Fast and Accurate Fish Classification from Underwater Video using You Only Look Once

Hasna Maudi Lathifah\textsuperscript{1}, Ledya Novamizanti\textsuperscript{1,2}, Syamsul Rizal\textsuperscript{1}

\textsuperscript{1}School of Electrical Engineering, Telkom University
Jl. Telekomunikasi Terusan Buah Batu Bandung 40257 Indonesia

\textsuperscript{2}ledyaldn@telkomuniversity.ac.id

Abstract. Indonesia is a maritime country and one of the largest archipelago countries in the world. Indonesian fisheries have many types of fish in stock, this causes difficulties in introducing fish species directly. This study designed a fish species classification system using the You Only Look Once (YOLO) architecture. YOLO is an object detection method using a convolutional network that will only be just once. Unlike the convolutional networks in general that spend thousands of networks to obtain an image with computing that is long enough. The architecture of this work using YOLO9000. The dataset consists of 6 classes, that is banded butterflyfish, blue tang surgeonfish, barred hamlet, blackside hawkfish, Arabian Picasso triggerfish, dan black margate grunt. System testing produces an accuracy of 92%, IoS 0.75, and 2.223 FPS using Adam optimizer. The proposed system model has good accuracy and fast detection time.

1. Introduction
Indonesia is a maritime country and one of the largest archipelago countries in the world. The vast area of Indonesian fisheries can increase the potential to improve the Indonesian economy, especially in the fisheries sector. In Indonesia, there are many species of fish. Different types have thousands of different species. The classification of fish species is not only seen based on the analysis of color patterns, but also from the shape characteristics of the fins, and so on [1]. After using the characteristics, various features are matched according to the reference characteristics. The process of classifying fish manually requires a long time and a level of human accuracy [2]. Therefore, we need a system that can recognize fish species automatically, and can shorten the time.

Fish Recognition processes the framework of recognition and feature extraction recognition in an informative manner. Feature extraction is divided into two methods, namely, supervised and unsupervised. In the supervised method, one fish has been determined beforehand to be used as a contour. Fish descriptors are labeled as appearance like texture. In the unsupervised method, mathematically, this algorithm will group features or fish that are almost the same in a certain area [3].

Along with the increasing ocean observation, scientists began to research the introduction of fish species with automatic [2]. Ranju Mandal proposed the classification of fish underwater video use deep learning [4]. The method used is Faster R-CNN with additional architectures. So it takes longer to process fish introduction. Sushil Kumar Mahapatra investigated algorithm Fuzzy C-Mean Multithreading for detecting fish underwater. This algorithm resulted in a high detection rate and...
showed an accuracy rate of 98.35% [5]. However, in this study, it is difficult to track objects moving underwater. Elhoseny proposed technology and applications using Machine Learning. These studies use a robot for automated immersion tools that try to take a potential underwater video with high-resolution video [6]. The data obtained is the only video with high resolution. Anak Agung Putra Adi Kesava Esa performs the classification of marine fish underwater video with methods Faster R-CNN. The study using deep learning to identify four types of fish with VGG-16 architecture and AlexNet and searching the best combination [7]. However, in that study, just classify four types of fish with an accuracy of 87.25%.

This research applies You Only Look Once (YOLO) in the classification of types of fish from the underwater video. There are six different types of fish. YOLO is an object detection method that simplifies the task generalize object recognition task more quickly because the convolutional network is only passed once in detecting objects [8]. The way YOLO works is dividing the image input into grids of S size where the S value is 7 with the image input size of 488×488. To calculate bounding boxes, YOLO implements two main process steps, namely, Intersect over Union (IoU) and Non-Maximum Suppression (NMS) [8].

2. Methods

2.1. You Only Look Once

YOLO is an object detection method that is suitable for realtime processing [9]. The object detection method is tasked to determine the location where the object is in an image and determine the object class [10]. The R-CNN method uses several steps to perform the task, so that performance is slow. This is so because each component needs to be trained separately, unlike YOLO, which uses a single neural network. YOLO will divide the input image into 7 × 7 grid cells to match the use of the PASCAL VOC dataset [9]. Each grid cell is responsible for providing predicted bounding boxes. Predicted bounding boxes will provide an object existence score and a prediction score for the class of objects detected.

The YOLO9000 architecture is a development of the YOLOv1 architecture, which has 24 convolutional layers and two fully connected layers (FC) [9]. In YOLO9000 softmax layer is added as the last output layer, which functions to do multi-class classification. YOLO9000 also adds an average pooling layer to its architecture. YOLO9000 uses only the convolutional layer and the pooling layer. Each layer has its own role. Figure 1 is the Architecture of the YOLO network [11].

![Figure 1. The architecture of the YOLO network.](image)

2.2. Convolutional Layers
Convolutional layers are convolution layers that function as filters for the detection and extraction of image features. In other words, convolutional layers are the main layer on CNN. The extracted image features are then processed by convolution operations. The results obtained are edge detection, corner detection, and so on.

The formula of Convolution as follows

$$h(x) = f(x) * g(x)$$

with $f(x)$ and $g(x)$ is an input image and filter/kernel windows [12],[13].

2.3. Pooling Layer

Pooling layers are used to reduce the size of features. It takes important information by comparing the values of the matrix elements in one window. Pooling layers use the pooling function, which is max pooling, and average pooling. The concept of max pooling is to create a new matrix with a smaller size by taking the strongest features available on a layer. In its implementation, max pooling will take the largest pixel to be arranged into a new matrix. The average pooling will take the average value on a layer [12].

2.4. Transfer Learning

Transfer learning is a method that utilizes models that have been trained on a dataset to solve similar problems. The solution is to use it as a starting point, modify and update its parameters so that it matches the new dataset [14],[15]. For example, in real life, when we can recognize a type of banded fish, we can also recognize other types of fish from the characteristics of the type of banded or previous fish. This can be done by searching a dataset about the types of fish in the process of use. Transfer learning is encouraged because there is prior knowledge to solve new problems quickly.

2.5. Optimizer

The optimizer is a parameter that is used to improve learning on the system by using the value of learning rate or a certain level of learning in object detection. It optimizes the process of deep learning CNN via layer deep enough. The learning rate value determines the system's ability to detect objects quickly or slowly. The size of the learning rate depends on the default value of each optimizer. Optimizer functions that are commonly used are Stochastic Gradient Descent (SGD) and Adaptive with momentum (Adam) [16]. SGD is an optimizer function with a good level of effectiveness in the process of machine learning to do deep learning. This optimizer function is stochastic or able to measure an event with unstable data. SGD in the machine learning process uses a learning rate or learning rate of 0.01. Adam is an optimizer function, which is an optimization of the SGD optimizer function. Adam has an advantage in the computational process that is better than SGD and is widely used for large amounts of data. The learning rate value used by Adam is 0.001 [16].

3. System Design

This section designed a system for classifying fish species from underwater videos using YOLO architecture. Figure 2 shows a flow diagram of a fish species classification system.

3.1 Dataset

The dataset is in the form of 6 types of fish, namely banded butterflyfish, blue tang surgeonfish, barred hamlet, blackside hawkfish, Arabian Picasso triggerfish, and black margate grunt. The dataset was obtained from fishdb with a total dataset of 220 images, with training data for 120 images and test data for 100 images.

3.2 YOLO Process

This research uses a pre train model that has been trained using the COCO and PASCAL VOC dataset. In YOLO training, the system performs predictions or performs functions of the YOLO model with the trained dataset. The training process in YOLO has three stages: (1) Resize the image by resizing the image to $448 \times 488$, (2) Run a convolutional network that is one way or single past on the image,
(3) Perform non-max suppression, i.e., limit the detection produced, with the number of bounding boxes to output only one bounding box based on the specified threshold. From the dataset that has been trained, a new image is obtained with a bounding box along with the image's json file. In the bounding box, there is also the class name of each object, which is the type of fish. The json file is the confidence score and coordinates of the bounding box. Weight parameters have been trained with the COCO 20 class dataset, which is then carried out transfer learning into 6 classes. Update weight is a pre-training weight that must be in accordance with the basic model used. If the pre-training weights do not match the basic model used, then the train cannot be done. This study uses the config model YOLO 6 class, with the basic model yolo.cfg. Figure 3 is an example of the results of the test image detection with bounding boxes.

![Flow diagram of a fish classification using YOLO](image_url)

**Figure 2.** Flow diagram of a fish classification using YOLO (a) Training, and (b) Testing.
**Figure 3.** Fish images with bounding box (a) arabian picasso, (b) banded, (c) barred hamlet, (d) blackside, (e) black margate, dan (f) blue tang.

### 3.3 Performance Parameter

In this study, the parameters used to measure system performance, namely accuracy, precision, and Intersect of Union (IoU).

- **Accuracy** is used to measure the truth in recognizing fish species. Mathematically, accuracy is calculated as follows

\[
\eta = \frac{B}{C} \times 100
\]

with \(\eta\) is the percentage accuracy, \(B\) is an accurate type of fish, and \(C\) is the total data [15-17].

- **Precision** is the calculation of the distance between the centroid ground-truth and the actual centroid bounding box. Precision is calculated as follows

\[
E = \sqrt{\left(C_{x_{GT}} - C_{x_{ac}}\right)^2 + \left(C_{y_{GT}} - C_{y_{ac}}\right)^2}
\]

with \(E\) is the precision parameter, \(C_{x_{GT}}\) is the \(x\) coordinate for the centroid groundtruth, \(C_{x_{ac}}\) is the \(x\) coordinate for the actual centroid, \(C_{y_{GT}}\) is the groundtruth, and \(C_{y_{ac}}\) is the \(y\) coordinate for the actual centroid [2].

- **Intersect of Union (IoU)** is a parameter to calculate how much prediction is done to produce overlapping with groundtruth. IoU is calculated using a formula

\[
IoU = \frac{\text{Area of Overlap}}{\text{Area of Union}}
\]

IoU requires two areas to be intersected and unionized. The two areas are the groundtruth bounding box area, which is the actual bounding box, and the area detected from the model built [2].

### 4. Result and Discussion

The results of the detection are in the form of bounding boxes in the fish object area. From the bounding box, calculations can be done to get system performance. In this study, it is implemented and compared two types of optimizers, namely Adam and Stochastic Gradient Descent Optimizer (SGD). The test scenario is carried out to find the best optimizer and threshold value for the performance parameters.

#### 4.1 System performance on IoU

This section tests the parameters of the IoU to measure the accuracy of the method in detecting fish objects. The IoU parameter will play an important role in the accuracy parameter. The measurement is based on two areas, namely between groundtruth bounding boxes and predicted bounding boxes. Groundtruth boxes with predicted bounding boxes get better when IoU approaches 1. The system can detect well if IoU \(\geq 0.5\). Figure 4 shows the performance of YOLO with Adam and SGD optimizers. Based on Fig. 4, in Adam and SGD optimizers, the best IoU is at the 0.1 thresholds. The IoU value will get better if it gets closer to 1. In Adam’s optimizer, when the 0.1 threshold is obtained, the IoU value is 0.75. In the SGD optimizer, when the threshold is 0.1, the IoU value is 0.69. The resulting IoU value is optimizer Adam, 0.06 points greater than the SGD optimizer. System performance continues to decline if the threshold value is enlarged to 0.9. The greater the threshold, the bounding box is thinning with confidence score close to 0. The smaller the threshold, the bounding box is thickening.
with a confidence score close to 1. However, when tested with a threshold of less than 0.1, the bounding box overlaps and does not match the groundtruth bounding box.

4.2. **System performance on precision**
This section tests the parameters of precision to test the level of centroid groundtruth proximity with centroid predicted bounding boxes. A system with good precision if precision approaches 0. The closer the centroid groundtruth is to the centroid predicted bounding boxes, the smaller the precision. Precision approaching 0 means the system is getting better. Figure 5 shows the performance of Adam and SGD optimizers on precision.

![Figure 5. The results of the optimizer on precision.](image)

Based on Fig. 5, in the Adam and SGD optimizers, the best precision is obtained at the threshold 0.9. The greater the threshold, the two centroids predicted bounding boxes will be closer to the centroid bounding boxes groundtruth. The smaller the threshold, the two centroids predicted bounding boxes are increasingly away from the groundtruth centroid bounding boxes. However, when tested with a threshold of more than 0.9, centroid predicted bounding boxes cannot be seen whether approaching or moving away from centroid bounding boxes groundtruth because bounding boxes are very thin.

4.3. **System performance on accuracy**
This section examines system performance in recognizing fish species against accuracy parameters. The output network that has been trained with the training dataset is then tested in recognizing an object test image. Figure 6 shows the performance of YOLO with Adam and SGD optimizer to parameter accuracy.

![Figure 6. The results of the optimizer on accuracy.](image)
Based on Fig. 6, the best accuracy is obtained when the threshold is 0.1. The accuracy obtained by Adam Optimizer is 2% better than the accuracy obtained by the SGD optimizer. In the Adam and SGD optimizers, the accuracy parameter is directly proportional to the IoU parameter. When the predicted bounding boxes fail to approach the groundtruth bounding boxes, the system cannot detect the object. So the system cannot classify fish. IoU parameters for Adam optimization and SGD optimization, get results that are directly proportional to accuracy. A system with high accuracy does not guarantee the best precision. The system is more intelligent in recognizing an object if the accuracy is higher. But the proximity between centroid groundtruth and centroid predicted bounding boxes is not necessarily close. The precision value between centroid groundtruth and centroid predicted bounding boxes will still be included in the calculation, even though the results of the precision are quite large.

5. Conclusion
This study proposes the classification of fish species from underwater videos using YOLO architecture. The system successfully classified six fish classes with IoU 0.75, 92% accuracy, 0.24 precision, and 2.223 FPS. The best configuration is obtained at threshold 0.1 with Adam Optimizer. Adam's optimization uses the adaptive learning rate calculation, so it is good to be applied to the classification of fish species by the YOLO method. The system precisely detects fish objects from a short distance compared to a distance.

6. Reference
[1] Andayani U, Wijaya A, Rahmat R F, Siregar B, Syahputra M F 2019 Fish Species Classification Using Probabilistic Neural Network IOP Conf. Series: Journal of Physics: Conf. Series 1235 012094
[2] Li X, Shang M, Hao J and Yang Z 2016 Accelerating fish detection and recognition by sharing CNNs with objectness learning Proc. Ocean (Shanghai) IEEE
[3] Chuang M.-C, Hwang J.-N and Williams K A 2016 Feature Learning and Object Recognition Framework for Underwater Fish Images IEEE Trans. Image Process. 25(4) 1-1
[4] Mandal R, Connolly R M, Schlacher T A and Stantic v B 2018 Assessing fish abundance from underwater video using deep neural networks Proc. Int. Jt. Conf. Neural Networks (Brazil) IEEE
[5] Mahapatra S K, Mohapatra S K, Mahapatra S and Tripathy S K 2016 A Proposed Multithreading Fuzzy C-Mean Algorithm for Detecting Underwater Fishes 2nd International Conference on Computational Intelligence and Networks (India) IEEE
[6] Elhoseny H, Elhoseny M and Riad A M 2018 The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2018) 723 Springer international publishing
[7] Kesava A A E., Raharjo J and Wibowo S A 2019 Klasifikasi Ikan Laut Menggunakan Video Bawah Laut dengan Metode Faster RCNN (Region Convolutional Neural Network) Skripsi (Bandung: Universitas Telkom)
[8] Redmon J, Divvala S, Girshick R and Farhadi 2016 A You only look once: Unified, realtime object detection Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 779
[9] Redmon J, Divvala S, Girshick R and Farhadi 2016 A You only look once: Unified, realtime object detection Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. pp 779–788
[10] Phadnis R, Mishra J and Bendale S 2018 Objects Talk - Object Detection and Pattern Tracking Using TensorFlow Proc. Int. Conf. Inven. Commun. Comput. Technol. pp 1216–1219
[11] Redmon, Divvala, Girshick R and Farhadi A 2016 You only look once: Unified, realtime object detection In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp 779–788
[12] Wulandari C D R, Wibowo S A and Novamizanti L 2019 Classification of diabetic retinopathy using statistical region merging and convolutional neural network *IEEE Asia Pacific Conference on Wireless and Mobile*

[13] Priatama M B, Novamizanti L, Aulia S, Candrasari E B 2020 Hand gesture recognition using discrete wavelet transform and convolutional neural network *Bulletin of Electrical Engineering and Informatics* 9(3)

[14] Yang Q and Pan S J 2010 A Survey on Transfer Learning *IEEE Trans. Knowl. Data Eng.* 22(10) pp 1345–1359

[15] Radiuk P M 2018 Impact of Training Set Batch Size on the Performance of Convolutional Neural Networks for Diverse Datasets *Inf. Technol. Manag. Sci* 20(1) pp 20–24

[16] Kingma D P and Ba J L 2015 Adam: A method for stochastic optimization *3rd Int. Conf. Learn. Represent.* pp 1–15