Optimizing and Gauging Model Performance with Metrics to Integrate with Existing Video Surveys

Jack H. Prior
Northern Gulf Institute
Mississippi State Univesity
Pascagoula, MS, 39567, USA
jack.prior@noaa.gov

Simegnew Yihunie Alaba
Electrical and Computer Engineering
Mississippi State Univeristy
Miss. State, MS, 39762, USA
sa1724@msstate.edu

Farron Wallace
NOAA NMFS SEFSC
PEM FATES
Galveston, TX, 77551, USA
farron.wallace@noaa.gov

Matthew D. Campbell
NOAA NMFS SEFSC
PEM Gulf and Carribean Reef Fish
Pascagoula, MS, 39567, USA
mathew.d.campbell@noaa.gov

Chiranjibi Shah
Northern Gulf Institute
Mississippi State Univesity
Starkville, MS, 39759, USA
cshah@ngi.msstate.edu

M M Nabi
Electrical and Computer Enginnering
Mississippi State Univesity
Miss. State, MS, 39762, USA
mn918@msstate.edu

Paul F. Mickle
Northern Gulf Institute
Mississippi State University
Stennis Space Center, MS, 39556, USA
pmickle@ngi.msstate.edu

Robert Moorhead
Northern Gulf Institute
Mississippi State Univesity
Starkville, MS, 39759, USA
rjm@gri.msstate.edu

John E. Ball
Electical and Computer Enginnering
Mississippi State Univesity
Miss. State, MS, 39762, USA
jeball@ece.msstate.edu

***Abstract*—Baited underwater video sampling is a common method to monitor fish populations, yet the data requirements associated with imagery leads to bottlenecks in productivity. Image analysis that incorporates automated methods through deep-learning models could provide solutions. These models have the potential to improve efficiency, and decrease the cost of producing information on fish populations and habitats. In order to reduce human intervention, these models must produce precise, accurate results. While methods for gauging model performance through metrics such as mean-average-precision are helpful during the model training process, evaluating the performance on years of survey data requires a different approach. An otolith age-reader comparison method has been adapted to compare automated counts to true counts. The metrics produced in this analysis are then compared across a span of the model confidence levels in order to find the optimal settings per species to filter output and improve processing speed. For most species, increasing annotations for model training results in better performance, however issues persist with occlusion, turbidity, schooling species, and cryptic/conspecific appearances. With focus on Red Snapper (*Lutjanus campechanus*), this process of evaluation was carried out with multiple years of video data to test for fidelity based on location, time, and environmental conditions. Identifying common failures and adapting active learning algorithms can lead to targeted training for more efficient models in the future. These quality assessment and quality control methods of evaluation provide a framework for tracking performance drift and integrating automated methods properly with existing surveys and manual video count protocols.**

Keywords—Fisheries, Automation, Machine Learning, Gulf of Mexico, BRUVs

I. Introduction

Fishery management necessitates the collection of information on fish species abundances, ages, weights, lengths, fecundity, mortality, trophic interactions, and habitat health [1]. Active management is consistently becoming more important as populations are impacted by multiple stressors such as fishing, climate change, habitat reduction, and decreases in water quality associated with anthropogenic sources. Camera based monitoring has become a common method to gain information in habitats that are difficult to sample, and with species where hook-bias might impact observation [2]. Optic sampling with BRUVs (Baited Remote Underwater Videos) is less invasive, and also collects habitat imagery to supplement ecosystem-based management practices. The drawback of this type of sampling is that the large amount of data leads to bottlenecks in productivity due to storage and manual processing time expenditures. Large-scale combined camera sampling efforts—like that of NOAA's Gulf Fishery Independent Survey of Habitat and Ecosystem Resources (GFISHER) [3], which incorporates data from over 1,000 hours of video collected from nearly 2,000 camera deployments across the Gulf of Mexico (GoM) shelf from Brownsville, Texas to the Florida Keys—results in hundreds of terabytes of data and thousands of hours of manual scrutiny at high expense (West Gulf of Mexico – WGoM; East Gulf of Mexico – EgoM).

To increase efficiency in post processing of marine video sampling, scientists have begun utilizing the advancements in computing through graphics processing unit (GPU) technology and artificial intelligence/machine learning (AI/ML) tools [4]. Advantages to this approach are that computing pipelines can run continuously, can reduce inter/intra observer human biases, and can leverage an ability for pattern recognition that may exceed our own. Further, the detection and classification of fish in each frame of video provides a means to generate many different types of metrics and statistical analyses. Therefore, these models are characteristically versatile for integrating with different surveys that historically include different methods of data interpretation such as counts of fish species that are

averaged across a video (*MeanN*) [5], the maximum count of a fish species seen in a single frame of a video (*MaxN*, the compared count for this study) [6,7], proportions of a video that a fish is present within (temporal comparisons) [8], or just simply the presence/absence of a species.

Detection and classification model performance are most commonly gauged by mean-average-precision (mAP). These mAP scores are a frame-based precision metric produced from a selected fraction of the training library, rather than precision across a large set of full-length, high resolution, and high frame-rate videos; and so, a high mAP score may not be truly reflective of a model's actual capacity to produce accurate count estimates for novel unlabeled video in natural conditions (i.e., annual survey collection). In-situ sampling faces challenges in fish tracking that occur due to fish occlusion, fish behavior, fish density, cryptic appearance, and variable water quality [9,10,11,12]. Moreover, a recent review reveals the need for standard metrics of accuracy between models that are applied in different sectors of fisheries management and research [13].

II. Related Work

Previous efforts [14] focused on the development of the Southeast Fisheries Science Center's (SEFSC) automated reef fish detection and classification models using a cascade-faster regional neural network algorithm [15], and annotation software created through cooperative effort by NOAA's Automated Image Analysis Strategic Initiative (full description of model architecture and current models available open-source at github.com/VIAME). These fish tracking models apply a single identification class and associated confidence threshold (CT) to the set of connected detection frames of individuals that pass through the field-of-view. Automated counts at the model CTs of 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 0.95 were used to filter model output. This CT is calculated as:

$$score_t(c) = \left(b + (1.0 - b) * \frac{\sum_{t=0}^{n} fish\_conf(t)}{n} \right) * \frac{\sum_{t=0}^{n} fish\_conf(t) * class\_conf(t,c)}{\sum_{t=0}^{n} fish\_conf(t)} \quad (1)$$

Variables are given as c = the class ID; n = total number of unique localizations along the frames of each track; *fish_conf(t)* = fish detection value for a particular state in track time t; *class_conf(t,c)* = classifier confidence value for class c at time t; b = posterior probability that a track is definitely a fish [default = 0.1]. Fig. 1. exhibits common imagery from the GFISHER video survey overlaid with automated annotations.

Fig. 1. Example of automated detections of *Lutjanus campechanus* obtained on the NOAA GFISHER / Reef Fish Video Survey

The model outputs were analyzed through an adaption of methods used to compare accuracy of human otolith age estimation across readers [16]. Analogously, a human fish counter can be compared to a computational fish counter across a wide variety of precision and accuracy metrics. The metrics produced in this analysis are the percent agreement of survey videos that models counts match human reader counts exactly, the ratio of videos where a fish was falsely identified by the model (false +), the ratio of videos where a fish was detected manually but not by the model (false -), the coefficient of variation between the human reader and the model counts (CV%), the linear regression of human counts vs. manual counts, and the percent of stations that have model counts within both one and two fish of the manual counts. In the previous analysis, a portion of a single survey (280 stations from the WGoM, >89.5° W, in 2021) was evaluated with these metrics using three different models in order to determine which model was the top performer per species, which species the models worked best to classify and count, and how increases in training data can lead to increases in performance. The optimal model CT is chosen based upon maximum percent agreement, minimum false positive detections, and minimum CV%. The linear regression and the percent of stations within one and two fish is calculated at this optimal CT. One goal of these efforts was also to establish a QA/QC (Quality Assessment/Quality Control) processes that integrates with current assessment models and can track performance drift over time, which is critical for a long-term standardized survey.

The most observed species across stations, Red Snapper (*Lutjanus campechanus*), is also the species with the most annotations in the training library (>206.5k individual frame localizations). While the model still may not have enough information to achieve full reliability, simply increasing the number of annotations past this point does not improve performance drastically, thus other factors are likely influencing accuracy and precision of the model. By comparing the performance of the model on Red Snapper across this larger dataset, which also spans a greater time than was previously tested, it may be possible to discern factors that have a larger

MARFIN Grant 20MFIH002, NOAA National Marine Fisheries Service NA21OAR4320190 to Northern Gulf Institute, NOAA Office of Oceanic and Atmospheric Research

influence on model performance, and thus direct future model training towards imagery where improvements must be made. This approach, combined with the active learning algorithm advances being developed in tandem with these models [16,17], could be a means to more efficiently improve precision. Efforts such as this will also improve efficiency by focusing effort on only the most useful annotations.

III. Methods

To further test limits of model reliability across time and space, the aforementioned analysis has been repeated using multiple sets of survey videos from the full sampling area in the GoM from multiple years (2019 & 2021). Stations that recorded over 100 of a single species for manual counts were removed from the analysis (an arbitrary limit deemed out of the realm of model performance, selected beyond a point where model counts do not match any manual counts). The multiple-year comparison brings the sample size for testing the model from the original 280 stations to 1,219 stations. In this larger dataset there are several factors of interest that can be examined for variable performance, including the regional location of video (EGoM vs. WGoM), the survey the video was collected from (different vessels or slight changes in sampling locations, species assemblages, and gear), the designated complexity of the habitat (based upon max relief, biotic, and abiotic factors), the designated category of habitat (based upon substrates and reef types), the total number of fish counted per video (sum of all *MaxN* counts of all species in a video), the total number of species observed (diversity), and by the relative visibility of the video (overall clarity and ability to see the horizon). Each of the considered factors is filtered into categories for comparisons that yield relatively similar sample sizes based upon the distribution of observations across video stations. For example, breakdown by region results in comparison of 558 stations from the WGoM and 661 from the EGoM. If certain factors show limited performance, then those categories of videos will be focused on for annotation towards future iterations of training.

IV. Results

The pooled 2019 & 2021 model performance at optimal CTs for Red Snapper in relation to the total fish counts and total number of observed species is displayed in Fig. 2, while relation to habitat and water quality is presented in Fig. 3. Stations with no fish are omitted from this analysis as agreement of zeros between manual counts and automated counts would cause inflated percent agreement values; however evaluation of all stations with total *MaxN* counts of 0 showed that false positives still occur up to the 95% threshold. In other words, fish are detected when none are present.

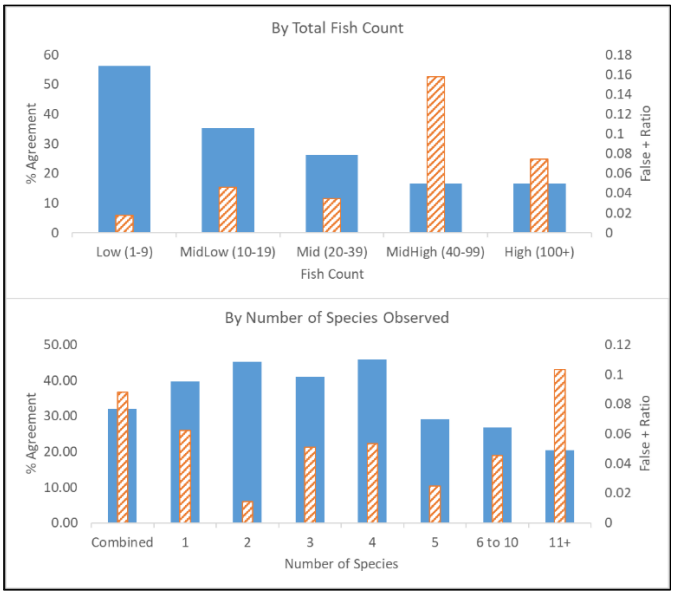

Fig. 2. Model performance for Red Snapper in terms of % agreement (solid blue bars, left axis) and false positive ratios (thin orange-hashed bars, right axis) when comparing effect of fish density and diversity across pooled stations from 2019 and 2021

The highest percent agreement was achieved when evaluating stations with less than 10 fish present, consistent with previous results from the 2021 WGoM dataset. High ratios of false positives are present beyond the third fish count bin (>40 fish). Higher agreement is achieved when four species of fish or less are observed in a video, yet accuracy is best when Red Snapper is not the only species present on screen.

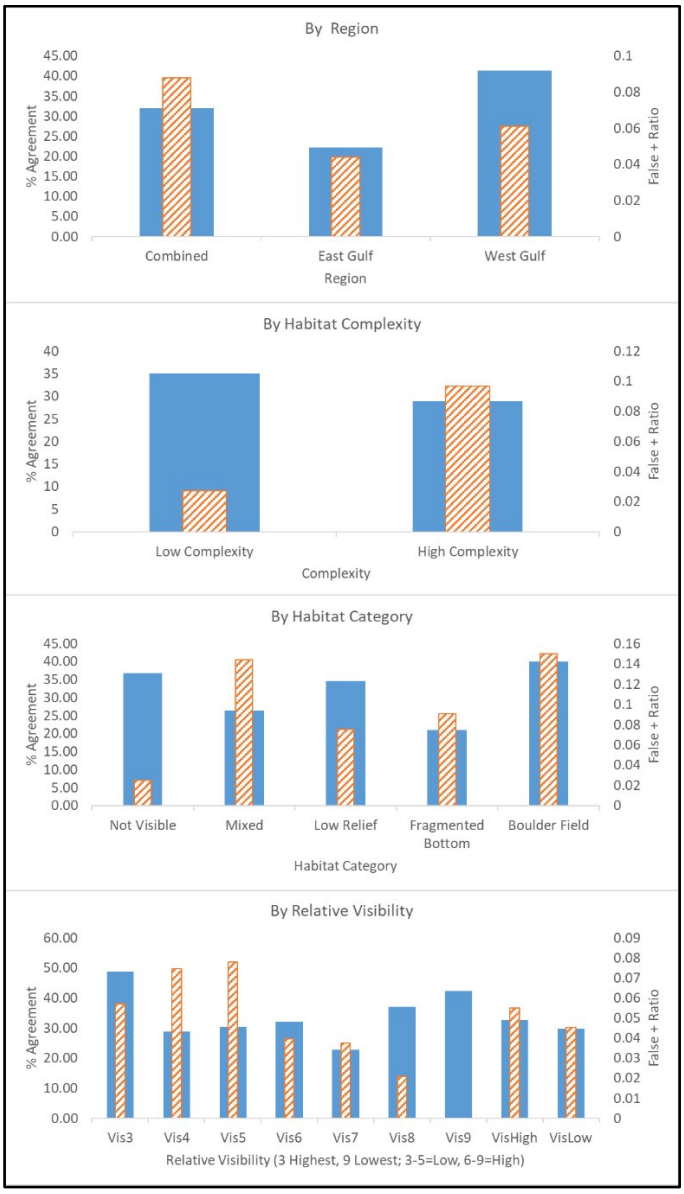

Fig. 3. Model performance in terms of % agreement (solid blue bars, left axis) and false positive ratios (thin orange-hashed bars, right axis) when comparing effects of regional location, habitat types, and water quality (visibility)

Model performance is doubled in terms of percent agreement in the WGoM, but high rates of false positives were evident across both regions. When breaking the full sampling universe down by habitat type there is an obvious difference between levels of false positives on low complexity stations (537 videos at complexities of one and two on a scale to eight), and high complexity stations (632 videos at complexity >2), although agreement is relatively similar. This habitat complexity score is based on the sum of the determined structural complexity (a score of one to five based on maximum habitat relief) and the determined biotic complexity (a score of one to five based on habitat and epifauna coverage). The more complicated habitat types (Mixed, Fragmented Bottom, and Boulder Fields) yielded higher rates of false positives than the less complex types (low relief, and non-visible). Other habitat categories were tested but were either under-represented in sample size, and/or the models yielded lower agreement than those displayed in Fig. 3. Performance was best on stations where the habitat was categorized as non-visible. This pattern was again evident against the scores of relative visibility, where there is a trend towards increased performance at the stations with the least visibility (Vis8 / Vis9), which display the lowest rates of false positives.

Optimal CTs of performance are consistently between 0.6 and 0.8 range as tradeoffs occur (Fig. 4) between maximum agreement and minimum false positives. This is the recommended range when using this iteration of the SEFSC model pack for novel video applications. This range is generally optimal for most species detected, however this paper is meant specifically to describe fine-tuning for Red Snapper detection.

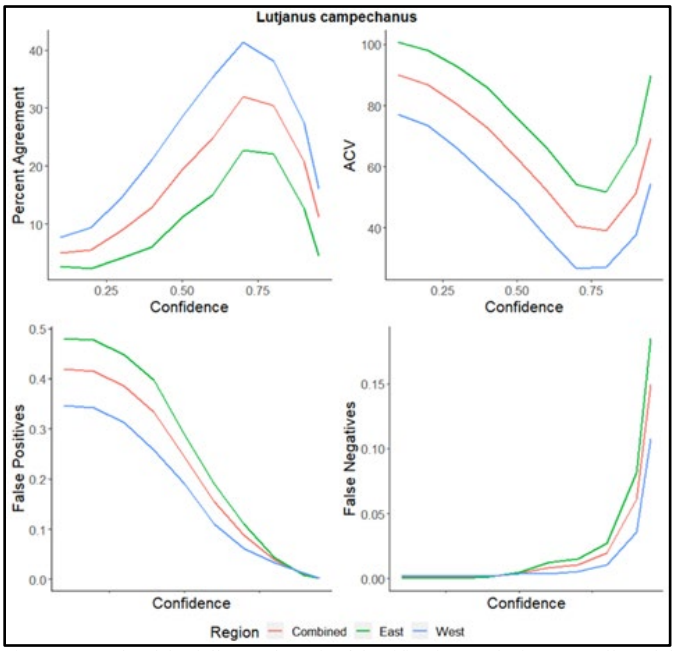

Fig. 4. Example of the full span detection/classification analysis across all model confidence thresholds for the regional comparison of performance (ACV = CV%)

V. Discussion

Many of the results showed expected trends towards increased difficulty of detection and classification with higher fish densities and diversity (Fig. 2), and with increasingly complex habitats (Fig. 3). Often this is a compounding issue because complex habitats generally yield the highest fish densities and diversity. This includes increased diversity within families leading to more instances of species with very similar morphologies. One unexpected outcome was that, while models had the highest agreement on the stations in the highest visibility category (Vis3), the rate of false positives decreased with decreased visibility (Fig. 3), leading to a trend of better CV% at low visibility. It is possible that higher turbidity provides a more constant background, and that clearer water allows more visibility of individuals and more complex habitat, leading to increased false positive detections. These results coincide with the higher performance in the more turbid WGoM, but it should also be noted that the models were also trained with more video data from the WGoM.

Image analysis on the combined survey data results in reliability of exact counts up to three Red Snapper on a frame at a time at the optimal confidence threshold of 0.7 (Fig. 5). This is a substantial decrease from the limit of nine Red Snapper that was determined for the 2021 WGoM video set. The red, open circles of Fig. 5 correspond to a one-sample t-test indicating that the mean of the manual counts (y-axis) are not equal to the corresponding mean of automated counts (x-axis), or if the difference in counts equals zero. Vertical bars represent the range of counts when the p-value is set to 0.05 for the one sample t-test. The dashed line corresponds to the theoretical ideal of a one-to-one relationship between automated counts and manual counts, and deviation from this line represents loss of reliability [19]. The linear regression is calculated as y=1.8x-1.2, (y = manual counts, x = automated counts; R^2=0.67). Correlation decreases at higher counts.

Similarly, models produced by Connelly et al. [11] for fish tracking had a limit of reliable fish counts up to three fish, but described mathematical corrections to allow for better estimations of higher fish counts. Fortunately, nearly 70% of all combined stations had counts of three Red Snapper or less. Agreement of exact counts only occurs at 32% of the combined stations, but 68% of counts were within one fish, and 77% of counts were within two fish of manual counts. Another characteristic of the model, as shown by the linear regression of all stations combined, is that fish are undercounted at optimal CTs, and are therefore providing conservative estimates of relative abundance. Despite the tendency to undercount, false negative detections are much less common than false positive detections until the highest CTs. In fact, Red Snapper shows one of the highest false positive rates of any species, possibly due to over-selection bias caused by the long-tail distribution of the training library [20,21].

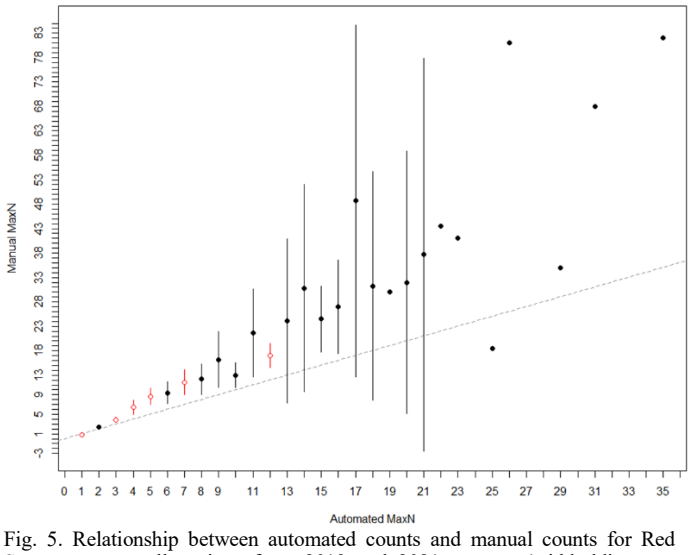

Fig. 5. Relationship between automated counts and manual counts for Red Snapper across all stations from 2019 and 2021 surveys (withholding any stations with counts >100 for any species).

VI. CONCLUSION

This broad-scale approach does not inform the model user/trainer on exactly which frames or videos to include in future training, but it allows the user to choose which dataset to apply active learning to, at a better resolution for that purpose. These active learning algorithms will be applied to data in an iterative process in order to determine exactly which frames can be beneficial for model performance. The active learning methods can be carried out at different scales for optimizing performance (e.g. a single station video, a selected region, or full years of survey data).

Improvements are needed in both the tracking model itself and the distribution of annotations to other species in the training library in order to reduce over-selection and false positive detections of Red Snapper that are the product of misidentifications of other fish species. Further efforts are being made using one-shot [20] and class-aware loss functions [21] to address the imbalance of the training library. Other future work involves integrating domain-shift adaptation algorithms that can be used to deal with changing environmental conditions, backgrounds, and species assemblages [22]. This process should also be repeated for other important commercial and recreationally fished species to see if the same trends hold true, or if each species requires a case-by-case approach to achieve the largest increases in performance. Sample sizes for comparison will increase with each year, and further factors should be considered (e.g., water transmissivity, maximum relief, or the presence of other species most commonly misclassified as Red Snapper). Moreover many of these effects are likely compounding and may require a multi-variate analytical approach, such as a GLM (General Linear Model) or GAM (General Additive Model), for even stronger targeting of beneficial imagery. For now the model capability is only strong enough to assist humans in evaluating imagery with low levels of complexity, fish density, and diversity; however, this automation should still be considered a powerful tool, because these types of videos constitute a large fraction of the sampling datasets each year.

ACKNOWLEDGMENT

Thanks to video readers, A. Paul Felts, J.R. Salisbury, K.R. Rademacher, B. Masarik, K. Overly, S. Thomas, A. Ravas, and others who participated in at-sea surveys to provide this data, especially NOAA research vessels *Southern Journey* and *Pisces*. Thanks to VIAME/Kitware tech support for helping develop models and annotation software. Thanks to NOAA innovation working group for collaborative efforts.

REFERENCES

[1] S. Jennings and M.J. Kaiser, "The effects of fishing on marine ecosystems," Adv. Mar. Biol., vol. 34, pp. 201-352, 1998. doi: 10.1016/S0065-2881(08)60212-6

[2] M. Cappo, E. Harvey, M. Shortis, "Counting and measuring fish with baited video techniques – an overview," Aust. Soc. Fish Biol., vol. 1, pp. 101-114, 2006 Workshop Proc.

[3] K.A. Thompson, T.S. Switzer, M.C. Christman, S.F. Keenan, C.L. Gardner, K.E. Overly, et al., "A novel habitat-based approach for combining indices of abundance from multiple fishery-independent surveys," Fish. Res., vol. 247, pp. 106178, 2022. doi: 10.1016/j.fishres.2021.106178

[4] A.T.M. van Helmond, L.O. Mortensen, K.S. Plet-Hansen, C.Ulrich, C. Needle, D. Osterwind, "Electronic monitoring in fisheries: lessons from global experiences and future opportunities," Fish, vol. 21, pp. 162-189, 2020. doi: 10.1111/faf.12425

[5] N. M. Bacheler and K.W. Shertzer, “Estimating relative abundance and species richness from video surveys of reef fishes,” Fish. Bull., vol. 113, pp. 15-26, 2015. doi: 10.7755/FB.113.1.2

[6] D.M. Ellis and E.E. DeMartini, “Evaluation of a video camera technique for indexing abundances of juvenile pink snapper, *Prisitpomoides filametosus*, and other Hawaiian insular shelf fishes,” Fish. Bull., vol. 93(1), pp. 67-77, 1995.

[7] M.D. Campbell, A.G. Pollack, C.T. Gledhill, T.S. Switzer, D.A. DeVries, “Comparison of relative abundance indices calculated from two methods of generation video count data,” Fish. Res., vol. 170, pp. 125-133, 2015. doi: 10.1016/j.fishres.2015.05.011

[8] I.G. Priede, P.M. Bagley, A. Smith, S. Creasey, N.R. Merrett, “Scavenging deep demersal fishes of the porcupine seabight, north-east Atlantic: observations by baited camera, trap and trawl,” J. Mar. Biol. Assoc. United Kingdom, vol. 74(3), pp. 481-498, 1994. doi: 10.1017/S0025315400047615

[9] S. Marini, E. Fanelli, V. Sbragaglia, E. Azzurro, J. Rio Fernandez, J. Aguzzi, “Tracking fish abundance by underwater image recognition,” Nat. Sci. Rep., vol. 8, 13748, 2018. doi: 10.1038/s41598-018-32089-8

[10] A. Salman, S. Siddiqui, F. Shafait, A. Mian, M. Shortis, K. Khurshid, et al., “Automatic fish detection in underwater videos by a deep neural network-based hybrid motion system,” ICES J. Mar. Sci., vol. 77(4), pp. 1295-1307, 2020. doi: 10.1093/icesjms/fsz025

[11] R.M. Connolly, D. Fairclough, E. Jinks, E. Ditria, G. Jackson, S. Lopez-Marcano, et al., “Improved accurracy for automated counting of fish in baited underwater videos for stock assessment,” Front. Mar. Sci., vol. 8, 2021. doi: 10.3389/fmars.2021.658135

[12] S. Lopez-Marcano, M.P. Turschwell, C.J. Brown, E.L. Links, D. Wang, R.M. Connolly, “Computer vision reveal fish behaviour through structural equation modelling of movement patterns,” Res. Square Prelim. Rep., pp 1-24, 2022. doi: 10.21203/rs.3.rs-1371027/v1

[13] J.C.A. Barbedo, “A review of the use of computer vision and aritifical intelligence for fish recognition, monitoring, and management,” Fishes, vol. 7, pp. 335, 2022. doi: 10.3390/fishes7060335

[14] J.H. Prior, M.D. Campbell, M. Dawkins, P.F. Mickle, R.J. Moorhead, S.Y. Alaba, et al., “Estimating precision and accuracy of automated video post-processing: A stop towards implementation of AI/ML for optics-based fish sampling,” Front. Mar. Sci., vol. 10(1150651), 2023. doi: 10.3389/fmars.2023.1150651

[15] Z. Cai and N. Vasconcelos, “Cascase r-CNN: Delving into high quality object detection,” IEEE/CVF Conference on Computer Vision and Pattern Rcognition, Salt Lake City, UT, USA, pp. 6154-6162, 2018. doi: 10.1109/CVPR.2018.00644

[16] S.E. Campana, “Accuracy, precision and quality control in age determination including a review of the use and abuse of age validation methods,” J. Fish Biol., vol. 59, pp. 197-242, 2001. doi: 10.1111/j.1095-8649.2001.tb00127.x

[17] S.Y. Alaba, C. Shah, M.M. Nabi, J. Ball, RJ. Moorhead, D. Han, et al., “Semi-supervised learning for fish species recognition,” Proc. SPIE 12543 Ocean Sensing and Monitoring XV, 125430P, 12 June 2023. https://doi.org/10.1117/12.2663422

[18] C. Shah, S.Y. Alaba, M.M. Nabi, R. Caillouet, J.H. Prior, M.D. Campbell, et al., “MI-AFR: multiple instance active learning-based approach for fish species recognition in underwater environments,” Proc. SPIE 125430N, 12 June 2023. https://doi.org/10.1117/12.2663404

[19] D. Ogle, “fishR vignette – precision and accuracy in ages,” Northland College, Ashland, WI, USA, 2013.

[20] J. Cai, Y. Wang, J.N. Hwang, “ACE: Ally complementary experts for solving long-tailed recognition in one-shot,” IEEE/CVF Internation Conference on Computer Vision (ICCV), pp. 1-10, 2021.

[21] S.Y. Alaba, M.M. Nabi, C. Shah, J.H. Prior, M.D. Campbell, F. Wallace, et al., “Class-aware fish species recognition using deep learning for an imbalance dataset,” Sensors, vol. 22(21), pp. 8268, 2022. doi: 10.3390/s22218268

[22] A. Zheng, J. Mei, F. Wallace, C. Rose, R. Hussein, J.N. Hwang, “Progressive mixup augmented teacher-student learning for unsupervised adaptation,” ICLR, 13 February 2023