Automated Analysis of Marine Video With Limited Data

Deborah Levy[1,2], Yuval Belfer[2], Elad Osherov[2], Eyal Bigal[1], Aviad P. Scheinin[1], Hagai Nativ[1], Dan Tchernov[1], and Tali Treibitz[1]

[1]Leon H. Charney School of Marine Sciences, University of Haifa
[2]Technion – Israel Institute of Technology

{dlrun14, yuvalbelfer, osherovelad, eyalbigal, shani.aviad, hagainativ}@gmail.com
{dtchernov, ttreibitz}@univ.haifa.ac.il

Abstract

Monitoring of the marine environment requires large amounts of data, simply due to its vast size. Therefore, underwater autonomous vehicles and drones are increasingly deployed to acquire numerous photographs. However, ecological conclusions from them are lagging as the data requires expert annotation and thus realistically cannot be manually processed. This calls for developing automatic classification algorithms dedicated for this type of data. Current out-of-the-box solutions struggle to provide optimal results in these scenarios as the marine data is very different from everyday data. Images taken under water display low contrast levels and reduced visibility range thus making objects harder to localize and classify. Scale varies dramatically because of the complex 3 dimensionality of the scenes. In addition, the scarcity of labeled marine data prevents training these dedicated networks from scratch. In this work, we demonstrate how transfer learning can be utilized to achieve high quality results for both detection and classification in the marine environment. We also demonstrate tracking in videos that enables counting and measuring the organisms. We demonstrate the suggested method on two very different marine datasets, an aerial dataset and an underwater one.

1. Introduction

Obtaining ecologically meaningful data on highly mobile species in the marine environment is challenging due to their spatiotemporal heterogeneity and occurrence at low densities. Commercial catch records provide some of the most spatially and temporally extensive points of reference that are available for stock assessment. However, fishery-dependent methods are considerably biased as they depend on catchability and an incentive to report [4, 29]. Several non-extractive approaches have been used or are currently emerging to replace fishery-dependent methods. These are considerably low-cost relative to their coverage of large water volumes and may inform about the abundance of marine

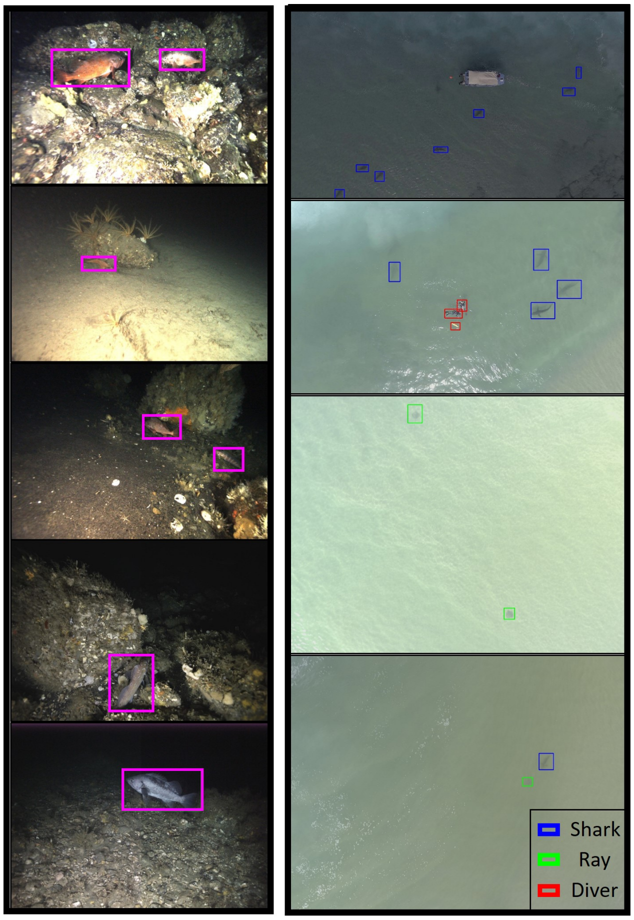

Figure 1. **Example results of our method.** [Left] Underwater detection of fish. [Right] Detection and classification into 3 classes (shark, ray, diver) of marine aerial photographs. The same network architecture works well for both datasets, with dataset specific training.

organisms. However, they necessitate further technological innovation considering the elusive nature of these species and the spatiotemporal dimensions of their ecology [29].

For example, aerial surveys, which usually employ piloted, qualified observers, are a recognized technique for abundance estimates in large marine areas [29, 46]. However, they are characterized by a number of drawbacks such as human risk, high costs, missed sightings, low resolution of location data, and the requirement to fly on airstrips [18]. Recent advances in the development of unmanned aerial vehicles (UAVs) have enabled research to compensate for these constraints and, by means of image analysis and photogrammetry, provided a tool to account for the number and size of individual animals within observed groups, thus yielding more reliable, quantitative estimates (e.g., [9, 14, 25]). Similarly, image analysis could benefit other fishery-independent methods at smaller spatial scales; underwater visual censuses (UVCs) are employed for the computation of relative abundance in a quick, repeatable, and cost-effective setup, yet may be affected by human presence impacts [16, 24]. However, image analysis in both aerial surveys and UVCs are performed post hoc and are labor-intensive to the point of impracticality. Thus, automatic object identification algorithms are required to produce a comprehensive, effortless technique for monitoring a variety of species using different technologies.

Object detection in the marine environment is different from object detection in natural, real world images, and is characterized by several prominent challenges, demonstrated in Fig. 2. First, the marine environment has a complex structure. This, combined with the organisms' tendency to move and hide creates many occlusion scenarios as well as a large variety of organism poses. Second, marine images are heavily affected by light attenuation and scattering, where the effect varies with object distance and thus changes across the scene. Third, surface caustics may cause spatial and temporal changes in lighting, which in turn reduce the classifier accuracy and make tracking the object harder. Fourth, the objects' scale can change dramatically.

Scale changes are very prominent also in images acquired by UAVs, where the UAV can fly in variable heights above the water. Additionally, UAV images may differ in the color of water, existence of stones, ripples, streams and light rays reflectance, all of which may impair classification results. In our work we also used several different methods for aerial data collection, where the viewing angle varied from vertical to side view, making the task even more challenging (examples in Fig. 3). Last, and maybe most important, to yield significant scientific value, marine images (both under and above water) have to be annotated by expert scientists. Thus, citizen science labeling tools cannot be used on them to aid in the labeling effort. This results in small amounts of data that are not sufficient for training Deep Convolutional Neural Networks (CNNs).

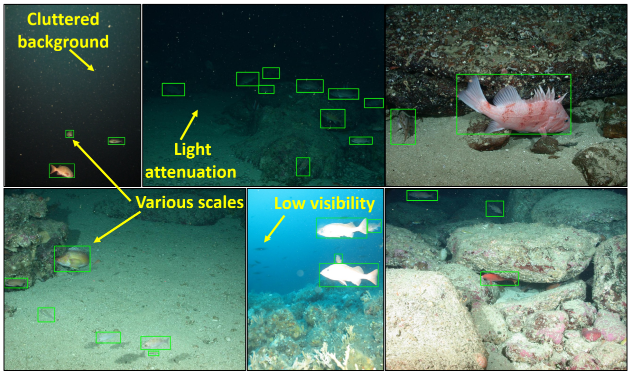

Figure 2. **Examples of the underwater fish detection challenges.** Examples from [7] and the AFSC dataset from this workshop's challenge. Data was manually annotated by us. The dataset images exhibit several detection challenges such as low visibility, various scales, cluttered background, and light attenuation.

CNNs [28] have recently exhibited remarkable performance in the tasks of image classification [26] and object detection [12, 13, 45]. This impressive performance builds on the availability of large amounts of annotated data [8], as well as parallel computational resources such as GPUs, and regularization techniques [15, 50, 55]. Nevertheless, CNNs, as well as other visual classification algorithms, struggle when dealing with challenging data such as partially occluded objects, lighting variations, inter and intra class variations and deformations [39, 41], that are common in the marine environment.

Here we demonstrate a method that is based on state-of-the-art CNN detector and an object tracker for detection and classification in marine videos. The method achieves high precision and recall despite the very small datasets used to train it (only hundreds of images) and works well on both our datasets – above and under water. Thus, it can be used by marine scientists to efficiently analyze the multitude of survey data that is being collected around the world.

2. Related Work

Current state-of-the-art solutions for underwater detection of marine species rely on either acoustic or optical measurements.

Acoustic methods vary from counting the echoes of sonar beams [2, 56] to more complex methods using only acoustic images [19]. Optical methods range from more traditional methods such as edge-based classifiers [47], to a fish detector and tracker [48] which combines an Adaptive Gaussian Mixture Model [59] detector and Continuously Adaptive Mean Shift Algorithm [11] tracker.

More recent works utilize Convolutional Neural Networks (CNNs) object detectors for this task. Marbug et al. [37] use a series of CNNs to create an underwater detec-

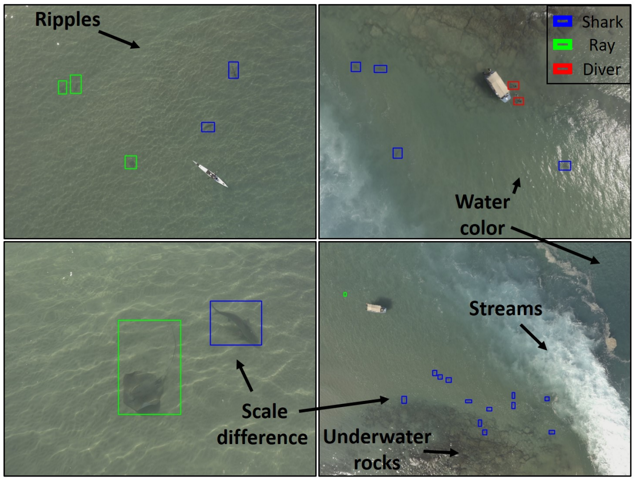

Figure 3. **Examples of the marine aerial dataset acquired by us.** As the UAV is flown in various heights above sea level, object scale becomes a serious challenge for detection. Labels are color-coded according to the legend. Note the various detection challenges such as scale difference, ripples, water color, underwater rocks, streams and viewing angle.

tor and classifier for benthic macrofauna, Li et al. [30] employed a Fast R-CNN [12] detector for fish detection while Sung et al. [51] used the YOLO detector [42].

Several works aimed at underwater fish tracking rather than detection. Some methods used statistical approaches such as Bayesian filtering [38] and maximum likelihood function for detection, covariance based tracker [49] with manually initialization. Zhou et al. [58] suggested to use a series of filters, such as Gabor filters, combined with projection segmentation, in order to segment and track a particular species of fish.

UAVs are increasingly used for 3D mapping of the marine environment [6, 54]. Automatic detection of marine organisms was demonstrated in a commercial product by *Westpac Little Ripper Lifesaver* [52], who claim to detect sharks with high accuracy in real-time.

3. Review of Relevant CNN Methods

CNNs which are multi-layer neural networks, with shared weights and local receptive fields are the current goto algorithm for object detection. These networks comprise of millions of parameters and are able to represent higher level semantic cues in the data, allowing a very accurate lower dimension embedding of the input data. This desirable attribute comes at a heavy price, requiring vast amounts of labeled training data in order to avoid overfitting of the training set. As large amount of labeled training data is usually hard to acquire, a possible solution, which we harnessed in this work, is to pre-train the network with available labeled data [8] and later fine tune the network to our specific task. Since the first part of the network layers extracts features such as edges, blobs, colors and textures, which usually generalize well to the domain of natural images [57], the fine tunning is usually done on the later layers of the network, which represent the non-linear combinations of the extracted features.

The depth of the network increases with the complexity level of the semantic cues the network can extract. This interesting attribute makes the training process harder, since training very deep networks tends to result in overfitting. In this work we chose to use Residual network topology [17], which is a deep convolutional network that uses identity mapping between the layers in order to avoid overfitting. Residual network provides better gradient flow through the network layers, making the training of deep networks possible and with small generalization error. Residual network is also characterized by another desired attribute, low number of parameters, which make the training process more robust to overfitting. This attribute is achieved by using only convolution layers with no fully connected layers at all. This topology is known as Fully Convolutional Network FCN [36]. Last, the FCN allows the use of varying input size, which is desirable in the task of object detection.

3.1. Object Detection using CNNs

State-of-the-art solutions for object detection using CNNs usually belong to one of two categories: *one-stage* detectors and *two-stage* detectors. Two-stage detectors, as the name suggests, split the detection problem into two consecutive stages where each is independent of the other. The first stage generates a sparse set of candidate object locations. This stage can be performed using either basic algorithms for region proposals [13] or using a dedicated region proposal CNN [45]. The second stage of the detector consists of a CNN that classifies each location candidate as one of the classes or as a background class. Those detectors achieve state-of-the-art accuracy scores on the COCO benchmark [33].

A major disadvantage of the two-stage detectors is their run-time execution, which is limiting in real-time applications. This encouraged the development of the one-stage detectors. These detectors classify all image regions at once as opposed to the two-stage detectors which classify a sparse set of location candidates. Current state-of-the-art one-stage detectors include YOLO [43] and SSD [35].

Despite the promising run-time performance achieved by one-stage classifiers, they generally lag behind the two-stage classifiers in terms of accuracy. Since one-stage detectors process a dense set of candidates, they inherently suffer from class imbalance, i.e., most of the candidates in both training and testing, belong to the background and only a small number of candidates contain recognizable objects.

To solve this Lin et al. introduced *RetinaNet* [32] in 2017. RetinaNet is a one-stage CNN detector that achieves

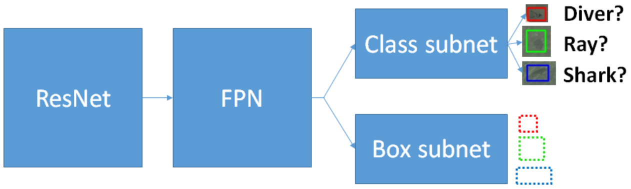

Figure 4. **RetinaNet architecture.** Feature pyramid net (FPN) on top of a residual network generates a feature map, which is then fed to two fully convolutional sub-networks, Class subnet for classification and Box subnet for regression (bounding box).

near state-of-the-art performance in terms of accuracy, comparable to the accuracy of state-of-the-art two-stage detectors. RetinaNet is comprised of a feature pyramid network (FPN) [31] on top of a residual network [17]. This part is then concatenated to two additional fully convolutional sub-networks [36], one for classifying anchor boxes and one for regressing from anchor boxes to ground-truth object boxes, as illustrated in Fig. 4.

Another novelty of the RetinaNet is the formulation of a unique loss function termed *focal loss*. This loss function is designed to handle the class imbalance problem by re-weighting easy examples with a lower weight allowing the training process to focus on harder examples, i.e. examples with low classification probability. This is controlled by a non-negative parameter γ, where the re-weighting is exponential with γ.

3.2. Object Tracking

The field of multiple object tracking (MOT) has developed in recent years. These algorithms usually focus on both high accuracy and low run-time. Several studies addressed this problem using Multiple Hypothesis Tracking (MHT) [22, 44], which provide high accuracy tracking but fails to achieve real-time performance.

The Simple Online Realtime Tracker (SORT) [5], is an MOT tracker based on a combination of Kalman filter [21] and the Hungarian algorithm [27]. We chose this algorithm since it achieved both high tracking accuracy and real-time performance. An important parameter of SORT is T_{lost}, the number of consecutive frames an object can be undetected before the tracking on it stops. The default is $T_{\text{lost}} = 1$.

4. Our Proposed Method

In this work, we chose to use the RetinaNet [32] as the object detector for the reasons detailed above. Our implementation was based on *Keras-RetinaNet* [34]. Note that this object detector receives still images as input, while some of the input data in our case is video streams. In this case having independent detections on each frame is not desirable for several reasons – detections can contradict, and it is not clear how to count the organisms even if the detections agree. Therefore, we use tracking in order to make the detections coherent, and to connect detections of the same object. This enables more accurate counting, and also gathering more comprehensive statistics about each organism, such as length. Recurrent neural networks can perform detection directly on videos [53]. However, they are computationally expensive and can sometimes converge to bad local minima, resulting in reduced detection accuracy. Therefore, we use the detection results on the still images as an input to an object tracker, that requires less computational effort than a recurrent neural network approach and provides more accuracy than a weight decay time window.

4.1. Detection Algorithm

As discussed earlier, a key advantage of RetinaNet compared to other one-stage detectors is the use of the focal loss function which handles both class imbalance and scale invariance, both are significant issues in marine datasets due to the low density of the marine organisms of interest. As we show in the results, RetinaNet can cope well with these issues even in the challenging marine data, both under and above water.

All image detection datasets that are used for training CNNs are composed of everyday scenes containing common objects and do not suit the unique distribution of marine images. Existing marine datasets usually contain a very small number of labeled images and thus cannot be used to train the network from scratch. We solved this issue using transfer learning [40]. In this process, we used a pre-trained residual network [17] trained with the ILSVRC12 (imagenet) dataset [8], and then continued to train the two detection subnets using the COCO dataset [33]. This allowed us to use relatively low number of labeled training images during the training process. In the case of aerial point-of-view, we used roughly 300 manually annotated images of sharks and rays. For the underwater network, we used roughly 500 manually annotated images of fish.

The training process included two stages. In both stages the loss function of the network is an average between focal-loss (with γ=2), classification loss (see Sec. 3.1), and a smooth L_1 for the bounding box regression. The first stage of the training enabled the update of the two detection sub-networks only (Fig. 4). The learning rate determined by ADAM [23] optimization algorithm, in order to maintain convergence. The network was trained in this manner for 40 epochs. Then during the second stage, we enabled the update of all the network weights, providing slightly better feature extraction in the earlier layers. The second stage took place for additional 40 epochs. This dual stage training process converged to higher accuracy than either of the two stages alone. We found images of 800×1333 resolution[1] to provide good balance between accuracy and training time.

[1] The original resolution of the fish images was in this order, whereas the original shark images were at 3000×4000.

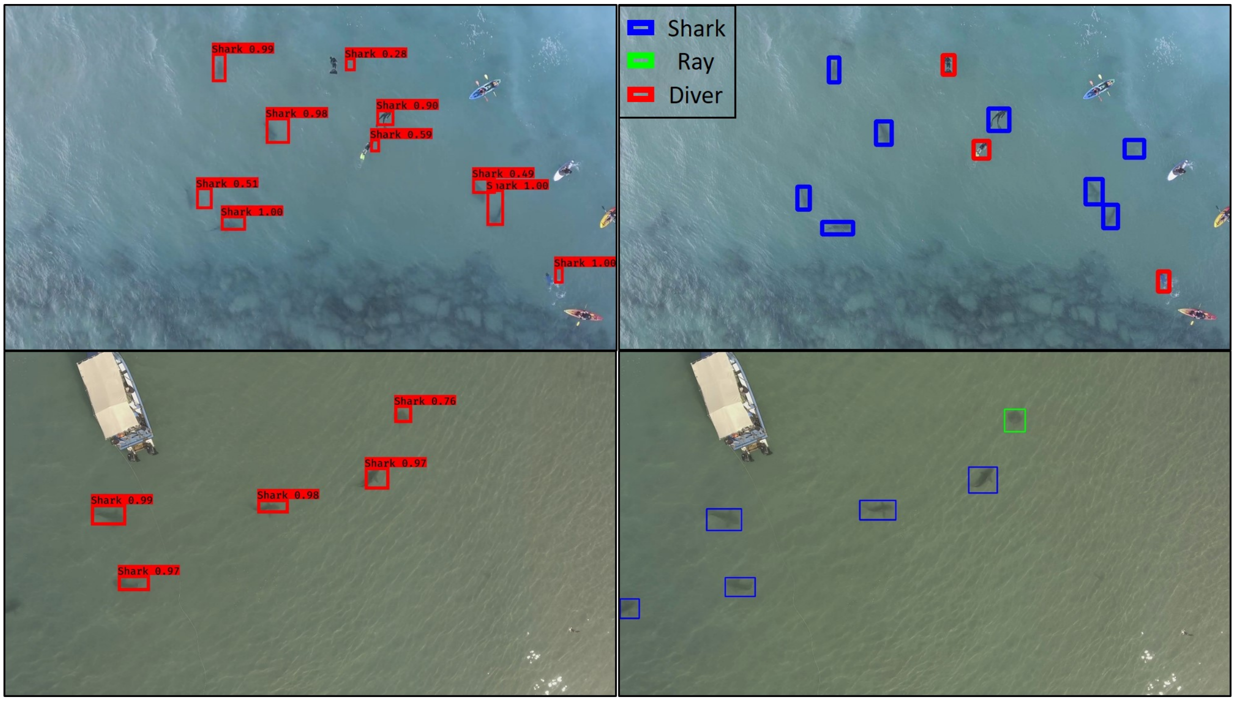

Figure 5. **Comparison with an aerial shark detection network based on YOLO topology.** [Left] Results of YOLO. [Right] Our results. Generally our algorithm provides better recall rate with less false positives.

In order to handle the small amounts of labeled images, we used vertical and horizontal flipping in both datasets.

4.2. Tracking

For tracking we used the SORT algorithm [5], fed with detections from the CNN in each frame. To overcome temporary mis-detections we changed T_{lost} to be 5 instead of the default value of 1.

5. Experiments and Analysis

5.1. Datasets

To demonstrate robustness of the proposed algorithm, we present results on two marine datasets which are of ecological importance. The same algorithm was used for both datasets but the networks were trained separately for each one. In both datasets, for detection, each object is labeled with both its class name and its bounding box coordinates.

The first is an underwater dataset which contains mostly images from *labeled fishes in the wild* [7] image dataset. The annotation in the original dataset is somewhat problematic as not all the fish instances in each frame are labeled. Needless to mention, such errors can impair the training process and induce errors during test. We addressed this issue by using only a subset of 463 images out of the entire dataset, which we manually re-annotated. To comprise a more heterogeneous dataset, we added 45 images from the AFSC dataset, which is provided in this workshop's challenge. The images were split into groups of 448 and 60 images for training and validation sets respectively (keeping the same training-validation relation in both datasets). Note that the dataset contains only fish labels with no regard to the fish species. The dataset contains challenging images with low visibility, different scales, and cluttered background (examples in Figs. 1,2).

Our second dataset was aerial footage collected at the coast of the Mediterranean sea in Israel using a vertical takeoff and landing UAV (DJI Phantom IV) at a shallow-water aggregation of sharks and rays. The UAV was either operated from the shore or deployed from a boat, constantly remaining within line-of-sight; see Fig. 5. Images were captured from a range of altitudes up to 50 meters above the survey area in different environmental conditions, i.e., sea-state, sun glare, cloud coverage and seawater turbidity (the latter dictated by runoff and discharge waters of a coastal stream and an electric power station nearby, respectively). The photographed objects were then labeled as Shark, Ray or Diver by expert scientists. The images were split into groups of 232 and 40 images for training and validation sets respectively.

5.2. Results

Figs. 2, 3 depict several example results of our method on challenging images. To evaluate them we compare our algorithm results to a YOLO network topology [43], also a one-stage detector. For the comparison on our aerial image

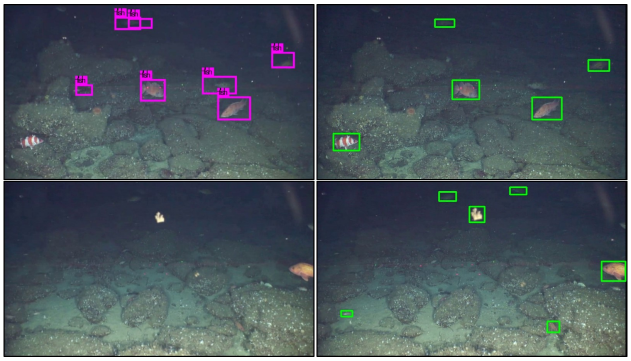

Figure 6. **Comparison with an underwater fish detection network based on YOLO topology.** [Left] Results of YOLO. [Right] Our results. Generally our algorithm provides better recall rate with less false positives.

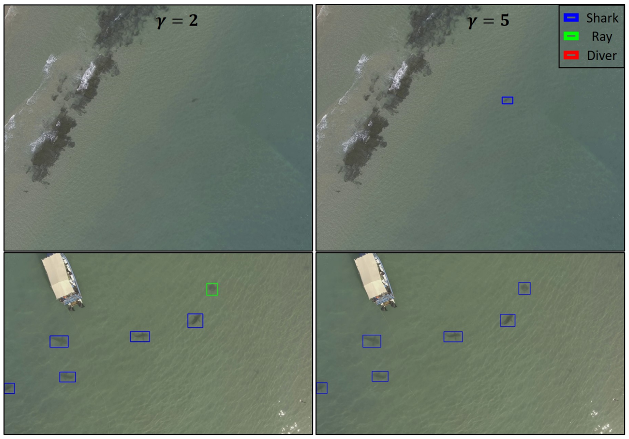

Figure 7. **The effect of the γ parameter on the results.** A higher value of γ provides better detections on aerial images that were taken from a very high altitude of 50m (top row). However, higher value of γ harmed the classification in lower altitudes, as seen in the bottom row (altitude of 28.5m).

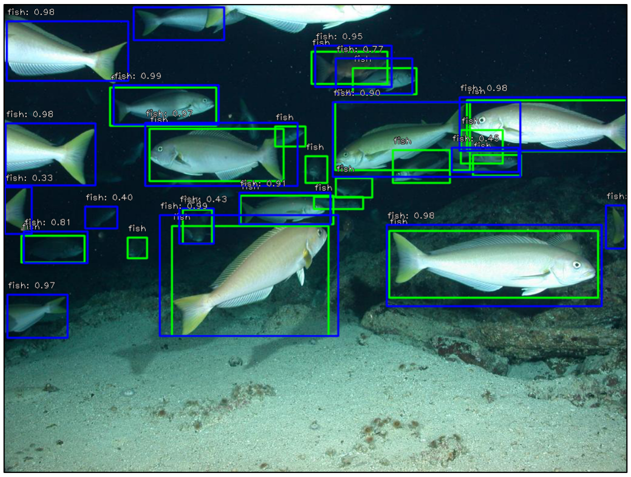

Figure 8. **Example of challenges in labeling.** A scene with manual annotations (green) and our detections (blue). Some fish were detected (in blue) but not annotated (no green frame). These cases lower the precision as they were counted as a false positive though the detection was actually correct.

detection network, we trained YOLO in the same way we trained RetinaNet with the same dataset. Our network had more accurate detections and classifications (Fig. 5). The comparison on the underwater fish detection was done with respect to the results published in [51] on the labeled fishes in the wild dataset [7]. Note that in [51] the training set was much larger than ours ($\sim$ 2000 images), but still we achieve higher accuracy (Fig. 6). Examining our results, we noticed that aerial images that were taken by the drone in higher altitudes, which results in images where the marine organisms are represented by a small amount of pixels, have bad detections or do not have detections at all. This is because these images are suffering from severe class imbalance, i.e., the ratio between background pixels and class object pixels differs with the altitude, where higher altitude images have significantly more background pixels (see Sec. 3.1). The proportion between We tried to solve this by increasing γ in the focal loss function, in order to reduce the influence of the background pixels. Indeed, the detections in the high altitude images improved, but the classification accuracy decreased. Results can be seen in Fig. 7.

To evaluate the results quantitatively, we calculated the average precision for each class. The average precision (AP) was calculated as the area under the precision-recall curve in the operating point, defined by the non-maximal suppression, i.e., the detection with the highest confidence. In order to have an independent test set, we manually annotated 41 additional images from the aerial videos and 42 images from the fish dataset. These images did not participate in the training (neither in the validation set) and were not taken from videos of which frames were taken to the training set. In the evaluation we consider only detections that have a confidence level above 0.4. We measure a successful detection where the classification label is correct and the intersection-over-union ratio is above 0.4. The average precision for each of the three classes in the aerial videos was 0.75 for the Shark class, 0.4 for Ray , and 0.25 for Diver. For the Fish class in the fish dataset it was 0.74. Note that these numbers were calculated on a small number of images because of the scarcity of the data and might not be fully representative. In addition, the annotations are not always ideal; see for example, Fig. 8, where some fish were detected but not annotated. This lowers the precision as it was counted as a false positive though the detection was actually correct. Moreover, there is a difference in the number of images per class in the aerial training set, where there are not as many divers and rays instances as opposed to sharks, which results in higher AP rate for the sharks.

Tracking results are demonstrated in Fig. 9. In many cases the tracker was able to estimate the location of organ-

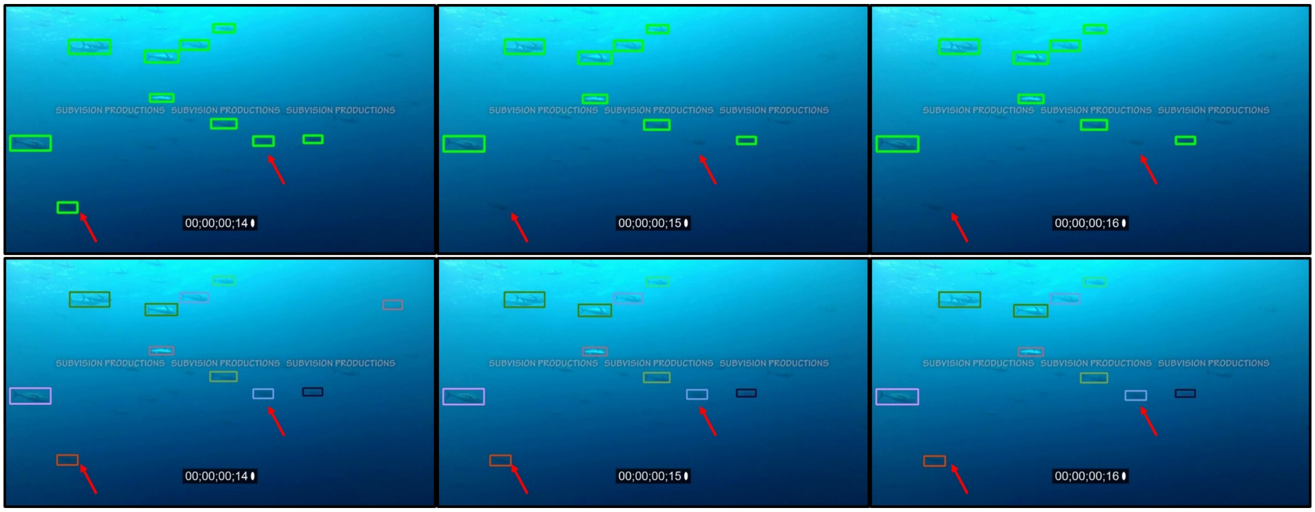
SUBVISION PRODUCTIONS SUBVISION PRODUCTIONS SUBVISION PRODUCTIONS
00;00;00;14
00;00;00;15
00;00;00;16

Figure 9. **The benefit of the tracking stage.** [Top] Objects that were detected in single consecutive frames. [Bottom] Tracking results on the same frames. The detection results on the single frames are the input to the object tracker. Objects that were tracked across frames are assigned the same color of the bounding box. The tracker was able to estimate the location of organisms that were not detected in the single images, increasing overall detection precision (see red arrows). In addition, the tracking enables discrimination between individuals of the same specie, which is extremely significant for extracting meaningful ecological measures.

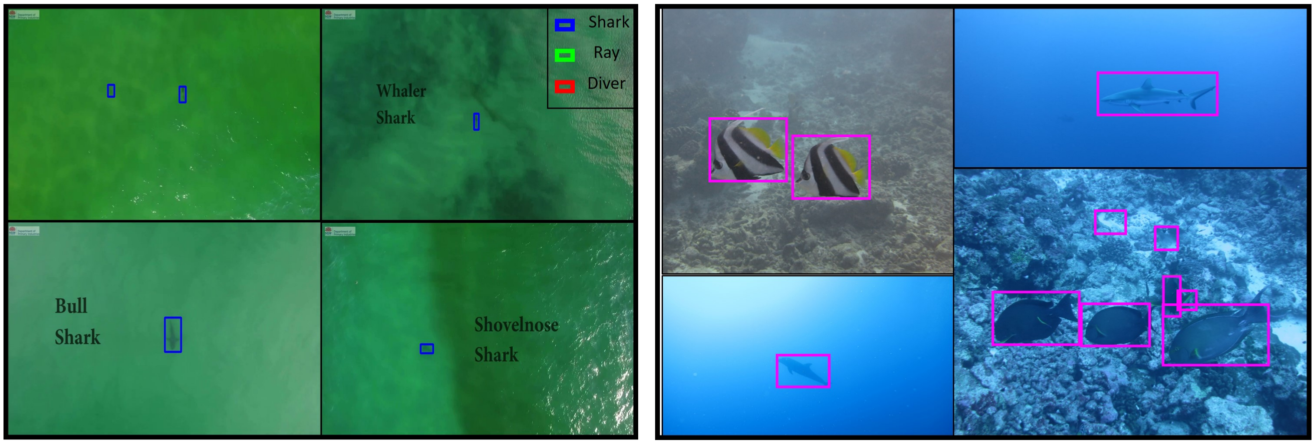

Figure 10. **Generalization of our algorithm on a different dataset than the training set.** [Left] The aerial test examples were taken in a different geographical location [10], exhibiting variations in shark species, water color and scale. Our network was still able to detect most sharks. [Right] The underwater test examples are different than the training images by several factors: fish species, water color, lighting conditions and scale. Though there are some false detections (e.g. in the bottom right scene), it seems that our network can rapidly yield initial detection results on various fish images.

isms that were not detected in the single images, increasing overall detection precision (pointed to by red arrows). In addition, the tracking enables discrimination between individuals of the same specie, which is extremely significant for extracting meaningful ecological measures. This tracker heavily depends on the detection accuracy and thus provides excellent results where the detections are correct.

Video results are available at the following playlist https://www.youtube.com/playlist?list=PLdv3FXVDfbcxzE3q5gMqXJ2Sd5vNCHOQn.

5.3. Generalization of the Detection Network

CNNs are known for their good generalization capability, where a network which was trained on a specific image and class distribution can sometime generalize well and provide descent accuracy for test images drawn from different image and class distribution. In this section we show examples of how well did the algorithm generalize to other distributions of data. The underwater dataset contains images taken with artificial illumination assistance, variations in depth, and different species of fish than training set. Most of the images are from the bottom of the ocean, where one can see the ground, which may introduce different image characteristics. As we wanted to examine our detection network ability to cope with images of different characteristics, such as natural illumination, different water color and distinctively different species of fish (like size and colors).

Likewise, all of our marine aerial dataset were taken in the same geographic location. In order to check generalization, we fed the network with a video that was taken in New South Wales, Australia [10]. The conditions in this video are vastly different from those in our training data in many aspects such as: shark species, water color and scale. Nevertheless the algorithm demonstrated good generalization; see Fig. 10 for both networks results.

6. Discussion

We demonstrated that transfer learning on state-of-the-art trained networks can yield excellent results on marine datasets that contain a very small amount of annotated data. This can become a very useful tool for automatic annotation, as marine researchers are only required to manually label a small number of photographs in order to initiate the automatic classification. We showed that this method is robust and can handle different types of data, and copes well with the unique challenges of marine images.

As for further directions, first we aim to test whether preprocessing using underwater image restoration methods such as [1, 3] can help improve the results. Second, we aim to test the performance of a tracker that requires only a single initial detection rather than detections in each frame [20]. The tracker in [20] operates only on single objects and therefore requires extension for multiple objects. Last, currently it is difficult to perform extensive evaluation of methods as there is no standard dataset available that is annotated well. Therefore we plan to release our aerial and underwater test sets for future comparison and evaluation.

Acknowledgements

This work was supported by the European Union's Horizon 2020 research and innovation program under grant agreement No. 773753 (SYMBIOSIS), the Leona M. and Harry B. Helmsley Charitable Trust, the Maurice Hatter Foundation, and the Technion Ollendorff Minerva Center for Vision and Image Sciences. Research at the University of Haifa was conducted at the Hatter Department for Marine Technologies and the Morris Kahn Marine Research Station, Department of Marine Biology.

References

[1] C. O. Ancuti, C. Ancuti, C. De Vleeschouwer, and P. Bekaert. Color balance and fusion for underwater image enhancement. *IEEE Trans. Image Processing*, 27, 2018. 8

[2] H. Balk and T. Lindem. Improved fish detection in data from split-beam sonar. *Aquatic Living Resources*, 13(5), 2000. 2

[3] D. Berman, T. Treibitz, and S. Avidan. Diving into haze-lines: Color restoration of underwater images. In *Proc. British Machine Vision Conference (BMVC)*, 2017. 8

[4] A. Bertrand and E. Josse. Acoustic estimation of longline tuna abundance. *ICES J. of Marine Science*, 57(4), 2000. 1

[5] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft. Simple online and realtime tracking. In *Proc. IEEE ICIP*, 2016. 4, 5

[6] E. Casella, A. Collin, D. Harris, S. Ferse, S. Bejarano, V. Parravicini, J. L. Hench, and A. Rovere. Mapping coral reefs using consumer-grade drones and structure from motion photogrammetry techniques. *Coral Reefs*, 36(1), 2017. 3

[7] G. Cutter, K. Stierhoff, and J. Zeng. Automated detection of rockfish in unconstrained underwater videos using haar cascades and a new image dataset: labeled fishes in the wild. In *Proc. IEEE Applications and Computer Vision Workshops (WACVW)*, 2015. 2, 5, 6

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. IEEE CVPR*, 2009. 2, 3, 4

[9] J. W. Durban, M. J. Moore, G. Chiang, L. S. Hickmott, A. Bocconcelli, G. Howes, P. A. Bahamonde, W. L. Perryman, and D. J. LeRoi. Photogrammetry of blue whales with an unmanned hexacopter. *Marine Mammal Science*, 32(4), 2016. 2

[10] N. D. Fisheries. Drones detect sharks at lennox head, nsw, 2017. 7, 8

[11] K. Fukunaga. *Introduction to statistical pattern recognition*. Academic press, 2013. 2

[12] R. Girshick. Fast R-CNN. In *Proc. IEEE ICCV*, 2015. 2, 3

[13] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proc. IEEE CVPR*, 2014. 2, 3

[14] M. E. Goebel, W. L. Perryman, J. T. Hinke, D. J. Krause, N. A. Hann, S. Gardner, and D. J. LeRoi. A small unmanned aerial system for estimating abundance and size of antarctic predators. *Polar Biology*, 38(5), 2015. 2

[15] I. J. Goodfellow, D. Warde-farley, M. Mirza, A. Courville, and Y. Bengio. Maxout networks. In *Proc. IEEE ICML*, 2013. 2

[16] E. Harvey, D. Fletcher, and M. Shortis. Estimation of reef fish length by divers and by stereo-video: a first comparison of the accuracy and precision in the field on living fish under operational conditions. *Fisheries Research*, 57(3), 2002. 2

[17] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *Proc. IEEE CVPR*, 2016. 3, 4

[18] A. Hodgson, N. Kelly, and D. Peel. Unmanned aerial vehicles (UAVs) for surveying marine fauna: a dugong case study. *PloS one*, 8(11), 2013. 2

[19] J. A. Holmes, G. M. Cronkite, H. J. Enzenhofer, and T. J. Mulligan. Accuracy and precision of fish-count data from a dual-frequency identification sonar (DIDSON) imaging system. *ICES J. of Marine Science*, 63(3), 2006. 2

[20] Z. Kalal, K. Mikolajczyk, and J. Matas. Tracking-learning-detection. *IEEE Trans. PAMI*, 34(7):1409–1422, 2012. 8

[21] R. E. Kalman. A new approach to linear filtering and prediction problems. *J. of Basic Engineering*, 1960. 4

[22] C. Kim, F. Li, A. Ciptadi, and J. M. Rehg. Multiple hypothesis tracking revisited. In *Proc. IEEE ICCV*, 2015. 4

[23] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[24] D. Kinzey and T. Gerrodette. Distance measurements using binoculars from ships at sea: accuracy, precision and effects of refraction. *J. of Cetacean Research and Management*, 5(2):159–172, 2003. 2

[25] J. J. Kiszka, J. Mourier, K. Gastrich, and M. R. Heithaus. Using unmanned aerial vehicles (uavs) to investigate shark and ray densities in a shallow coral lagoon. *Marine Ecology Progress Series*, 560, 2016. 2

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. IEEE NIPS*, 2012. 2

[27] H. W. Kuhn. The HUNGARIAN method for the assignment problem. *Naval Research Logistics (NRL)*, 2(1-2), 1955. 4

[28] B. B. Le Cun, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Handwritten digit recognition with a back-propagation network. In *Proc. IEEE NIPS*, 1990. 2

[29] T. B. Letessier, P. J. Bouchet, and J. J. Meeuwig. Sampling mobile oceanic fishes and sharks: implications for fisheries and conservation planning. *Biological Reviews*, 92(2), 2017. 1, 2

[30] X. Li, M. Shang, H. Qin, and L. Chen. Fast accurate fish detection and recognition of underwater images with fast r-cnn. In *Proc. MTS/IEEE OCEANS*, 2015. 3

[31] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *Proc. IEEE CVPR*, 2017. 4

[32] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. *arXiv preprint arXiv:1708.02002*, 2017. 3, 4

[33] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Proc. ECCV*, 2014. 3, 4

[34] E. Liscio, H. Gaiser, and D. V. Maarten. Keras retinanet, 2017. `https://github.com/fizyr/keras-retinanet`. 4

[35] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. SSD: Single shot multibox detector. In *Proc. ECCV*, 2016. 3

[36] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proc. IEEE CVPR*, 2015. 3, 4

[37] A. Marburg and K. Bigham. Deep learning for benthic fauna identification. In *Proc. MTS/IEEE OCEANS*, 2016. 2

[38] E. F. Morais, M. F. M. Campos, F. L. Padua, and R. L. Carceroni. Particle filter-based predictive tracking for robust fish counting. In *Proc. IEEE Brazilian Symp. Computer Graphics and Image Processing*, 2005. 3

[39] E. Osherov and M. Lindenbaum. Increasing cnn robustness to occlusions by reducing filter support. In *Proc. IEEE ICCV*, 2017. 2

[40] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Trans. on knowledge and data engineering*, 2010. 4

[41] B. Pepik, R. Benenson, T. Ritschel, and B. Schiele. What is holding back convnets for detection? In *German Conf. on Pattern Recognition*, 2015. 2

[42] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *Proc. IEEE CVPR*, 2016. 3

[43] J. Redmon and A. Farhadi. YOLO9000:better, faster, stronger. *Proc. IEEE CVPR*, 2017. 3, 5

[44] D. Reid. An algorithm for tracking multiple targets. *IEEE transactions on Automatic Control*, 1979. 4

[45] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proc. IEEE NIPS*, 2015. 2, 3

[46] W. D. Robbins, V. M. Peddemors, S. J. Kennelly, and M. C. Ives. Experimental evaluation of shark detection rates by aerial observers. *PloS one*, 9(2), 2014. 2

[47] G. Shrivakshan, C. Chandrasekar, et al. A comparison of various edge detection techniques used in image processing. *IJCSI Int. J. of Computer Science Issues*, 9(5), 2012. 2

[48] C. Spampinato, Y.-H. Chen-Burger, G. Nadarajan, and R. B. Fisher. Detecting, tracking and counting fish in low quality unconstrained underwater videos. *VISAPP*, 2008. 2

[49] C. Spampinato, S. Palazzo, D. Giordano, I. Kavasidis, F.-P. Lin, and Y.-T. Lin. Covariance based fish tracking in real-life underwater environment. In *VISAPP*, 2012. 3

[50] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *IJML*, 15, 2014. 2

[51] M. Sung, S.-C. Yu, and Y. Girdhar. Vision based real-time fish detection using convolutional neural network. In *Proc. MTS/IEEE OCEANS*, 2017. 3, 6

[52] TheLittleRipper. Autonomous shark detecting UAV, 2017. 3

[53] S. Tripathi, Z. C. Lipton, S. Belongie, and T. Nguyen. Context matters: Refining object detection in video with recurrent neural networks. *arXiv preprint arXiv:1607.04648*, 2016. 4

[54] D. Ventura, M. Bruno, G. J. Lasinio, A. Belluscio, and G. Ardizzone. A low-cost drone based application for identifying and mapping of coastal fish nursery grounds. *Estuarine, Coastal and Shelf Science*, 171, 2016. 3

[55] L. Wan, M. Zeiler, S. Zhang, Y. L. Cun, and R. Fergus. Regularization of neural networks using dropconnect. In *Proc. IEEE ICML*, 2013. 2

[56] Y. Xie, G. Cronkite, and T. J. Mulligan. A split-beam echosounder perspective on migratory salmon in the fraser river: a progress report on the split-beam experiment at mission, bc, in 1995. *Pacific Salmon Commission*, 11, 1997. 2

[57] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *Proc. ECCV*, 2014. 3

[58] J. Zhou and C. M. Clark. Autonomous fish tracking by rov using monocular camera. In *Proc. IEEE Canadian Conf. Computer and Robot Vision*, 2006. 3

[59] Z. Zivkovic. Improved adaptive gaussian mixture model for background subtraction. In *IEEE Proc. ICPR*, 2004. 2