An enhanced YOLOv5 model for fish species recognition from underwater environments

Chiranjibi Shah[a], Simegnew Yihunie Alaba[b], M M Nabi[b], Jack Prior[a,c], Matthew D Campbell[c], Farron Wallace[d], John E. Ball[b], and Robert Moorhead[a]

[a]Northern Gulf Institute, 2 Research Blvd., Mississippi State University, Starkville, MS 39759, USA
[b]Department of Electrical and Computer Engineering, James Worth Bagley College of Engineering, Mississippi State University, Miss. State, MS 39762, USA
[c]National Marine Fisheries Services, Southeast Fisheries Science Center, 3209 Frederic Street, Pascagoula, MS 39567, USA
[d]NOAA Fisheries, 4700 Avenue U, Galveston, TX 77551, USA

ABSTRACT

Species recognition is an important aspect of video based surveys, which support stock assessments, inspecting the ecosystem, handling production management, and protecting endangered species. It is a challenging task to implement fish species detection algorithms in underwater environments. In this work, we introduce the YOLOv5 model for the recognition of fish species that can be implemented as an object detection model for analyzing multiple fishes in a single image. Moreover, we have modified the depth scale of different layers in the backbone of the YOLOv5 model to obtain improved results on fish species recognition. In addition, we have implemented a transformer block in the backbone network and introduced a class balance loss function to obtain enhanced performance. It can perform fish species recognition as an object detection approach by classifying each of the fish species in addition to localizing for the estimation of the position and size of the fish in an image. Experiments are conducted on the fine-grained and large-scale reef fish dataset that we have obtained from the Gulf of Mexico – the Southeast Area Monitoring and Assessment Program Dataset 2021 (SEAMAPD21). The experimental results demonstrate that an enhanced YOLOv5 model can yield better detection results in comparison to YOLOv5 for underwater fish species recognition.

Keywords: object detection, underwater fish species, enhanced YOLOv5

1. INTRODUCTION

Fish species identification is an important aspect of fisheries management and environmental monitoring. Precise identification of fish species is crucial for various purposes, such as identifying endangered species, determining the best time and size for harvesting, monitoring ecosystems, and creating efficient production management systems.[1,2] Legal constraints on fishing methods make accurate fish species recognition even more critical, especially for threatened or endangered species. The traditional methods for identifying fish species require significant human labor, time, and can disrupt the normal behavior of fish. These conventional approaches pose challenges for maintaining high-quality data for managing sustainability in fisheries, monitoring federal fisheries, assessing fish populations, and identifying different fish species. However, the use of deep learning technology can provide robust models for fish species identification, which can significantly reduce costs, time, and improve identification accuracy.

Manual count and identification of species can be replaced with machine vision solutions given equal or improved precision. Various methods for fish detection exist, such as lidar,[3,4] sonar,[5] and RGB imaging.[6] RGB

Further author information: (Send correspondence to John E. Ball)
Chiranjibi Shah: E-mail: cshah@ngi.msstate.edu,
John E. Bal: E-mail: jeball@ece.msstate.edu

Ocean Sensing and Monitoring XV, edited by Weilin Hou, Linda J. Mullen,
Proc. of SPIE Vol. 12543, 125430O · © 2023 SPIE
0277-786X · doi: 10.1117/12.2663408

imaging is the preferred option in clear water for species identification because it allows for easy identification of fish based on their color, texture, and geometry. Moreover, it is cost effective, lightweight, and does not harm the fish habitat. In previous research, video frames[7,8] were analyzed individually to detect any object in the frame. In recent times, various camera systems have been used to track the stock of fish and determine the sustainability of marine ecology; however, they all suffer from the same time-intensive manual processing bottlenecks.[9] Recently developed deep learning techniques can be used to gather information about marine ecology by detection and classification. However, the underwater environment presents challenges, including low light and turbid conditions, occlusion, low images and videos resolution, and difficulty distinguishing fish from the background. The movement of fish also introduces shape variations and occlusion issues, making it challenging to identify and detect underwater fish species. In computer vision, deep learning has been widely used to solve a variety of issues, including detection, localization, estimation, and classification.[10–13] Several machine learning (ML) and deep learning algorithms have been developed to categorize fish species. For instance, Jager *et al.*[14] employed AlexNet architecture for feature extraction and multiclass SVM for classification, whereas hierarchical features and support vector machine (SVM) are used for fish classification.

You Only Look Once (YOLO)[15,16] is a popular object detection model that can be applied to identify fish, particularly for videos. YOLO is a single-shot detection model, which means that it processes the entire image or video frame in a single pass and predicts the locations and classes of objects in that image or video. YOLO is designed to be very fast, making it appropriate for real-time applications, and when trained on large datasets, it can achieve high accuracy. The YOLO model is trained using a collection of labeled images and is based on a convolutional neural network (CNN) architecture. The model learns to predict bounding boxes, or where the fish appears in an image or video, as well as the corresponding class probabilities for each box. (i.e., the fish species). A loss function is used to train the model, penalizing incorrect predictions and enticing it to get better at making predictions over time.

One advantage of YOLO is that it can be trained on a dataset of labeled fish images to distinguish between various fish species. This enables the model to adjust to various fish populations and environmental conditions. In order to speed up processing, YOLO is also highly parallelizable, which enables it to run concurrently on multiple GPUs. But because YOLO is a single-stage detection network, it is less precise than two-stage networks like Faster R-CNN. Additionally, like other object detection models, YOLO may have trouble identifying fish in low lighting or when they are partially obscured. However, YOLO remains a popular and effective choice for fish classification and detection tasks. Jocher et. al[17] originally introduced YOLOv5 based technique for object dection in MS COCO public datasets, and Jung et. al[18] implemented YOLOv5 for object detection in drone images. However, abovementioned techniques lack proper implementation for fish species recognition in underwater environments with highly imbalanced distributions of species. We were able in improving the performance of original YOLOv5, by modifications in the backbone of original YOLOv5, for fish species recognition in underwater environments.

Our main contributions can be summarized as follows.

1. We have modified the depth scale of different layers in the backbone of the YOLOv5 model to obtain improved results on fish species recognition.
2. Incorporated the transformer block in the backbone network of the YOLOv5-based approach.
3. Introduced the class balance (CB) loss function to get better classification and localization performance for an highly imbalanced dataset.

2. RELATED WORK

Deep learning (DL) techniques have quickly gained popularity and are now being successfully applied in several industries, including the fishery industry.[19] Information and data processing in smart fish farming is impacted by DL in a variety of ways, including new opportunities and challenges. Aquaculture makes extensive use of DL techniques for tasks like live fish identification, species classification, behavioral analysis, feeding selection, size or biomass estimation, and water quality forecasting. There are several DL approaches that are particularly useful for fish datasets, and the DL model can be developed based on that specific task.

In particular, the YOLO algorithm has been increasingly used for fish identification and detection tasks. One of the earliest studies using YOLO for this purpose was conducted by Xu *et al.* (2018),[20] who developed a model and used three very distinct datasets captured at water power sites to train YOLO to identify fish in underwater video. The mean average precision (mAP) score obtained after training and testing using examples from all three datasets was 0.5392. These findings suggest that distinct techniques are required to develop a trained model generalizable to new data sets, such as those found in practical applications. By performing an image enhancement, Liu *et al.* (2018)[21] were able to produce depth information that would be useful for many vision algorithms and sophisticated image editing. They developed a underwater fish detection and tracking strategies combining YOLOv3 algorithm with a parallel correlation filter. They demonstrated online fish detection and tracking on the NVIDIA Jetson TX2, enabling a fast system and rapid experimentation.

The Southeast Area Monitoring and Assessment Program Dataset (SEAMAPD21) was used in a recent study where YOLOv4 was employed to identify and categorize various fish species. The outcomes demonstrated that the YOLOv4 model was successful in locating and identifying various fish species in the SEAMAPD21 dataset.[22] Alaba *et al.*[10] used two feature extraction networks, MobileNetv3-large and VGG16, to extract features from the images and a single-shot multi-box detector (SSD) to classify the species and detect the location of the fish in the image for the SEAMAPD21 dataset.

Sung *et al.*[23] presented a CNN model based on the YOLO algorithm for real-time fish detection using underwater vision. The validity and precision of the proposed method were examined using actual fish video images. The network achieved a classification accuracy of 93%, a predicted bounding box intersection over union of 0.634, and a fish detection rate of 16.7 frames per second. It outperformed a fish detector that employs a sliding window algorithm and a classifier that was trained using a histogram of oriented gradient features and a support vector machine. Jalal *et al.*[24] proposed a two-step DL method to identify and classify temperate fishes. In the first step, they used the YOLO object detection method to identify each fish in the image, regardless of its species or sex. In the second step, they used a CNN with the Squeeze-and-Excitation structure to detect every single fish in the image without any prior filtering. In order to reduce the effect of the limited samples issue, they applied transfer learning. It improves the overall classification accuracy.

Overall, these studies demonstrate the potential of DL particularly the YOLO-based approach for fish identification and detection tasks, with promising results in terms of accuracy and processing speed. However, there are still challenges to be addressed, such as dealing with the complex and varied underwater environment and improving the generalization ability of the models to handle new and unseen fish species.

3. PROPOSED METHOD

3.1 A YOLOv5 technique for fish species recognition from underwater environments

Yolov5 is a single-stage network that can solve classification and localization problems simultaneously. YOLOv5 has different models, such as YOLOv5s, YOLOv5m, and YOLOv5l.[17,18] The basic principle is the same but is classified depending on the memory usage. For YOLOv5l, the backbone network consists of convolutional layers in addition to the cross-stage partial (CSP) connections with multiple residual layers inside them as 3xC3-6xC3-9xC3-3xC3. The Neck consists of image features, obtained by combining and mixing various layers, that can be delivered for the prediction. Finally, the Head makes predictions on image features obtained from the Neck by utilizing steps of class and box predictions.

3.2 An enhanced YOLOv5 approach for fish species recognition from underwater environments

We have selected YOLOv5l as a baseline and modified the architecture to get an enhanced performance for fish species detection in underwater environments. Backbone is the network that can combine features of images with various convolutional neural networks.

In the YOLOv5l backbone, the CSP layer C3 has a 3-6-9-3 number of layers, respectively. In YOLOv5enh, we modified the depth scale of backbones as 7xC3-15xC3-15xC3-7xtrans by replacing the final C3 layer with a transformer (trans) block as shown in Fig 1. Moreover, we introduced the class balance (CB) loss function[25] to enhance the detection performance in the SEAMAPD21, an imbalanced dataset. The Class Balance (CB) terms

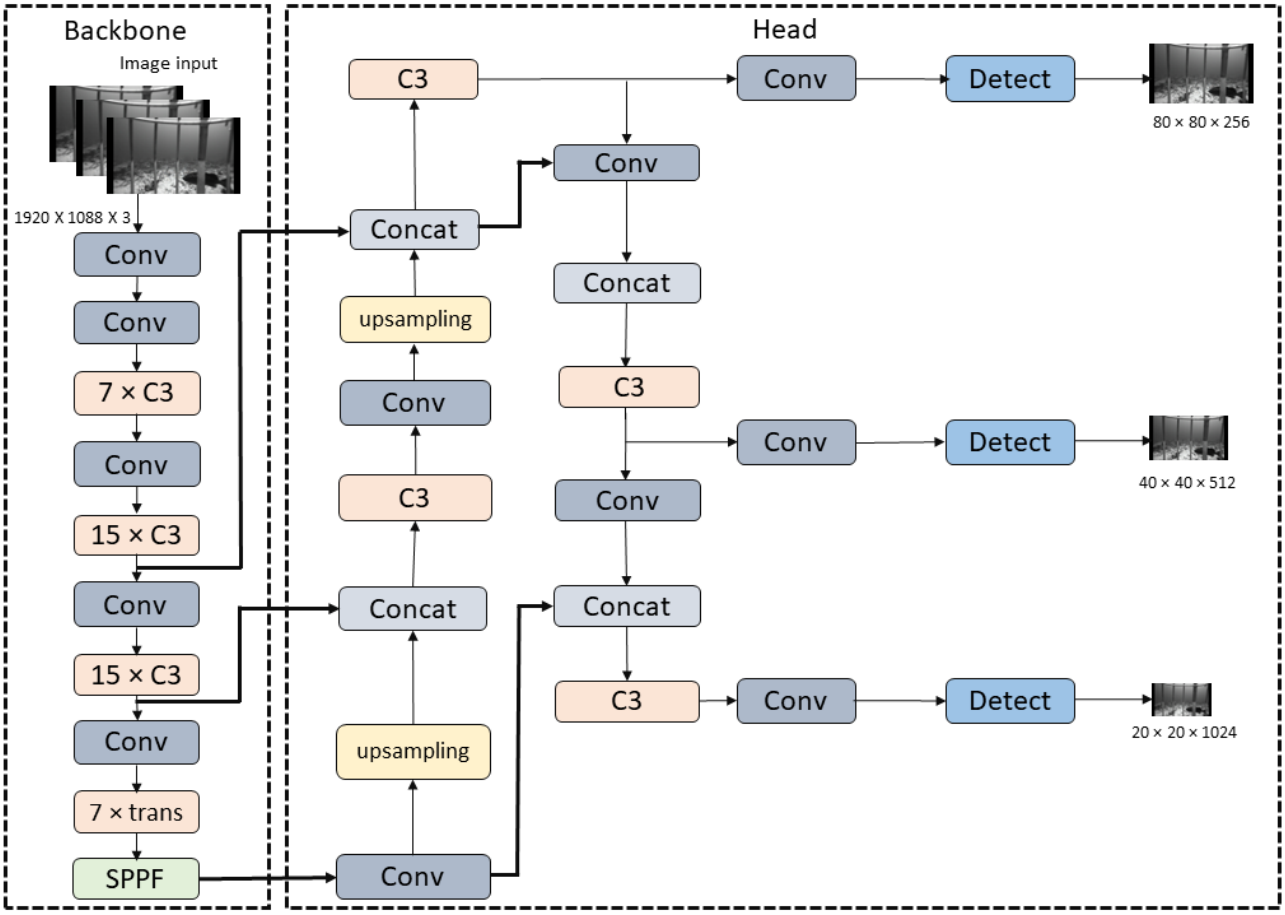

Figure 1. An Enhanced YOLOv5 based architectures for fish species detection on Pascagoula data (SEAMAPD21)

can be incorporated into localization loss and classification loss terms. The classification loss with the CB term can be estimated as:

$$Ls_{clscb} = CB_{los} Ls_{cls}(h, cl) = \frac{1-\Gamma}{1-\Gamma^{ns_y}} Ls_{cls}(h, cl)\,, \tag{1}$$

where $CB_{los} = \frac{1-\Gamma}{1-\Gamma^{ns_y}}$ is the regularization term for the class balance loss function,ns_y is the number of instances per species, Γ is the hyperparameter for class balance loss, Ls_{cls} represents the classification loss estimated with the focal loss, h represents ground truth and cl is the predicted output.

Similarly, The localization loss with the CB term can be estimated as:

$$Ls_{loccb} = CB_{los} Ls_{loc}(q, t) = \frac{1-\Gamma}{1-\Gamma^{ns_y}} Ls_{loc}(q, t))\,, \tag{2}$$

where Ls_{loc} denotes the localization loss, q is the prediction box, and t is the ground truth box.

3.3 Dataset

The large-scale reef fish data obtained from the Gulf of Mexico - the Southeast Area Monitoring and Assessment Program Dataset 2021 (SEAMAPD21)[26] consists of 130 distinct classes of fish species in underwater environments with 28,319 total images. However, some species are very small in number and the model is influenced by samples with more species per class. This is an underwater fishery-independent dataset and it is difficult to detect fish in such a low-resolution environment consisting of indistinguishability between images and background. The

ratio of 70/15/15 is used for train, validation, and test set respectively. The mean average precision (mAP) is evaluated on the test set.

4. EXPERIMENTAL RESULTS

4.1 Implementation details

We used PyTorch 1.13.0 to train the proposed Yolov5enh and utilized an NVIDIA A100-SXM GPU to train and test all the models. We trained the proposed approach from scratch and utilized the Stochastic Gradient Descent (SGD) as a model optimizer. A total of 300 epochs were used to train. The learning rate utilized is 0.01, $\Gamma = 0.5$, and the long side of the input image is 1920 leading to a batch size of 32.

4.2 Performance

The widely used mean average precision (mAP) is used to measure the performance of the proposed YOLOv5enh and other existing versions of YOLOv5. The mean average precision $mAP_{0.5-0.95}$ uses average of the Intersection Over Union (IOU) values varying from 0.5 to 0.95, and $mAP_{0.5}$ uses I0U of 0.50. As shown in Table 1, the mAP is used to measure the performance of the proposed YOLOv5enh-based approach in comparison to others. Some species have a small number of samples, for instance, less than 10, and may not have sufficient samples for training, valid, and test sets. 126 species are thus used to estimate mAP. It can be observed that the proposed YOLOv5enh, with $mAP_{0.5}$ of 85.2% and $mAP_{0.5-0.95}$ of 56.6%, outperforms all other compared versions, such as YOLOv5s, YOLOv5m, and YOLOv5l for fish species detection in underwater environments.

Moreover, the size of the network in terms of the number of parameters in millions (M) and number of calculations (GFLOPS) is presented in Table 1 for comparing the effect of the proposed approach with existing yolov5-based techniques. It can be observed that the number of parameters (61.30M) and GFLOPS(151.0) for yolov5enh is higher compared to other yolov5s, yolov5m, and yolov5l. However, accuracy in terms of mAP is also high for yolov5enh in comparison to others. In addition, Figure 2 shows a comparison of proposed yolov5enh with other existing versions, such as yolov5s, yolov5m, and yolov5l in terms of mAP, number of calculations (GFLOPS), and number of parameters in millions (M). The number of parameters is shown by the radii of circles. Although the GFLOPS and parameters required by YOLOv5enh are higher than those needed by YOLOv5l, the gain in performance in terms of mAP compensates for that.

Table 1. mean average precision (mAP(%)) for 126 species on Pascagoula data (SEAMAPD21).

method	$mAP_{0.5}$	$mAP_{0.5:0.95}$	Parameters	GFLOPS
YOLOv5s	78.7	48.7	7.37M	17.1
YOLOv5m	80.1	51.4	21.37M	49.5
YOLOv5l	80.9	53.0	46.80M	109.9
YOLOv5enh	**85.2**	**56.6**	61.30M	151.0

Table 2. Ablation analysis.

method	$mAP_{0.5}$	$mAP_{0.5:0.95}$
YOLOv5enh(w/o CB and w/o trans)	82.4	53.4
YOLOv5enh(with CB and and w/o trans)	83.1	54.0
YOLOv5enh	**85.2**	**56.6**

Table 2 shows an ablation study on the proposed technique YOLOv5enh. When the trans block, as shown in the backbone section of Fig 1, is removed, yolov5enh with class balance (CB) and without transformer block (with CB and w/o trans) shows less accuracy i.e. $mAP_{0.5}$ of 83.1% and $mAP_{0.5-0.95}$ of 54.0%. Moreover, when the class balance loss function is removed from YOLOv5enh, the accuracy of YOLOv5enh (w/o CB and w/o

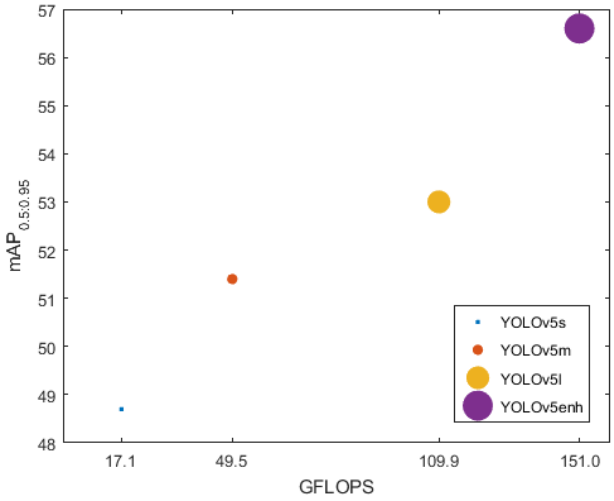

Figure 2. Comparing the performance of proposed YOLOv5enh and other YOLOv5-based approaches based on GFLOPS, and parameters in million(M), shown by radii of the circles in Pascagoula data (SEAMAPD21).

trans) decreases to $\text{mAP}_{0.5}$ of 82.4% and $\text{mAP}_{0.5-0.95}$ of 53.4%. However, YOLOv5enh(w/o CB and w/o trans) outperforms the existing YOLOv5s, YOLOv5m, and YOLOv5l.

Figures 3 and 4 show detection maps for fish species, in underwater environments with SEAMAPD21, with the existing YOLOv5l and the proposed YOLOv5enh, respectively. It can be observed that YOLOv5enh has better detection results in comparison to YOLOv5l.

5. CONCLUSIONS

In this work, we have introduced an enhanced YOLOv5 (YOLOv5enh) based approach for fish species recognition in underwater environments. Experiments are conducted on an underwater fish species dataset obtained from the Gulf of Mexico - the Southeast Area Monitoring and Assessment Program Dataset 2021 (SEAMAPD21) - to illustrate the superiority of the proposed YOLOv5enh. In order to improve the performance of the method, the depth scale in the backbone has been modified. As the dataset has a class imbalance issue, the class balance loss function is introduced to give more weight to species with fewer samples. It can reweight the loss depending upon the inverse number of samples in each class species to minimize the effect of approach biased towards the dominant class. Besides that, a transformer block is incorporated in the backbone network to get an enhanced performance for fish species recognition in comparison to the existing YOLOv5-based techniques. We demonstrate different YOLO model comparisons to illustrate the performance improvements. The improved YOLOv5enh model is able to achieve $\text{mAP}_{0.5}$ of 85.2% and $\text{mAP}_{0.5-0.95}$ of 56.6% with 61.30M parameters. As the model size is large, the computational complexity is also high. For real-world applications, we may need to consider a small network but compromise the accuracy for that.

ACKNOWLEDGMENTS

This work was supported by awards NA16OAR4320199 and NA21OAR4320190 to the Northern Gulf Institute at Mississippi State University from NOAA's Office of Oceanic and Atmo- spheric Research, U.S. Department of Commerce. Authors are thankful for the source of funding.

REFERENCES

[1] Chang, C., Fang, W., Jao, R.-C., Shyu, C., and Liao, I.-C., "Development of an intelligent feeding contrkrizhevsky2009learningoller for indoor intensive culturing of eel," *Aquacultural engineering* **32**(2), 343–353 (2005).

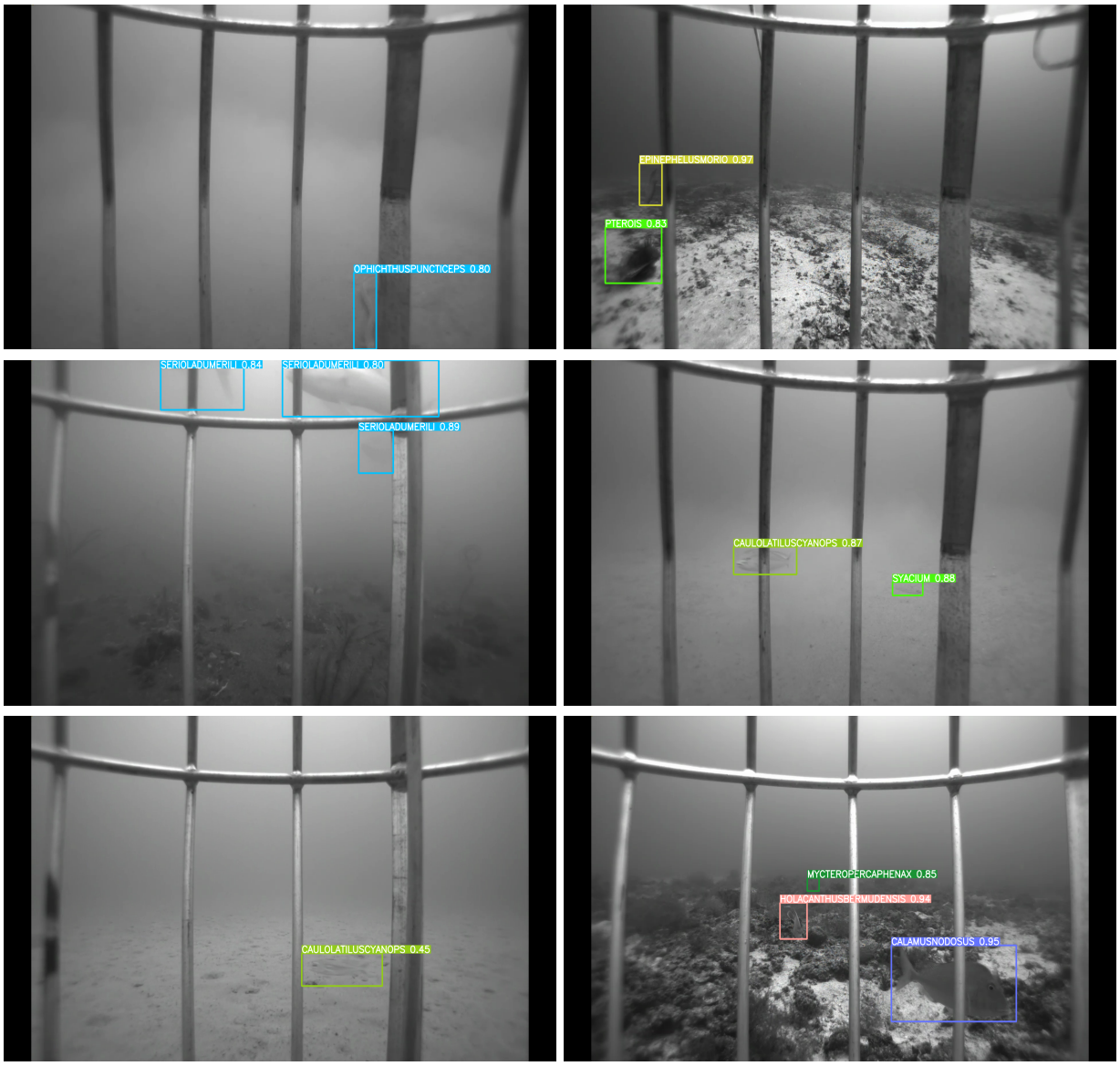

Figure 3. Detection maps of YOLOv5l on Pascagoula data (SEAMAPD21).

[2] Cabreira, A. G., Tripode, M., and Madirolas, A., "Artificial neural networks for fish-species identification," *ICES Journal of Marine Science* **66**(6), 1119–1129 (2009).

[3] Churnside, J. H., Wells, R., Boswell, K. M., Quinlan, J. A., Marchbanks, R. D., McCarty, B. J., and Sutton, T. T., "Surveying the distribution and abundance of flying fishes and other epipelagics in the northern gulf of mexico using airborne lidar," *Bulletin of Marine Science* **93**(2), 591–609 (2017).

[4] Jalali, M. A., Ierodiaconou, D., Monk, J., Gorfine, H., and Rattray, A., "Predictive mapping of abalone fishing grounds using remotely-sensed lidar and commercial catch data," *Fisheries research* **169**, 26–36 (2015).

[5] Boswell, K. M., Wilson, M. P., and Cowan Jr, J. H., "A semiautomated approach to estimating fish size, abundance, and behavior from dual-frequency identification sonar (didson) data," *North American Journal of Fisheries Management* **28**(3), 799–807 (2008).

[6] Villon, S., Chaumont, M., Subsol, G., Villéger, S., Claverie, T., and Mouillot, D., "Coral reef fish detection and recognition in underwater videos by supervised machine learning: Comparison between deep learning

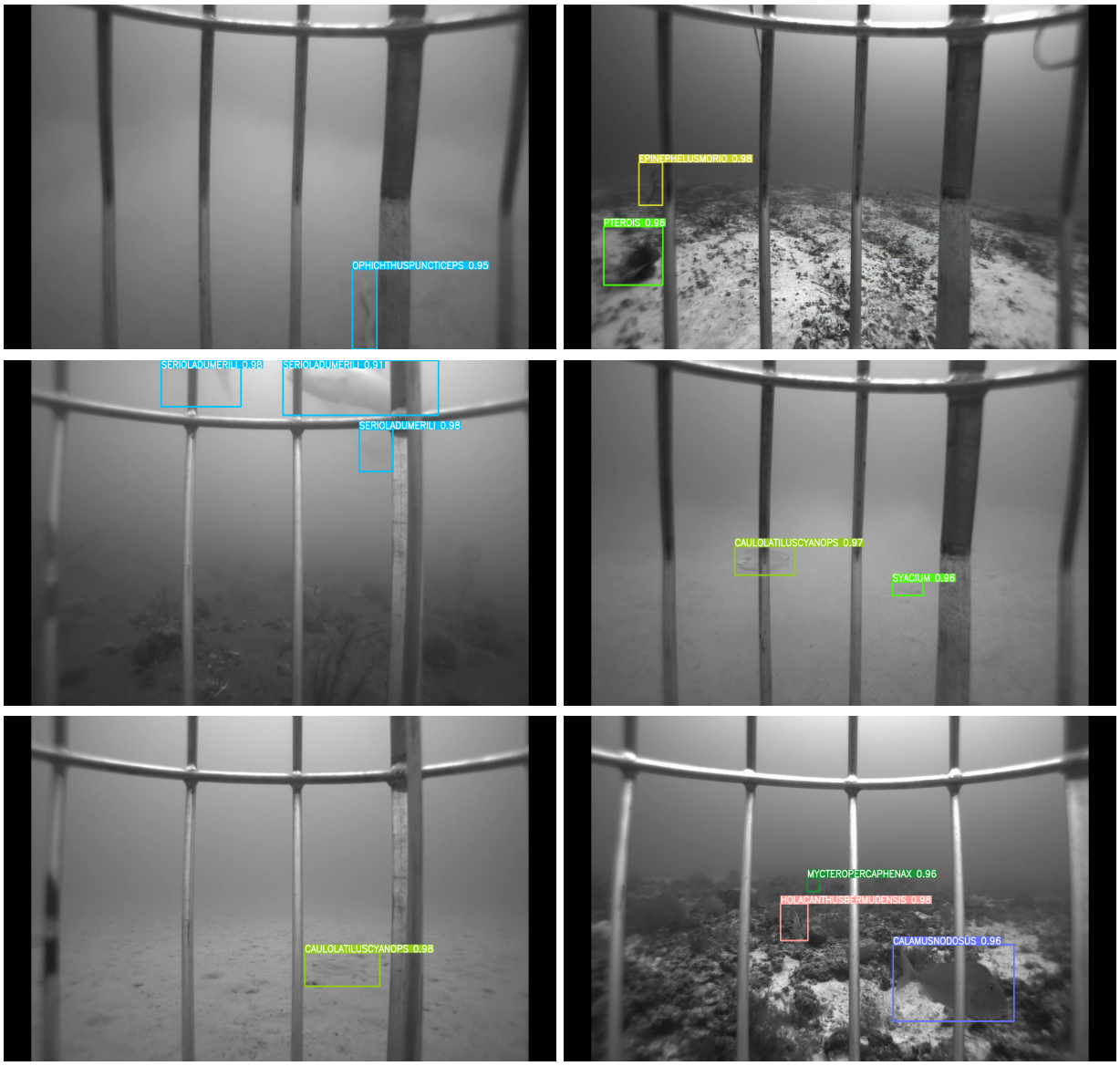

Figure 4. Detection maps of proposed YOLOv5enh on Pascagoula data (SEAMAPD21).

and hog+ svm methods," in [*International Conference on Advanced Concepts for Intelligent Vision Systems*], 160–171, Springer (2016).

[7] Bicknell, A. W., Godley, B. J., Sheehan, E. V., Votier, S. C., and Witt, M. J., "Camera technology for monitoring marine biodiversity and human impact," *Frontiers in Ecology and the Environment* **14**(8), 424–432 (2016).

[8] Morshed, M., Nabi, M., and Monzur, N., "Frame by frame digital video denoising using multiplicative noise model," *Int. J. Technol. Enhanc. Emerg. Eng. Res* **2**, 1–6 (2014).

[9] Shortis, M. and Abdo, E. H. D., "A review of underwater stereo-image measurement for marine biology and ecology applications," *Oceanography and marine biology* , 269–304 (2016).

[10] Alaba, S. Y., Nabi, M., Shah, C., Prior, J., Campbell, M. D., Wallace, F., Ball, J. E., and Moorhead, R., "Class-aware fish species recognition using deep learning for an imbalanced dataset," *Sensors* **22**(21), 8268 (2022).

[11] Nabi, M., Senyurek, V., Gurbuz, A. C., and Kurum, M., "A deep learning-based soil moisture estimation in conus region using cygnss delay doppler maps," in [*IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*], 6177–6180, IEEE (2022).

[12] Islam, F., Nabi, M., and Ball, J. E., "Off-road detection analysis for autonomous ground vehicles: a review," *Sensors* **22**(21), 8463 (2022).

[13] Alaba, S. Y. and Ball, J. E., "Deep learning-based image 3d object detection for autonomous driving," *IEEE Sensors Journal* (2023).

[14] Jäger, J., Rodner, E., Denzler, J., Wolff, V., and Fricke-Neuderth, K., "Seaclef 2016: Object proposal classification for fish detection in underwater videos.," in [*CLEF (working notes)*], 481–489 (2016).

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

[16] Wang, C.-Y., Bochkovskiy, A., and Liao, H.-Y. M., "Scaled-YOLOv4: Scaling cross stage partial network," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*], 13029–13038 (June 2021).

[17] Jocher, G., Stoken, A., Borovec, J., NanoCode012, Chaurasia, A., TaoXie, Changyu, L., V, A., Laughing, tkianai, yxNONG, Hogan, A., lorenzomammana, AlexWang1900, Hajek, J., Diaconu, L., Marc, Kwon, Y., oleg, wanghaoyang0106, Defretin, Y., Lohia, A., ml5ah, Milanko, B., Fineran, B., Khromov, D., Yiwei, D., Doug, Durgesh, and Ingham, F., "ultralytics/yolov5: v5.0 - YOLOv5-P6 1280 models, AWS, Supervise.ly and YouTube integrations," (Apr. 2021).

[18] Jung, H.-K. and Choi, G.-S., "Improved yolov5: Efficient object detection using drone images under various conditions," *Applied Sciences* **12**(14) (2022).

[19] Yang, X., Zhang, S., Liu, J., Gao, Q., Dong, S., and Zhou, C., "Deep learning for smart fish farming: applications, opportunities and challenges," *Reviews in Aquaculture* **13**(1), 66–90 (2021).

[20] Xu, W. and Matzner, S., "Underwater fish detection using deep learning for water power applications," in [*2018 International conference on computational science and computational intelligence (CSCI)*], 313–318, IEEE (2018).

[21] Liu, S., Li, X., Gao, M., Cai, Y., Nian, R., Li, P., Yan, T., and Lendasse, A., "Embedded online fish detection and tracking system via yolov3 and parallel correlation filter," in [*OCEANS 2018 MTS/IEEE Charleston*], 1–6, IEEE (2018).

[22] Boulais, O., Alaba, S. Y., Ball, J. E., Campbell, M., Iftekhar, A. T., Moorehead, R., Primrose, J., Prior, J., Wallace, F., Yu, H., et al., "Seamapd21: A large-scale reef fish dataset for fine-grained categorization," in [*Proceedings of the FGVC8: The Eight Workshop on Fine-Grained Visual Categorization, Online*], **25** (2021).

[23] Sung, M., Yu, S.-C., and Girdhar, Y., "Vision based real-time fish detection using convolutional neural network," in [*OCEANS 2017-Aberdeen*], 1–6, IEEE (2017).

[24] Jalal, A., Salman, A., Mian, A., Shortis, M., and Shafait, F., "Fish detection and species classification in underwater environments using deep learning with temporal information," *Ecological Informatics* **57**, 101088 (2020).

[25] Cui, Y., Jia, M., Lin, T.-Y., Song, Y., and Belongie, S., "Class-balanced loss based on effective number of samples," in [*2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*], 9260–9269 (2019).

[26] Boulais, O., Alaba, S. Y., Ball, J. E., Campbell, M., Iftekhar, A. T., Moorehead, R., Primrose, J., Prior, J., Wallace, F., Yu, H., et al., "Seamapd21: a large-scale reef fish dataset for fine-grained categorization," *The Eight Workshop on Fine-Grained Visual Categorization* (2021).