# REAL-TIME MARINE ANIMAL DETECTION USING YOLO-BASED DEEP LEARNING NETWORKS IN THE CORAL REEF ECOSYSTEM

Jiageng Zhong[1], Ming Li[1]*, Jiangying Qin[1], Yuxin Cui[2], Ke Yang[3], Hanqi Zhang[1]

[1] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Luoyu Road NO.129, Wuhan, China - (zhongjiageng, lisouming, jy_qin, hqzhang)@whu.edu.cn

[2] Hongyi Honor College of Wuhan University, Bayi Road NO. 299, Wuhan, China. yuxincui@whu.edu.cn

[3] School of Water Resources and Hydropower Engineering, Wuhan University, Wuhan, China - YyangkeK@whu.edu.cn

**Commission III, WG III/1**

**KEY WORDS:** underwater images, object detection, deep learning, neural networks, coral reef, YOLO.

**ABSTRACT:**

In recent years, with the advancement of marine resources and environment research, the ecological functions of reef-building coral reef ecosystems distributed in warm shallow waters of the ocean are being continuously discovered and valued by people. It is important for ecosystem protection to monitor the population of marine animals. Besides, many projects of Autonomous Underwater Vehicle (AUV) also need technology to perceive and understand environment information in real-time for better decision-making. Therefore, marine animal detection has become a challenge for researchers to study nowadays. Deep neural network models have been used to solve fish-related tasks and gained encouraging achievements, but there are still many problems in this field. In this paper, several YOLO-based methods are chosen for comparison. Experiment results indicate that these methods can recognize the marine animals in coral reef quickly and accurately. Finally, several recommendations for model improvement according to assessment results are presented.

## 1. INTRODUCTION

In recent years, with the advancement of marine resources and environment research, the ecological functions of coral reef ecosystems distributed in warm shallow waters of the ocean are being continuously discovered and valued by people (Modasshir and Rekleitis, 2020). Because it is closely related to global climate warming and the development of marine resources, the demand for its dynamic and normalized monitoring is increasingly urgent. And in the coral reef ecosystem, marine animals play an important part in maintaining the balance of various components and their existence usually indicates that the ecosystem is in a good condition (Riansyah et al. 2018), so it is important for ecosystem protection to monitor the population of marine animals. Furthermore, as there are more and more projects of Autonomous Underwater Vehicle (AUV) in order to perform different underwater tasks, such as marine organism capturing, ecological surveillance and biodiversity monitoring (Kumar et al., 2018), there is an increasing need to develop technology on how to perceive and understand environment information in real-time for better decision-making. Therefore, marine animal detection has become a challenge for researchers to study nowadays (Berg et al., 2022).

In application, we hope that the marine animal detector can satisfy the demand for both precision and efficiency. In other word, a robust and lightweight detector is required. Various approaches have been followed for marine animal detection. Earlier in the decades ago, shallow learning architectures have been used for fish detection. But limited to their generalization capability, they are unable to extract high-level features robustly from complicated changing environment and exhibit low performance in real-world scenarios. As the deep learning technology improves, deep neural network models have been used to solve fish-related tasks (Moniruzzaman et al. 2017). One of the most commonly used variants of deep neural networks is the convolutional neural network (CNN) which can learn the inner features automatically and detect objects accurately. However, there are still some problems in this field. First, it is much harder to obtain underwater images, and the annotation task is costly. The labelled dataset of underwater object detection is extremely limited (Hashisho et al., 2019), hence data augmentation is needed to make model focus more on semantic information (Shorten et al., 2021). Second, the contradiction between the precision and speed becomes even more critical. Third, scale-changing objects and complicated condition bring great challenges to the detection model. Among varied methods, YOLO-based methods (Redmon et al., 2016) are representative and have good performance on both precision and speed, so that they have been at the centre of attention for many researchers.

You Only Look Once (YOLO) is a viral and widely used algorithm (Sultana et al., 2020) and has many derivative algorithms. Inspired by the GoogLeNet (Szegedy et al., 2015) model for image classification, the first YOLO version (YOLOv1) was proposed by Redmon et al. (Redmon et al., 2016) to create a one step process involving detection and classification. Unlike the previous traditional methods, YOLO can make bounding box predictions and class predictions simultaneous, so that it would meet the need of both accuracy and speed. In 2017, YOLOv2 (Redmon and Farhadi, 2017) was proposed to improve the speed of object detection while keeping the detection accuracy. Its primary network is a new classification model named Darknet-19, and anchor box mechanism is added to increase the accuracy of the network. It is deeper than YOLOv1 and keeps the real-time detection speed. However, its detection accuracy is still low for small or dense objects, and YOLOv3 (Redmon and Farhadi, 2018) was proposed to solve this problem. In order to solve the vanishing gradient problem of deep networks and preserve fine-grained features for small object detection, YOLOv3 applies a residual skip connection and uses an up-sampling and concatenation method. It has done a great job and left every object detection algorithm behind in terms of speed and accuracy (Gani et al., 2021). However, the original author of YOLO

announced to discontinue his research in the computer vision field after the release of YOLOv3. And YOLOv4 (Bochkovskiy et al., 2020) was published by a Russian developer named Alexey Bochkovskiy in 2020. A series of experiments with many of the most advanced innovation ideas of computer vision were performed for each part of the architecture of YOLOv4. One month later, researcher Glenn Jocher and his Ultralytics LLC research department published YOLOv5 (Ultralytics, 2020) with a few differences and improvements. YOLOv5 is written in Python programming language instead of C as in previous versions, so that it is easier to install. According to relevant researches, YOLOv5 has proved higher performance than YOLOv4 under certain circumstances (Thuan, 2021). Inspired by these researches, many improved algorithms have been proposed. Baidu released PP-YOLO (Long et al.,2020me) based on YOLOv3, and then proposed its improved version named PP-YOLOv2 (Huang et al., 2021). YOLOX (Ge et al., 2021) achieves a better trade-off between speed and accuracy due to advanced detection techniques, i.e., decoupled head, anchor-free, and advanced label assigning strategy. YOLOS (You Only Look at One Sequence) (Fang et al., 2021) is a series of object detection models based on the vanilla Vision Transformer (Dosovitskiy et al., 2020). Scaled-YOLOv4 (Wang et al., 2021) proposes a network scaling approach that modifies not only the depth, width, resolution, but also structure of the network. YOLOR (You Only Learn One Representation) (Wang et al., 2021) used a unified network that encoded implicit and explicit knowledge to predict the output, it can perform multitask learning such as object detection. Recently, Poly-YOLO (Hurtik et al., 2022) can achieve the same precision as YOLOv3, but it is three times smaller and twice as fast. Besides, some researchers (Kim et al., 2020) tried to convert pre-trained deep networks to Spiking Neural Networks (SNNs) and achieved suitable precision on non-trivial datasets.

YOLOv3 (Redmon and Farhadi, 2018) and YOLOv5 (Ultralytics, 2020) are classical networks in YOLO series, and YOLOR (Wang et al., 2021) is one of the representative neural network models. Therefore, in this paper, above three methods are chosen for comparison. Considering the lack of high-resolution underwater image datasets, an underwater object detection dataset containing thousands of instances is constructed by this paper, on which these methods are all tested. The results of the experiment show that these methods can recognize the marine animals in coral reef quickly and accurately. Due to the different structure of networks, they have different performances, and functions of their components from different perspectives are analyzed and discussed in this paper.

## 2. METHODOLOGY

### 2.1 YOLOv3

For deep neural networks, more layers usually mean more accuracy. But deeper layers are likely to lose fine-grained features, YOLOv2 often struggled with small object detections even if it is deeper. ResNet (He et al., 2016) brought the idea of skip connections to help the activations to propagate through deeper layers without gradient vanishing. On the basis of these studies, YOLOv3 (Redmon and Farhadi, 2018) was proposed. For performing feature extraction, a new network named Darknet-53 is applied. It is a hybrid approach between the network used in YOLOv2, Darknet-19, and that residual network stuff. As shown in Table 1, the network uses successive 3*3 and 1*1 convolutional layers but has some shortcut connections as well. According to the authors, this new network is much more powerful than Darknet-19 but still more efficient than ResNet-101 or ResNet-152.

| | Type | Filters | Size | Output |
|---|---|---|---|---|
| | Convolutional | 32 | 3*3 | 256*256 |
| | Convolutional | 64 | 3*3/2 | 128*128 |
| 1* | Convolutional | 32 | 1*1 | |
| | Convolutional | 64 | 3*3 | |
| | Residual | | | 128*128 |
| | Convolutional | 128 | 3*3/2 | 64*64 |
| 2* | Convolutional | 64 | 1*1 | |
| | Convolutional | 128 | 3*3 | |
| | Residual | | | 64*64 |
| | Convolutional | 256 | 3*3/2 | 32*32 |
| 8* | Convolutional | 128 | 1*1 | |
| | Convolutional | 256 | 3*3 | |
| | Residual | | | 32*32 |
| | Convolutional | 512 | 3*3/2 | 16*16 |
| 8* | Convolutional | 256 | 1*1 | |
| | Convolutional | 512 | 3*3 | |
| | Residual | | | 16*16 |
| | Convolutional | 1024 | 3*3/2 | 8*8 |
| 4* | Convolutional | 512 | 1*1 | |
| | Convolutional | 1024 | 3*3 | |
| | Residual | | | 8*8 |
| | Avgpool | | Global | |
| | Connected | | 1000 | |
| | Softmax | | | |

**Table 1**. Darknet-53 (Redmon and Farhadi, 2018).

### 2.2 YOLOv5

YOLOv5 (Ultralytics, 2020), which is proposed by Ultralytics LLC, is the latest product of the YOLO architecture series. There are several state-of-art innovations in the field of computer vision at that time being applied, so that it can achieve high precision and high speed. And its model is small, indicating that it is suitable for development to the embedded devices to implement real-time object detection. The YOLOv5 architecture contains four architectures, specifically named YOLOv5s, YOLOv5m, YOLOv5l and YOLOv5x, respectively. The main difference among them is that the amount of feature extraction modules and convolution kernels in the specific location of the network is different (Yan et al., 2021). YOLOv5s architecture is chosen for comparison in this paper.

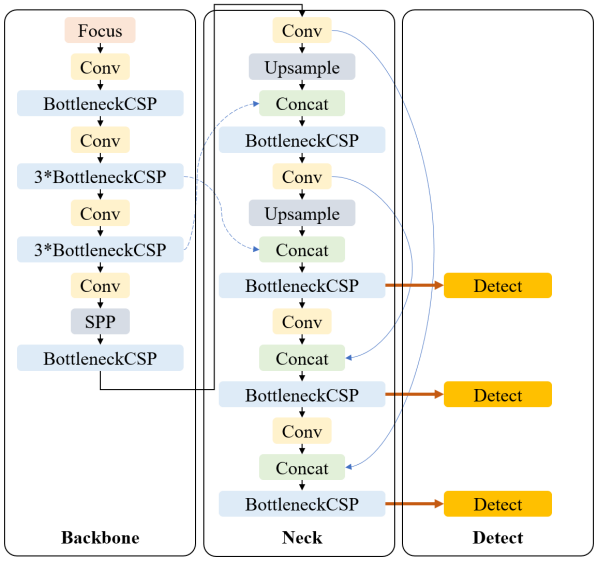


**Figure 1**. Architecture of YOLOv5s (Yan et al., 2021).

The YOLOv5s mainly consists of three components, including backbone network, neck network and detect network. Its architecture is shown in Figure 1. The BottleneckCSP module is mainly composed of a Bottleneck module with a convolutional layer, and the SPP module means spatial pyramid pooling. The backbone network is to aggregate different fine-grained images and form image features. The neck network is to aggregate feature image features. And the detect network is used to output the final detection.

### 2.3 YOLOR

YOLOR (You Only Learn One Representation) (Wang et al., 2021) applies a network that can learn a general representation by integrating implicit knowledge and explicit knowledge. As YOLOR has variants with multiple network depth and height scales, YOLOR-P6 and YOLOR-W6 are considered for the study. YOLOR-P6 has same architecture as YOLOv4-P6-light. The architecture topology of YOLOv4-P6-light is shown in Figure 2(a). It uses Stem D (Figure 2(b)) which is called focus layer, and base channels are set as {128, 256, 384, 512, 640}. It is noted that ReOrg (re-organization) in Stem D is a kind of down-sampling modules. All Mish activation in YOLOv4-P6-light is replaced by SiLU activation. As for YOLOR-W6, it is wider YOLOR-P6, and base channels are set as {128, 256, 512, 768, 1024}.

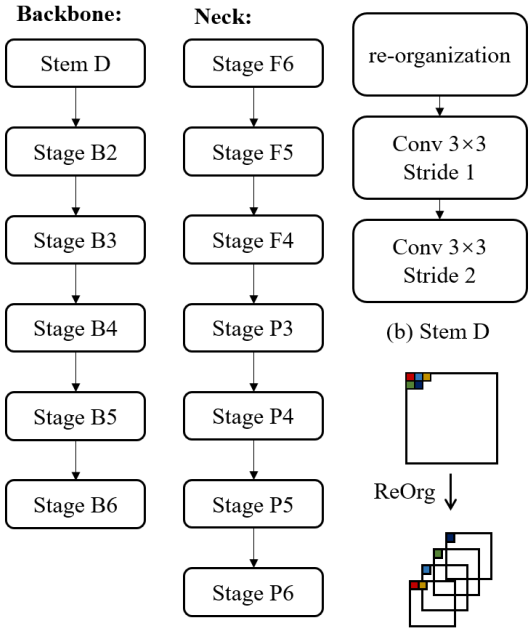


**Figure 2**. The illustration of YOLOR (Wang et al., 2021).

## 3. EXPERIMENTS AND DISCUSSION

### 3.1 Data

Due to the lack of high-resolution underwater image datasets, an underwater object detection dataset containing thousands of instances is created. The example images from the dataset are shown in Figure 3. The objects are classified into two main categories: fish and turtle.

Figure 4 shows the 2D histogram of bounding box pixel coordinates. It can be seen that bounding boxes are distributed uniformly in images. In addition, most of objects are fish whose size is usually small, while almost all big objects are turtles.

**Figure 3**. Example images from the dataset.

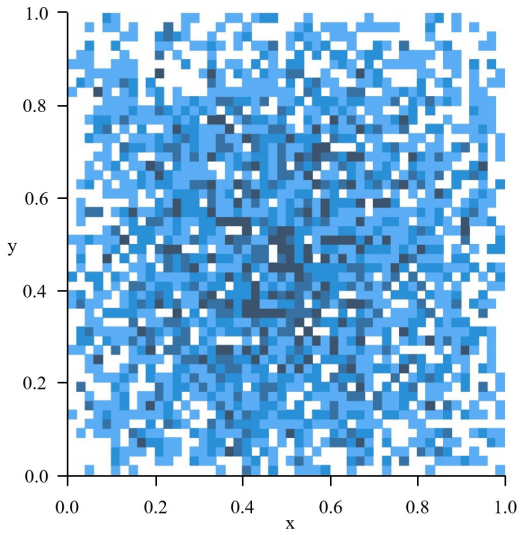


**Figure 4**. 2D histogram of bounding box pixel coordinates.

### 3.2 Training details

Three networks are all trained for 300 epochs with image resizing to 640*640. Models are trained on a Nvidia Geforce RTX 3060 GPU with PyTorch implementation, and an 80/20 training/test split is used on the dataset.

### 3.3 Experiment results

The importance of an object detection algorithm is weighted by how accurate and quickly it can detect objects. For accuracy evaluation, the detection Average Precision (AP) is calculated in this paper, because this is the actual metric for object detection. The results can be seen in Table 2. The third column shows the AP when the IoU threshold is 50% ($AP_{50}$). And the fourth column shows the AP averaged over 10 different thresholds from 50% to 95% ($AP_{50:95}$).

| Model | Category | $AP_{50}$ | $AP_{50:95}$ |
|---|---|---|---|
| YOLOv3 | fish | 0.842 | 0.542 |
| | turtle | 0.995 | 0.902 |
| YOLOv5 | fish | 0.768 | 0.464 |
| | turtle | 0.995 | 0.771 |
| YOLOR-P6 | fish | 0.794 | 0.487 |
| | turtle | 0.995 | 0.995 |
| YOLOR-W6 | fish | 0.805 | 0.512 |
| | turtle | 0.995 | 0.902 |

**Table 2**. Precision of three models.

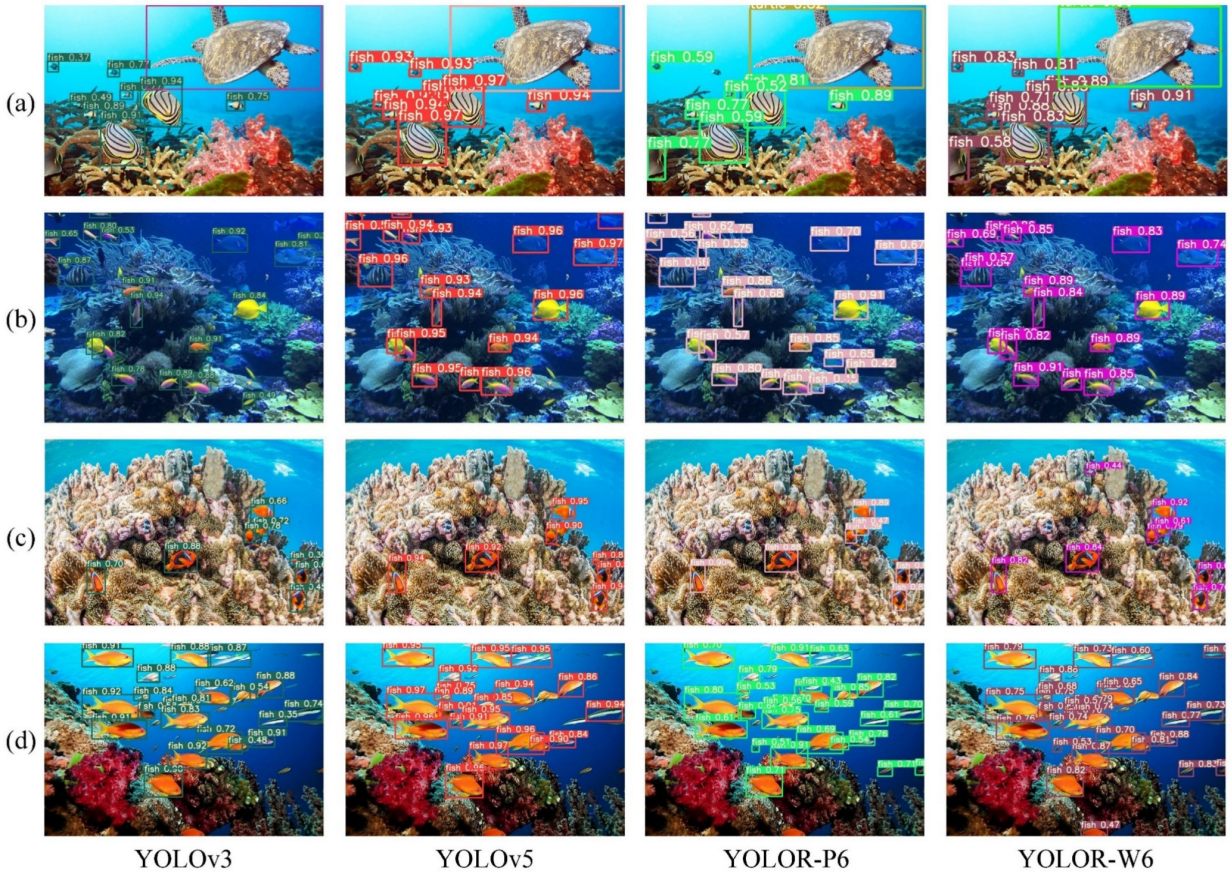


**Figure 5**. Examples of marine animal detection.

To evaluate speed of three models, their average execution time is measured and shown in Table 3. All these methods can perform real-time marine animal detection.

| Model | Average Execution Time (ms) |
|---|---|
| YOLOv3 | 24.3 |
| YOLOv5 | 13.3 |
| YOLOR-P6 | 22.4 |
| YOLOR-W6 | 23.4 |

**Table 3**. Precision of three models.

According to experimental results, YOLOv5 is the most efficient, and YOLOv3 has the best performance as it achieves highest AP. As for YOLOR, YOLOR-W6 has higher precision with a very small amount of additional cost.

Several examples of marine animal detection are displayed in Figure 5, and each row corresponds to a scene. From these examples, it can be seen that YOLOv5 sometimes misses objects, such as fish in Scene (b) and Scene (c). YOLOv3 and YOLOR-W6 can nearly detect all objects with high confidence.

### 3.4 Discussion

It can be found from the experiment results that deeper network does not necessarily perform better. This may mainly because the scale of the dataset is small. Deep networks, such as YOLOR-P6 and YOLOR-W6, can achieve the state-of-the-art performance on the large-scale dataset. But for such a small dataset, they probably suffer from overfitting. Conversely, a simpler network, i.e. YOLOv3, can fit better and get higher-precision detection.

To solve the problem of insufficient data, on the one hand, pretraining process plays a critical role, especially for the task with a small dataset. It seems better to pretrain an object detection model on a large-scale dataset, such as COCO (Lin et al., 2014). On the other hand, data augmentation is a solution. Considering deep model severely suffers from domain shift, there is a new augmentation method Water Quality Transfer (WQT) which is proposed to enlarge the dataset and increase domain diversity has achieved promising performance of domain generalization (Liu et al., 2020).

## 4. CONCLUSION

In conclusion, object detection using several YOLO-based methods to real-time identify marine animals in the coral reef ecosystems has been done successfully. These methods have properties of high efficiency and accuracy, so that they can provide other tasks with reliable processing results in high frame rate. Based on this, this paper evaluates effects of different structure of network models, and discuss the solution for limited underwater dataset. This work can give reference for research of real-time marine animal detection.

## ACKNOWLEDGEMENTS

This research was funded by the National Key R&D Program of China, grant numbers 2018YFB0505400, the National Natural Science Foundation of China (NSFC), grant number 41901407 and the College Students' Innovative Entrepreneurial Training Plan Program, Research on visual navigation, perception and localization algorithm of unmanned underwater vehicle/robot (UUV).

## REFERENCES

Berg, P., Santana Maia, D., Pham, M. T., Lefèvre, S., 2022. Weakly Supervised Detection of Marine Animals in High Resolution Aerial Images. *Remote Sensing*, 14(2), 339.

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

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint* arXiv:2010.11929.

Fang, Y., Liao, B., Wang, X., Fang, J., Qi, J., Wu, R., Niu, J., Liu, W., 2021. You only look at one sequence: Rethinking transformer in vision through object detection. *Advances in Neural Information Processing Systems*, 34.

Gani, M. O., Kuiry, S., Das, A., Nasipuri, M., Das, N., 2021. Multispectral object detection with deep learning. In International Conference on Computational Intelligence in Communications and Business Analytics. pp. 105-117.

Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J., 2021. Yolox: Exceeding yolo series in 2021. *arXiv preprint* arXiv:2107.08430.

Hashisho, Y., Albadawi, M., Krause, T., von Lukas, U. F., 2019. Underwater color restoration using u-net denoising autoencoder. In 2019 11th International Symposium on Image and Signal Processing and Analysis, pp. 117-122.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770-778.

Huang, X., Wang, X., Lv, W., Bai, X., Long, X., Deng, K., Dang, Q., Han, S., Liu, Q., Hu, X., Yu, D., Ma, Y., Yoshie, O., 2021. PP-YOLOv2: A practical object detector. *arXiv preprint* arXiv:2104.10419.

Hurtik, P., Molek, V., Hula, J., Vajgl, M., Vlasanek, P., Nejezchleba, T., 2022. Poly-YOLO: higher speed, more precise detection and instance segmentation for YOLOv3. *Neural Computing and Applications*, 1-16.

Hutchings, L., Augustyn, C. J., Cockcroft, A., Van der Lingen, C., Coetzee, J., Leslie, R. W., Tarr, R. J., Oosthuizen, H., Lipinski, M. R., Roberts, M. R., Wilke, C., Crawford, R., Shannon, L. J., Mayekiso, M. (2009). Marine fisheries monitoring programmes in South Africa. South African Journal of Science, 105(5), 182-192.

Kim, S., Park, S., Na, B., Yoon, S., 2020. Spiking-yolo: spiking neural network for energy-efficient object detection. In Proceedings of the AAAI Conference on Artificial Intelligence . pp. 11270-11277.

Kumar, G. S., Painumgal, U. V., Kumar, M. C., Rajesh, K. H. V., 2018. Autonomous underwater vehicle for vision based tracking. *Procedia computer science*, 133, 169-180.

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollar, P., Zitnick, C. L., 2014, September. Microsoft coco: Common objects in context. In European conference on computer vision. pp. 740-755.

Liu, H., Song, P., Ding, R., 2020. WQT and DG-YOLO: towards domain generalization in underwater object detection. *arXiv preprint* arXiv:2004.06333.

Long, X., Deng, K., Wang, G., Zhang, Y., Dang, Q., Gao, Y., et al., 2020. PP-YOLO: An effective and efficient implementation of object detector. *arXiv preprint* arXiv:2007.12099.

Modasshir, M., Rekleitis, I., 2020. Enhancing coral reef monitoring utilizing a deep semi-supervised learning approach. In 2020 IEEE International Conference on Robotics and Automation. pp. 1874-1880.

Moniruzzaman, M., Islam, S. M. S., Bennamoun, M., Lavery, P., 2017. Deep learning on underwater marine object detection: A survey. In International Conference on Advanced Concepts for Intelligent Vision Systems. pp. 150-160.

Redmon, J., Divvala, S., Girshick, R., 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.

Redmon, J., & Farhadi, A., 2017. YOLO9000: better, faster, stronger. In Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7263-7271.

Redmon, J., Farhadi, A., 2018. Yolov3: An incremental improvement. *arXiv preprint.* arXiv:1804.02767.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Riansyah, A., Hartono, D., Kusuma, A. B, 2018. Ikan Kepe–kepe (Chaetodontidae) sebagai bioindikator kerusakan perairan ekosistem terumbu karang Pulau Tikus. *Majalah Ilmiah Biologi Biosfera: A Scientific Journal*, 35(2), 103-110.

Shorten, C., Khoshgoftaar, T. M., Furht, B., 2021. Text data augmentation for deep learning. *Journal of big Data*, 8(1), 1-34.

Sultana, F., Sufian, A., Dutta, P., 2020. A review of object detection models based on convolutional neural network. *Intelligent computing: image processing based applications*, 1-16.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1-9.

Thuan, D, 2021. Evolution Of Yolo Algorithm And Yolov5: The State-Of-The-Art Object Detection Algorithm. Bachelor's Thesis, Oulu University of Applied Sciences, Oulu, Finland.

Ultralytics. yolov5. Available online: https://github.com/ultralytics/yolov5 (10 March 2022).

Wang, C. Y., Bochkovskiy, A., Liao, H. Y. M., 2021. Scaled-yolov4: Scaling cross stage partial network. In Proceedings of the IEEE/cvf conference on computer vision and pattern recognition. pp. 13029-13038.

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

Yan, B., Fan, P., Lei, X., Liu, Z., Yang, F., 2021. A real-time apple targets detection method for picking robot based on improved YOLOv5. *Remote Sensing*, 13(9), 1619.