Marine-tree: A Large-scale Hierarchically Annotated Dataset for Marine Organism Classification

Tanya Boone-Sifuentes
Deakin University, Geelong, Australia
tboonesifuentes@deakin.edu.au

Asef Nazari
Deakin University, Geelong, Australia
asef.nazari@deakin.edu.au

Imran Razzak
UNSW, Sydney, Australia
imran.razzak@unsw.edu.au

Mohamed Reda Bouadjenek
Deakin University, Geelong, Australia
reda.bouadjenek@deakin.edu.au

Antonio Robles-Kelly
Deakin University, Geelong, Australia
antonio.robles-kelly@deakin.edu.au

Daniel Ierodiaconou
Deakin University, Geelong, Australia
daniel.ierodiaconou@deakin.edu.au

Elizabeth S. Oh
University of Tasmania, Australia
elizabeth.oh@utas.edu.au

ABSTRACT

This paper presents *Marine-tree*, a large-scale hierarchical annotated dataset for marine organism classification. Marine-tree contains more than 160k annotated images divided into 60 classes organised in a hierarchy-tree structure using an adapted CATAMI (Collaborative and Automated Tools for the Analysis of Marine Imagery and video) classification scheme. Images were meticulously collected by scuba divers using the RLS (Reef Life Survey) methodology and later annotated by experts in the field. We also propose a hierarchical loss function that can be applied to any multi-level hierarchical classification model, which takes into account the *parent-child* relationship between predictions and uses it to penalize inconsistent predictions. Experimental results demonstrate that *Marine-tree* and the proposed hierarchical loss function are a good contribution for both research in underwater imagery and hierarchical classification.

CCS CONCEPTS

• **Computing methodologies → Supervised learning by classification**.

KEYWORDS

Hierarchical Image Classification, CNNs, Deep Learning, Marine Image Classification.

ACM Reference Format:
Tanya Boone-Sifuentes, Asef Nazari, Imran Razzak, Mohamed Reda Bouadjenek, Antonio Robles-Kelly, Daniel Ierodiaconou, and Elizabeth S. Oh. 2022. *Marine-tree*: A Large-scale Hierarchically Annotated Dataset for Marine Organism Classification. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management (CIKM '22), October 17–21, 2022, Atlanta, GA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3511808.3557634

CIKM '22, October 17–21, 2022, Atlanta, GA, USA

ACM ISBN 978-1-4503-9236-5/22/10…$15.00
https://doi.org/10.1145/3511808.3557634

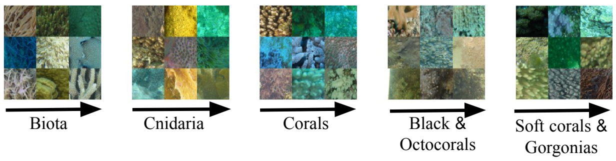
1 INTRODUCTION

Classifying underwater life is a paramount task for marine scientists [1, 2]. It is of vital importance to understand the associated anthropogenic effects on marine habitats if we want to preserve the natural ecosystem as well as to protect it from potential threats [3]. To ensure systematic and integrated ecological monitoring, marine biologists require to periodically obtain images and videos from different locations of subtidal and intertidal reef communities [4]. These images are normally obtained using Baited Remote Underwater Videos (BRUVs), Unmanned Aerial Vehicles (UAVs) [5], Remotely Operated Vehicle (ROVs) or scuba diver-derived photography. Thanks to these technological advancements, it is possible to obtain a large amount of images at high resolution and higher volumes. Nevertheless, for these images to be useful, it is necessary to label each organism on the image. This poses challenges that need to be addressed: (i) the annotation process is tedious and time-consuming [6], (ii) only experts in the field can label these images correctly due to the complex taxonomic structure of marine organisms, and (iii) the speed of labelling is often below the speed of obtaining new images resulting in a backlog of processing. This often results in a sub-sample of images being processed and a bottleneck between data capture and information delivery to inform the state of the marine environment.

In order to support marine researchers, some authors have proposed to automate this labelling process with Machine Learning algorithms, specifically, using Convolutional Neural Nets (CNNs) [7–9]. However, CNNs need to be trained with large amounts of data to achieve reliable results. This is an obstacle because most available marine datasets have a reduced number of images and classes, which often don't reflect the complex marine environment, thus limiting the data available for training. Additionally, marine images are impacted by light attenuation and quality of the water column impacting illumination with high variability making imagery challenging for automated classification approaches. These images result in poor quality due to weather conditions, tide and selective absorption of light [10, 11].

The aim of this paper is to present a novel marine dataset for marine Multi-level Hierarchical Classification (MLHC) of benthic substrate and invertebrates, which we call *Marine-tree* annotated up to a five-level of hierarchy as shown in Figure 1. *Marine-tree* due to

(a) Examples of images in *Marine-tree*.



(b) A snapshot of one root-to-leaf branch.

Figure 1: Examples of *Marine-tree* images.

its properties such as flexibility and variety is an important contribution to existing marine image datasets. Additionally, we propose a hierarchical loss function and evaluate its performance when applied with state-the-art MLHC methods. The key contributions of this work are: (i) a publicly available marine dataset (*Marine-tree*) containing more than 160K images labelled up to five levels of hierarchy divided in two according to climate (Temperate and Tropical), (ii) a benchmark evaluation of *Marine-tree* using state-of-the-art MLHC methods, and (iii) a hierarchical loss function that enforces consistency of predictions.

2 DATASET

In this section, we provide a detailed description of the *Marine-tree* dataset[1].

2.1 Source

Marine-tree dataset was generated from images provided by the School of Life and Environmental Science (LES) at Deakin University, Warnambool. Images were obtained through Reef Life Survey (RLS) [2], which has become a common approach for monitoring subtidal reef communities. The RLS program has surveyed more than 3,300 sites around the world by a combination of scientists and trained citizens. RLS divers use the underwater visual census (UVC) technique to record the structure and abundance of the fish and invertebrate communities along a 50 meter (m) transect line laid along a depth contour of reef. To capture habitat data, series of digital photoquadrats are collected at 2.5 m intervals along the same transect. Photoquadrats are taken from approximately 50 cm above the seabed (usually sufficient to encompass an area of approximately $0.3m \times 0.3m$). A flash with a diffuser, or a strobe, is used to improve the colour spectrum of the photograph unless there is sufficient ambient light (at very shallow depths) or suspended particulate matter causes too much back scatter. Raw images are RGB and have a resolution of around $3,000 \times 4,000$ pixels. These photoquadrats (PQs) are later annotated with points classified using an adapted CATAMI classification scheme [12]. This classification provides annotations for morphologically distinct algae, corals, sessile invertebrates, and substratum types to be produced and stored for later analysis. Figure 2 is a world map representing the distribution of the locations where images were captured.

[1]Dataset: https://github.com/tboone91/Marine-tree
[2]Website: https://reeflifesurvey.com/

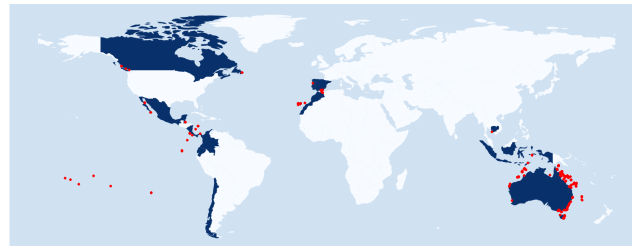


Figure 2: Location of RLS campaigns: Red dots represent approximate location which belongs to the countries in blue.

2.2 Annotation process

Images were catalogued and uploaded to the online annotation platform Squidle+ [3]. Points were overlaid on the images in the user interface and then annotated by scientific experts according to the class of benthos that each point intersects. Exemplar images for each class in the classification scheme are viewable in Squidle+ to reference during the annotation process. As the underwater conditions and the quality of the camera used to capture the photoquadrat is not controlled, image quality varies. Brightness and contrast can be adjusted manually during the annotation process in the user interface of the annotation platform to aid classification of the points, however changes are not saved to the image file. Additionally, images sometimes contain a type of occlusion such as the survey transect line, in which case the point is moved so that it no longer overlays the occlusion. If the point cannot be assigned a class, due to poor image quality, it is discarded.

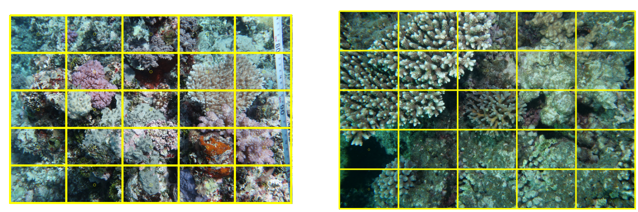


Figure 3: An example of two RLS diver photoquadrats used to build the dataset. The images are divided in a 5x5 grid in the cropping process and the cells with annotations are kept.

2.3 Dataset construction & analysis

After gathering the images, the first step was to crop each object around its annotation provided in the RLS datasheet. Each image would provide a maximum of 25 crops (most images had around 6 annotations). We constructed a grid of 5×5 as shown in Figure 3. We saved the crops that were annotated and appended the coordinates on the image filename. We eliminated crops that were too blurry, were not annotated up to level three or had too much debris resulting in a total of 161,180 cropped images. Figure 1a shows examples of cropped images.

The next step was to assign the annotation labels following the CATAMI classification scheme. Crops that were not labelled until level three, were discarded. Figure 1b shows the maximum levels of hierarchy depth that can be assigned to a single image, and Figure

[3]Website: https://squidle.org/

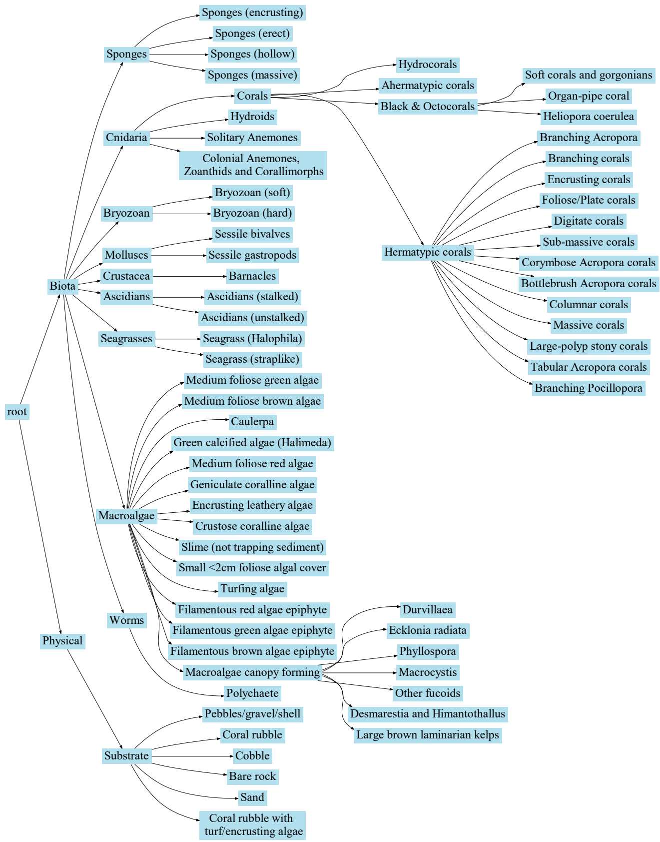


Figure 4: *Marine-tree* dataset classification scheme[5].

4 shows the taxonomy of *Marine-tree* [4]. As mentioned before, the total number of images were divided in two subsets according to the climate they belong (Tropical or Temperate). Some species were more abundant in, or restricted to one climate only. Temperate is approximately double the size of Tropical. The cropped image size depends on the size of the original image but they are around 600x400 pixels. In our experiments, we need to consider a generic taxonomic that is complete to all levels. This led us to apply padding, where if an image is labelled only at the third level, we propagate this label in the fourth and fifth level. Table 1 shows statistics of *Marine-tree*.

Table 1: Number of images and classes per level and partition.

Hierarchy level	Temperate	Tropical	Combined
	118,637 Images	42,548 Images	161,185 Images
	Classes		
ℓ_1	2	2	2
ℓ_2	10	9	10
ℓ_3	37	34	38
ℓ_4	44	38	46
ℓ_5	50	52	60

[4]In this paper we make use of the word taxonomy as a general term knowing that Figure 4 represents a custom classification scheme.
[5]https://github.com/tboone91/Marine-tree/blob/main/taxonomy_tree.pdf

Table 2: An overview of annotated marine image datasets.

Dataset name	Environment	Recording type	Classes	Images	Labeling	Hiera.?
F4K - Species [17]	Reef	Stationary	23	27,350	Masks	N
J-EDI [18]	Deep sea	ROV	-	1,500,000	Image level	N
HabCam [19]	Shelf Sea	Towing	-	2,500,000	Bounding box	N
Croatian Fish Dataset [20]	Deep Sea	Various	12	794	Bounding box	N
QUT Fish Dataset [21]	Controlled	Various	468	3,960	Bounding Box	N
BENTHOZ-2015 [14]	Reef	AUV	148	9,974	Points	Y
Tasmania CPC [15]	Reef	AUV	19	1258	Points	Y
Moorea LB [16]	Reef	Photoquadrat	9	2,055	Points	N
Brackish Dataset [22]	Brackish strait	Stationary	6	14,518	Bounding box	N
Marine-tree Combined	Various	Photoquadrat	60	161,185	Image level	Y
Marine-tree Tropical	Various	Photoquadrat	52	42,548	Image level	Y
Marine-tree Temperate	Various	Photoquadrat	50	118,637	Image level	Y

2.4 Comparison with other datasets

Classification of benthic organisms is of critical importance for the preservation and monitoring of marine habitats. Corals are vital for protecting coastal areas and many species rely on them as a source of food or protection [13]. Several marine datasets have been created for fish classification but only a few for benthic organisms, in which we can find BENTHOZ-2015 [14], Tasmania Coral Point Count [15] and The Moorea Labeled Corals [16]. One important characteristic about BENTHOZ-2015 and Tasmania Coral Point Count is that they're labelled hierarchically but their downside is that they contain a very limited number of images. Table 2 summarizes the aforementioned marine datasets along with our proposed *Marine-tree* dataset (combined, tropical, and temperate).

3 EXPERIMENTAL EVALUATION

In this section, we evaluate the performance of several MLHC models on *Marine-tree*.

3.1 Experimental setup

Hierarchical Baselinesi pero Models: To benchmark *Marine-tree*, we have implemented seven state-of-the-art MLHC models: (1) ***n*-nets:** n independent networks for each hierarchy level; (2) ***n*-outs:** a single network with n output layers; (3) **B-CNN:** Branch-CNN [23]; (4) **B-CNN_v2:** a variant of B-CNN, which takes the ReLu activation of every branch output and uses them as the input of the next Fully-Connected layer; (5) **Bi-CNN:** Bilinear-CNN described in [24]; (6) **MLPH:** Multi-linear Pooling with Hierarchy described in [25]; (7) **Flat classifier**: flat classification approach that consists of completely ignoring the class hierarchy, typically predicting only classes at the leaf nodes and use this prediction to infer lower levels.

Hierarchical Loss: It is important for an MLHC model to ensure consistency between predictions across all hierarchical levels. One way to achieve this is by penalizing inconsistent predictions during training. We propose to accomplish this by adding a penalty term to the multi-class cross entropy loss function as follows:

$$\frac{1}{m}\sum_{j=1}^{m}\sum_{i=1}^{n}\mathcal{L}(\hat{y}_{\ell_i}^{(j)}, y_{\ell_i}^{(j)}) - \lambda \log(\check{y}_{\ell_i}^{(j)}\, \hat{y}_{\ell_i}^{(j)}) \qquad (1)$$

where $\mathcal{L}(\bullet, \bullet)$ denotes the cross-entropy function, $\hat{y}_{\ell_i}^{(j)}$ and $y_{\ell_i}^{(j)}$ are respectively the prediction and ground truth for for training example j and level ℓ_i in the taxonomy, $\check{y}_{\ell_i}^{(j)}$ is the predicted probability value for the true parent of the predicted $\hat{y}_{\ell_i}^{(j)}$, and λ is a hyperparameter to tune. We note that the penalty term of the loss

Table 3: Performance comparison of the used MLHC models on *Marine-tree* dataset.

Network	HP	HR	HF1	Acc@ℓ_1	Acc@ℓ_2	Acc@ℓ_3	Acc@ℓ_4	Acc@ℓ_5	Cons.	EM
Tropical										
n-nets	0.5538	0.6514	0.5915	0.8473	0.6474	0.5000	0.4514	0.4197	0.5294	0.2803
n-outs	0.5534	0.5935	0.5693	0.8427	0.6457	0.4770	0.4376	0.4120	0.7568	0.3561
B-CNN[23]	0.5549	0.6424	0.5897	0.8155	0.67042	0.5129	0.4329	0.4131	0.5018	0.2605
B-CNN_v2	0.4858	0.5148	0.4975	0.8224	0.6191	0.4074	0.3451	0.2973	0.7683	0.2697
Bi-CNN	0.5471	0.5936	0.5652	0.8434	0.6474	0.4730	0.4283	0.3977	0.73497	0.3347
MLPH [25]	0.5995	0.6165	0.6066	0.8533	0.6807	0.5336	0.4892	0.4553	0.8813	0.4208
Flat classifier	0.5568	0.5568	0.5568	0.8071	0.6420	0.4781	0.4425	0.4143	1	0.4143
Temperate										
n-nets	0.6899	0.7522	0.7148	0.9217	0.8511	0.6329	0.5683	0.5689	0.6592	0.4653
n-outs	0.6940	0.7157	0.7028	0.9192	0.8468	0.6204	0.5579	0.5560	0.8573	0.5238
B-CNN[23]	0.6939	0.7389	0.7121	0.9138	0.8544	0.6325	0.5569	0.5646	0.7182	0.4833
B-CNN_v2	0.6494	0.6627	0.6550	0.9172	0.8463	0.5488	0.4808	0.4812	0.8722	0.4629
Bi-CNN	0.6879	0.7160	0.6990	0.9143	0.8426	0.6130	0.5557	0.5525	0.8277	0.5179
MLPH [25]	0.721	0.7300	0.7249	0.9285	0.8612	0.6433	0.5918	0.5909	0.9279	0.5722
Flat classifier	0.7086	0.7086	0.7086	0.9190	0.8459	0.6289	0.5749	0.5745	1	0.5745
Combined										
n-nets	0.6425	0.711	0.6697	0.8968	0.7978	0.5811	0.5233	0.5129	0.6329	0.4121
n-outs	0.6590	0.6986	0.6748	0.9046	0.8064	0.5831	0.5328	0.5259	0.7759	0.4682
B-CNN [23]	0.6453	0.7029	0.6685	0.8841	0.7934	0.5813	0.5111	0.5141	0.6422	0.4107
B-CNN_v2	0.5896	0.6037	0.5956	0.8975	0.7654	0.4916	0.42398	0.4100	0.8545	0.3962
Bi-CNN	0.6549	0.7029	0.6739	0.8944	0.8005	0.5831	0.5303	0.5248	0.7331	0.4530
MLPH [25]	0.6729	0.6867	0.6787	0.9033	0.8025	0.5948	0.5436	0.5366	0.8972	0.5116
Flat classifier	0.6617	0.6617	0.6617	0.8956	0.7915	0.5735	0.5260	0.5217	1	0.5217

Table 4: Performance of hierarchical baseline models using proposed loss on *Marine-tree* (Combined).

Network	HP	HR	HF1	Acc@ℓ_1	Acc@ℓ_2	Acc@ℓ_3	Acc@ℓ_4	Acc@ℓ_5	Cons.	EM
Combined										
n-nets_HL	0.6797	**0.7446**	**0.7053**	0.9082	0.8102	**0.6174**	**0.5769**	**0.5670**	0.6665	0.4657
n-outs_HL	0.6497	0.6666	0.6566	0.8932	0.7887	0.5648	0.5142	0.5065	0.8846	0.4808
MLPH_HL	**0.6887**	0.6979	0.6927	**0.9090**	**0.8146**	0.6143	0.5641	0.5552	**0.9238**	**0.5369**
B-CNN_v2_HL	0.5941	0.6080	0.6000	0.8946	0.7844	0.4865	0.4253	0.4129	0.8713	0.3977

function aims to maximize both $\check{y}_{\ell_i}^{(j)}$ and $\hat{y}_{\ell_i}^{(j)}$ to guarantee consistent predictions while not being applied for ℓ_1 – first level of the taxonomy.

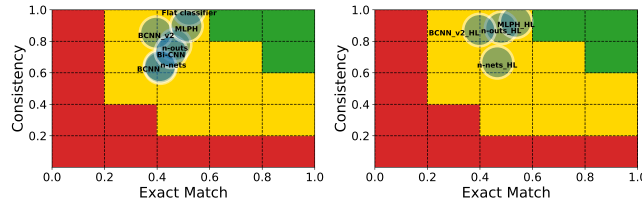
Evaluation metrics: We report the performance of the analyzed MLHC models using Hierarchical metrics [26, 27], Accuracy per level, Consistency (average of consistent predictions w.r.t. taxonomy), and Exact Match (average of predictions that are both correct and consistent).

Implementation details[6]**:** The MLHC models were implemented using either VGG16 or VGG19 [28] (according to original implementation) pretrained on the ImageNet dataset [29]. We used an image size of 64x64 pixels while fine-tuning the pre-trained models. All models were implemented using Tensorflow/Keras [30] and performed on a Linux Ubuntu 18.04.1 LTS Dual Intel(R) Xeon(R) Silver 4114 CPU @2.20GHz with a GPU NVIDIA Tesla V100. We used a 12% stratified seeded split for the validation and test from the training set, a batch size of 128, and Adam optimizer [31] with a learning rate of 1e-4.

3.2 Results

Performance: Table 3 shows performance of the above MLHC models using Hierarchical Precision (HP), Hierarchical Recall (HR), Hierarchical F1-Score (HF1), Accuracy per level (ℓ_1 to ℓ_5), Consistency (Cons.) and Exact Match (EM) for *Marine-tree* (Combined) and its subsets. To assess the effectiveness of the hierarchical loss (HL) described in Equation 1, we performed experiments on *Marine-tree* Combined dataset for *n*-nets, *n*-outs, MLPH, and B-CNN_v2. The obtained results are described in Table 4.

[6]Code repository: https://github.com/tboone91/Marine-tree



(a) MLHC models. **(b) MLHC models + Hierarchy loss.**

Figure 5: Scatter plots of Exact Match vs Consistency on *Marine-tree* (Combined).

Discussion: In light of our experimental results, we make the following key remarks: (1) The Flat classifier and MLPH show the best performance in terms of Consistency and Exact Match but MLPH reports higher HF1 and higher accuracy on the last level; (2) For Tropical, MLPH reports high Consistency but Exact Match remains low (≈ 40%). This is due to the architecture being unable to get higher accuracy for the last levels; (3) Overall, *n*-nets and BCNN are the models that perform the poorest. *n*-nets is the only model that doesn't share weights with other levels; (4) B-CNN_v2 performs better than BCNN for all subsets. (5) Regarding the loss function, *n*-outs increases Consistency by 10% but the accuracy is slightly lower. *n*-nets, MLPH, and B-CNN_v2 report improvement in all metrics, and *n*-nets is the model that benefited the most from this loss function.

Finally, Figure 5 shows a plot of Exact Match vs. Consistency. Ideally, in MLHC, we are looking for a model with both high Consistency and high Exact Match. We observe that while our hierarchy allows improvement both in terms of accuracy and consistency, there remain a lot of margin in terms of improvement in particular to design a top-down divide-and-conquer strategy to constraint predictions.

4 CONCLUSIONS & FUTURE WORK

In this work, we described a large-scale dataset for marine classification. Additionally, a benchmark evaluation of the dataset is performed using several MLHC models. Also, we presented a novel loss function that tackles the problem of inconsistency between predictions that would penalize inconsistent predictions. Our results show that our proposed hierarchical loss function improve the performance, specially Consistency and Hierarchical F1-Score in comparison with the baselines. Results demonstrate that there is a room to develop new effective and efficient algorithms to improve MLHC. Future work involve testing the method described in [32] on the *marine-tree* dataset and investigating MLHC for clustering in Information Retrieval [33, 34].

Acknowledgements: The image and annotation repository used was provided by Understanding Marine Imagery (UMI) a facility within Australia's Integrated Marine Observing System (IMOS). IMOS is enabled by the National Collaborative Research Infrastructure Strategy (NCRIS). It is operated by a consortium of institutions as an unincorporated joint venture, with the University of Tasmania as Lead Agent.

REFERENCES

[1] Lin Li, Bo Dong, Eric Rigall, Tao Zhou, Junyu Dong, and Geng Chen. Marine animal segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[2] D Ierodiaconou, J Monk, A Rattray, L Laurenson, and VL Versace. Comparison of automated classification techniques for predicting benthic biological communities using hydroacoustics and video observations. *Continental Shelf Research*, 31(2):S28–S38, 2011.

[3] Daniel Ierodiaconou, L Laurenson, S Burq, and M Reston. Marine benthic habitat mapping using multibeam data, georeferencedvideo and image classification techniques in victoria, australia. *Journal of Spatial Science*, 52(1):93–104, 2007.

[4] D Ierodiaconou, S Wines, P Carnell, P Tinkler, B Allan, J Carey, M Young, S Howe, J Pocklington, et al. An enhanced signs of healthy parks monitoring program for victoria's marine national parks and marine sanctuaries: Point addis marine national park. parks victoria technical series 114. front cover: Port phillip heads marine national park photo: Museum victoria. Technical Report 114, Parks Victoria Technical Series, 2020.

[5] Richard Zavalas, Daniel Ierodiaconou, David Ryan, Alex Rattray, and Jacquomo Monk. Habitat classification of temperate marine macroalgal communities using bathymetric lidar. *Remote Sensing*, 6(3):2154–2175, 2014.

[6] Bastiaan J Boom, Phoenix X Huang, Jiyin He, and Robert B Fisher. Supporting ground-truth annotation of image datasets using clustering. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pages 1542–1545. IEEE, 2012.

[7] Zheng Cao, Jose C Principe, Bing Ouyang, Fraser Dalgleish, and Anni Vuorenkoski. Marine animal classification using combined cnn and hand-designed image features. In *OCEANS 2015-MTS/IEEE Washington*, pages 1–6. IEEE, 2015.

[8] Huimin Lu, Yujie Li, Tomoki Uemura, Zongyuan Ge, Xing Xu, Li He, Seiichi Serikawa, and Hyoungseop Kim. Fdcnet: filtering deep convolutional network for marine organism classification. *Multimedia tools and applications*, 77(17):21847–21860, 2018.

[9] Cheng Siong Chin, JianTing Si, Anthony S Clare, and Maode Ma. Intelligent image recognition system for marine fouling using softmax transfer learning and deep convolutional neural networks. *Complexity*, 2017, 2017.

[10] Chongyi Li, Chunle Guo, Wenqi Ren, Runmin Cong, Junhui Hou, Sam Kwong, and Dacheng Tao. An underwater image enhancement benchmark dataset and beyond. *IEEE Transactions on Image Processing*, 29:4376–4389, 2019.

[11] Shaojian Song, Jingxu Zhu, Xiuhua Li, and Qingbao Huang. Integrate msrcr and mask r-cnn to recognize underwater creatures on small sample datasets. *IEEE Access*, 8:172848–172858, 2020.

[12] Franziska Althaus, Nicole Hill, Renata Ferrari, Luke Edwards, Rachel Przeslawski, Christine HL Schönberg, Rick Stuart-Smith, Neville Barrett, Graham Edgar, Jamie Colquhoun, et al. A standardised vocabulary for identifying benthic biota and substrata from underwater imagery: the catami classification scheme. *PloS one*, 10(10):e0141039, 2015.

[13] NS Barrett and Jacquomo Monk. Progress towards a nationally integrated benthic biodiversity monitoring program for australia's marine realm. Technical report, Institute for Marine and Antarctic Studies, University of Tasmania, 2021.

[14] Michael Bewley, Ariell Friedman, Renata Ferrari, Nicole Hill, Renae Hovey, Neville Barrett, Ezequiel M Marzinelli, Oscar Pizarro, Will Figueira, Lisa Meyer, et al. Australian sea-floor survey data, with images and expert annotations. *Scientific data*, 2(1):1–13, 2015.

[15] Australian centre for field robotics. tasmania coral point count. http://marine.acfr.usyd.edu.au/datasets/. Accessed: 2021-07-19.

[16] Oscar Beijbom, Peter J Edmunds, David I Kline, B Greg Mitchell, and David Kriegman. Automated annotation of coral reef survey images. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1170–1177. IEEE, 2012.

[17] Robert B Fisher, Yun-Heh Chen-Burger, Daniela Giordano, Lynda Hardman, Fang-Pang Lin, et al. *Fish4Knowledge: collecting and analyzing massive coral reef fish video data*, volume 104. Springer, 2016.

[18] Tomoki Sasaki, Shuko Azuma, Shoko Matsuda, Anri Nagayama, Moritaka Ogido, Hideaki Saito, and Yasunori Hanafusa. Jamstec e-library of deep-sea images (j-edi) realizes a virtual journey to the earth's unexplored deep ocean. In *AGU Fall Meeting Abstracts*, volume 2016, pages IN53C–1911, 2016.

[19] Northeast fisheries science center, 2021: Habitat mapping camera (habcam). https://www.fisheries.noaa.gov/inport/item/27598. Accessed: 2021-07-19.

[20] Jonas Jäger, Marcel Simon, Joachim Denzler, Viviane Wolff, Klaus Fricke-Neuderth, and Claudia Kruschel. Croatian fish dataset: Fine-grained classification of fish species in their natural habitat. *Swansea: Bmvc*, 2015.

[21] Kaneswaran Anantharajah, ZongYuan Ge, Chris McCool, Simon Denman, Clinton Fookes, Peter Corke, Dian Tjondronegoro, and Sridha Sridharan. Local inter-session variability modelling for object classification. In *IEEE Winter Conference on Applications of Computer Vision*, pages 309–316. IEEE, 2014.

[22] Malte Pedersen, Joakim Bruslund Haurum, Rikke Gade, and Thomas B Moeslund. Detection of marine animals in a new underwater dataset with varying visibility. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 18–26, 2019.

[23] Xinqi Zhu and Michael Bain. B-cnn: branch convolutional neural network for hierarchical classification. *arXiv preprint arXiv:1709.09890*, 2017.

[24] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear cnn models for fine-grained visual recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 1449–1457, 2015.

[25] Yuqi Huo, Yao Lu, Yulei Niu, Zhiwu Lu, and Ji-Rong Wen. Coarse-to-fine grained classification. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1033–1036, 2019.

[26] Aris Kosmopoulos, Ioannis Partalas, Eric Gaussier, Georgios Paliouras, and Ion Androutsopoulos. Evaluation measures for hierarchical classification: a unified view and novel approaches. *Data Mining and Knowledge Discovery*, 29(3):820–865, 2015.

[27] Svetlana Kiritchenko, Stan Matwin, Richard Nock, and A. Fazel Famili. Learning and evaluation in the presence of class hierarchies: Application to text categorization. In Luc Lamontagne and Mario Marchand, editors, *Advances in Artificial Intelligence*, pages 395–406, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[28] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[29] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[30] Bo Pang, Erik Nijkamp, and Ying Nian Wu. Deep learning with tensorflow: A review. *Journal of Educational and Behavioral Statistics*, 45(2):227–248, 2020.

[31] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[32] Tanya Boone-Sifuentes, Mohamed Reda Bouadjenek, Imran Razzak, Hakim Hacid, and Asef Nazari. A Mask-based Output Layer for Multi-level Hierarchical Classification. In *Proceedings of the 31th ACM International Conference on Information & Knowledge Management*, CIKM '22, New York, NY, USA, 2022. Association for Computing Machinery.

[33] Mohamed Reda Bouadjenek, Scott Sanner, and Yihao Du. Relevance- and interface-driven clustering for visual information retrieval. *Information Systems*, 94:101592, 2020.

[34] Mohamed Reda Bouadjenek and Scott Sanner. Relevance-Driven Clustering for Visual Information Retrieval on Twitter. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, CHIIR '19, page 349–353, New York, NY, USA, 2019. Association for Computing Machinery.