture contextual information across varying object sizes—a feature critical for segmenting riverine waste with high size diversity [36]. In contrast, FCN's reliance on single-scale upsampling [38] limits its ability to resolve small or fragmented objects, as evidenced by its lower mIoU (0.782 vs. 0.811 on D1). Furthermore, DeepLabv3+'s decoder module refines segmentation boundaries using low-level features, reducing edge artifacts and improving precision (0.900 vs. 0.874 on D1). The visual results (Figure 6) confirm this, with DeepLabv3+ producing sharper delineations of both small and large debris. The performance gap between datasets (e.g., mIoU of 0.832 on D2 vs. 0.811 on D1 for DeepLabv3+) may reflect differences in dataset complexity. D2's higher scores suggest objects with more distinct boundaries or fewer occlusions, whereas D1's cluttered river scenes likely challenge both models. However, DeepLabv3+'s consistent superiority across datasets highlights ASPP's robustness to scale variation, aligning with findings by Chen et al. [36].

IV. CONCLUSION

This research was carried out to assess the performance of modern DL architectures for object detection and semantic segmentation for riverine waste detection, which is a key environmental monitoring challenge. From our results, we observe that DeepLabv3+ outperforms FCN consistently in semantic segmentation because of ASPP and a boundary-refining decoder to combat scale variation and edge errors intrinsic to complicated waste-filled scenes. For object detection, YOLOv7x performed well in Dataset D1, and YOLOv5m recorded better mAP@50 for Dataset D2. The findings add to the collective aspiration of automating environmental monitoring and provide an agenda to weigh the optimization of accuracy, efficiency, and adaptability of waste detection systems. Future research should focus on leveraging large-scale datasets from multiple rivers across different regions of the country and exploring advanced DL approaches to further enhance waste detection and support sustainable environmental management.

V. Acknowledgments

The authors would like to express sincere thanks to The Government of Nepal and District Administration Office, Kathmandu for providing us an opportunity for data collection. In addition, we would like to express sincere thanks to Suraj Paudel of Airlift Technology for their kind support in data collection. We sincerely thank Sneha Subedi and Sahil Rai from the Advanced College of Engineering and Management, Nepal, for assisting in data annotation.

References

[1] Lourens J. J. Meijer, Tim H. M. van Emmerik, Ruud J. van der Ent, Christian Schmidt, and Laurent Lebreton. More than 1000 rivers account for 80% of global riverine plastic emissions into the ocean. Science Advances, 7, 2021.

[2] Sunil Kumar Karn and Hideki Harada. Surface water pollution in three urban territories of nepal, india, and bangladesh. Environmental management, 28:483–496, 2001.

[3] Swastik Ghimire, Nishan Pokhrel, Susmita Pant, Tunisha Gyawali, Apekshya Koirala, Bandita Mainali, Michael J Angove, and Shukra Raj Paudel. Assessment of technologies for water quality control of the bagmati river in kathmandu valley, nepal. Groundwater for Sustainable Development, 18:100770, 2022.

[4] Dinesh Chandra Devkota and Kunio Watanabe. Impact of solid waste on water quality of bishnumati river and surrounding areas in kathmandu, nepal. Journal of Nepal Geological Society, 31:19–24, 2006.

[5] Kinga Wieczorek, Anna Turek, Małgorzata Szczesio, and Wojciech M Wolf. A holistic approach to the spatiotemporal variability investigation of the main river water quality–the importance of tributaries. Science of The Total Environment, 906:167588, 2024.

[6] Pallavi Dubey, John Jackman, Gül E. Kremer, and Paul Kremer. A probabilistic model to estimate automated and manual visual inspection errors. In Kyoung-Yun Kim, Leslie Monplaisir, and Jeremy Rickli, editors, Flexible Automation and Intelligent Manufacturing: The Human-Data-Technology Nexus, pages 685–695, Cham, 2023. Springer International Publishing.

[7] S. Ham, Y. Oh, K. Choi, and I. Lee. Semantic segmentation and unregistered building detection from uav images using a deconvolutional network. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLII-2:419–424, 2018.

[8] Tianlong Jia, Zoran Kapelan, Rinze de Vries, Paul Vriend, Eric Copius Peereboom, Imke Okkerman, and Riccardo Taormina. Deep learning for detecting macroplastic litter in water bodies: A review. Water research, 231:119632, 2023.

[9] Xia Yao, Ni Wang, Yong Liu, Tao Cheng, Yongchao Tian, Qi Chen, and Yan Zhu. Estimation of wheat lai at middle to high levels using unmanned aerial vehicle narrowband multispectral imagery. Remote. Sens., 9:1304, 2017.

[10] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 779–788, 2015.

[11] Nisha Maharjan, Hiroyuki Miyazaki, Bipun Man Pati, Matthew N. Dailey, Sangam Shrestha, and Tai Nakamura. Detection of river plastic using uav sensor data and deep learning. Remote. Sens., 14:3049, 2022.

[12] Shijun Pan, Keisuke Yoshida, Afia S. Boney, and Satoshi Nishiyama. The application of drone-assisted deep learning technology in riverbank garbage detection. Journal of Japan Society of Civil Engineers, Ser. B1 (Hydraulic Engineering), 2022.

[13] Maiyatat Nunkhaw and Hitoshi Miyamoto. An image analysis of river-floating waste materials by using deep learning techniques. Water, 16(10), 2024.

[14] Gilroy Aldric Sio, Dunhill Guantero, and Jocelyn Villaverde. Plastic waste detection on rivers using yolov5 algorithm. In 2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT), pages 1–6, 2022.

[15] Ferdinandus Fidel Putra and Yulius Denny Prabowo. Low resource deep learning to detect waste intensity in the river flow. Bulletin of Electrical Engineering and Informatics, 2021.

[16] Nur Athirah Zailan, Muhammad Mokhzaini Azizan, Khairunnisa Hasikin, Anis Salwa Mohd Khairuddin, and Uswah Khairuddin. An automated solid waste detection using the optimized yolo model for riverine management. Frontiers in Public Health, 10, 2022.

[17] Xingshuai Yang, Jingyi Zhao, Li Zhao, Haiyang Zhang, Li Li, Zhanlin Ji, and Ivan Ganchev. Detection of river floating garbage based on improved yolov5. Mathematics, 2022.

[18] Feng Lin, T., Hou, Qiannan Jin, and Aiju You. Improved yolo based detection algorithm for floating debris in waterway. Entropy, 23, 2021.

[19] M. Tharani, Abdul Wahab Amin, Fezan Rasool, Mohammad Maaz, Murtaza Taj, and A. S. Muhammad. Trash detection on water channels. In International Conference on Neural Information Processing, 2021.

[20] Colin van Lieshout, Kees van Oeveren, Tim van Emmerik, and Eric O. Postma. Automated river plastic monitoring using deep learning and cameras. Earth and Space Science, 7, 2020.

[21] Qijie Zhao, Tao Sheng, Yongtao Wang, Zhi Tang, Ying Chen, Lingyi Cai, and Haibin Ling. M2det: A single-shot object detector based on multi-level feature pyramid network. In AAAI Conference on Artificial Intelligence, 2018.

[22] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39:1137–1149, 2015.

[23] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2818–2826, 2015.

[24] Manuel Alberto Cordova Neira, Allan da Silva Pinto, Christina Carrozzo Hellevik, Saleh Abdel-Afou Alaliyat, Ibrahim A. Hameed, Hélio Pedrini, and Ricardo da Silva Torres. Litter detection with deep learning: A comparative study. Sensors (Basel, Switzerland), 22, 2022.

[25] Nishat Mahmud Roshni, Meshkatul Arefin, Kazi Masfiqul Alam Joy, Marjanul Hassan, and Shashwata Karmakar. Trash detection in an aquatic environment. PhD thesis, Brac University, 2024.

[26] Najiha‘Izzaty Mohd Yusof, Ali Sophian, Hasan Firdaus Mohd Zaki, Ali Aryo Bawono, Arselan Ashraf, et al. Assessing the performance of yolov5, yolov6, and yolov7 in road defect detection and classification: a comparative study. Bulletin of Electrical Engineering and Informatics, 13(1):350–360, 2024.

[27] Boris Gaparovic, Goran Maua, Josipa Rukavina, and Jonatan Lerga. Evaluating yolov5, yolov6, yolov7, and ´ yolov8 in underwater environment: Is there real improvement? 2023 8th International Conference on Smart and Sustainable Technologies (SpliTech), pages 1–4, 2023.

[28] Kyriaki Kylili, Alessandro Artusi, and Constantinos Hadjistassou. A new paradigm for estimating the prevalence of plastic litter in the marine environment. Marine pollution bulletin, 173 Pt B:113127, 2021.

[29] Shin'ichiro Kako, Shohei Morita, and Tetsuya Taneda. Estimation of plastic marine debris volumes on beaches using unmanned aerial vehicles and image processing based on deep learning. Marine pollution bulletin, 155:111127, 2020.

[30] J. Mifdal, N. Longépé, and M. Rußwurm. Towards detecting floating objects on a global scale with learned spatial features using sentinel 2. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, V-3-2021:285–293, 2021.

[31] Hongjie Deng, Daji Ergu, Fangyao Liu, Bo Ma, and Ying Cai. An embeddable algorithm for automatic garbage detection based on complex marine environment. Sensors (Basel, Switzerland), 21, 2021.

[32] Gordana Jakovljevic, Miro Govedarica, and Flor Álvarez-Taboada. A deep learning model for automatic plastic mapping using unmanned aerial vehicle

(uav) data. Remote. Sens., 12:1515, 2020.

[33] Àlex Solé Gómez, Leonardo Scandolo, and Elmar Eisemann. A learning approach for river debris detection. Int. J. Appl. Earth Obs. Geoinformation, 107:102682, 2022.

[34] Shijun PAN, Keisuke YOSHIDA, and Takashi KOJIMA. Comprehensive analysis of on-site riparian waste pollution: A case study on the hyakken river basin. Intelligence, Informatics and Infrastructure, 5(1):98–103, 2024.

[35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. ArXiv, abs/1505.04597, 2015.

[36] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In European Conference on Computer Vision, 2018.

[37] Patrick Nicholas Hadinata, Djoni Simanta, Liyanto Eddy, and Kohei Nagai. Multiclass segmentation of concrete surface damages using u-net and deeplabv3+. Applied Sciences, 2023.

[38] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 3431–3440, 2014.

[39] Lixiang Shi and Guofang Liu. Application of deeplabv3+ network model in garbage detection and classification. International Journal of Wireless and Mobile Computing, 24(3-4):380–389, 2023.

[40] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. ArXiv, abs/1804.02767, 2018.

[41] TorchVision maintainers and contributors. Torchvision: Pytorch’s computer vision library. https://github. com/pytorch/vision, 2016.

[42] supervisely. Accessed: 2024-12-29.

[43] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 7464–7475, 2022.

[44] Haiying Liu, Fengqian Sun, Jason Jianjun Gu, and Lixia Deng. Sf-yolov5: A lightweight small object detection algorithm based on improved feature fusion mode. Sensors (Basel, Switzerland), 22, 2022.