**Underwater Trash Detection**

By

**Rohan Alroy.B (20BAI1245)**

**Shivaprakash S J (20BAI1205)**

**Raagul Bharatwaj K( 20BAI1003)**

***A project report submitted to***

**Dr. Maheswari S**

**Associate Professor**, **School of Computer Science and Engineering**

***in partial fulfilment of the requirements for the course of***

**CSE3043 - Video Analytics**

in

**B. TECH., COMPUTER SCIENCE AND ENGINEERING**



**Vellore Institute of Technology, Chennai**

**Vandalur – Kelambakkam Road**

**Chennai – 600 127**

**APRIL 2023**

**Chennai**

**Abstract**

Underwater trash, including plastic waste and discarded fishing gear, is a growing problem with significant ecological and economic impacts. The accumulation of underwater trash can harm marine ecosystems by entangling marine animals, spreading pollutants, and altering habitats. Additionally, underwater trash can also pose navigation hazards for ships and impact the livelihoods of local communities who rely on fishing and tourism. We present an examination of the utility of YOLOv5, a cutting-edge object detection algorithm, for underwater object detection. The underwater environment presents a set of distinctive difficulties for computer vision algorithms, including low illumination, turbid water, and restricted visibility. To address these difficulties, we will modify YOLOv5 to include tailored modifications to accommodate the distinctive underwater imaging conditions. An improvement in accuracy and efficiency provided by YOLOv5 has the potential to reduce the duration and resources required for underwater surveys and evaluations, thus reducing the ecological impact of these activities and reducing disturbance to marine ecosystems.   
**Keywords:** YOLO, YOLOv5, Object detection, real time detection, marine conservation

**1. Introduction:**

The build-up of rubbish in the world's oceans is a serious environmental problem. This waste may have detrimental long-term effects on the ecology and marine life. As a result, finding and removing undersea waste is a critical duty that calls for creative solutions.

Deep learning models have recently demonstrated considerable promise for resolving a range of computer vision issues, including object detection. In this study, we suggest detecting and underwater trash using deep learning algorithms. In order to support cleanup efforts and keep track of the health of ocean environments, this project aims to create an accurate and effective system that can identify garbage in underwater videos. Data gathering, data pre-processing, model creation and training, evaluation and optimization, and deployment are some of the steps that the project will have. Obtaining a dataset of underwater movies including various kinds of trash is part of the data gathering process. Cleaning and preparing the data for model training will be part of the data pre-processing stage.

We will train our deep learning models using YOLO and transfer learning during the model design and training stages. The models will be trained using the data gathered to find and classify any underwater rubbish. We will assess the models' performance on a test dataset and refine the models to increase their accuracy during the assessment and optimization stage. The deployment phase will culminate with the integration of the optimized models into a system suitable for usage in practical applications.

In summary, this research has the potential to have a big influence on ocean conservation. We want to create a useful tool that can aid in clean-up operations and contribute to the preservation of the oceans by using deep learning techniques to underwater trash detection. The study will also improve the field of deep learning research by shedding light on how deep learning may be applied to solve problems in the real world.

This project aims to develop a computer vision mode, based of YOLO algorithm, that can accurately identify Waste in underwater video feeds. It should also be able to identify biolife and the parts of the robot it is attached, to ensure that, it only focuses on waste materials.

The motivation behind this project is as follows. Traditional approaches to object detection include using models like RCNN and faster RCNN. These approaches provide good results but are not able to do real-time detection. This is highly disadvantageous when trying to create an application that can be deployed on a drone or other autonomous devices. These devices need real-time detection to make decisions. YOLOv5 is a cutting-edge algorithm that is an improvement over the original You Only Look Once algorithm. The model can be divided into three components: the backbone, neck, and head. The backbone extracts features from the input frame. The neck creates feature pyramids; this is essential for scaling objects. Finally, the head predicts bounding boxes along with the class the object belongs to. YOLOv5 is faster and produces better results than the original YOLO algorithm. The ability of YOLOv5 to do real-time detection makes it very useful for underwater trash detection. A well-trained YOLOv5 model can be deployed on an underwater drone and be used to actively monitor trash present in the water. Real-time detection also helps enable potential applications that are able to interact with and collect the waste present in the water

**2. Literature Survey:**

In the paper, Hong, Jungseok, Michael Fulton, and Junaed Sattar. "Trashcan: A semantically-segmented dataset towards visual detection of marine debris." , they propose the usage of Mask R-CNN and Faster R\_CNN with a ResNetXt-101-FPN backbone. Both these models were pre-trained on the coco dataset, on15k and 10k iterations respectively. They have followed two methods, one is training the models with an instance segmented version and one with a object-detection version. They see that the Faster R-CNN with the instance segmented version works best as per the results from their findings. As a future work they have decided to work on getting a larger dataset. Or a more complex model.

Papageorgiou and Poggio propose a Support Vector Machine that uses the Haar wavelet transform to detect objects. They propose that an image can be represented as an overcomplete dictionary of Haar wavelet features. Wavelets are used because they are capable of capturing features about the shape and interior structure of the object. This representation is also quite robust to noise and provides a rich description of the image. A sliding window-based approach is used to create representations of parts of the image to identify objects of interest. The proposed method performed well for the detection of faces and vehicles.

Hu et al. propose the use of relational networks for object detection. Many of the current deep learning methods for recognizing objects do not make use of relations between objects for the detection of objects. The approach takes inspiration from attention modules in NLP and proposes a similar architecture for computer vision applications. A set of objects is processed together, and the relation between each object with respect to the whole object set is calculated. In this proposed methodology, instance recognition and duplicate removal are done by using the relational model.

Szegedy et al. propose the use of Deep Neural Networks for object detection and making object bounding boxes. The proposed methodology is to be used to classify and localize objects. A DNN-based regression is used to produce the bounding box. The approach makes use of multiple masks to localize objects present in the image. The DNN tries to estimate the upper-left corner and lower-right corner of the bounding box that localizes the image. The DNN used for the classification of an object can be done through the use of pretrained model weights, but for the localization model, full fine tuning of the DNN network is required.

Redmon et al. propose an innovative approach known as You Only Look Once (YOLO). In this approach, a single CNN predicts the bounding box and classifies the object simultaneously. An image is divided into a SxS grid where each grid predicts B bounding boxes and outputs the confidence that a box has an object in it. Each grid cell also predicts the class probabilities, which show the probability of the grid containing an object of a particular class. YOLO has performed well in the task of object detection and is also able to perform real-time detection with speeds of 45 fps.

Yu et al. propose the use of the Intersection over Union loss function for the prediction of bounding boxes. This approach aims to improve the localization capability of classifiers. The IoU loss function aims to maximize the overlap of the predicted bounding box with the ground truth and also regress the bounding box variables together instead of individually regressing them. This loss function is an alternative to the L2 loss function. To test out the performance of the proposed function, a neural network based on the VGG 16 architecture was used. On evaluating the performance, it has been observed that IoU is more efficient and also better at detecting objects.

Kong et al. propose an alternative to region proposals-based models called HyperNet. HyperNet combines feature maps generated by the different layers of the CNN into one feature called the Hyper Feature. This feature is used to make 100 proposals for a given image. The combining of feature maps is done through different sampling strategies, which include pooling, deconvolutional operations, and normalization. The HyperFeature is then used in the region proposal generation. Due to the fact that it produces only 100 proposed regions, the efficiency of HyperNet is very high. It is able to produce a speed of 5 fps, which shows that the model can be used for real-time detection.

Zoph et al. discuss data augmentation strategies that can be used for object detection problems. Image transformation strategies need to take into consideration bounding boxes when being applied, for example when applying transformations like shearing, the bounding box must also be changed. Transformations can also be done within the bounding box itself instead of the entire image. Image transformations include rotation, changing brightness, shearing, and rotating only within a bounding box. A policy of selecting K transformations from N transformations is proposed when augmenting an image. Finding a good set of transformations can be considered an optimization problem that can be solved for a given dataset.

The research paper "Exploring plain vision transformer backbones for object detection" by Li, Yanghao, et al. presents an exploration of using plain vision transformer backbones for object detection. The authors focus on using transformer backbones, which have become increasingly popular in computer vision tasks such as image classification, for object detection. The paper presents a comprehensive analysis of the potential benefits of using transformer backbones for object detection, including improved feature representation, enhanced context modeling, and greater computational efficiency. The authors conducted a series of experiments to evaluate the performance of plain vision transformer backbones on several popular object detection benchmarks. They compare the results to state-of-the-art object detection algorithms and provide a detailed analysis of the strengths and weaknesses of using transformer backbones for object detection. The results of the experiments indicate that plain vision transformer backbones have the potential to achieve comparable or even superior performance compared to traditional convolution neural network (CNN) backbones. The authors also discuss the limitations of using transformer backbones for object detection and identify areas for future research. Overall, this research provides valuable insights into the application of transformer backbones for object detection and highlights the potential benefits of using this approach. The authors’ contribution is a contribution to the growing body of literature on the use of transformers for computer vision tasks and will likely spur further research in this area.

The paper "Video Swin Transformer" by Liu, Ze, et al. presents a new approach to video object detection using the Swin Transformer architecture. The authors propose using a Swin Transformer, a variant of the transformer architecture that has been successful in image classification tasks, for video object detection. The authors argue that the Swin Transformer architecture can effectively capture both spatial and temporal information in video frames, which is critical for accurate video object detection. The authors conduct a series of experiments to evaluate the performance of the Video Swin Transformer on several popular video object detection benchmarks. They compare the results to state-of-the-art video object detection algorithms and provide a detailed analysis of the strengths and weaknesses of the Video Swin Transformer. The results of the experiments indicate that the Video Swin Transformer can achieve superior performance compared to traditional video object detection algorithms that use convolutional neural networks (CNNs). The authors also discuss the limitations of the Video Swin Transformer and identify areas for future research. Overall, this research provides valuable insights into the application of the Swin Transformer architecture for video object detection. The authors’ contribution is a significant step forward in the field of video object detection and highlights the potential of the Swin Transformer for video-related computer vision tasks. The findings of this study will likely inspire further research in this area and help to advance the state of the art in video object detection.

**3. Proposed Methodology**

**3.1. Introduction**

Object detection in the underwater environment is a complex problem due to several factors such as lighting conditions, water turbidity, and object occlusion. In recent years, deep learning-based approaches such as YOLOv5 have shown great potential in object recognition tasks. The aim of this proposed methodology is to develop a YOLOv5 model to detect and classify objects in an underwater film of plastic/garbage, bio-animals or ROV objects. The proposed method consists of the following steps:

**3.2. Data preparation**

A diverse dataset of underwater footage is collected from various sources, including ROVs and underwater cameras. Domain experts manually annotate the dataset to label objects as plastic/debris, biological life or ROV. The annotated dataset is divided into a training, validation and testing set. Data augmentation techniques such as random truncation, inversion, and rotation are applied to the dataset to improve model reliability.

**3.3. Model architecture**

The YOLOv5 architecture was chosen for this study due to its high accuracy and efficiency in object recognition tasks. The architectures of body, neck and head of the model are chosen based on the characteristics of the data set. Hyperparameter tuning is performed using a grid search method to find the best configuration for the model. Hyperparameters include learning rate, weight reduction, set size and number of periods.

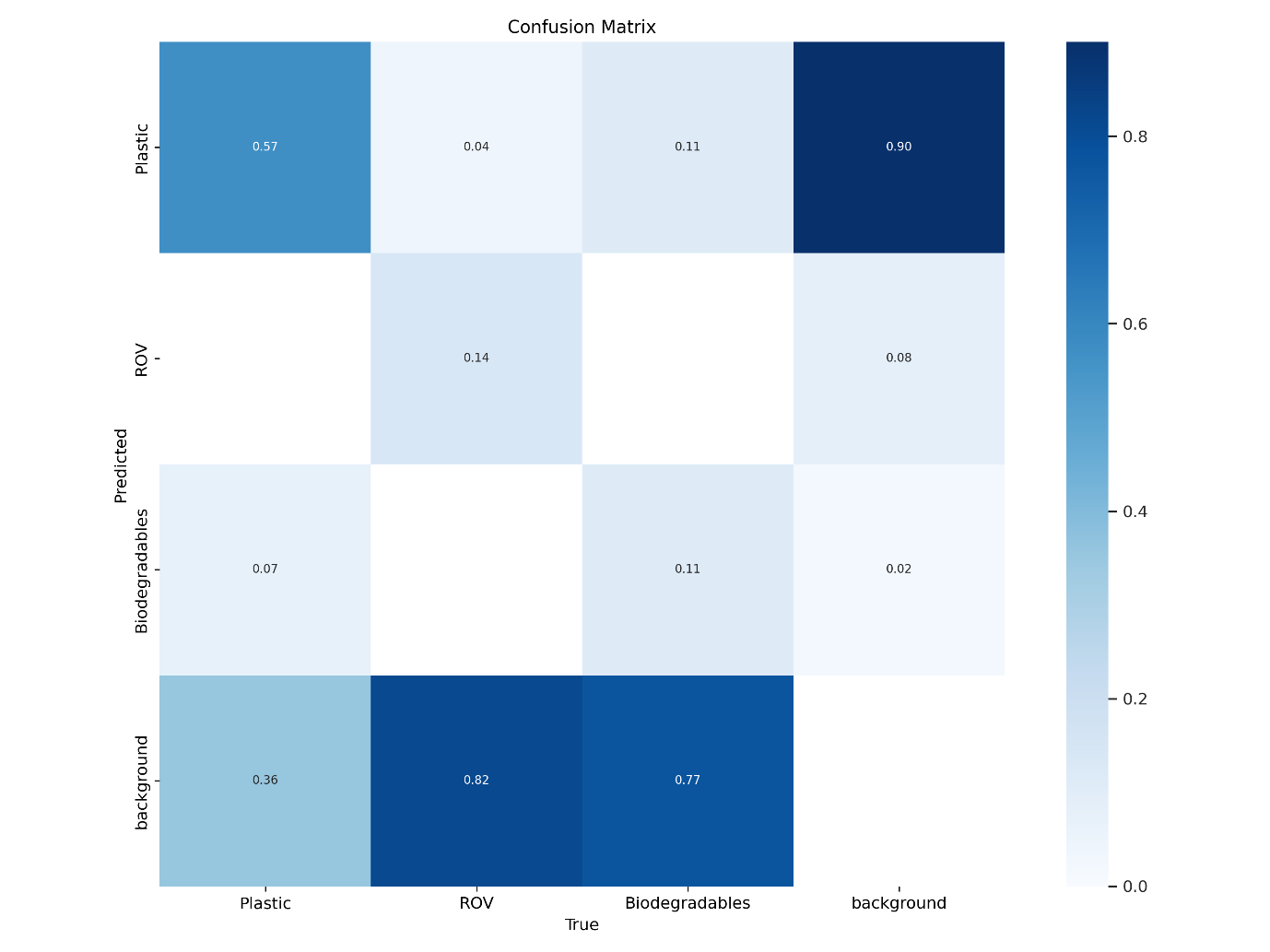
**3.4. training**

The YOLOv5 model is trained on an annotated dataset using deep learning. The Adam optimization tool is used. A multi-step training strategy is used to improve the performance of the model. The loss functions used in the exercise include a combination of cross entropy and focus loss.

**3.5 Evaluation**

The proposed method is evaluated with a test set of material. Model performance is measured by precision, accuracy, recall, F1 score and mAP.

**4. Results and discussion:**



Chart

Description automatically generated

Graphical user interface, chart

Description automatically generated

Graphical user interface, chart, line chart

Description automatically generated

A screenshot of a computer

Description automatically generated with low confidence

**5. Conclusion:**

In this project, we aimed to develop a computer vision model that can accurately identify different objects and organisms in underwater footage. Specifically, our focus was on detecting garbage, biolife, and robot arms. Our model leveraged state-of-the-art deep learning techniques, such as yolov5 which is a transfer learning model, to achieve high accuracy in object detection.

We trained and evaluated our model on a diverse set of underwater footage, captured from various sources and locations. Our model was able to successfully identify garbage, biolife, and robot arms in underwater environments with a decent degree of accuracy. Additionally, we conducted several experiments to fine-tune our model's parameters, and achieved even better results by optimizing the learning rate and regularization.

The applications of this project are numerous. The identification of garbage in underwater footage can help support efforts to remove waste and pollution from our oceans and protect marine life. The identification of biolife can help monitor the health of marine ecosystems and understand how different species interact with each other. The identification of robot arms can aid in the inspection and maintenance of underwater infrastructure, such as pipelines and cables.

While our model achieved high accuracy, there are still limitations that should be considered. For instance, our model may not be able to detect all types of garbage or biolife, particularly if they are rare or not well-represented in the training data. Additionally, our model may struggle in low-light or murky underwater environments where visibility is poor.

In conclusion, our computer vision model for identifying garbage, biolife, and robot arms in underwater footage has the potential to be a valuable tool for marine conservation, scientific research, and industrial applications. We hope that this project inspires further research in this area and helps contribute to a better understanding of our planet's underwater environments.

**6. References:**

Papageorgiou, C. and Poggio, T., 2000. A trainable system for object detection. International journal of computer vision, 38, pp.15-33.

Hu, H., Gu, J., Zhang, Z., Dai, J. and Wei, Y., 2018. Relation networks for object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3588-3597).

Szegedy, C., Toshev, A. and Erhan, D., 2013. Deep neural networks for object detection. Advances in neural information processing systems, 26.

Redmon, J., Divvala, S., Girshick, R. and Farhadi, A., 2016. You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779-788).

Yu, J., Jiang, Y., Wang, Z., Cao, Z. and Huang, T., 2016, October. Unitbox: An advanced object detection network. In Proceedings of the 24th ACM international conference on Multimedia (pp. 516-520).

Kong, T., Yao, A., Chen, Y. and Sun, F., 2016. Hypernet: Towards accurate region proposal generation and joint object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 845-853).

Zoph, B., Cubuk, E.D., Ghiasi, G., Lin, T.Y., Shlens, J. and Le, Q.V., 2020. Learning data augmentation strategies for object detection. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16 (pp. 566-583). Springer International Publishing.

Li, Yanghao, et al. "Exploring plain vision transformer backbones for object detection." Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX. Cham: Springer Nature Switzerland, 2022.

Liu, Ze, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. "Video swin transformer." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 3202-3211. 2022.