Evaluating multiple YOLO deep learning models for detecting fish

1st Faerie Mattins
School of Computer Science Engineering
Vellore Institute of Technology
Chennai, India
fr679611@dal.ca

2nd Chris Whidden
Faculty of Computer Science
Dalhousie University
Halifax, Canada
cwhidden@dal.ca

***Abstract*—Ocean covers about three-fourths of the surface of the world, providing habitat for 34,000 different species of fish. In order to ensure that all species live in a secure and healthy habitat, it is crucial to keep track of all them. Fish have been detected manually in the past using a variety of techniques that required a lot of human labour and had an adverse effect on the fish. Hence, the deep learning models have been employed in order to safely detect fishes using object detection methods. In this research, the popularly used object detection model, You Only Look Once (YOLO), is used to detect fish. Three versions of YOLO, namely YOLOv4, YOLOv5 and YOLOR are used in this research to detect fish and determine which model works best on videos of fish with 5 species.**

***Index Terms*—YOLOv4, YOLOv5, YOLOR, object detection**

I. Introduction

Fish are a vital part of marine ecosystems that keep the food chain in balance. However, due to inadequate living conditions—caused by climate change, pollution, food demand, etc.—their numbers have declined. Therefore, it is crucial to keep track of fish and their species to prevent extinction and to investigate strategies to boost their population. Various techniques have been used in the past to detect fish. One way is to use implanted tags and receivers. To research at-risk species that cannot be damaged or tagged, or to track vast numbers of fish, a different strategy is required.

Therefore, a safer way is needed to monitor fish without endangering them. For this purpose, deep learning models are used to detect and analyze fish using the data collected from sensors. Kandimalla et al. [1] has used YOLOv3 and Mask -RCNN model to detect fish for eight species of fish using a public high resolution DIDSON imaging sonar dataset. They acquired a good mAP of 0.73 and 0.62 with an IOU threshold of 0.4 for YOLOv3 and Mask-RCNN respectively. To assess the consequences of seismic oil and gas exploration on commercial fishing, Morris et al. [2] created 100 inexpensive underwater camera systems and gathered 3,500+ hours of video. They created a numerical database for statistical analysis using R-CNN to automatically recognise and count Atlantic cod. Though 94% of the labelled animals could be found with success, their technique can only distinguish between a few distinct species. Wageeh et al. [3] combined an enhancement algorithm based on retinex (MSR) with YOLO object detection algorithm. They used one species of fish in their training dataset consisting of 2000 images. Their technique has shown to be more effective in finding fish in murky water. They suggested clustering algorithms to be used for unlabeled dataset which can be implemented in real-time. The previous work shows that, it is important to develop an object detection model which provides a higher mAP and is trained upon a big dataset with multiple numbers of species.

In this work, 3 versions of the YOLO model are used, YOLOv4, YOLOv5 and YOLOR, to detect fish. The YOLO model is pre-trained using the COCO dataset [4], which doesn't have the object fish. Hence, a dataset developed for fish detection, FishCLEF-2015, is used and custom weights are generated. All the models are trained using fish data and their results are evaluated.

II. Methodology

A. Dataset

In this work, the dataset from FishCLEF-2015 [5] is used. This dataset is unbalanced in the number of instances of fish species. Hence, from the total of 15 fish species in the dataset, only the top 5 fish species with the highest number of fishes in the dataset were considered in this research for simplicity. This is given in Table I. The training data contained 20 videos and annotated XML files of fish. Individual frames were extracted from the videos and saved as images. For each image, a text file was created which contained the annotation of each image. An empty text file was created for images which didn't have any annotations. Fig.1 shows an example of an annotated image of a fish. Three folders were created, namely, train, valid and test. Each of these folders had two sub-folders named images and labels. The labels folder consist of the annotated text files. The models were trained on this refined data. During analysis, another class called None was taken to account for images where the model does not detect any fish. This class was included for simplicity and to find the true positives in a simpler manner. It is essential for the model to detect only fishes, so its equally important for the model to differentiate when the frame is empty or if it has fish. Hence, the None class plays a vital role in this research.

TABLE I
DATASET DESCRIPTION

Class ID	Species	Training	Testing
0	Empty frames	2698	698
1	Chaetodon Lununatus	999	1178
2	Dascyllus Aruanus	894	987
3	Dascyllus Reticulatus	2678	3031
4	Pempheris Vanicolensis	999	376
5	Plectrogly-Phidodon Dickii	737	700

Fig. 1. Annotated image of a fish.

B. YOLOv4

The YOLOv4 model for real-time object identification was developed by Alexey Bochkovskiy et al [6]. YOLOv4 incorporates the latest bag of freebies (BoF) and a number of bag of specials (BoS) to considerably increase the detector's accuracy and object detection accuracy at the cost of interference and price of training. YOLOv4 is based on the Darknet trained on the COCO dataset.

C. YOLOv5

YOLOv5, by Jocher G et al. [7], uses PyTorch rather than the original Darknet. The two main advancements are bounding box anchors and mosaic data augmentation. Mosaic data augmentation combines four images into one in specific ratios. This enables the model to recognise objects at a much smaller scale.

D. YOLOR

You Only Learn One Representation (YOLOR), by Chien-Yao Wang et al [8], differs from other YOLO versions due to its architecture, author, and model framework. YOLOR fuses explicit and implicit knowledge to carry out tasks utilising a single image representation. It obtains implicit and explicit information from the shallow and deep layers, respectively. The model joins the two representations to create a single representation that can then be applied to a variety of applications.

E. Configuration

The YOLO models need to be configured before training. Max batch is the maximum batch size for a iteration, batch size is the number of images in a batch, subdivisions shows the division of a batch, and steps denote the iterations where the learning rate is updated. The learning rate and scales are 0.00261 and (.1,.1) for each of the three models, respectively. Here, all the models have a confidence score of 25%. This can be viewed in Table II.

TABLE II
CONFIGURATION OF YOLO MODELS

Model	batch size	subdivisions	max_batches	steps
YOLOv4	64	24	2000	1600, 1800
YOLOv5	64	8	500500	400000,450000
YOLOR	64	8	500500	400000,450000

F. Evaluation metrics

The Intersection over union (IoU) is a parameter used in object detection systems, that calculates the difference between ground truth and predicted bounding boxes. The model predicts and removes bounding boxes for objects based on their threshold value. IoU ranges from 0 to 1, which indicates no and complete overlap, respectively. The IoU taken in this research is 0.5. The calculation of IoU is given in Eq. (1).

$$\text{IOU} = \text{Area of Overlap}/\text{Area of Union} \tag{1}$$

In this research, mean Average Precision (mAP) is taken as the evaluation metrics. mAP is derived from two components: precision and recall. Precision evaluates the accuracy of the predictions, whereas recall assesses how well the positives are identified. The average precision (AP) is defined as the area under the precision-recall curve. AP is always between 0 and 1 with 1 indicating perfect precision and recall. The formula for AP is given in Eq. (2).

$$\text{AP} = \int_0^1 p(r)dr \tag{2}$$

Where p(r) indicates the value of precision measured at recall r. AP denotes a detection accuracy category, while AP is the average detection accuracy of numerous categories. mAP is represented as given in Eq. (3).

$$\text{mAP} = 1/N \sum_{i=1}^{N} AP_i \tag{3}$$

III. RESULTS

The three YOLO models were trained on the fish dataset consisting of 5 species. Table III shows the mAP of each model upon training. It can be seen that YOLOR gives the highest mAP of 89.5, following which YOLOv5 gave a comparatively good result of 83.9 mAP. Nonetheless, YOLOv4 produced the least result of 28.34 mAP.

Despite the fact that YOLOv4 has the least mAP, it is the model which is able to detect the most number of fishes. In the confusion matrix shown in Fig.2, 0 to 5 represents the class ID as mentioned in Table I. The confusion matrix indicates that YOLOv4, YOLOv5, and YOLOR each exhibit 5808 true positives, 848 true positives, and 799 true positives, respectively. This behavior of the YOLOv4 model can be

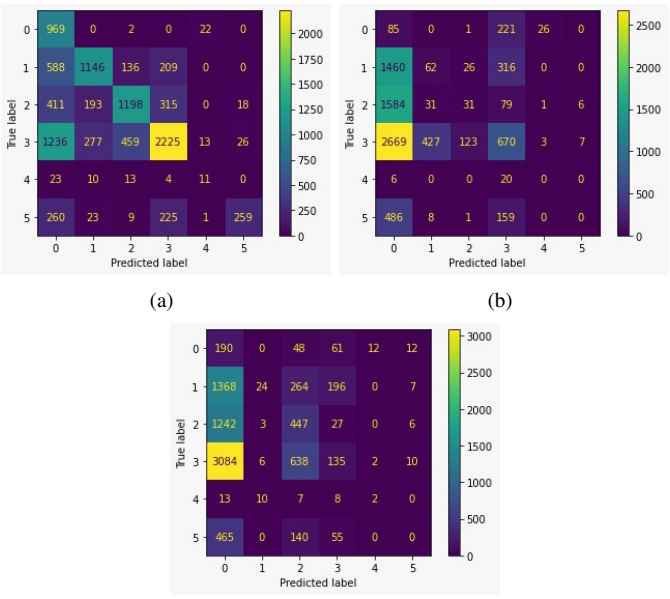

Fig. 2. Confusion Matrix of (a) YOLOv4 (b) YOLOv5 (c) YOLOR

attributed to its high sensitivity due to the low confidence score. For example, the ground truth of an image contains two fish of the same species. The YOLOv4 model performs the correct prediction for the two fish but it also predicts two other fish of the same species to be present in the frame. Even if the prediction is right for the given frame, the model fails in locating the fish, thus, resulting in the high classification metrics but low mAP. When the same image is given to the YOLOv5 and YOLOR model, they locate the correct number of fish but fail to determine the species of the fish. This can be seen in the poor classification metrics but relatively high mAP. Another reason for the poor species identification is due to the imbalance in the training and testing data sets. It can be noticed that the model is trained with more number of empty frames and hence overfits, but the testing data for the empty frames is relatively low. This could attribute to the reason why models like YOLOv5 and YOLOR labeled fish as empty frames in many scenarios as seen in the confusion matrix.

TABLE III
MAP OF YOLO MODELS

Model	mAP at IoU=0.5
YOLOv4	28.34
YOLOv5	83.9
YOLOR	89.5

IV. Conclusion

Three YOLO models—YOLOv4, YOLOv5, and YOLOR—were employed in this study to identify various fish species from the FishCLEF-2015 dataset. Five different fish species were used to train the model, which had an IoU of 0.5 and a confidence score of 0.25. As can be seen, YOLOR produced the highest results with a mAP of 89.5, while YOLOv4 produced the lowest results with a mAP of 28.34. Even with YOLOv4's low mAP, this model was able to properly identify more fish. Due of its high sensitivity and low confidence score, the YOLOv4 model exhibits this behaviour. Additionally, because of the low confidence level, all of the models had greater false positive rates. An informative comparison has been produced for all the three models for which further fine-tuning is required.

V. Future Work

The next step is to raise the confidence level to up to 0.50 and assess the outcomes. Increasing the amount of training data may improve per-species performance and allow training with more species. Unlabelled data datasets can be used with clustering methods. Also, multiple augmentation techniques can be performed. Upon training using more epochs, there is a possibility that the model might give better results. Validation tests can be implemented to show the connection between the knowledge gained from the training set and the capacity to recognise the test set at regular intervals. This might provide some information about how different algorithms manage imbalanced data sets. More evaluation metrics like Mean Average Recall can be implemented. The dataset could be re-balanced for better analysis. Moreover, the model can be trained from scratch.

VI. Acknowledgement

I would like to thank the Mitacs Globalink Research Internship program for facilitating me with this opportunity to come to Dalhousie University as a visiting researcher and present this work. I would also like to thank Dalhousie University for giving me this opportunity to present my work.

References

[1] Kandimalla, V., Richard, M., Smith, F., Quirion, J., Torgo, L., & Whidden, C. (2022). Automated detection, classification and counting of fish in fish passages with deep learning. Frontiers in Marine Science, 2049.

[2] Morris, C., Barnes, J., Schornagel, D., Whidden, C., & Lamontagne, P. (2021). Machine Learning Analysis of Underwater Video: Measuring Effects of Seismic Surveying on Groundfish Resources off the Coast of Newfoundland, Canada. Journal of Ocean Technology, 16(3) 57-63.

[3] Wageeh, Y., Mohamed, H.E.D., Fadl, A., Anas, O., ElMasry, N., Nabil, A. and Atia, A., 2021. YOLO fish detection with Euclidean tracking in fish farms. Journal of Ambient Intelligence and Humanized Computing, 12(1), pp.5-12.

[4] 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).

[5] Joly A., Goeau H., Glotin H., Spampinato C., Bonnet P., Vellinga W.-P., Planquè R., Rauber A., Palazzo S., Fisher R., and others, LifeCLEF 2015: multimedia life species identification challenges, International Conference of the Cross-Language Evaluation Forum for European Languages, pp. 462-483, Springer, 2015.

[6] Bochkovskiy, A., Wang, C.Y. and Liao, H.Y.M., 2020. Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934.

[7] Jocher G, Nishimura K, Mineeva T, Vilariño R. yolov5. Code repository. 2020 May.

[8] Wang, C.Y., Yeh, I.H. and Liao, H.Y.M., 2021. You only learn one representation: Unified network for multiple tasks. arXiv preprint arXiv:2105.04206.