A Deep Learning Approach to Plastic Bottle Waste Detection on the Water Surface using YOLOv6 and YOLOv7

Naufal Laksana Kirana
Universitas Brawijaya, Indonesia
naufallaksana@student.ub.ac.id

Diva Kurnianingtyas
Universitas Brawijaya, Indonesia
divaku@ub.ac.id (corresponding author)

Indriati
Universitas Brawijaya, Indonesia
indriati.tif@ub.ac.id

Received: 28 September 2024 | Revised: 8 October 2024, 17 October 2024, and 28 October 2024 | Accepted: 29 October 2024

 DOI: https://doi.org/10.48084/etasr.8592

ABSTRACT

Deep learning is a branch of machine learning with many layers, such as the You Only Look Once (YOLO) method. From various versions of YOLO, YOLOv6 and YOLOv7 are considered more prominent because they achieve high Mean Average Precision (mAP) values. Both versions of YOLO have been implemented into various problems, especially in the waste detection problem. Plastic bottle waste is one of the most common types of waste that pollutes Indonesian waters. This study aims to solve this problem by helping to sort waste in surface waters by applying YOLOv6 and YOLOv7. FloW-Img was used, obtained on request from the Orcaboat website. The dataset consists of 500,000 bottle objects in 2,000 images. The YOLOv6 and YOLOv7 models were evaluated using mAP and running time. The results show that YOLOv6 and YOLOv7 can handle bottle waste detection well, with mAP values of 0.873 and 0.512, respectively. In addition, YOLOv6 (4.21 m/s) has a higher detection speed than YOLOv7 (13.7 m/s). However, in tests with images that do not have bottle objects, YOLOv7 provides better detection accuracy and consistency results, making it more suitable for real-world applications that demand high accuracy in environments with much visual noise.

Keywords-artificial intelligence; neural networks; computer vision; optimization; waste detection

I. INTRODUCTION

Deep learning is a branch of machine learning that consists of many layers between input and output. Deep learning models can achieve exceptional performance, sometimes exceeding that of humans [1-2]. You Only Look Once (YOLO) is one of the world's most popular deep-learning methods for object detection and is considered a significant and extensive method for object identification [3, 4]. YOLO has several versions. YOLOv1 is a model for object detection using a neural network, where an input image is divided into a grid that detects objects within it. YOLOv3 uses the Darknet-53 architecture by introducing multiscale detection to detect objects of various sizes. YOLOv4 is easy to use with various features for training, validation, and inference, and can be used in various applications. YOLOv5 includes various features updated from YOLOv4. YOLOX is a model with optimization characteristics from YOLOv4 [5]. PP-YOLOE uses several optimization techniques and improvements from YOLO that focus on improving efficiency and accuracy [6]. YOLOv6 performs focused updates with speed and efficiency, achieving a balance between accuracy and inference speed, and can be used for real-time object detection [7, 8]. YOLOv7 has improvements in architecture and performs efficiently in computational matters [9, 10]. YOLOv8 is an update of the YOLOv5 model by increasing the detection speed to be more efficient when used in real-time [11]. In [12], YOLOv6 achieved higher mean Average Precision (mAP) than other YOLO models. Each YOLO version has differences, including the YOLOv6 and YOLOv7 series. All YOLO models can be utilized in various aspects, such as real-time inspection of fire safety equipment, skin lesion detection, item detection, and vehicle detection on public roads [8, 13-15].

YOLO can also be used in waste detection. Garbage is one of the problems often faced by countries. Indonesia has a water surface area of 6,400,000 km^2. The vastness of Indonesia's water area makes waste on the water surface a difficult-to-address problem. Waste consists of organic and inorganic waste. According to data from the Ministry of Environment and Forestry, Indonesia is polluted by 35 million tons of waste per year, with the most significant waste being plastic waste, such as plastic bottles. The waste is spread over land and water and ends in the sea. Therefore, waste in Indonesian waters must be resolved, and the Indonesian government has set five strategies to tackle marine debris. The five strategies implemented in the National Action Plan (NAP) for handling marine debris are [16, 17]:

1. Organizing a national movement to socialize the negative impacts of marine debris.
2. Controlling waste in watersheds and controlling plastic waste from the upstream and downstream industrial sectors.
3. Managing plastic waste from marine transportation activities, activities in marine tourism, marine and fisheries, and coastal and small islands.
4. Diversify funding schemes outside of the state budget/APBD, strengthen institutions, and improve the effectiveness of supervision and law enforcement.
5. Encourage innovation in managing and overcoming marine debris pollution through research and development.

Based on data from the National Action Plan for Marine Debris Management, Indonesia reduced 35.36% of plastic waste in the sea, from 615,675 tons in 2018 to only 408,885 tons in 2022. The government aimed to reduce plastic waste in Indonesian marine waters by 70% by 2025 [16-18]. Therefore, an acceleration is needed to achieve the objective. Based on the five strategies above, the waste sorting stage is critical to distinguish between living organisms and bottle waste in the water area. This sorting can be performed by detecting bottle waste objects in the water. Table I shows previous studies on the detection of marine debris. However, the YOLOv6 and YOLOv7 models have not previously been used to solve the problem of waste in water. This study aims to help sort bottle waste in surface waters by examining two YOLO versions to help Indonesia's waste problem.

TABLE I. PREVIOUS STUDIES

Study	Issue	Method	Results
[19]	Underwater garbage	YOLOv4	Detection speed: 66.67 fps, mAP: 95.099%
[20]	Water surface garbage	YOLOv3	Detection speed: 18.47 fps, mAP: 91.43%
[21]	Waste detection	EfficientDet-D2, EfficientDet-B2	Average precision: 70%, Classification accuracy: 75%
[22]	Water classification	EfficientNet-B0, EfficientNet-B3	Accuracy: 84% using B0

II. IMAGE DETECTION MODELING

This study followed the following stages, as shown in Figure 1: dataset collection, dataset exploration, modeling, and model evaluation. The modeling process has three stages: data preprocessing, training model design, and testing YOLOv6 and YOLOv7.

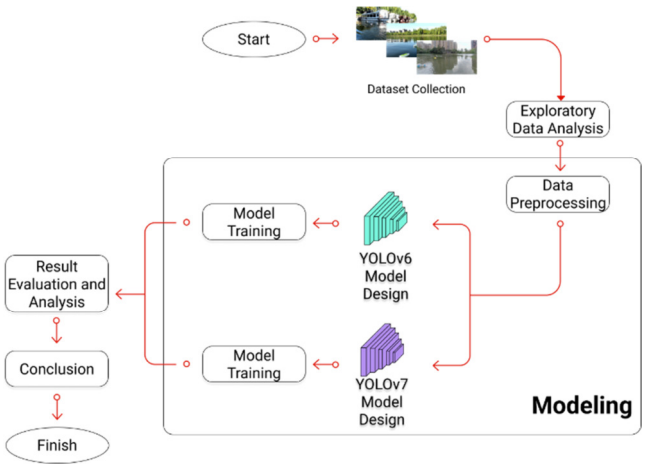

Fig. 1. Research method.

A. Dataset

The data used for modeling is an image dataset of bottle waste on the water surface comprising 500,000 bottle objects in 2000 images. The data was obtained by requesting the Orcaboat company to provide access to the data from its website [23, 24].

B. Preprocessing

The data consists of images with two different sizes, namely 1280×720 and 1280×640 px. As shown in Figure 2, some adjustments were made by converting the images to 640×640 px size with bounding box labels on each object in the image. The data was divided into training and validation sets with a 60:40 ratio.

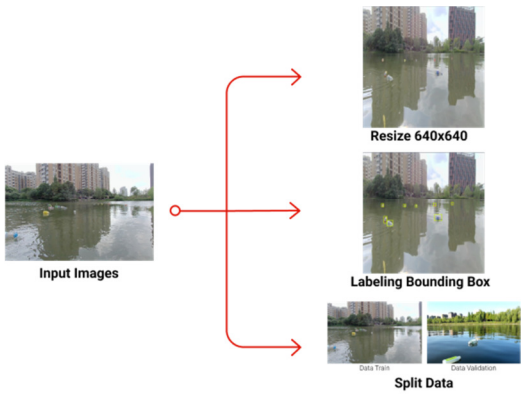

Fig. 2. Data preprocessing.

C. YOLOv6 and YOLOv7 Modeling

YOLO is a set of algorithms created for real-time object detection. YOLO has many versions, including YOLOv1-3

[11, 25, 26], which are the pioneers of the YOLO series. YOLOv4 [4] reorganized the detection architecture into the backbone, neck, and head. More recent efficient detection models in the YOLO series are YOLOv5, YOLOX, PP-YOLOE, YOLOv6, YOLOv7, and YOLOv8. YOLO performs well in detecting normal-sized objects but cannot detect small-sized objects [27].

The YOLO architecture starts with the input image passing through multiple convolution layers with various filter sizes (e.g., 7×7, 3×3, and 1×1) that serve to extract features from an image (see Figure 3). Each convolution layer is followed by a pooling or subsampling operation that reduces the spatial dimensions and helps the network to capture more complex features. This architecture involves several additional deep convolution layers to capture features at multiple scales. After the last convolution layer, the feature results are extracted and compiled by the complete connection layer, which results in the final prediction of bounding box coordinates and object classification. The YOLO architecture is designed to efficiently perform object detection in a single stage, thus enabling a speedy detection process [28].

YOLOv6 consists of three parts: the head, the backbone, and the neck. YOLOv6 is updated in the detector neck with Bidirectional Concatenation (BiC) [7]. The neck and backbone structures have been replaced with Rep-PAN and EfficientRep [26]. The YOLOv6 architecture, shown in Figure 4, consists of three main parts: the backbone, neck, and head. In the backbone section, marked with a red box, several blocks, such as RepVGG Block and SimCSPPF Block, are used to extract features from the input image. Meanwhile, the neck, marked with a green box, combines and refines the characteristics using a combination of blocks such as the BiC module, 3×3 Conv, 1×1 Conv, and Concatenate. Finally, the head section, marked with a blue box, uses the efficient decoupled head to process the features generated by the neck and produce the output as a final prediction for object detection. This architecture is designed to improve efficiency and accuracy in object detection tasks. In addition, it simplifies the SPPF block into a SimCSPSPFF block. YOLOv6 also uses Anchor-Aided Training (AAT), which is both anchor-based and anchor-free [29, 30].

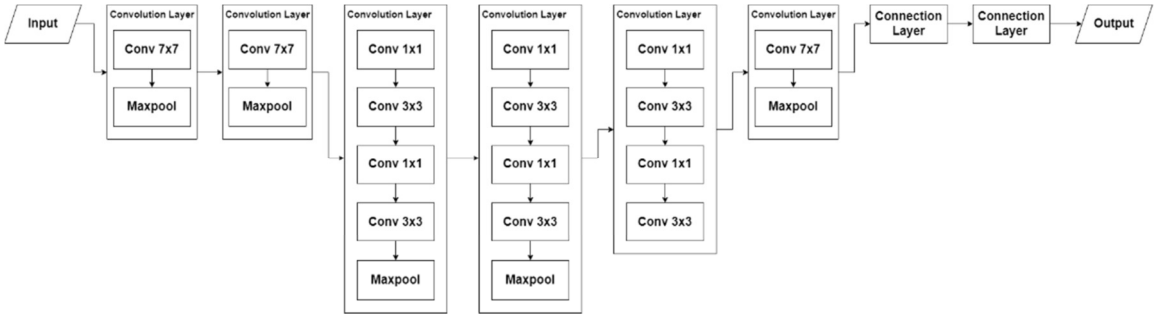

Fig. 3. YOLO architecture.

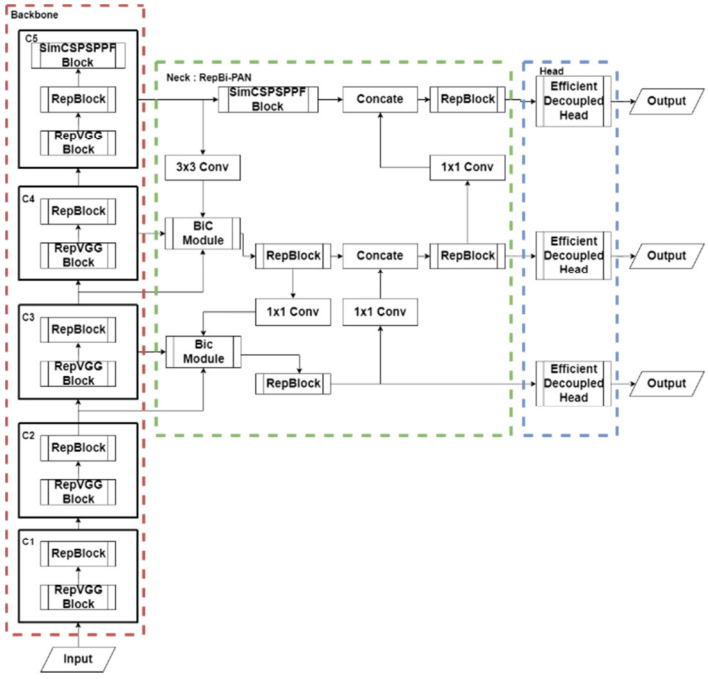

Fig. 4. YOLOv6 architecture.

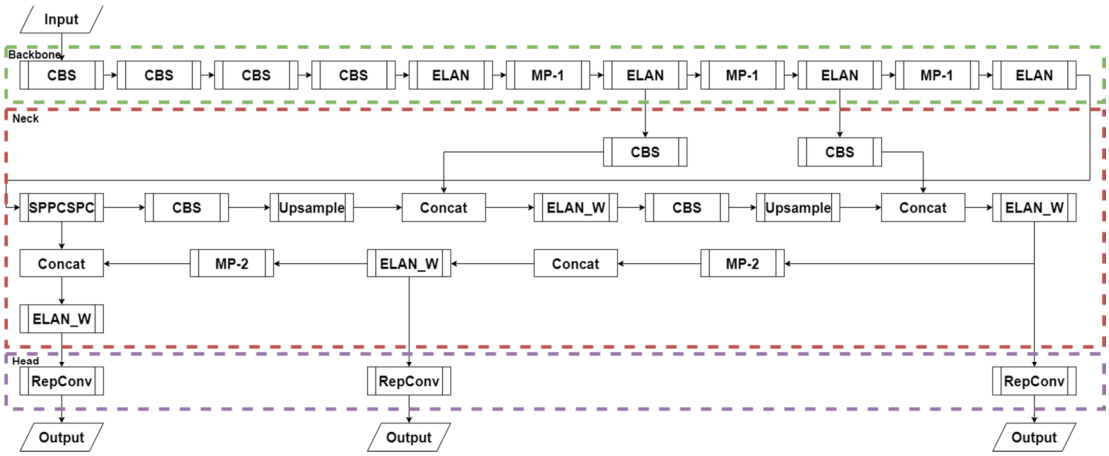

Fig. 5. YOLOv7 architecture.

YOLOv7 is a relatively new development of YOLO. YOLOv7 also comprises three parts, as shown in Figure 5: the backbone, neck, and head. YOLOv7 is an update of the previous YOLO series on the backbone architecture for lower power consumption. In the backbone section, marked with a green box, a series of blocks such as Convolution BatchNorm SiLU (CBS), ELAN, and MaxPooling (MP) are used to extract features from the input image. The neck section, marked with a red box, processes the features using a combination of blocks such as SPPCSPC, upsample, concatenation, and ELAN_W to enhance the network's ability to capture multi-scale information. Finally, the head section, marked with a purple box, uses RepConv to process the features of the neck and generate the final prediction for object detection. This architecture is designed to improve object detection efficiency and accuracy, focusing on optimizing feature processing at multiple scales [31, 32].

D. Evaluation Method

Evaluation metrics are a way to measure the performance of a particular model. Evaluation can be performed using the mean value of Average Precision (mAP) [33, 34]. The mAP value is obtained from the Average Precision (AP) from a set of data. AP is a key performance indicator that eliminates the dependence on selecting a single threshold value. The AP value is obtained from the area under the precision-recall curve. AP summarizes the precision-recall curve into a single scalar value. The precision-recall curve is created to visually depict the best threshold. The precision-recall curve plots the precision value against the recall for a threshold value. Precision and recall values depend on True Positives (TP), False Positives (FP), and False Negatives (FN), which can be calculated using (1) for recall and (2) for precision. The determination of these values depends on the predicted label compared to the ground truth and the Intersection over Union (IoU) value between the two bounding boxes, i.e., ground truth and prediction [33].

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (1)$$

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

The threshold probability distribution is determined between 0 and 1 to classify the bounding box. The prediction is declared correct if the label class on the predicted bounding box is the same as the ground truth bounding box or the IoU value is higher than the threshold value. The following values are obtained based on the IoU, threshold, and label class on the predicted and ground truth bounding boxes. TP is obtained when the model predicts that the bounding box is in a favorable value position, and it is true. FP is obtained when the model predicts that the bounding box is in a positive value position, but that is wrong. FN is obtained when the model does not predict that the bounding box is in a particular negative position and is false. True Negative (TN) is obtained when the model does not predict that the bounding box is in a negative position, and it is true. Thus, mAP obtained from AP, which is sourced from precision and recall, can be calculated using (3) for AP and (4) for mAP [35].

$$\text{AP} = \sum_{k=0}^{k=n-1}\big(\text{R}(k) - \text{R}(k-1)\big)\text{P}(k) \quad (3)$$

$$\text{mAP} = \frac{1}{n}\sum_{i=1}^{n}\text{AP}_i \quad (4)$$

III. RESULTS AND DISCUSSION

A. Exploratory Data Analysis (EDA)

EDA is a step taken to analyze data and aims to thoroughly understand the data characteristics before data preprocessing. Table II shows the values of the FloW-Img data features. The table contains count, Mean, Std, Q1, Q2, Q3, Min, and Max values. These values are obtained from width, height, saturation, contrast, and the most significant object size in the FloW-Img data, which consists of 2000 data points. The Width variable has a constant value of 1280 px for all data, indicating no variation in width. In contrast, Height varies with a mean of 689.64 pixels and a standard deviation of 38.83 pixels, indicating slight variation in image height. The median and first and third quartile values for Height are 720 px and 640 px, indicating that most images have similar heights.

More significant variations exist for the saturation, contrast, and largest object size variables. Saturation has a mean of 52.88 with a standard deviation of 38.16, which shows that the saturation level varies quite widely between images, with a

minimum value of 13 and a maximum of 195.22. Contrast has an average of 54.21 with a standard deviation of 18.36, which also shows considerable variation in the contrast of the images. The largest object size has a considerable variation, with a mean value of 405,189.49 pixels and a standard deviation of 244,193.97 pixels, indicating that the size of the largest object in the image varies widely, ranging from 22,274.5 pixels to almost 920,000 pixels. The quartile values also show that half of the data has the largest object below about 320,000 pixels, while 25% of the data has the largest object with a size above 585,822.12 pixels.

TABLE II. RESULTS OF EXPLORATORY DATA ANALYSIS

	Width	Height	Saturation	Contrast	Largest object's size
Count	2000	2000	2000	2000	2000
Mean	1280	689.64	52.88	54.21	405189.49
Std	0	38.83	38.16	18.36	244193.97
Min	1280	640	13	16.41	22274.5
25%	1280	640	24.84	39.73	215135
50%	1280	720	34.69	53.87	320070.25
75%	1280	720	78.56	68.33	585822.12
max	1280	720	195.22	95.38	919505

Figure 6 shows a heat map of the correlation between four image attributes, namely height, saturation, contrast, and largest object size. Each section of this heat map shows the correlation between two attributes, with colors ranging from dark blue (strong negative correlation) to dark red (strong positive correlation). For example, height has a robust negative correlation with saturation (-0.82) and contrast (-0.75), as indicated by the dark blue color. This correlation means that as the height of the image increases, the saturation level and contrast tend to decrease. Height shows a moderate positive correlation with the largest object size (0.57), indicating that images with larger, more significant objects also tend to have more considerable heights. There is a positive correlation between saturation and contrast (0.63) and a negative correlation between contrast and the largest object size (-0.76). These visualizations make it easy to identify relationships between variables, which is important for model development.

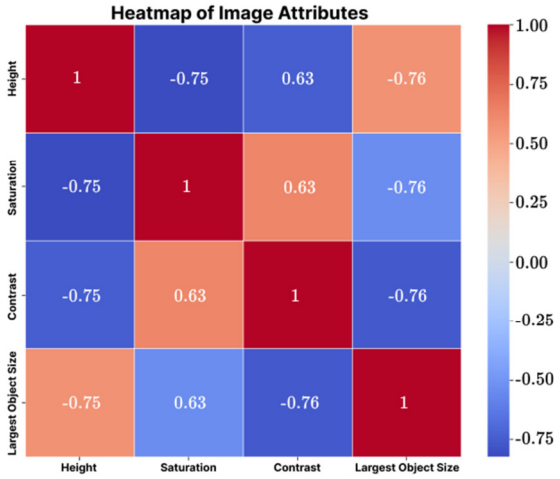

Fig. 6. Heatmap for dataset features.

B. YOLOv6 Results

YOLOv6 was trained and evaluated by performing 10 to 100 epochs to determine the resulting mAP value. The mAP score was obtained using (4). According to Table III, increasing epochs tends to increase the mAP value of the model by 0.873 in 100 epochs. Meanwhile, as seen in Figure 7, the recall curve decreases drastically with increasing confidence score. This indicates that although YOLOv6 exhibits high sensitivity with good recall at initial detection, its accuracy is compromised as the confidence threshold increases. Furthermore, the precision curve shows a steadily rising curve as the confidence score increases, with precision reaching a maximum value of 1.0 at a confidence of around 0.921 for all object classes. Despite the reduction in detection coverage, YOLOv6 can predict objects very accurately when only high-confidence predictions are taken. The combination of the precision recall curve for YOLOv6 with a mAP@0.5 value around 0.873 reinforces these results, showing that while the precision is quite good, there is a gap in the recall, indicating that the model tends to miss several objects when prioritizing higher accuracy.

C. YOLOv7 Results

YOLOv7 was also trained and tested until epoch 100 to find the best mAP value. According to Table III, YOLOv7 obtained the highest mAP value of 0.512 at epoch 100. However, based on the precision, recall, and precision-recall curves in Figure 7, YOLOv7 has a pattern similar to YOLOv6 in the recall curve, with a gradual decrease as confidence increases. The main difference is the recall of YOLOv7 starting at a slightly lower level than YOLOv6, but the decline is more consistent and less steep at higher confidence values. Thus, YOLOv7 is more selective and stable in maintaining object detection at higher confidence thresholds. Furthermore, the precision curve of YOLOv7 shows a similar increase to that of YOLOv6 but with less fluctuation at high confidence levels. Hence, YOLOv7 can predict objects with good precision, although there is still a slight uncertainty in some predictions when the model confidence approaches the highest threshold. Looking at the YOLOv7 precision-recall graph, the value of mAP@0.5 is much lower than YOLOv6 at 0.512. Although YOLOv7 is more selective and has a lower detection error tolerance, it could have missed some critical object detections that the higher confidence threshold might not have identified.

TABLE I. MAP MODEL YOLOV6 AND YOLOV7

No	Number of epochs	mAP of YOLOv6	mAP of YOLOv7
1	10	0.767	0.032
2	20	0.800	0.148
3	30	0.851	0.136
4	40	0.840	0.246
5	50	0.862	0.354
6	60	0.865	0.409
7	70	0.869	0.421
8	80	0.863	0.471
9	90	0.865	0.494
10	100	0.873	0.512

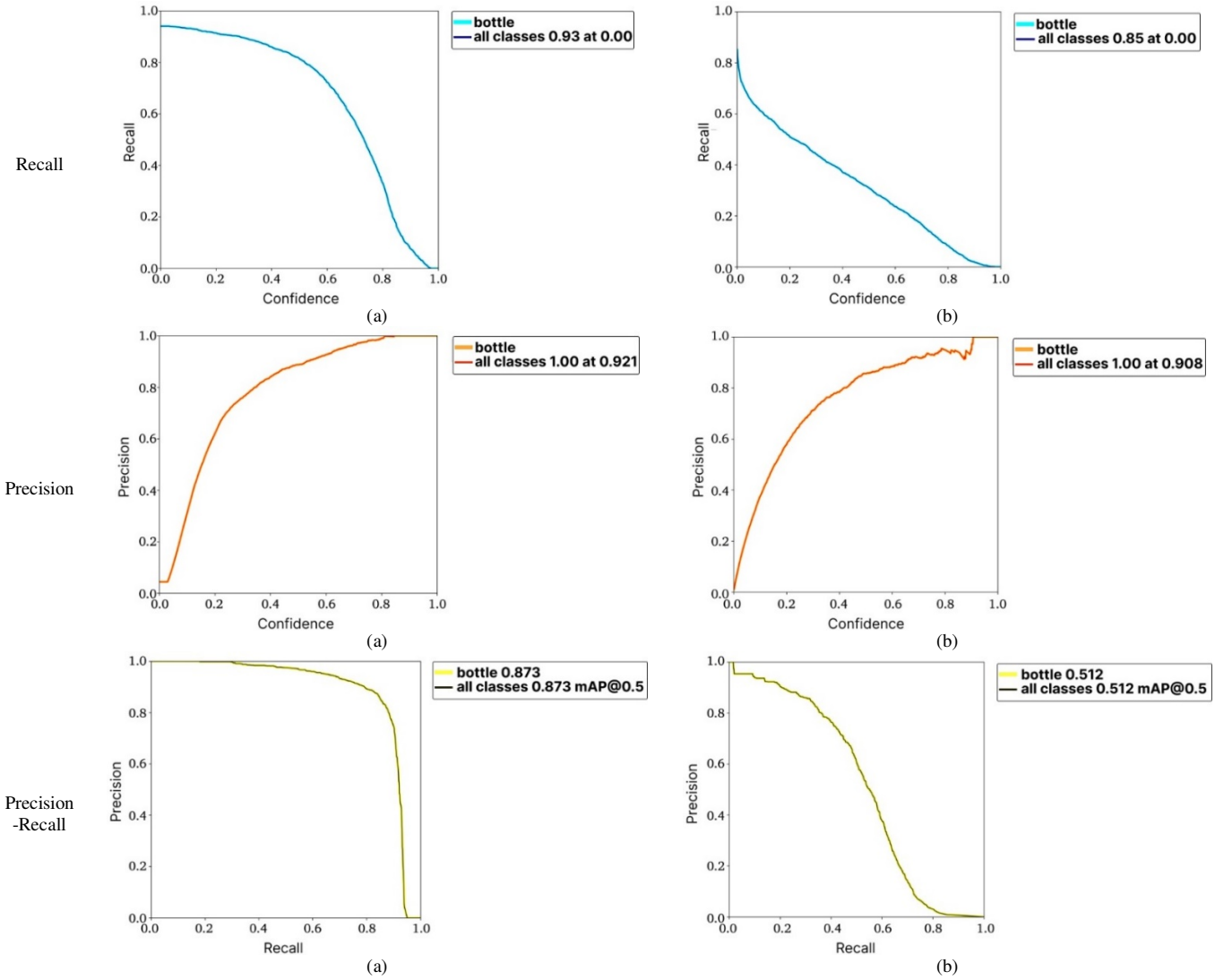

Fig. 7. Curves of recall, precision, and precision-recall on (a) YOLOv6 and (b) YOLOv7.

D. Comparison YOLOv6 and YOLOv7

YOLOv6 and YOLOv7 achieved the best results at the 100th epoch. YOLOv6 (0.873) had higher mAP results than YOLOv7 (0.512). These results are due to YOLOv6's better architecture in feature extraction compared to YOLOv7. In addition, the YOLOv6 neck can combine features better to produce more information for object detection.

TABLE II. MODEL TESTING RESULT

Model	Epoch	mAP	Inference speed (ms)	FPS	Detection success whole object
YOLOv6	100	0.873	4.21	237	Success
YOLOv7	100	0.512	13.7	72	Unsuccessful

According to Table IV, in addition to mAP, YOLOv6 can outperform YOLOv7 in detection speed. YOLOv6 (4.21 ms) can detect objects more effectively in real-time object detection than YOLOv7 (13.7 ms). The difference in inference speed is quite far for YOLOv7. YOLOv6 successfully detects all objects present when inference is made. Compared to the results in [27] with a mAP of 48.1%, YOLOv6 has a much better mAP value.

In the water environment image, there is a significant comparison between the object detection performance of YOLOv6 and YOLOv7 in finding floating bottles floating on water, as shown in Figure 8. YOLOv6 produces more detections with six bounding boxes with varying degrees of confidence. The confidence score generated varies from 0.49 to 0.86, with some detections below the standard threshold of 0.6. In contrast, YOLOv7 detected only four bottle objects with a more consistent confidence score ranging from 0.41 to 0.75. Despite the smaller number of detections, the YOLOv7 model tends to be more selective, with a more stable confidence score distribution closer to a higher threshold. YOLOv7 can reduce the number of irrelevant detections due to water reflection or other visual distractions. In addition, YOLOv7's more stable confidence score values indicate that the model is better at maintaining detection accuracy in more complex environments.

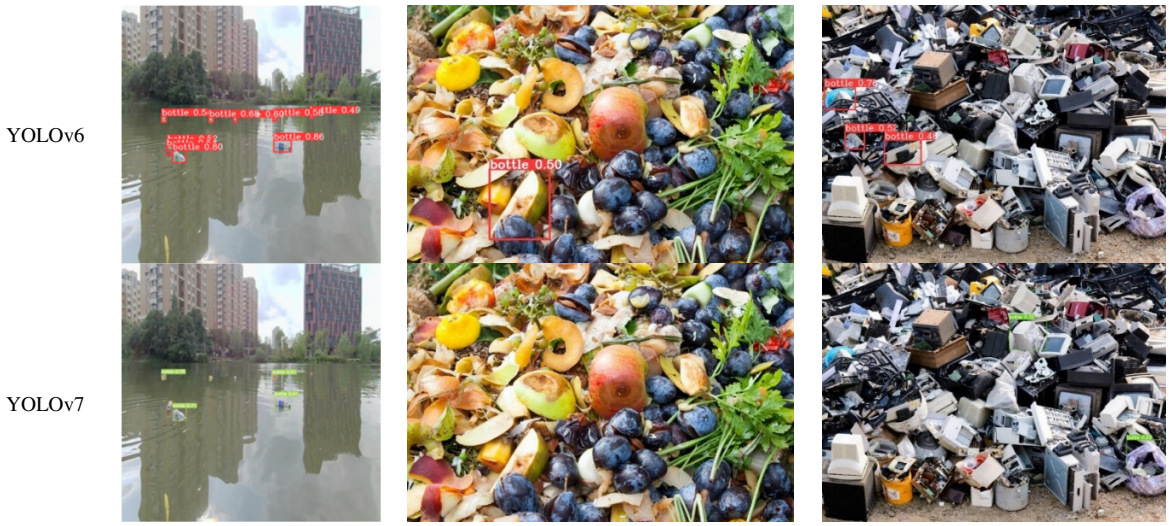

Fig. 8. Comparison of YOLOv6 and YOLOv7 on non-water images: (a) water surface, (b) organic waste, and (c) electronic waste.

IV. DISCUSSION

This study used YOLOv6 and YOLOv7 to detect bottles on the water surface. The objective is to help sort out waste, especially plastic bottles that pollute the water the most. Figure 8 shows a comparison of YOLOv6 and YOLOv7 in detecting bottle objects on the water surface. Overall, YOLOv6 can detect more objects, but many may be irrelevant or have low confidence. YOLOv6 shows that it is more sensitive to noise in the visual environment. Meanwhile, YOLOv7 is more robust, precise, and reliable, especially in terms of selectivity and noise reduction in detecting objects with high confidence. YOLOv7 is more suitable for real-world scenarios with complex environments such as water. However, based on the precision, recall, and precision-recall curves in Figure 7, the performance difference between YOLOv6 and YOLOv7 clarifies the trade-off between detection sensitivity and prediction accuracy. YOLOv6 is better at finding more objects with high recall (0.93), while YOLOv7 shows improvement in reducing false detections with more stable precision (0.908).

In addition, this study attempted to detect bottle objects in organic waste and electronic waste images, testing how well YOLOv6 and YOLOv7 can detect bottle objects on out-of-scope datasets (see Figure 8). In organic waste images, YOLOv6 had a confidence score of 0.50. In contrast, YOLOv7 did not detect any bottle in this image, and the number of detections was zero. This shows that YOLOv6 is not entirely confident that the object is a bottle since the confidence threshold is in the middle (0.5). YOLOv7 explicitly ignores objects visually similar to bottles but not relevant targets. YOLOv6 is more susceptible to visual noise than YOLOv7, which is reflected in confidence scores in the low to medium range.

Regarding electronic waste images, YOLOv6 detected more bounding boxes with varying confidence scores of 0.76, 0.52, and 0.48. In contrast, YOLOv7 detects objects with a more consistent confidence score of 0.32 to 0.29. These values tend to be low because YOLOv7 selectively detects only objects that are more likely to be correct to reduce detection errors. Therefore, comparing these numbers reinforces the conclusion that YOLOv7 is superior in reliability, especially in environments with many visual distraction elements. This statement is consistent with [12], which states that YOLOv7 performs better in object detection with more structured and high-contrast visual patterns such as road damage. However, in [35], YOLOv7 showed higher recall values with precision values, mAP@0.5, and mAP@0.5:0.95 achieving lower values. Hence, the performance of YOLOv7 is highly influenced by the type of object and background complexity. Therefore, YOLOv7 may perform lower in detecting more apparent objects with sharper contours, such as weapons, than in environments with high visual noise and dynamic backgrounds, such as water.

V. CONCLUSION

Deep learning technology development, especially the YOLO model, can help with various problems in object detection, especially waste. Indonesia faces a great challenge in water surface waste. Implementing YOLO for object detection, especially plastic bottle waste, can help the waste sorting process. YOLOv6 and YOLOv7 are reliable methods to detect waste on water surfaces. Based on precision, recall, and precision-recall testing on a water area image dataset, YOLOv6 achieved a mAP of 0.873 with an inference speed of 4.21 ms (237 fps). Meanwhile, the YOLOv7 model achieved a mAP of 0.512 with an inference speed of 13.7 ms (72 fps). Although the YOLOv7 model has a lower mAP than YOLOv6, it does not make detection errors on many plastic bottle waste objects.

More detailed results were also evidenced in the presence of non-watersphere images of organic and electronic waste. YOLOv7 consistently showed improved precision, with fewer but more accurate detections. In contrast, YOLOv6 tends to make more detections but has a lower confidence score, showing weakness in dealing with visual noise. YOLOv7's ability to detect objects more precisely can be a significant reference in real-world applications requiring high accuracy.

However, due to the limitations of the method used and the use of only water surface images in the dataset, future developments are needed to make a real contribution to solving water waste problems, especially in Indonesia.

REFERENCES

[1] U. Diaa, "A Deep Learning Model to Inspect Image Forgery on SURF Keypoints of SLIC Segmented Regions," *Engineering, Technology & Applied Science Research*, vol. 14, no. 1, pp. 12549–12555, Feb. 2024, https://doi.org/10.48084/etasr.6622.

[2] L. Muflikhah, W. F. Mahmudy, and D. Kurnianingtyas, *Machine Learning*. Malang, Indonesia: Universitas Brawijaya Press, 2023.

[3] P. Jiang, D. Ergu, F. Liu, Y. Cai, and B. Ma, "A Review of Yolo Algorithm Developments," *Procedia Computer Science*, vol. 199, pp. 1066–1073, Jan. 2022, https://doi.org/10.1016/j.procs.2022.01.135.

[4] T. Saidani, "Deep Learning Approach: YOLOv5-based Custom Object Detection," *Engineering, Technology & Applied Science Research*, vol. 13, no. 6, pp. 12158–12163, Dec. 2023, https://doi.org/10.48084/etasr.6397.

[5] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO Series in 2021." arXiv, Aug. 06, 2021, https://doi.org/10.48550/arXiv.2107.08430.

[6] S. Xu *et al.*, "PP-YOLOE: An evolved version of YOLO." arXiv, Dec. 12, 2022, https://doi.org/10.48550/arXiv.2203.16250.

[7] C. Li *et al.*, "YOLOv6 v3.0: A Full-Scale Reloading." arXiv, Jan. 13, 2023, https://doi.org/10.48550/arXiv.2301.05586.

[8] A. Alayed, R. Alidrisi, E. Feras, S. Aboukozzana, and A. Alomayri, "Real-Time Inspection of Fire Safety Equipment using Computer Vision and Deep Learning," *Engineering, Technology & Applied Science Research*, vol. 14, no. 2, pp. 13290–13298, Apr. 2024, https://doi.org/10.48084/etasr.6753.

[9] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada, Jun. 2023, pp. 7464–7475, https://doi.org/10.1109/CVPR52729.2023.00721.

[10] E. Iren, "Comparison of YOLOv5 and YOLOv6 Models for Plant Leaf Disease Detection," *Engineering, Technology & Applied Science Research*, vol. 14, no. 2, pp. 13714–13719, Apr. 2024, https://doi.org/10.48084/etasr.7033.

[11] F. Hermens, "Automatic object detection for behavioural research using YOLOv8," *Behavior Research Methods*, vol. 56, no. 7, pp. 7307–7330, Oct. 2024, https://doi.org/10.3758/s13428-024-02420-5.

[12] N. I. M. Yusof, A. Sophian, H. F. M. Zaki, A. A. Bawono, A. H. Embong, and A. Ashraf, "Assessing the performance of YOLOv5, YOLOv6, and YOLOv7 in road defect detection and classification: a comparative study," *Bulletin of Electrical Engineering and Informatics*, vol. 13, no. 1, pp. 350–360, Feb. 2024, https://doi.org/10.11591/eei.v13i1.6317.

[13] Z. Zhang, X. Lu, G. Cao, Y. Yang, L. Jiao, and F. Liu, "ViT-YOLO:Transformer-Based YOLO for Object Detection," in *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, Montreal, BC, Canada, Oct. 2021, pp. 2799–2808, https://doi.org/10.1109/ICCVW54120.2021.00314.

[14] G. Widjaja and S. L. Gunawan, "Dampak Sampah Limbah Rumah Tangga Terhadap Kesehatan Lingkungan," *Zahra: Journal of Health and Medical Research*, vol. 2, no. 4, pp. 266–275, Nov. 2022.

[15] P. di L. K. L. Hidup, "Kementerian Lingkungan Hidup dan Kehutanan," *Jakarta: Kementerian Lingkungan Hidup dan Kehutanan*, 2021.

[16] N. P. Purba *et al.*, "Marine debris in Indonesia: A review of research and status," *Marine Pollution Bulletin*, vol. 146, pp. 134–144, Sep. 2019, https://doi.org/10.1016/j.marpolbul.2019.05.057.

[17] P. Agamuthu, S. Mehran, A. Norkhairah, and A. Norkhairiyah, "Marine debris: A review of impacts and global initiatives," *Waste Management & Research*, vol. 37, no. 10, pp. 987–1002, Oct. 2019, https://doi.org/10.1177/0734242X19845041.

[18] "Kementerian Kelautan dan Perikanan," 2024. [Online]. Available: https://kkp.go.id/download-pdf-akuntabilitas-kinerja/akuntabilitas-kinerja-pelaporan-kinerja-2024-pusdik-lkj-triwulan-i.pdf.

[19] M. Tian, X. Li, S. Kong, L. Wu, and J. Yu, "A modified YOLOv4 detection method for a vision-based underwater garbage cleaning robot," *Frontiers of Information Technology & Electronic Engineering*, vol. 23, no. 8, pp. 1217–1228, Aug. 2022, https://doi.org/10.1631/FITEE.2100473.

[20] X. Li, M. Tian, S. Kong, L. Wu, and J. Yu, "A modified YOLOv3 detection method for vision-based water surface garbage capture robot," *International Journal of Advanced Robotic Systems*, vol. 17, no. 3, May 2020, Art. no. 1729881420932715, https://doi.org/10.1177/1729881420932715.

[21] S. Majchrowska *et al.*, "Deep learning-based waste detection in natural and urban environments," *Waste Management*, vol. 138, pp. 274–284, Feb. 2022, https://doi.org/10.1016/j.wasman.2021.12.001.

[22] M. Malik *et al.*, "Waste Classification for Sustainable Development Using Image Recognition with Deep Learning Neural Network Models," *Sustainability*, vol. 14, no. 12, Jan. 2022, Art. no. 7222, https://doi.org/10.3390/su14127222.

[23] Y. Cheng *et al.*, "FloW: A Dataset and Benchmark for Floating Waste Detection in Inland Waters," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 10933–10942, https://doi.org/10.1109/ICCV48922.2021.01077.

[24] "Flow-Img," *OrcaUBoat*. https://orca-tech.cn/en/datasets/FloW/FloW-Img.

[25] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement." arXiv, Apr. 08, 2018, https://doi.org/10.48550/arXiv.1804.02767.

[26] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 779–788, https://doi.org/10.1109/CVPR.2016.91.

[27] M. Tharani, A. W. Amin, M. Maaz, and M. Taj, "Attention Neural Network for Trash Detection on Water Channels." arXiv, Jul. 09, 2020, https://doi.org/10.48550/arXiv.2007.04639.

[28] T. Diwan, G. Anirudh, and J. V. Tembhurne, "Object detection using YOLO: challenges, architectural successors, datasets and applications," *Multimedia Tools and Applications*, vol. 82, no. 6, pp. 9243–9275, Mar. 2023, https://doi.org/10.1007/s11042-022-13644-y.

[29] C. Y. Wang and H. Y. M. Liao, "YOLOv1 to YOLOv10: The fastest and most accurate real-time object detection systems." arXiv, Aug. 18, 2024, https://doi.org/10.48550/arXiv.2408.09332.

[30] M. M. Moussa, R. Shoitan, Y. I. Cho, and M. S. Abdallah, "Visual-Based Children and Pet Rescue from Suffocation and Incidence of Hyperthermia Death in Enclosed Vehicles," *Sensors*, vol. 23, no. 16, Jan. 2023, Art. no. 7025, https://doi.org/10.3390/s23167025.

[31] L. Zhao and M. Zhu, "MS-YOLOv7:YOLOv7 Based on Multi-Scale for Object Detection on UAV Aerial Photography," *Drones*, vol. 7, no. 3, Mar. 2023, Art. no. 188, https://doi.org/10.3390/drones7030188.

[32] Z. Chen, C. Liu, V. F. Filaretov, and D. A. Yukhimets, "Multi-Scale Ship Detection Algorithm Based on YOLOv7 for Complex Scene SAR Images," *Remote Sensing*, vol. 15, no. 8, Jan. 2023, Art. no. 2071, https://doi.org/10.3390/rs15082071.

[33] E. Sediyono, Suhartono, and C. Nivak, "Measuring the Performance of Ontological Based Information Retrieval from a Social Media," in *2014 European Modelling Symposium*, Pisa, Italy, Oct. 2014, pp. 354–359, https://doi.org/10.1109/EMS.2014.15.

[34] J. Revaud, J. Almazan, R. Rezende, and C. D. Souza, "Learning With Average Precision: Training Image Retrieval With a Listwise Loss," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South), Oct. 2019, pp. 5106–5115, https://doi.org/10.1109/ICCV.2019.00521.

[35] O. E. Olorunshola, M. E. Irhebhude, and A. E. Evwiekpaefe, "A Comparative Study of YOLOv5 and YOLOv7 Object Detection Algorithms," *Journal of Computing and Social Informatics*, vol. 2, no. 1, pp. 1–12, Feb. 2023, https://doi.org/10.33736/jcsi.5070.2023.