MI-AFR: Multiple instance active learning based approach for fish species recognition in underwater environments

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

[a]Northern Gulf Institute, Mississippi State University, Starkville, MS 39759, USA
[b]Department of Electrical and Computer Engineering, James Worth Bagley College of Engineering, Mississippi State University, Starkville, 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

Video surveys are commonly used to monitor the abundance and distribution of managed species to support management. However, considerable effort, time, and cost are required for human review and automated fish species recognition provides an effective solution to remove the bottleneck of post-processing. Implementing fish species detection techniques for underwater imagery is a challenging task. In this work, we present the Multiple Instance Active-learning for Fish-species Recognition (MI-AFR), which is formulated as an object detection-based approach to perform localization and classification of fish species. It can select the most informative fish images from unlabeled sets by estimating the uncertainty of unlabeled images by using adversarial classifiers trained on labeled sets. Moreover, we have analyzed the improved performance of MI-AFR by considering different backbone networks as a trade-off between speed and accuracy. For experiments, we have used the fine-grained and large-scale reef fish dataset obtained from the Gulf of Mexico – the Southeast Area Monitoring and Assessment Program Dataset 2021 (SEAMAPD21). The experimental results illustrate that the superiority of the proposed method can establish a solid foundation for active learning in fish species recognition, especially with a small number of labeled sets.

Keywords: active learning, fish species recognition, object detection, MI-AFR, underwater

1. INTRODUCTION

NOAA's Southeast Fisheries Science Center has been conducting video surveys in the Gulf of Mexico for more than 30 years. These surveys provide information that scientists use to estimate population status of ecologically and economically important species over long-time frames. These surveys are vital to maintain the health of the snapper-grouper fisheries of the Gulf of Mexico. While video surveys have a growing interest in NOAA Fisheries and the Southeast Fisheries Science Center, processing the video remains a slow and expensive manual process.

The identification and numeration of fish species is a crucial aspect of processing video based fish surveys. It is essential to accurately and efficiently recognize fish species in order to identify all species, supervise ecosystems, and build effective controlling systems.[1,2] Determination of population status relies on survey information precise identification of fish species is particularly important, especially when their productivity may be at risk. Furthermore, traditional human-based methods of identification is labor intensive and not timely, which can greatly delay management decisions. This creates challenges for sustainably managing the fishing production, observing national fisheries, assessing fish populations, and identifying at risk species of fish. Therefore, using deep learning (DL) based models to identify fish species[3,4] is a promising approach that could reduce costs and time and improve identification accuracy.

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, 125430N · © 2023 SPIE
0277-786X · doi: 10.1117/12.2663404

Implementing a machine vision approach is one possible solution for replacing the manual system. Several approaches, such as sonar,[5] lidar,[6] and RGB[7] imaging, are used to identify fish, but RGB imaging is preferred due to its ability to detect species based on their shape, texture, and color. Generally, images are extracted from a video sequence and processed further.[8,9] This method is cost-effective, lightweight, and does not harm fish habitats. Various camera-based technologies have been utilized to monitor fish stocks and sustainability in the ecosystem.[10,11] Numerous DL-based methods for identifying and categorizing objects can be used to gather information on marine ecology. However, the underwater environment presents several challenges due to low light conditions and limited image resolution, making it harder to distinguish fish from the background. Additionally, fish movement and density result in images of fish in various poses and introduce occlusion issues, making underwater fish species localization and classification challenging.

DL has been extensively utilized in computer vision to tackle various problems such as detection, localization, estimation, and classification.[12–14] However, its application in agriculture and marine ecosystems is limited. Various machine learning (ML) and DL techniques have been introduced to classify fish species. For instance, Huang *et al.*[15] used hierarchical features and support vector machines (SVM) to classify fish, while Jager *et al.*[16] utilized the AlexNet deep-learning model for feature extraction and multiclass SVM for classification. Zhuang *et al.*[17] classified fish data using pre- and post-processing with an advanced DL approach.

However, the majority of existing DL techniques have been developed and trained on simple classification datasets that contained a single fish in each image. This does not reflect the reality of marine environments, where multiple fish are often present in an individual image, making it challenging to apply simple classification networks. To address this issue, the Southeast Area Monitoring and Assessment Program Dataset 2021 (SEAMAPD21)[11] was developed to contain several fish in a single image, which is more representative of natural habitats.

Instance-level active learning methods have also been introduced, specifically for object detection, that choose the most informative samples for detector training by observing instance-level uncertainty. Yuan et al.[18] introduced multiple instance active learning for object detection (MI-AOD) for the widely used PASCAL VOC and MS COCO public datasets."

The MI-AOD[19] creates an instance uncertainty learning module that predicts instance uncertainty of the unlabeled set using the difference between two adversarial instance classifiers trained on the labeled set. Inspired by this, we have extended the work for fish species detection in underwater environments. The MI-AOD approach can be adapted for fish detection and classification by using fish images as input data and training the detection network with fish species labels. The instance uncertainty learning (IUL) module can be used to learn the uncertainty of the fish instances in the input images. This is achieved by training two instance classifiers, each of which predicts the likelihood of the fish being present in the image. The discrepancy between the predictions of the two classifiers is used to estimate the instance-level uncertainty of each fish instance in the image. The image-level uncertainty is then obtained by incorporating a multiple instance learning (MIL) module that treats each image as an instance bag and performs instance uncertainty re-weighting (IUR) by evaluating instance appearance consistency across images.

In this paper, we have introduced Multiple Instances Active-learning for Fish-species recognition (MI-AFR) in underwater environments. To use this approach for fish detection and classification, one would first need to collect a large dataset of fish images, annotated with species labels. The dataset would be split into labeled and unlabeled sets, with the labeled set used for training the detection network and the unlabeled set used for selecting informative images for detector training using the MI-AFR approach. The MI-AFR approach would be applied alliteratively to the unlabeled set to select the most informative images for training the detection network. A small set of initial labeled images along with the large unlabeled set of images can be used for the MI-AFR. The resulting network could then be used for fish detection and classification on new, unseen images.

In MI-AFR, different backbone networks, such as ResNet-50[20] and VGG-16[21] are implemented with SSD[22] detection head for active learning. Experiments are conducted on the SEAMAPD21. Experimental results illustrate the superiority of MI-AFR with the VGG512 backbone network, in terms of mean average precision (mAP), on such a challenging dataset.

2. RELATED WORK

The classification of fish in computer vision is a well-researched problem. which involves identifying and categorizing fish species based on their resemblance to representative specimen images. In the past, hand-crafted feature-generation techniques have been used for identifying fish,[23] but these methods have limitations, such as low accuracy and inability to scale with data. Deep learning (DL) approaches outperform these shallow learning methods due to their deep layer architecture and extensive data support. Researchers have proposed different DL-based methods for fish classification,[24,25] including fish detection and recognition in underwater videos.[26,27] With the help of three incredibly diverse datasets captured at actual water power sites, the You Only Look Once (YOLO) deep learning[28] model was trained to identify fish in underwater video. The mAP score obtained after evaluating samples from all three datasets was 0.5392. Without using pre-filtering, Jalal *et al.*[29] proposed a two-stage deep-learning method to detect and classify temperate fishes. The initial step was to localize each fish in an image regardless of their species and sex. For this case, they employed the YOLO object detection method. In the next step, they adopted a Convolutional Neural Network (CNN) with the Squeeze-and-Excitation (SE) structure to classify each single fish in the image without using any filtering. Transfer learning was used to overcome the limited number of temperate fish training samples and increase the classification accuracy. In the review article,[30] a comprehensive review based on computer vision model for fish detection under unique application scenarios such as high noise, illumination variations, low contrast, fish deformation, frequent occlusion, and dynamic background is presented. They presented image acquisition based on 2D and 3D systems. Additionally, many fish detection techniques were categorized based on their appearance, motion, and algorithms. They also discussed applications of fish detection and publicly available open-source datasets. Alshdaifat *et al.*[31] presented a novel framework for fish instance segmentation in underwater videos. The proposed approach is made up of four primary stage for better recognition: 1) pre-processing phases to eliminate outside factors in the videos for better fish identification in underwater videos, 2) deep learning approach implementation for improved fish detection, 3) enhanced detection of multiple fish based on the Region Proposal Network (RPN) architecture, and 4) application of a dynamic instance segmentation technique.

The major challenge in the SEAMAPD21 dataset is that it is underwater data with low light conditions and various levels of turbidity. Sometimes it is difficult for a human to find the fish in the images. Another big challenge for this huge dataset is annotating/labeling the dataset. It requires time and effort to manually annotate the fish. Additionally, the dataset has a class-imbalance problem which means some classes have a higher number of samples and some classes have fewer samples. In order to solve the problem, Alaba *et al.*[4] proposed a class-aware loss method that takes into account the inverse of the number of samples in each class to solve the class imbalance problem. They also incorporate the class-balanced loss into the object detection problem to re-weight the classification and localization losses. The proposed method is evaluated on the SEAMAPD21. Two popular feature extraction networks are utilized such as MobileNetv3-large and VGG16, and the single-shot multibox detector (SSD) detection is as detection head for regression and classification task. The results show that the proposed method outperformed other state-of-the-art methods in terms of accuracy, speed, and robustness to class imbalance.

In this work, we present an active learning approach that considers multiple instances for fish species recognition. This approach helps to minimize the data annotation cost while keeping the detection and classification accuracy similar to typical object detection algorithms. We apply the MI-AOD method that can select the most informative instances from images to improve the detection performance of training.[18]

3. PROPOSED METHOD

3.1 MI-AFR: Multiple Instance Active-learning for Fish-species Recognition in Underwater Environments

MI-AFR consists of a module that can learn uncertainty to predict uncertainty in the image for unlabeled sets by training the labeled set in an adversarial manner.[18] As shown in figure 1, initial small labeled set of images H_L^0 with corresponding labels Y_L^0 will be used to select the most informative image samples H_S^0 from an unlabeled set H_U^0 such that $H_L^1 = H_L^0 + H_S^0$. For active learning, model can be trained and samples can be selected by repeating the cycles such that the labeled set attains the budget of annotation. Each image can have large

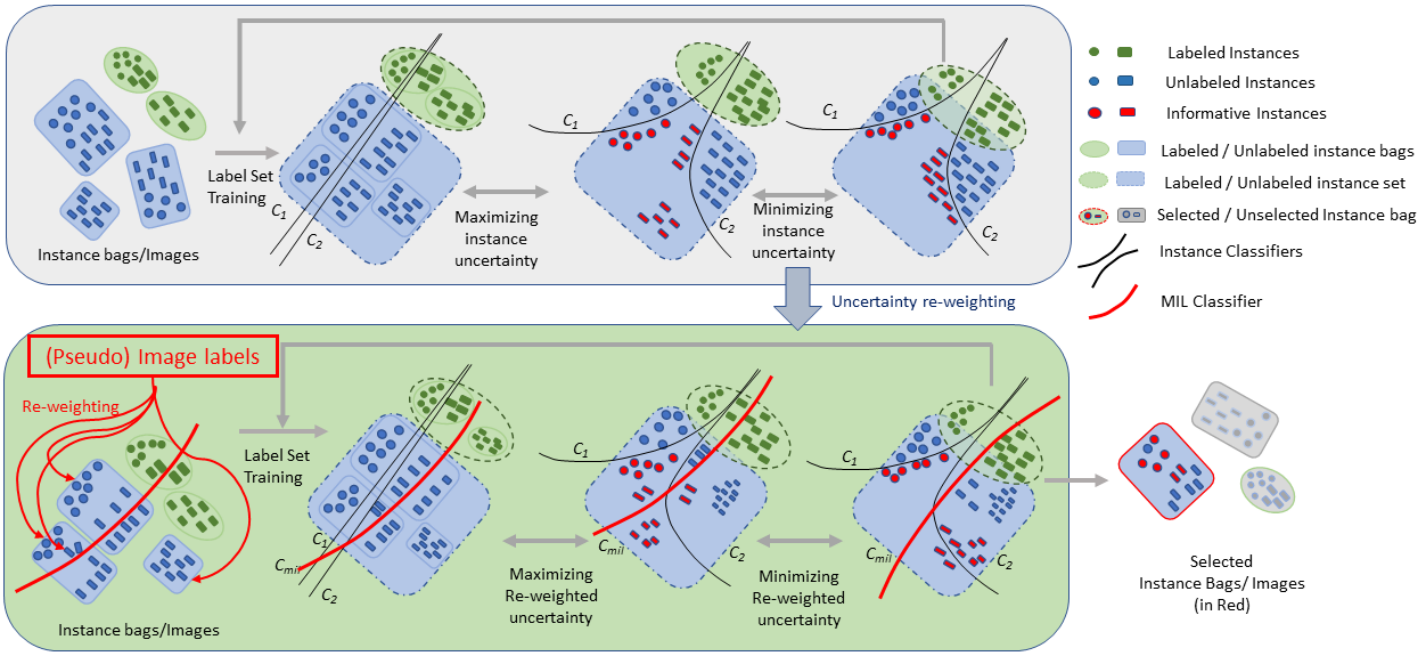

Figure 1. Multiple Instance Active-learning for Fish-species Recognition (MI-AFR)

number of instances that can be used to learn uncertainty of instances in unlabeled set. By aligning the labeled set and unlabeled set, MI-AFR can highlight most informative instances from unlabeled set for the learning of uncertainty. Moreover, MI-AFR can incorporate multiple instance learning (MIL) classifier in both labeled set and unlabeled set for calculating the uncertainty of image with the re-weighting of uncertainty in instances. MIL can utilize the classification loss to re-weight the uncertainty of instances and minimize the noisy instances by considering an instance bag for each image.

3.2 Learning Uncertainty of Instances

Using SSD head for the detection,[22] two classifiers (c_1 and c_2) are designed for discrepancy or instance uncertainty classification and a regressor (d_r) is formulated for the bounding box prediction. For object detection, each of image h can be denoted with various instances as $\{h_n, n = 1, ..., M\}$ for M number of instances in an image. For training the detection approach, detection loss can be optimized from the labeled set as:

$$det_l(h) = \sum_n (CE(\hat{y}_n^{c_1}, y_n^{cl}) + CE(\hat{y}_n^{c_2}, y_n^{cl}) + SmL1(\hat{y}_n^{d_r}, y_n^{lo})) \,, \tag{1}$$

where CE is the cross entropy loss function in classifiers, $SmL1$ is the smooth L1 regression loss function for bounding box,[32,33] $\hat{y}_n^{c_1}$=$c_1(h)$,$\hat{y}_n^{c_2}$=$c_2(h)$, and $\hat{y}_n^{d_r}$=$d_r(h)$ represents result of predictions for classifiers and localizer with y_n^{cl} and y_n^{lo} being the ground-truths for class and bounding box labels.

To reduce prediction discrepancy between small labeled set and large unlabeled set, instance uncertainty among instances can be maximized as:

$$I_{max} = \sum_{h \epsilon H_L} det_l(h) - \sum_{h \epsilon H_U} \gamma . dis_l(h) \,, \tag{2}$$

where dis_l=$\sum_n (\hat{y}_n^{c_1} - \hat{y}_n^{c_2})$ and γ is the regularization parameter

To get closer alignment among labeled and unlabeled samples, Instance uncertainty among instances can be minimized as:

$$I_{min} = \sum_{h \epsilon H_L} det_l(h) + \sum_{h \epsilon H_U} \gamma . dis_l(h) \,, \tag{3}$$

3.3 Re-weighting Uncertainty of Instances

The image classification loss,Im_{cls}, can be used to learn multiple instances as:

$$Im_{cls}(h) = -\sum_{ct}(y_{ct}^{cl} log \sum_{n} \hat{y}_{n,ct}^{cl} + (1 - y_{ct}^{cl}) log(1 - \sum_{h} \hat{y}_{n,ct}^{cl}))\,, \tag{4}$$

where $\hat{y}_{n,ct}^{cl}$ is the predicted classification score for multiple instance learning (MIL) classifier such that

$$\hat{y}_{n,ct}^{cl} = \frac{exp^{\hat{y}_{n,ct}^{c_{mil}}}}{\sum_{ct} exp^{\hat{y}_{n,ct}^{c_{mil}}}} \cdot \frac{exp^{(\hat{y}_{n,ct}^{c_1} + \hat{y}_{n,ct}^{c_2})/2}}{\sum_{ct} exp^{(\hat{y}_{n,ct}^{c_1} + \hat{y}_{n,ct}^{c_2})/2}}\,, \tag{5}$$

Following relation can be used to suppress small classification scores by estimating discrepancies among instances by updating Eq. (2).

$$max_{\widetilde{L}}(h) = \sum_{h \epsilon H_L} (det_l(h) + Im_{cls}(h)) - \sum_{h \epsilon H_U} \gamma . \widetilde{dis}_l(h)\,, \tag{6}$$

where $\widetilde{dis}_l(h)$ is the reweighted instance uncertainty.

Similarly, the following relation can be used to update Eq. (3).

$$min_{\widetilde{L}}(h) = \sum_{h \epsilon H_L} (det_l(h) + Im_{cls}(h)) + \sum_{h \epsilon H_U} \gamma . \widetilde{dis}_l(h)\,, \tag{7}$$

3.4 Backbone Networks and Detection Head

In active learning for fishery with MI-AFR, we have introduced different backbone networks along with varying input resolution for fish analysis. As a feature extraction method, different backbone networks, such as VGG,[21] EfficinetNet,[34] MobileNet,[35] and ResNet[20] have been widely used. The VGG can provide better detection performance although it may require more parameters. We have also implemented ResNet50 as a backbone network in MI-AFR to get competitive performance as VGG with relatively less computational cost.

For detection, the SSD detection head[22] is utilized for regression of bounding box and prediction of fish species. At varying scales and aspect ratios, SSD can detect the head at four object locations.[22,36] With this property, large and small scale objects can be detected properly. To reduce the Intersection Over Union (IOU) value, the non maximal suppression is used. IOU can estimate the overlap between true prediction and original groundtruth. A higher value IOU can give more accurate prediction. In addition, for both VGG and ResNet backbone networks, we have used the SSD detection head. Input resolution is changed from 300×300 to 512×512 in both networks for detecting performance at different scales.

3.5 Dataset

The large-scale reef fish SEAMAPD21[11] dataset has been used for the experiment. In total, it has 130 distinct classes of fish species in an underwater environments with 28,319 images. However, some of the species are very small in number and the model is influenced by samples with more species per class. For species with lower occurrence, it is difficult to obtain a sufficient train-test ratio required for an active learning based technique. In order to avoid intraclass similarity among different fish species and to reduce the problem of less samples of rare species, we have tried to reduce the redundancy due to similar species and select the species with higher number of occurrences. For this purpose, we have selected 51 distinct class species of fish from the SEAMAPD21 data. 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.

4. EXPERIMENTAL RESULTS

4.1 Experimental Settings

We used the VGG-16[21] and ResNet-50[20] backbone networks with the SSD[22] detection head. In addition, the input resolution is varied from 300×300 to 512×512 to observe the improved performance at varying scales in different backbone networks. For the active learning based technique called MI-AFR on the SEAMAP21 data, initially 5 (%) of the labeled samples were selected from the training set. In each active learning cycle, 2.5 (%) of the fish images were selected from the remaining unlabeled pool until the labeled pool reached 17.5 (%) of the training set. Then, detection results in terms of mAP was evaluated on test set. We selected IOU of 0.45 to estimate mAP over 51 classes of fish species. In each active learning cycle, a learning rate of 0.001 and a mini-batch size of 4 were used.Parameter γ mentioned in Eq. (2), (3), (6), and (7) was set to 0.5.

4.2 Performance Analysis

As shown in Table 1, performance in terms of mAP estimated by averaging the average precision of 51 different fish species can be observed for the VGG-16 and ResNet backbone networks with the SSD detection head. For the input resolution of 300×300, the VGG-16 backbone network with the SSD head outperforms the ResNet-50 with the SSD head by 7.4%, 2.3%, 3.3%, and 1.2% while using 5%, 7.5%, 10%, and 12.5% labeled samples. However, for a higher number of active learning cycles when the labeled set becomes 15% and 17.5 %, the ResNet-50 with the SSD detection head outperforms the VGG-16 with the SSD detection head by 0.9% and 2.87% respectively. Similarly, for the input resolution of 512×512, the VGG-16 backbone network with the SSD detection head outperforms the ResNet-50 backbone network with the SSD detection head by 8.08%, 11.4%, 3.73%, 5.67%, 4.56%, and 2.8% while using 5%, 7.5%, 10%, 12.5%, 15%, and 17.5% labeled samples respectively. Moreover, the VGG-16 backbone with the SSD detection head having the input resolution of 512×512 shows better performance than all other backbone networks and input resolutions with labeled samples from 5% to 17.5%.

Table 1. mAP (%) for MI-AFR on Pascagoula data (SEAMAPD21) for varying backbone networks

Model	Backbone	mAP(%) on ratio (%) of labeled samples					
		5	7.5	10	12.5	15	17.5
SSD300	ResNet-50	37.90	44.10	45.70	49.80	52.00	55.17
	VGG-16	45.30	46.40	49.00	51.00	51.10	52.30
SSD512	ResNet-50	46.90	47.40	56.37	58.83	61.14	63.00
	VGG-16	54.98	58.80	60.10	64.50	**65.70**	**65.80**

Figure 2 shows a graphical representation of the performance for the varying backbone networks and input resolutions in Table 1.

Table 2 shows the number of parameters and inference time in terms of frames per second (FPS) for different backbone networks and input resolutions. For fair comparison, all the experiments were conducted on a single node of a NVIDIA A100-SXM GPU. The VGG-16 with SSD512 has a higher number of parameters and inference time in terms of FPS compared to VGG-16 with SSD300, ResNet-50 with SSD300, and ResNet-50 with SSD512.

Table 2. Inference time of MI-AFR on Pascagoula data (SEAMAPD21) for different backbone networks

Model	Backbone	Parameters	Frames per second (FPS)
SSD300	ResNet-50	42.94M	36
	VGG-16	44.33M	18
SSD512	ResNet-50	43.67M	28
	VGG-16	47.1M	15

For qualitative analysis, detection results on different backbone networks with varying input resolutions are shown in Figures 3- 6. Figure 3 shows qualitative outputs for MI-AFR with ResNet50 backbone network

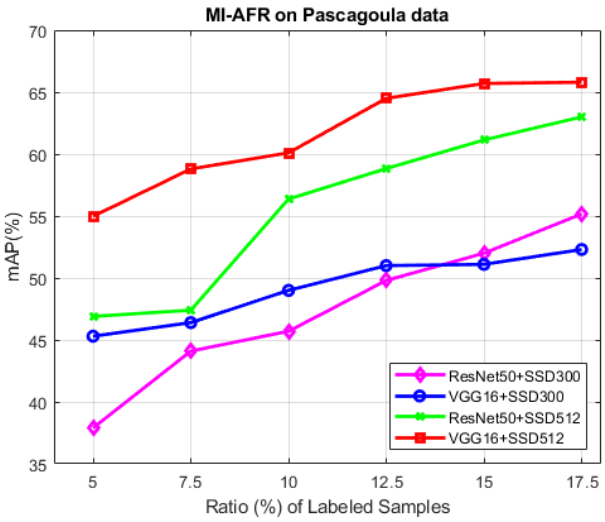

Figure 2. Performance comparison of MI-AFR on Pascagoula data (SEAMAPD21) for different backbone networks

and SSD300 detection head on SEAMAPD21, Figure 4 demonstrates qualitative outputs for MI-AFR with VGG-16 backbone network and SSD300 detection head on SEAMAPD21, Figure 5 lists qualitative outputs for MI-AFR with ResNet-50 backbone network and SSD512 detection head on SEAMAPD21, and Figure 6 shows qualitative outputs for MI-AFR with VGG-16 backbone network and SSD512 detection head on SEAMAPD21. It can be observed that the VGG-16 with SSD512 (Figure 6) shows better detection results for fish species in an underwater environments compared to VGG-16 with SSD300 (Figure 4), ResNet-50 with SSD300 (Figure 3), and ResNet-50 with SSD512 (Figure 5).

5. CONCLUSIONS

In conclusion, we utilized multiple instance active-learning for fish species localization and detection in the SEAMAPD21 dataset. Different backbones such as VGG16 and ResNET-50 were applied with multiple resolutions. We have seen that the VGG16 with the SSD512 model provides the best perfommace compared to the other applied models. However, this model is sufficiently large, requiring more computational power and thus inhibiting real-time processing. This is the trade-off between accuracy and computational power.

The proposed MI-AFR approach has several advantages. This method addresses the issue of limited labeled data by selecting informative images from the unlabeled set to enhance the training of object detection models. It incorporates both discrepancy learning and MIL to reduce the distribution bias and improve the generalization ability of the model. It leverages the instance-level uncertainty to perform instance uncertainty re-weighting and highlight the representative instances while suppressing the noisy ones. It can be fine-tuned with labeled data to improve the model's performance and further reduce false positives.

However, there are also some potential disadvantages to consider. The MI-AFR approach requires a pre-trained detection network as a base model, which may not be readily available or may require significant computational resources to train from scratch. The iterative process of instance uncertainty learning and re-weighting may require a significant amount of time and resources to select the most informative images for detector training. The performance of MI-AFR may be limited by the quality of the initial base model, the amount and quality of the labeled and unlabeled data, and the specific fish species being detected. Overall, the MI-AFR approach has the potential to improve the performance of fish detection and classification by leveraging unlabeled data and reducing the distribution bias, but it also requires careful consideration of the limitations and potential challenges in the implementation.

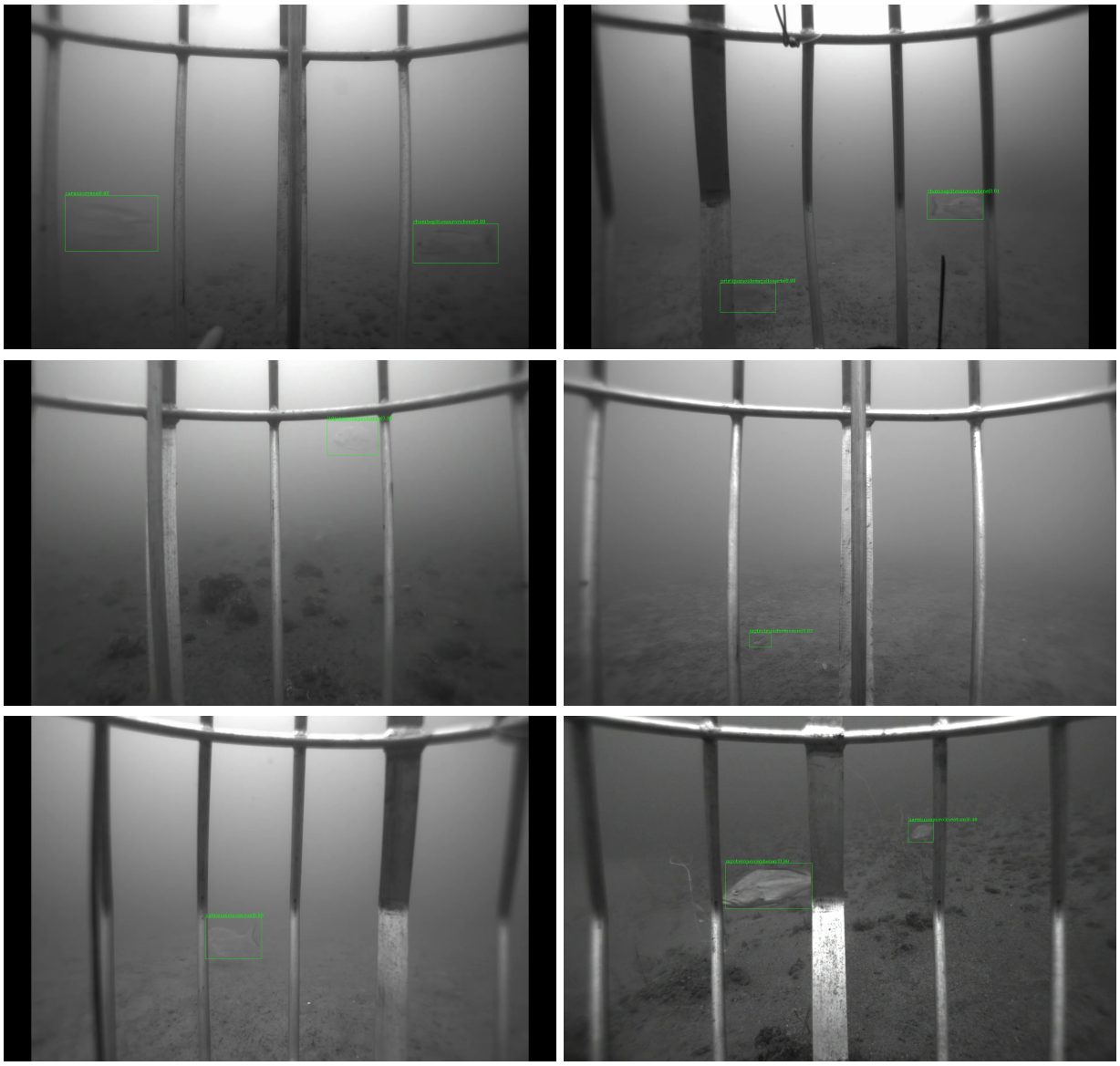

Figure 3. Detection images of MI-AFR with ResNet50 backbone network and SSD300 detection head on Pascagoula data (SEAMAPD21)

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 Atmospheric 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).

[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).

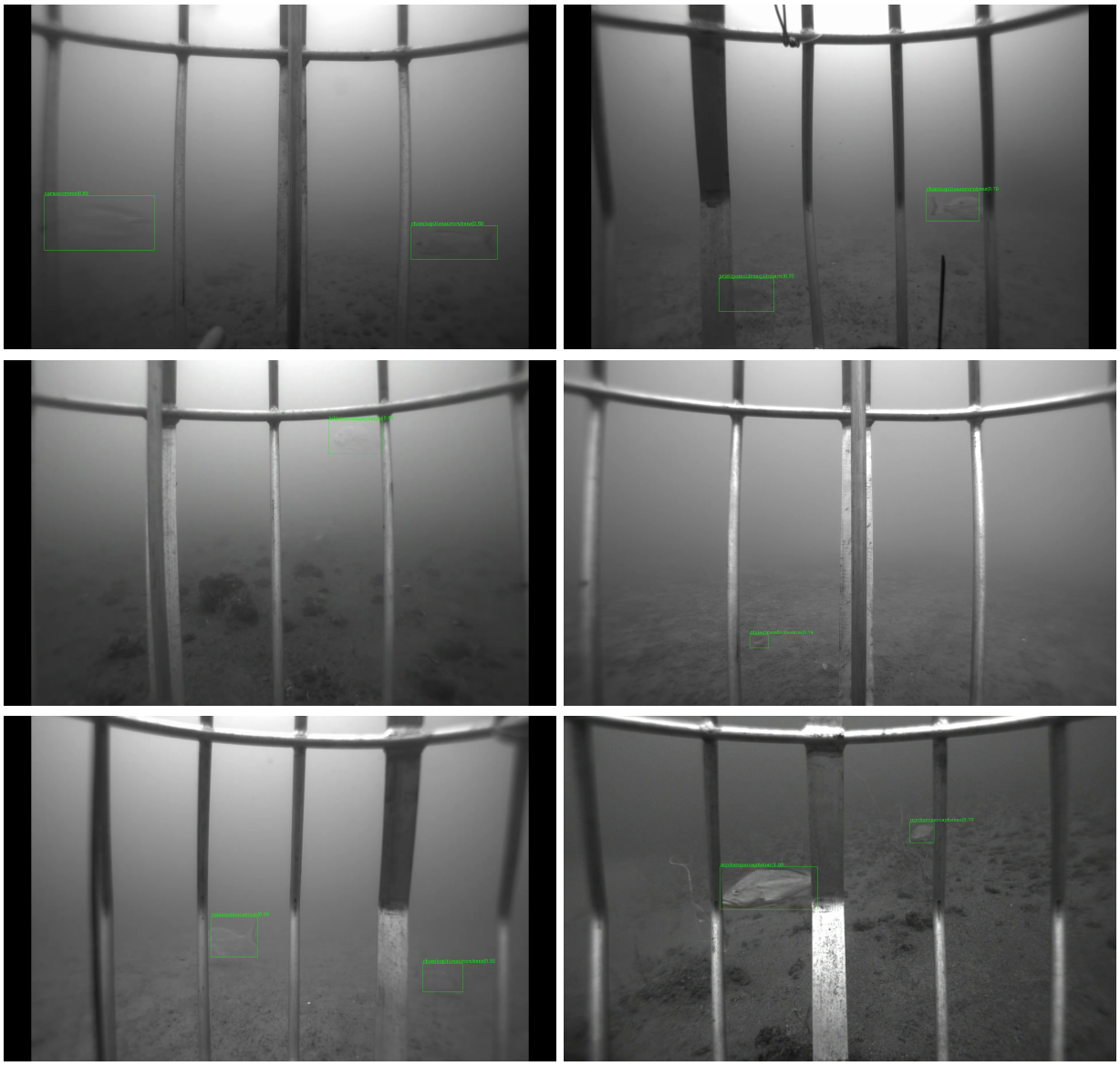

Figure 4. Detection images of MI-AFR with VGG-16 backbone network and SSD300 detection head on Pascagoula data (SEAMAPD21)

[3] 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).

[4] 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).

[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] 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).

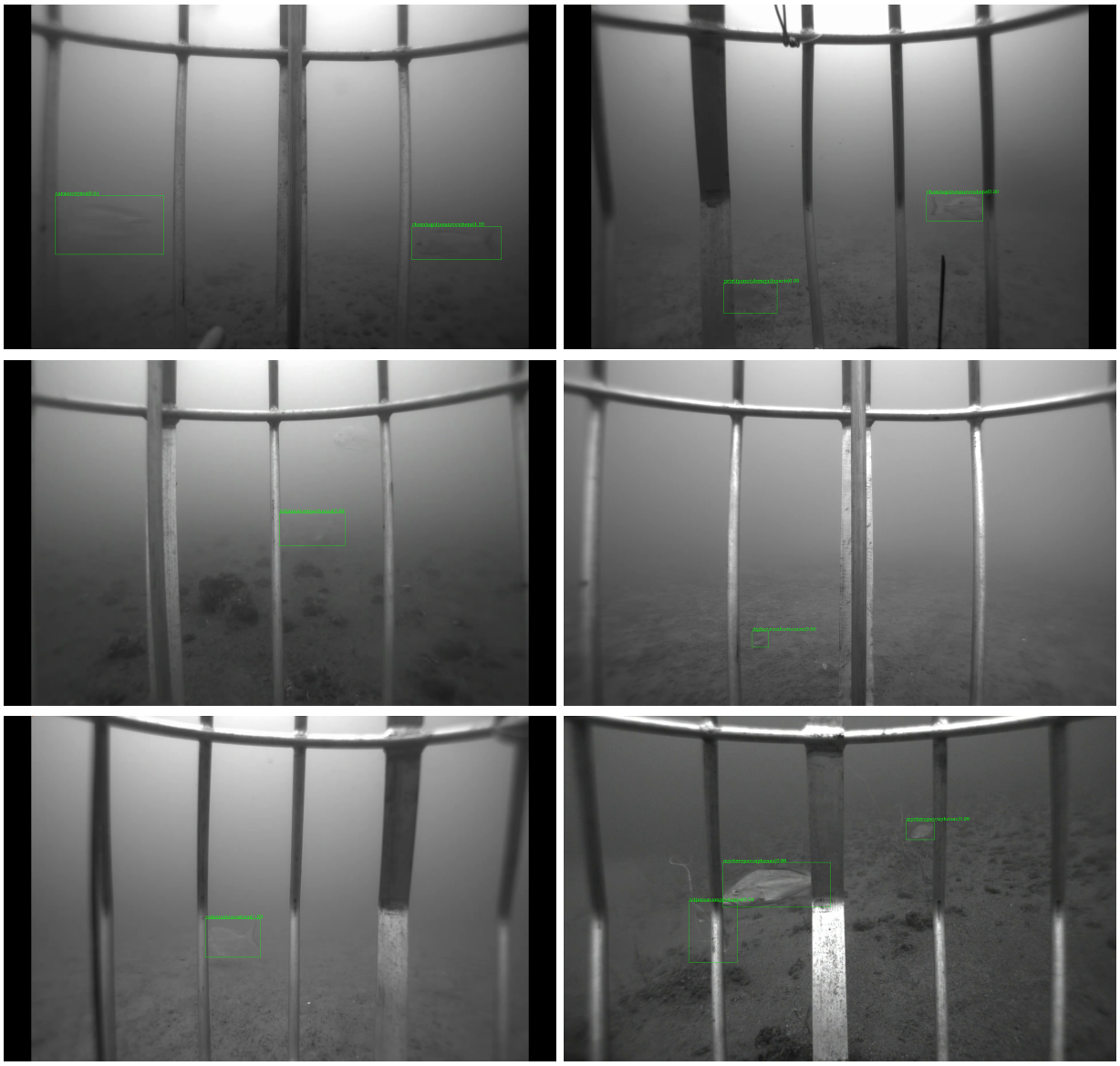

Figure 5. Detection images of MI-AFR with ResNet50 backbone network and SSD512 detection head on Pascagoula data (SEAMAPD21)

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

[8] Sirohey, S., Rosenfeld, A., and Duric, Z., "A method of detecting and tracking irises and eyelids in video," *Pattern recognition* **35**(6), 1389–1401 (2002).

[9] 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).

[10] Gilby, B. L., Olds, A. D., Connolly, R. M., Yabsley, N. A., Maxwell, P. S., Tibbetts, I. R., Schoeman, D. S., and Schlacher, T. A., "Umbrellas can work under water: Using threatened species as indicator and management surrogates can improve coastal conservation," *Estuarine, Coastal and Shelf Science* **199**, 132–140 (2017).

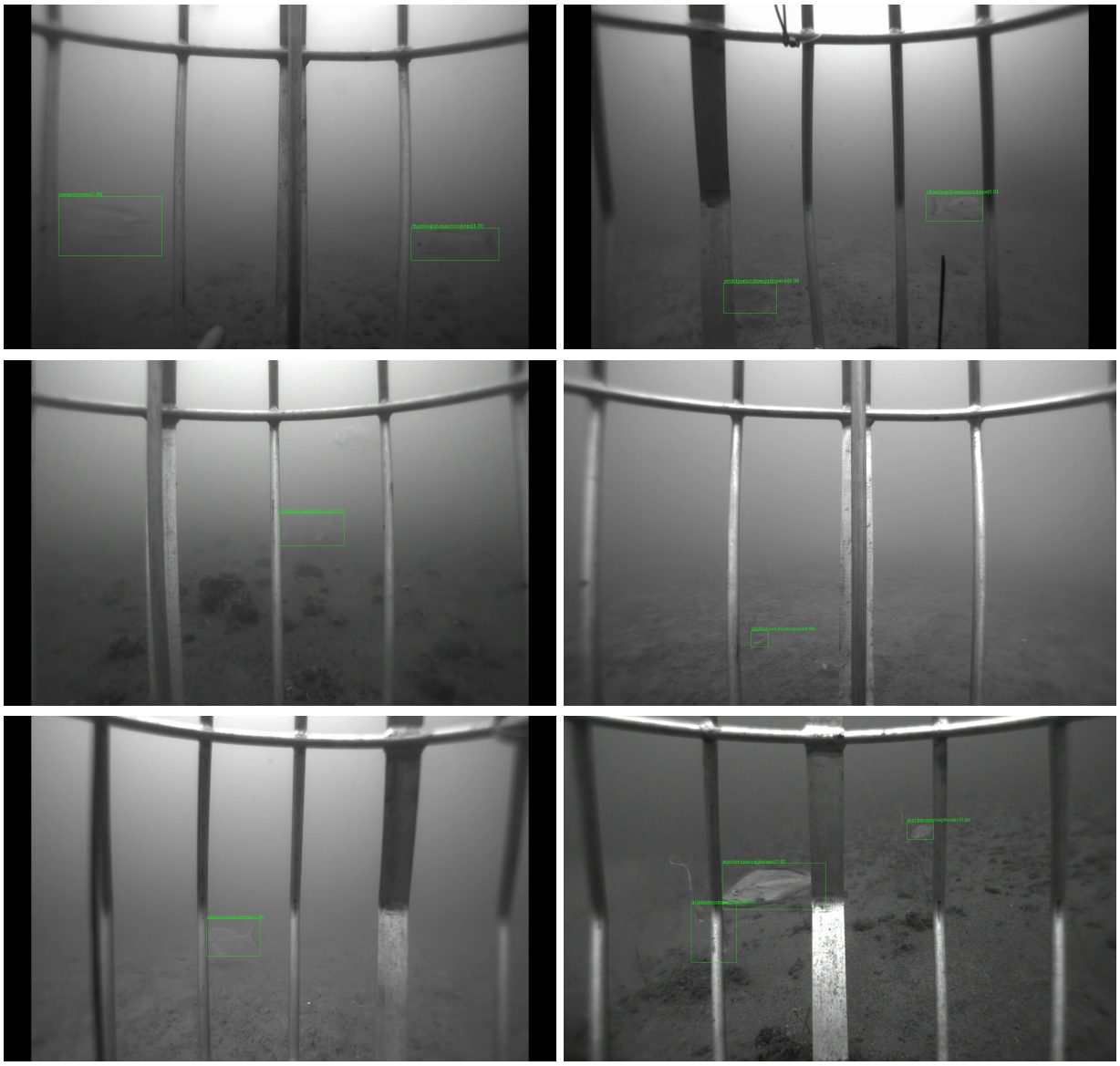

Figure 6. Detection images of MI-AFR with VGG-16 backbone network and SSD512 detection head on Pascagoula data (SEAMAPD21)

[11] 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).

[12] Zhao, M., Chang, C. H., Xie, W., Xie, Z., and Hu, J., "Cloud shape classification system based on multi-channel cnn and improved fdm," *IEEE Access* **8**, 44111–44124 (2020).

[13] Nabi, M., Senyurek, V., Gurbuz, A. C., and Kurum, M., "Deep learning-based soil moisture retrieval in conus using cygnss delay–doppler maps," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **15**, 6867–6881 (2022).

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

[15] Huang, P. X., Boom, B. J., and Fisher, R. B., "Underwater live fish recognition using a balance-guaranteed optimized tree," in [*Asian Conference on Computer Vision*], 422–433, Springer (2012).

[16] 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).

[17] Zhuang, P., Xing, L., Liu, Y., Guo, S., and Qiao, Y., "Marine animal detection and recognition with advanced deep learning models," in [*CLEF*], (2017).

[18] Yuan, T., Wan, F., Fu, M., Liu, J., Xu, S., Ji, X., and Ye, Q., "Multiple instance active learning for object detection," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*], 5330–5339 (June 2021).

[19] Yuan, T., Wan, F., Fu, M., Liu, J., Xu, S., Ji, X., and Ye, Q., "Multiple instance active learning for object detection," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 5330–5339 (2021).

[20] He, K., Zhang, X., Ren, S., and Sun, J., "Deep residual learning for image recognition," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 770–778 (2016).

[21] Simonyan, K. and Zisserman, A., "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556* (2014).

[22] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C., "Ssd: Single shot multibox detector," in [*European conference on computer vision*], 21–37, Springer (2016).

[23] Ravanbakhsh, M., Shortis, M., Shaifat, F., Mian, A. S., Harvey, E., and Seager, J., "An application of shape-based level sets to fish detection in underwater images.," in [*GSR*], (2014).

[24] Chen, G., Sun, P., and Shang, Y., "Automatic fish classification system using deep learning," in [*2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI)*], 24–29, IEEE (2017).

[25] Chhabra, H. S., Srivastava, A. K., and Nijhawan, R., "A hybrid deep learning approach for automatic fish classification," in [*Proceedings of ICETIT 2019: Emerging Trends in Information Technology*], 427–436, Springer (2020).

[26] Spampinato, C., Chen-Burger, Y.-H., Nadarajan, G., and Fisher, R. B., "Detecting, tracking and counting fish in low quality unconstrained underwater videos.," *VISAPP (2)* **2008**(514-519), 1 (2008).

[27] Mandal, R., Connolly, R. M., Schlacher, T. A., and Stantic, B., "Assessing fish abundance from underwater video using deep neural networks," in [*2018 International Joint Conference on Neural Networks (IJCNN)*], 1–6, IEEE (2018).

[28] 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).

[29] 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).

[30] Yang, L., Liu, Y., Yu, H., Fang, X., Song, L., Li, D., and Chen, Y., "Computer vision models in intelligent aquaculture with emphasis on fish detection and behavior analysis: a review," *Archives of Computational Methods in Engineering* **28**, 2785–2816 (2021).

[31] Alshdaifat, N. F. F., Talib, A. Z., and Osman, M. A., "Improved deep learning framework for fish segmentation in underwater videos," *Ecological Informatics* **59**, 101121 (2020).

[32] Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P., "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **42**(2), 318–327 (2020).

[33] Ren, S., He, K., Girshick, R., and Sun, J., "Faster r-cnn: Towards real-time object detection with region proposal networks," in [*Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*], *NIPS'15*, 91–99, MIT Press, Cambridge, MA, USA (2015).

[34] Tan, M. and Le, Q., "Efficientnet: Rethinking model scaling for convolutional neural networks," in [*International conference on machine learning*], 6105–6114, PMLR (2019).

[35] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H., "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861* (2017).

[36] Alaba, S. Y. and Ball, J. E., "Wcnn3d: Wavelet convolutional neural network-based 3d object detection for autonomous driving," *Sensors* **22**(18) (2022).