Overview of the ImageCLEFcoral 2020 Task: Automated Coral Reef Image Annotation

Jon Chamberlain[1], Antonio Campello[2], Jessica Wright[1], Louis Clift[1], Adrian Clark[1] and Alba García Seco de Herrera[1]

[1] School of Computer Science and Electronic Engineering, University of Essex, Colchester, UK
[2] Wellcome Trust, UK
Corresponding author: jchamb@essex.ac.uk

Abstract. This paper presents an overview of the ImageCLEFcoral 2020 task that was organised as part of the Conference and Labs of the Evaluation Forum - CLEF Labs 2020. The task addresses the problem of automatically segmenting and labelling a collection of underwater images that can be used in combination to create 3D models for the monitoring of coral reefs. The data set comprises 440 human-annotated training images, with 12,082 hand-annotated substrates, from a single geographical region. The test set comprises a further 400 test images, with 8,640 substrates annotated, from four geographical regions ranging in geographical similarity and ecological connectedness to the training data (100 images per subset). 15 teams registered, of which 4 teams submitted 53 runs. The majority of submissions used deep neural networks, generally convolutional ones. Participants' entries showed that some level of automatically annotating corals and benthic substrates was possible, despite this being a difficult task due to the variation of colour, texture and morphology between and within classification types.

Keywords: ImageCLEF, image annotation, image labelling, classification, segmentation, coral reef image annotation, marine image annotation

1 Introduction

Coral reef systems are delicate natural environments, formed of highly complex non-uniform structures that support the biodiversity found in tropical coral reefs. Coral reefs also form a vital source of income and food for over 500 million people, providing ecological goods and services such as food, coastal protection, new biochemical compounds, and recreation with an estimated value of around \$352,000 ha^{-1} y^{-1} [1].

However, there has been a steady decline in coral reefs in recent years [2]. Coral reefs are threatened by global stressors such as climate change and subsequent extreme weather events, as well as by local anthropogenic threats such

as overfishing and destructive fishing, watershed pollution, and reef removal for coastal development. Currently, more than 85% of the reefs within the Coral Triangle region are at risk of disappearing [3, 4].

Coral reef community composition is an essential element for monitoring reef health and the importance of automated data collection, 3D analysis and large-scale data processing are increasingly being recognised [5]. In 2017, Chamberlain et al. at the University of Essex developed a novel multi-camera system to scale up previous data capture approaches [6] by acquiring imagery from several viewpoints simultaneously. Results showed that accurate data models were created in a fraction of the time and complex structures were more accurately reconstructed. The increasing use of large-scale modelling of environments has driven the need to have such models labelled, with annotated data essential for machine learning techniques to automatically identify areas of interest, assess community composition and monitor phase shifts within functional groups.

The composition of marine life on a coral reef varies globally. Within the Coral Triangle, a region that encloses more than 86,500km^2 of coral reef area and includes the world's highest marine biodiversity, there are over 76% of all coral species and more than 3,000 fish species [3]. The Western Indian Ocean, and more specifically the Northern Mozambique Channel (NMC), is a centre of high diversity for hard corals and reef fauna [7] and forms an evolutionary distinct region within the Indian Ocean, but the diversity shows high resemblance with the diversity found in the Coral Triangle region. Coral reef fauna from the Caribbean within the Atlantic Ocean, is strongly delineated from (and shows low affinity with biodiversity found in) the Indian Ocean [8].

Geographically distinct regions can contain the same species or genera with entirely different morphological features and traits. The variety in both environmental conditions and competitive niche filling can lead to changes in phenotypic expression, which makes the task of identifying them difficult without an extensive training image set.

As part of ImageCLEF 2019 [9], the ImageCLEFcoral task required participants to automatically annotate and localise a collection of images with types of benthic substrate, such as hard coral and sponge. The training set and test sets contained images from the same coral reef [10].

Participants' entries showed that some level of automatically annotating corals and benthic substrates was possible, despite this being a difficult task due to the variation of colour, texture and morphology between and within classification types.

This year, as part of ImageCLEF 2020 [11], the volume of training data was increased and there were four subsets of test data ranging in geographical similarity and ecological connectedness to the training data. The intention was not only to assess how accurately the images could be annotated, but also how transferable the algorithms were between datasets collected from different geographical regions with different community compositions.

2 Tasks

The annotation task is different from other image classification and marine substrate classification tasks [12–14]. Firstly, the images are collected using low-cost action cameras (approx. £200 per camera) with a fixed lens and firing on time-lapse or extracted as stills from video. The effect of this on the imagery is that there is some blurring, the colour balance is not always correct (as the camera adjusts the white balance automatically based on changing environmental variables) and final image quality is lower than what could be achieved using high-end action cameras or DSLRs. However, the images can be used for reconstructing a 3D model and therefore have useful information in the pipeline. Low cost cameras were used to show this approach could be replicated affordably for future projects.

Following the success of the first edition of the ImageCLEFcoral task [10], in 2020 participants were again asked to devise and implement algorithms for automatically annotating regions in a collection of images containing several types of benthic substrate, such as hard coral or sponge. The images were captured using an underwater multi-camera system developed at the Marine Technology Research Unit at the University of Essex (MTRU), UK[3].

The ground truth annotations of the training and test sets were made by a combination of marine biology MSc students at the University of Essex and experienced marine researchers. All annotations were double checked by an experienced coral reef researcher. The annotations were performed using a web-based tool, initially developed in a collaborative project with London-based company Filament Ltd and subsequently extended by one of the organisers. This tool was designed to be simple to learn, quick to use and allows many people to work concurrently (full details are presented in the ImageCLEFcoral 2019 overview [10]).

The overall task comprises two subtasks:

- *Subtask 1*: Coral reef image annotation and localisation;
- *Subtask 2*: Coral reef image pixel-wise parsing.

In the "coral reef image annotation and localisation" subtask, the annotation is a bounding box, with sides parallel to the edges of the image, around identified features. In the "coral reef image pixel-wise parsing" subtask, participants submit a series of boundary image coordinates which form a single polygon around each identified feature (these polygons should not have self-intersections). Participants were invited to make submissions for either or both tasks.

As in the first edition, algorithmic performance is evaluated on the unseen test data using the popular intersection over union metric from the PASCAL VOC[4] exercise. This computes the area of intersection of the output of an algorithm and the corresponding ground truth, normalizing that by the area of their union to ensure its maximum value is bounded.

[3] https://essexnlip.uk/marine-technology-research-unit/
[4] http://host.robots.ox.ac.uk/pascal/VOC/

3 Collection

The data set comprises 440 human-annotated training images, with 12,082 substrates, from the Wakatobi Marine Reserve, Indonesia; this is the complete training and test sets as used in the ImageCLEFcoral 2019 task. The test set comprises a further 400 test images (see Figure 1), with 8,640 substrates annotated, from four geographical regions, 100 images per subset:

1. Wakatobi Marine Reserve, Indonesia – the same location as the training images;
2. Spermonde archipelago, Indonesia – geographically similar location to the training set;
3. Seychelles, Indian Ocean – geographically distinct but ecologically connected coral reef;
4. Dominica, Caribbean – geographically and ecologically distinct rocky reef.

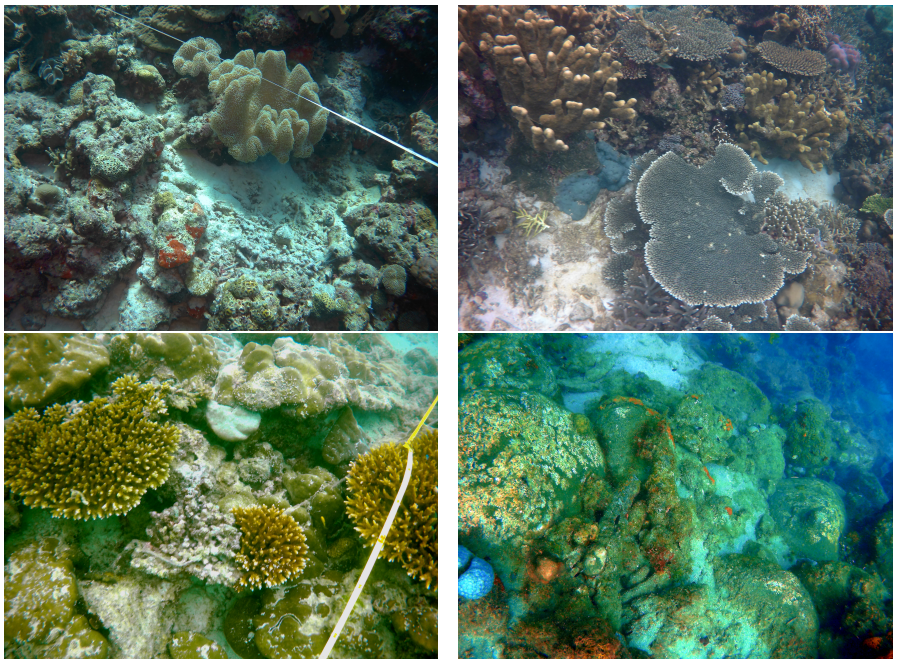

Fig. 1. Representative images from the 4 regions in the test dataset: same location as the training set (upper left); geographically similar (upper right); geographically distinct but ecologically connected (lower left); geographically and ecologically distinct (lower right).

The images are part of a monitoring collection and therefore many have a tape measure running through a portion of the image. As in 2019, the data

Table 1. Distribution of classified pixels for training data and different subsets of test data: same location, similar location, ecologically similar, ecologically distinct, and the four test sets combined.

Substrate	Training	Same	Similar	Eco_similar	Distinct	Combined
algae_macro_or_leaves	0.12	0.07	0.10	0.03	8.41	2.15
fire_coral_millepora	0.03	0.01	0.00	0.00	0.01	0.01
hard_coral_boulder	2.91	1.57	4.00	15.7	0.88	5.54
hard_coral_branching	1.93	2.66	14.79	4.34	0.00	5.45
hard_coral_encrusting	0.82	1.33	3.15	0.01	1.50	1.50
hard_coral_foliose	0.18	0.21	0.15	0.47	0.00	0.21
hard_coral_mushroom	0.10	0.05	0.00	0.00	0.00	0.01
hard_coral_submassive	0.40	0.40	12.54	0.11	0.01	3.26
hard_coral_table	0.03	0.10	5.35	0.00	0.00	1.36
soft_coral	8.69	7.09	0.03	0.01	0.00	1.78
soft_coral_gorgonian	0.26	0.14	0.00	0.00	0.00	0.04
sponge	1.42	1.63	0.36	0.01	5.05	1.79
sponge_barrel	0.30	0.10	0.00	0.00	1.95	0.53
unclassified	82.81	84.66	59.53	79.33	82.19	76.37

set comprises an area of underwater terrain. Many images contain the same ground features captured from different viewpoints. Each image contains some of the same thirteen types of benthic substrates as in 2019, namely hard coral — branching, submassive, boulder, encrusting, table, foliose, mushroom; soft coral; gorgonian sea fan (soft coral); sponge; barrel sponge; fire coral (millepora); algae (macro or leaves).

The test set from the same area as the training set will give an indication as to how well a submitted algorithm can localise and classify marine substrate, i.e., the maximum performance. We hypothesise that performance will deteriorate with other test subsets as the composition, morphology and identifying features of the substrate change and exhibit less similarity with the training data.

3.1 Collection Analysis

An important consideration when testing across the datasets is that the benthic composition will be different in the different locations, in addition to different species and morphologies being present and the total coverage of benthic fauna (represented by the total coverage of pixels in an image).

Analysis shows that the community distribution is similar in the same location test dataset to the training dataset, both in terms of structure and cover. The similar location test dataset shows a much higher distribution of hard corals and lower distribution of soft corals and sponge, with considerably higher coverage, indicative of a healthy coral reef. The geographically distinct but ecologically connected test set had a high distribution of hard corals in composition and similar coverage, indicative of a recovering coral reef. The geographically and ecologically distinct had a higher distribution of sponge and algae, commonly found in Caribbean reefs that suffer human and environmental impacts, and higher coverage indicative of a phase shift away from hard coral towards a sponge/algae dominated reef (see Table 1).

Table 2. Participating groups in ImageCLEFcoral task in 2020. Participants marked with a star participated also in 2019.

Team	Institution	# Runs T1	# Runs T2
FAV ZČU PiVa [18]	University of West Bohemia, Czechia	10	10
FAV ZČU CV [19]	University of West Bohemia, Czechia	2	1
HHUD* [20]	Heinrich-Heine-Universität Duesseldorf, Germany	10	0
FHD [21]	University of Applied Sciences and Arts Dortmund, Germany	10	10

4 Evaluation Methodology

The task was evaluated using the methodology of previous ImageCLEF annotation tasks [15, 16], which follows a PASCAL style metric of intersection over union (IoU). We used the following two measures:

MAP 0.5 IoU: the localised Mean Average Precision (MAP) for each submitted method using the performance measure of IoU >=0.5 of the ground truth;

MAP 0 IoU: the image annotation average for each method in which the concept is detected in the image without any localisation.

In addition, to further analyse the results per types of benthic substrate, the measure *accuracy per class* was used [17], in which the segmentation accuracy for a substrate was assessed using the number of correctly labelled pixels of that substrate, divided by the number of pixels labelled with that class (in either the ground truth labelling or the inferred labelling).

$$\text{agreement per class} = \frac{\#\text{ true positives}}{\#\text{ false positives} + \#\text{ false negatives} + \#\text{ true positives}}$$

5 Results

In 2020, 15 teams registered for the second edition of the ImageCLEFcoral task. Four individual teams submitted 53 runs. Table 2 gives an overview of all participants and their runs. There was a limit of at most 10 runs per team and subtask.

5.1 Subtask 1: Coral Reef Image Annotation and Localisation

Table 3 presents the performance of the participants on the coral reef image annotation and localisation subtask 1.

Table 4 presents the performance (Intersection over Union) of individual runs broken down by class. 32 runs were submitted in this subtask by 4 teams. No individual run performed highest in all classes; however, HHU and FHD

performed well across multiple classes. The highest IoU score (0.512) was for the *soft_coral* class from FAV ZČU PiVa.

Table 5 presents the pixel accuracy per location, per team, across classes for Subtask 1. No individual team performed best across all classes. The highest pixel accuracy scores were 0.5925 in the *hard_coral_branching* class from FHD and 0.5116 in the *soft_coral* class by FAV ZČU PiVa. Overall performance is best with the same location test subset; however, the accuracy of *hard_coral_branching* in the ecologically similar region was very good.

Table 3. The run performance ($MAP\ 0.5\ IoU$ and $MAP\ 0\ IoU$) of Subtask 1.

Run id	team	$MAP\ 0.5\ IoU$	$MAP\ 0\ IoU$
68143	FAV ZČU PiVa	0.582	0.853
67863	FAV ZČU PiVa	0.565	0.851
68094	FAV ZČU PiVa	0.53	0.825
68145	FAV ZČU PiVa	0.517	0.814
67539	FAV ZČU CV	0.49	0.822
68181	FHD	0.457	0.775
68188	FHD	0.44	0.725
67862	FAV ZČU PiVa	0.439	0.774
68187	FHD	0.424	0.729
68182	FHD	0.422	0.762
68146	FAV ZČU PiVa	0.415	0.747
68186	FHD	0.41	0.73
68183	FHD	0.405	0.759
68201	HHU	0.392	0.806
67914	FHD	0.391	0.72
68184	FHD	0.388	0.707
67919	FHD	0.383	0.703
68138	FAV ZČU PiVa	0.377	0.721
68185	FHD	0.369	0.722
67858	FAV ZČU PiVa	0.357	0.712
68093	FAV ZČU PiVa	0.349	0.709
67857	FAV ZČU PiVa	0.347	0.728
68202	HHU	0.323	0.753
68198	HHU	0.313	0.702
68205	HHU	0.303	0.727
68196	HHU	0.28	0.684
68212	HHU	0.263	0.663
68197	HHU	0.245	0.628
67558	FAV ZČU CV	0.243	0.664
68213	HHU	0.233	0.644
68178	HHU	0.01	0.206
68179	HHU	0.01	0.274

5.2 Subtask 2: Coral Reef Image Pixel-wise Parsing

Table 6 presents the performance of the participants on the coral reef image pixel-wise parsing subtask.

Table 7 presents the performance (Intersection over Union) of individual runs broken down by class. 21 runs were submitted in this subtask by 3 teams. No individual run performed highest in all classes; however, runs by FHD had the highest performance in all but one class (*hard_coral_submassive*). The highest IoU

Table 4. Coral reef image annotation and localisation performance in terms of the Intersection over Union (IoU) per benthic substrate for Subtask 1

Run id	Team	algae-macro-or-leaves	fire-coral-millepora	hard-coral-boulder	hard-coral-branching	hard-coral-encrusting	hard-coral-foliose	hard-coral-mushroom	hard-coral-submassive	hard-coral-table	soft-coral	soft-coral-gorgonian	sponge	sponge-barrel
68213	HHU	0.016	0	0.171	0.306	0.066	0.097	0.15	0.038	**0.042**	0.359	0.082	0.12	0.089
68212	HHU	0.012	0	0.218	0.327	**0.09**	0.105	0.231	**0.067**	0.034	0.445	0.059	0.121	0.134
68205	HHU	**0.019**	0	0.093	0.091	0.023	0.056	0.081	0.024	0.017	0.185	0	0.036	0.039
68202	HHU	0.007	0	0.16	0.247	0.053	0.115	0.154	0.032	0.026	0.314	0.037	0.067	0.082
68201	HHU	0.016	0	0.052	0.142	0.005	0.003	0.155	0.002	0	0.144	0	0.019	0.023
68198	HHU	0.002	0	0.221	0.316	0.077	0.119	0.183	0.041	0.029	0.462	0.037	0.107	0.115
68182	FHD	0	0	0.199	**0.35**	0.005	0.012	0.187	0	0	0.484	0	0.104	0.064
68183	FHD	0	0	0.209	0.337	0	0.022	0.197	0	0	0.456	0	0.07	0.098
68197	HHU	0.008	0	0.105	0.137	0.022	0.072	0.149	0.011	0.021	0.399	0	0.045	0.028
68196	HHU	0.005	0	0.159	0.3	0.056	0.056	0.128	0.015	0.022	0.367	**0.088**	0.123	0.053
68188	FHD	0.006	0	0.249	0.301	0.032	0.182	0.397	0.004	0.035	0.453	0.057	0.088	0.135
68187	FHD	0.009	0	0.256	0.306	0.038	0.192	0.402	0.007	0.039	0.48	0.061	0.099	0.134
68186	FHD	0.008	0	**0.257**	0.322	0.03	**0.195**	**0.42**	0	0.031	0.485	0.076	0.097	**0.142**
68185	FHD	0	0	0.185	0.319	0.039	0.131	0.156	0	0.02	0.423	0.046	0.084	0.119
68184	FHD	0.006	0	0.246	0.323	0.052	0.122	0.25	0.003	0.041	0.467	0.067	0.109	0.102
68181	FHD	0	0	0.179	0.316	0	0	0.288	0	0	0.472	0	0.12	0.02
68179	HHU	0	0	0	0.037	0	0	0	0	0	0.146	0	0	0
68178	HHU	0	0	0.017	0.014	0.005	0.001	0	0.002	0	0.056	0	0.009	0
68146	FAV ZåU PiVa	0	0	0.133	0.181	0.046	0.13	0.182	0.011	0	0.452	0	0.103	0.049
68145	FAV ZåU PiVa	0	0	0.105	0.123	0.021	0.038	0.122	0	0	0.387	0	0.093	0.014
68143	FAV ZåU PiVa	0	0	0.054	0.089	0.008	0.009	0.109	0.002	0	0.29	0	0.065	0.01
68138	FAV ZåU PiVa	0.001	0	0.159	0.211	0.052	0.149	0.204	0.016	0	0.462	0	0.113	0.062
68094	FAV ZåU PiVa	0	0	0.108	0.127	0.02	0.038	0.121	0.001	0	0.393	0	0.087	0.004
68093	FAV ZåU PiVa	0.001	0	0.206	0.3	0.08	0.147	0.22	0.017	0.009	0.465	0.082	0.117	0.067
67919	FHD	0	0	0.222	0.243	0.049	0	0	0	0	0.45	0	0.118	0
67914	FHD	0	0	0.227	0.259	0.042	0.086	0.194	0	0	0.474	0	**0.13**	0.057
67863	FAV ZåU PiVa	0	0	0.103	0.104	0.01	0.001	0.134	0.002	0	0.338	0	0.07	0.004
67862	FAV ZåU PiVa	0	0	0.176	0.219	0.038	0.101	0.211	0.006	0.008	0.464	0.033	0.106	0.03
67858	FAV ZåU PiVa	0	0	0.22	0.297	0.057	0.1	0.315	0.032	0.012	0.508	0.047	0.11	0.089
67857	FAV ZåU PiVa	0	0	0.221	0.306	0.06	0.105	0.32	0.034	0.015	**0.512**	0.044	0.111	0.09
67558	FAV ZåU CV	0	0	0.216	0.228	0.048	0.031	0.139	0	0	0.413	0.026	0.097	0.064
67539	FAV ZåU CV	0	0	0.155	0.259	0.047	0.094	0.096	0	0.015	0.475	0.028	0.057	0.085

scores were 0.545 for the *soft_coral* class and 0.505 for the *hard_coral_mushroom* class from FHD.

Table 8 presents the pixel accuracy per location, per team, across classes for Subtask 2. FHD performed highest in all classes except *hard_coral_submassive*. The highest pixel accuracy scores were 0.718 in the *hard_coral_branching* class, 0.562 for the *sponge_barrel* class, 0.547 for the *hard_coral_boulder* class and 0.556 for the *soft_coral* class from FHD. Overall performance was best with the same location test subset, with the exception of the *hard_coral_branching* class which was identified considerably more accurately within the ecologically similar test set. This is a good indication that transfer learning may at least be possible in some classes of substrate.

Table 5. Pixel accuracy per location, per team, Subtask 1, selecting the highest performance per class of all runs submitted by the participant.

Dataset	Team	algae-macro-or-leaves	fire-coral-millepora	hard-coral-boulder	hard-coral-branching	hard-coral-encrusting	hard-coral-foliose	hard-coral-mushroom	hard-coral-submassive	hard-coral-table	soft-coral	soft-coral-gorgonian	sponge	sponge-barrel
Same	FAV ZCU PiVA	0.0606	0.0000	**0.4481**	0.3156	**0.1457**	0.3606	0.3075	**0.1129**	0.1458	**0.5116**	**0.2078**	0.2222	0.2718
	FAV ZâU CV	0.0000	0.0000	0.4426	0.3802	0.1437	0.2932	0.1426	0.0000	0.1311	0.4903	0.1766	0.1457	0.1677
	FHD	**0.2488**	0.0000	0.4293	0.3873	0.1136	**0.3984**	**0.4056**	0.0018	0.2917	0.4999	0.1491	**0.2251**	**0.4875**
	HHU	0.0782	0.0000	0.4344	0.3122	0.1209	0.2098	0.2073	0.0840	**0.3094**	0.4527	0.1331	0.2051	0.2813
Similar	FAV ZCU PiVA	0.0000	0.0000	0.1632	0.1628	0.1007	0.0119	0.0000	0.0387	0.0097	0.0001	0.0000	0.0355	0.0000
	FAV ZâU CV	0.0000	0.0000	0.0664	0.1456	0.0112	0.0000	0.0000	0.0000	0.0051	0.0003	0.0000	0.0291	0.0000
	FHD	0.0000	0.0000	0.1911	0.3124	0.0392	0.0153	0.0000	0.0098	0.0291	0.0018	0.0000	0.0805	0.0000
	HHU	0.0081	0.0000	0.0886	0.2596	0.0237	0.0000	0.0000	0.0559	0.0040	0.0007	0.0000	0.0227	0.0000
Eco_Similar	FAV ZCU PiVA	0.0000	0.0000	0.2504	0.5147	0.0000	0.0182	0.0000	0.0000	0.0000	0.0000	0.0000	0.0066	0.0000
	FAV ZâU CV	0.0000	0.0000	0.2635	0.5901	0.0000	0.0330	0.0000	0.0000	0.0000	0.0011	0.0000	0.0016	0.0000
	FHD	0.0000	0.0000	0.2715	**0.5925**	0.0000	0.0161	0.0000	0.0000	0.0000	0.0024	0.0000	0.0102	0.0000
	HHU	0.0019	0.0000	0.2372	0.4885	0.0035	0.0031	0.0000	0.0000	0.0000	0.0002	0.0000	0.0004	0.0000
Distinct	FAV ZCU PiVA	0.0000	0.0000	0.0560	0.0000	0.0184	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0877	0.0447
	FAV ZâU CV	0.0000	0.0000	0.1720	0.0000	0.0350	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0727	0.0426
	FHD	0.0030	0.0000	0.1899	0.0000	0.0142	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0724	0.0912
	HHU	0.0415	0.0000	0.1534	0.0000	0.0471	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.1678	0.0653

FHD performed well in the pixel accuracy but not as well when considering the MAP scores and this may be indicative of their approach identifying large polygons well but missing many of the smaller polygon objects.

6 Discussion

FAV ZČU CV [19] worked with two neural networks for the first task, SSD [22] and a Mask R-CNN [23]; for the second task, they worked with only the latter. Both of these used the implementation in Keras [24], pre-trained on the Pascal VOC 2007 dataset [25].

They partitioned the training data into distinct training and validation sets containing rough 85% and 15% of the total number of training images. As some types of coral were relatively rare in the training set, there were as few as 16 instances for training and 3 for validation. To train neural networks, more data are clearly needed, so they augmented the images with horizontal and vertical flips, resizing and Gaussian blurring. They also noted that some of the image had a blueish tint while others featured a greenish one and simulated these effects too.

For training SSD, all training images were resized to 512×512, while for Mask R-CNN they were reduced to 1024×1024. It was found that Mask R-CNN detects many more bounding boxes than SDD, most of which are false positives: of the regions detected, 44.7% were true positives with the former, while 71.3% was achieved with the latter. In terms of average precision, figures as high as 62.17% were achieved (SSD for barrel sponges) but five coral classes were not found by either.

Table 6. The run performance ($MAP\ 0.5\ IoU$; and $MAP\ 0\ IoU$) of Subtask 2.

Run id	**team**	$MAP\ 0.5\ IoU$	$MAP\ 0\ IoU$
67864	FAV ZČU PiVa	0.678	0.845
68139	FAV ZČU PiVa	0.664	0.842
68095	FAV ZČU PiVa	0.629	0.817
68142	FAV ZČU PiVa	0.624	0.813
68144	FAV ZČU PiVa	0.617	0.807
68147	FAV ZČU PiVa	0.507	0.727
68190	FHD	0.474	0.715
68137	FAV ZČU PiVa	0.47	0.701
67968	FHD	0.469	0.708
67965	FHD	0.453	0.72
67964	FHD	0.449	0.717
67856	FAV ZČU PiVa	0.441	0.694
67967	FHD	0.435	0.695
68092	FAV ZČU PiVa	0.434	0.689
67963	FHD	0.433	0.694
68192	FHD	0.424	0.668
68191	FHD	0.416	0.692
68140	FAV ZČU PiVa	0.407	0.675
67969	FHD	0.376	0.629
68189	FHD	0.371	0.632
67620	FAV ZČU CV	0.304	0.602

Table 7. Coral reef image pixel-wise parsing performance in terms of the Intersection over Union (IoU) per benthic substrate type for Subtask 2.

Run id	Team	algae-macro-or-leaves	fire-coral-millepora	hard-coral-boulder	hard-coral-branching	hard-coral-encrusting	hard-coral-foliose	hard-coral-mushroom	hard-coral-submassive	hard-coral-table	soft-coral	soft-coral-gorgonian	sponge	sponge-barrel
68192	FHD	0.01	0	0.305	0.387	0.092	0.223	**0.505**	0.009	0.075	**0.545**	0.023	0.13	0.175
68191	FHD	0	0	0.222	0.333	0.009	0.132	0.255	0	0.021	0.49	0	0.085	0.116
68190	FHD	0	0	0.296	0.362	0.009	0.11	0.456	0	0.051	0.52	0.018	0.086	0.147
68189	FHD	0.01	0	0.294	0.338	0.072	0.124	0.245	0.003	0.059	0.522	0.061	0.133	**0.177**
68147	FAV ZČU PiVA	0	0	0.135	0.184	0.044	0.129	0.184	0.009	0	0.453	0	0.095	0.049
68144	FAV ZČU PiVA	0	0	0.109	0.125	0.02	0.041	0.122	0	0	0.394	0	0.092	0.014
68142	FAV ZČU PiVA	0	0	0.106	0.123	0.019	0.04	0.139	0	0	0.403	0	0.087	0.014
68140	FAV ZČU PiVA	0.001	0	0.203	0.283	0.075	0.148	0.226	0.012	0.01	0.443	0.055	0.113	0.079
68139	FAV ZČU PiVA	0	0	0.057	0.091	0.007	0.007	0.108	0.001	0	0.305	0	0.06	0.01
68137	FAV ZČU PiVA	0.001	0	0.162	0.213	0.05	0.148	0.199	0.013	0	0.456	0	0.102	0.064
68095	FAV ZČU PiVA	0	0	0.113	0.128	0.019	0.041	0.121	0	0	0.403	0	0.085	0.004
68092	FAV ZČU PiVA	0.001	0	0.21	0.293	0.077	0.128	0.225	0.013	0.01	0.462	0.055	0.109	0.071
67969	FHD	0.008	0	**0.321**	0.382	**0.093**	**0.275**	0.45	0.019	**0.087**	0.527	**0.074**	0.14	0.171
67968	FHD	**0.009**	0	0.307	0.342	0.043	0.213	0.435	0.006	0.048	0.544	0.047	0.113	0.158
67967	FHD	0	0	0.249	0.311	0.018	0.073	0.177	0	0	0.517	0	0.111	0.104
67965	FHD	0	0	0.286	0.296	0.014	0.102	0.226	0	0	0.522	0	0.105	0.11
67964	FHD	0	0	0.297	**0.398**	0.011	0.073	0.318	0	0	0.533	0	0.125	0.071
67963	FHD	0	0	0.276	0.303	0.058	0.149	0.19	0	0	0.538	0.05	**0.16**	0.018
67864	FAV ZČU PiVA	0	0	0.105	0.108	0.01	0.001	0.137	0	0	0.349	0	0.067	0.004
67856	FAV ZČU PiVA	0	0	0.228	0.287	0.055	0.104	0.318	**0.026**	0.014	0.498	0.051	0.099	0.091
67620	FAV ZČU CV	0	0	0.212	0.222	0.046	0.033	0.138	0	0	0.434	0.023	0.094	0.064

Table 8. Pixel accuracy per location, per team, Subtask 2, selecting the highest performance per class of all runs submitted by the participant.

Dataset	Team	algae-macro-or-leaves	fire-coral-millepora	hard-coral-boulder	hard-coral-branching	hard-coral-encrusting	hard-coral-foliose	hard-coral-mushroom	hard-coral-submassive	hard-coral-table	soft-coral	soft-coral-gorgonian	sponge	sponge-barrel
Same	FAV ZCU PiVA	0.042	0.000	0.500	0.338	0.156	0.358	0.341	**0.099**	0.132	0.502	0.273	0.203	0.291
	FHD	**0.251**	0.000	**0.547**	0.409	**0.192**	**0.420**	**0.496**	0.071	**0.417**	**0.556**	**0.422**	**0.242**	**0.562**
	FAV ZČU CV	0.000	0.000	0.427	0.264	0.113	0.068	0.107	0.000	0.000	0.460	0.191	0.078	0.115
Similar	FAV ZCU PiVA	0.000	0.000	0.228	0.202	0.096	0.013	0.000	0.024	0.010	0.000	0.000	0.048	0.000
	FHD	0.000	0.000	0.319	0.440	0.145	0.014	0.000	0.023	0.087	0.003	0.000	0.119	0.000
	FAV ZČU CV	0.000	0.000	0.097	0.160	0.006	0.000	0.000	0.000	0.000	0.000	0.000	0.015	0.000
Eco_similar	FAV ZCU PiVA	0.000	0.000	0.297	0.543	0.000	0.020	0.000	0.000	0.000	0.000	0.000	0.004	0.000
	FHD	0.000	0.000	0.407	**0.718**	0.000	0.010	0.000	0.000	0.000	0.003	0.000	0.014	0.000
	FAV ZČU CV	0.000	0.000	0.267	0.192	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.000
Distinct	FAV ZCU PiVA	0.000	0.000	0.057	0.000	0.013	0.000	0.000	0.000	0.000	0.000	0.000	0.088	0.049
	FHD	0.000	0.000	0.244	0.000	0.030	0.000	0.000	0.000	0.000	0.000	0.000	0.103	0.099
	FAV ZČU CV	0.000	0.000	0.237	0.000	0.014	0.000	0.000	0.000	0.000	0.000	0.000	0.070	0.046

Interestingly, both trained models performed better on the unseen test imagery than on the images they had retained for validation, and by a fairly large margin. The best mean average precision obtained was 49%, for localization using SSD. This phenomenon is particularly surprising given that the test set contains imagery from ocean regions not present in the training set: the designers of the dataset did not anticipate that this would be the case. It is a particularly promising result given that the ultimate aim of the research is to equip marine biologists and ecologists with a recognition system that can be taken anywhere in the world and expected to work.

FAV ZČU PiVa [18] also employed a Mask R-CNN but included a number of refinements in their training; they believe it is this set of refinements that led to the improvements in performance they achieved.

In forming their validation set, this team selected every eleventh image and substituted some of them so that the training and validation sets had similar distributions. As with other teams, they augmented the provided training set, using similar transformations as [19].

The underlying approach was transfer learning, and several 'backbone' networks were examined, including ResNet-50 and Inception-ResNet-V2. One refinement employed was a 'pseudo-labelling' approach inspired by [26], using a trained network to label untrained test data with weak labels. 'Accumulated gradient normalisation' [27] is credited as providing a considerable improvement in performance. An ensemble approach was ultimately used, in which multiple networks classified the same input and their majority vote yielded the final classification.

HHUD [20] explored two approaches. The first was a refined and improved version of the approach the took for the 2019 exercise while the second was based around RetinaNet [28].

The team used 80% of the identified regions for training and 20% for validation, again swapping individual regions between the two sets until they exhibited similar distributions. The difficulties inherent in underwater photography due to the severe attenuation of the red end of the spectrum were considered and RD [29] was ultimately demonstrated to be the more effective.

The team used a version of Yolo [30], though they suffered from some difficulties in the training data as initially released which meant that the annotations were inconsistent; there was not enough time to re-train after these were identified and corrected. The constraints on image size with their GPU-based implementation is also thought to have an effect. RetinaNet was also used, comprising a feature pyramid network based on ResNet [31], a regressor and a classifier.

The authors also explored more classical approaches. In 2019, a k-NN classifier was used; this year, it was enhanced with PCA was used to identify the best features and a naïve Bayes approach for locating and classifying substrates. It was found that the combination of PCA and naïve Bayes classifier improved performance – though despite this, the neural approaches still out-performed classical ones.

The authors' best performance was achieved using a ensemble of RetinaNet and Yolo v3, using RD-enhanced images for training. The authors' paper [20] has an interesting discussion on the interplay between thresholds, training epochs and performance.

One of the key aspects the dataset creators were keen to explore was whether the training dataset, which was acquired from a single coral reef, made it possible for trained classifiers to perform well on data sourced from geographically distinct reefs. In this case, this 'geographic generalization' was not found, though the number of test images from the different geographical regions was quite small.

FHD [21] went to some lengths to counter the attenuation of red illumination in the images and the blurriness of some of them, achieving impressive visual improvements in some cases. Further improvement was obtained by enhancement in HSV space based on the notion of Rayleigh scattering.

The classification architecture was again based around Mask R-CNN, implemented using Keras and TensorFlow and with Resnet 101 pre-trained on the COCO dataset [32], with the training images reduced to 1536×1536 pixels. The training data were augmented using similar transformations to the other groups. As expected, data augmentation reduced over-fitting. Colour correction led to poorer mean average precision values but better average accuracy. It was observed that the models do not detect objects as well as some other groups' submissions but those that are detected are classified very well.

Interestingly, the authors found their algorithms' performance on subtask 1 (bounding boxes) could be improved simply by re-defining their bounding boxes. This is really an indication that bounding boxes are a poor way of describing the output of processing that involves both segmentation and classification, exacerbated by the extended nature of some types of coral. This suggests that bounding boxes should not form part of ImageCLEFcoral in future years.

The analysis of the results in this paper explores the interplay between the performance measures used and the relative rankings of results. It is not known of course whether these apparent performance differences are statistically significant but this is an area that the designers of the imageCLEFcoral task will explore in future releases.

The approach taken by [18] proved to be the most effective as their approach yielded the highest scores, as measured by mean average precision, for both tasks: their submission 8 won the annotation and localisation task, while their submission 2 won the pixel-wise parsing task with scores of about 0.58 and 0.68 respectively, a significant improvement on the best that was achieved in the 2019 exercise where the equivalent figures were 0.24 and 0.04 respectively — though the above mentioned inconsistencies between image and annotation present in the 2019 dataset will have affected these figures. The authors consider the increased size of the training set in the 2020 exercise played an important part in the improvements in performance that they were able to achieve.

The MAP 0.5 IoU score from FAV of 0.582 over the entire test set is excellent, bearing in mind both the difficulty of the problem and that the problem involved 13 classes, some of which are sparsely represented. There is a signifi-

cant peroformance margin before the best run from the second-placed team, FAV ZČU CV, and the other teams' best submissions, which are closely spaced. FAV also made the best-ranked submission for $MAP\ 0\ IoU$ but the other teams' best-scoring submissions are much closer to this. However, the best-scoring submission for $R\ 0.5\ IoU$ does not yield the highest accuracy of all the submissions. Clearly then, there is some inconsistency in the evaluation measures employed — and this is more of an indication that the performance evaluation measures in widespread use in the vision research community are imperfect.

It is interesting to review the scores obtained from the four categories of test data. For the geographic regions which are similar in nature performance is generally similar. However, performance drops off for other regions, showing that the differences present in the imagery affect the ability to classify the substrates. This shows how difficult it will be to develop a system for marine biologists to automatically classify substrate without significant training resources (i.e., labelled datasets) from that area.

For the pixel-wise parsing task, the $MAP\ 0.5\ IoU$ score of the best-placed team, FAV, is actually higher than for the bounding box task, showing that their approach is able to identify the boundaries of the image features somewhat better than those of the other teams. This makes the performance gap between first- and second-placed teams somewhat larger than for the first task. Again, the best-scoring run in terms of $MAP\ 0.5\ IoU$ is not the best in terms of accuracy.

7 Conclusions

The results of the 2020 coral exercise demonstrate how effective modern deep neural networks are at a range of problems: a performance approaching 70% for a 13-class problem is excellent. The results show that the best pixel-wise parsing technique out-performed the best bounding box one, suggesting that future exercises should concentrate on pixel-wise parsing. There are always difficulties with overlapping bounding boxes and other types of feature in the background of bounding boxes which together reduce the value of that type of annotation.

It is clear that there are genuine performance differences between the four geographical categories of test images described above. This is an important practical problem for coral annotation, as well as for vision systems in general. We anticipate future coral annotation tasks will explore ways to overcome this difficulty. Close examination of the ground truth annotations for the pixel-parsing task shows that annotators tend to place the bounding polygons just outside the boundaries of the features being annotated. We are considering producing other annotations that lie within feature boundaries and encourage teams in a future exercise to train the same architecture with both, then see which works best. That would give us the opportunity to learn something about how annotations should be produced.

The fact that different measures rank-order the different runs differently does not come as a surprise but does show how difficult it is to devise a simple measure that encapsulates performance well. There is clearly research to be done

in this regard. Although there are performance differences between the runs, there is no indication as to whether they are statistically significant or not. This analysis shall be explored in future work. Bearing in mind the point made about performance measures in the previous paragraph, it will be especially interesting to ascertain whether different performance measures yield statistically-significant but inconsistent results.

Acknowledgments

The authors would like to thank those teams who have expended substantial amounts of time and effort in developing solutions to this task. The images used in this task were able to be gathered thanks to funding from the University of Essex and the ESRC Impact Acceleration Account, as well as logistical support from Operation Wallacea. We would also like to thank the MSc Tropical Marine Biology students who participated in the annotation of the test set and Dr Van Der Ven and Dr McKew for facilitating their internship.

References

1. Moberg, F., Folke, C.: Ecological goods and services of coral reef ecosystems. Ecological Economics **29**(2) (1999) 215–233
2. De'ath, G., Fabricius, K.E., Sweatman, H., Puotinen, H.: The 27-year decline of coral cover on the Great Barrier Reef and its causes. Proceedings of the National Academy of Sciences **109** (2012) 17995–17999
3. Burke, L., Reytar, K., Spalding, M., Perry, A.: Reefs at risk revisited. https://pdf.wri.org/reefs_at_risk_revisited.pdf (2012)
4. Hoegh-Guldberg, O., Poloczanska, E.S., Skirving, W., Dove, S.: Coral reef ecosystems under climate change and ocean acidification. Frontiers in Marine Science **4** (2017) 158
5. Obura, D.O., Aeby, G., Amornthammarong, N., Appeltans, W., Bax, N., Bishop, J., Brainard, R.E., Chan, S., Fletcher, P., Gordon, T.A.C., Gramer, L., Gudka, M., Halas, J., Hendee, J., Hodgson, G., Huang, D., Jankulak, M., Jones, A., Kimura, T., Levy, J., Miloslavich, P., Chou, L.M., Muller-Karger, F., Osuka, K., Samoilys, M., Simpson, S.D., Tun, K., Wongbusarakum, S.: Coral reef monitoring, reef assessment technologies, and ecosystem-based management. Frontiers in Marine Science **6** (2019) 580
6. Young, G.C., Dey, S., Rogers, A.D., Exton, D.: Cost and time-effective method for multi-scale measures of rugosity, fractal dimension, and vector dispersion from coral reef 3d models. PLOS ONE **12**(4) (04 2017) 1–18
7. Obura, D.: The diversity and biogeography of western indian ocean reef-building corals. PLOS ONE **7**(9) (2012) 1–14
8. Veron, J., Stafford-Smith, M., DeVantier, L., Turak, E.: Overview of distribution patterns of zooxanthellate scleractinia. Frontiers in Marine Science **1** (2015) 81
9. Ionescu, B., Müller, H., Péteri, R., Dicente Cid, Y., Liauchuk, V., Kovalev, V., Klimuk, D., Tarasau, A., Ben Abacha, A., Hasan, S.A., Datla, V., Liu, J., Demner-Fushman, D., Dang-Nguyen, D.T., Piras, L., Riegler, M., Tran, M.T., Lux, M., Gurrin, C., Pelka, O., Friedrich, C.M., García Seco de Herrera, A., Garcia, N., Kavallieratou, E., del Blanco, C.R., Cuevas Rodríguez, C., Vasillopoulos, N., Karampidis,

K., Chamberlain, J., Clark, A., Campello, A.: ImageCLEF 2019: Multimedia retrieval in medicine, lifelogging, security and nature. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 10th International Conference of the CLEF Association (CLEF 2019), Lugano, Switzerland, LNCS Lecture Notes in Computer Science, Springer (September 9-12 2019)
10. Chamberlain, J., Campello, A., Wright, J.P., Clift, L.G., Clark, A., García Seco de Herrera, A.: Overview of ImageCLEFcoral 2019 task. In: CLEF2019 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org (2019)
11. Ionescu, B., Müller, H., Péteri, R., Abacha, A.B., Datla, V., Hasan, S.A., Demner-Fushman, D., Kozlovski, S., Liauchuk, V., Cid, Y.D., Kovalev, V., Pelka, O., Friedrich, C.M., de Herrera, A.G.S., Ninh, V.T., Le, T.K., Zhou, L., Piras, L., Riegler, M., l Halvorsen, P., Tran, M.T., Lux, M., Gurrin, C., Dang-Nguyen, D.T., Chamberlain, J., Clark, A., Campello, A., Fichou, D., Berari, R., Brie, P., Dogariu, M., Ştefan, L.D., Constantin, M.G.: Overview of the ImageCLEF 2020: Multimedia retrieval in medical, lifelogging, nature, and internet applications. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Volume 12260 of Proceedings of the 11th International Conference of the CLEF Association (CLEF 2020)., Thessaloniki, Greece, LNCS Lecture Notes in Computer Science, Springer (September 22-25 2020)
12. Schoening, T., Bergmann, M., Purser, A., Dannheim, J., Gutt, J., Nattkemper, T.W.: Semi-automated image analysis for the assessment of megafaunal densities at the Arctic deep-sea observatory HAUSGARTEN. PLoS ONE **7**(6) (2012)
13. Culverhouse, P., Williams, R., Reguera, B., Herry, V., González-Gil, S.: Do experts make mistakes? A comparison of human and machine identification of dinoflagellates. Marine Ecology Progress Series **247** (2003) 17–25
14. Beijbom, O., Edmunds, P.J., Kline, D.I., Mitchell, B.G., Kriegman, D.: Automated annotation of coral reef survey images. In: Proceedings of the 25th IEEE Conference on Computer Vision and Pattern Recognition (CVPR'12), Providence, Rhode Island (June 2012)
15. Gilbert, A., Piras, L., Wang, J., Yan, F., Ramisa, A., Dellandrea, E., Gaizauskas, R.J., Villegas, M., Mikolajczyk, K.: Overview of the ImageCLEF 2016 scalable concept image annotation task. In: CLEF Working Notes. (2016) 254–278
16. Gilbert, A., Piras, L., Wang, J., Yan, F., Dellandrea, E., Gaizauskas, R.J., Villegas, M., Mikolajczyk, K.: Overview of the ImageCLEF 2015 scalable image annotation, localization and sentence generation task. In: CLEF Working Notes. (2015)
17. Everingham, M., Eslami, S.M.A., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International Journal of Computer Vision **111**(1) (January 2015) 98–136
18. Pixek, L., n ü Rí ha, A., s Zita, A.: Coral reef annotation, localisation and pixel-wise classification using mask-rcnn and bag of tricks. In: CLEF2020 Working Notes. CEUR Workshop Proceedings, Thessaloniki, Greece, CEUR-WS.org <http://ceur-ws.org> (September 22-25 2020)
19. Gruber, I., Straka, J.: Automatic coral detection using neural networks. In: CLEF2020 Working Notes. CEUR Workshop Proceedings, Thessaloniki, Greece, CEUR-WS.org <http://ceur-ws.org> (September 22-25 2020)
20. Bogomasov, K., Grawe, P., Conrad, S.: Enhanced localization and classification of coral reef structures and compositions. In: CLEF2020 Working Notes. CEUR Workshop Proceedings, Thessaloniki, Greece, CEUR-WS.org <http://ceur-ws.org> (September 22-25 2020)

21. Arendt, M., kert, J.R., ngel, R.B., Brumann, C., Friedrich, C.M.: The effects of colour enhancement and iou optimisation on object detection and segmentation of coral reef structures. In: CLEF2020 Working Notes. CEUR Workshop Proceedings, Thessaloniki, Greece, CEUR-WS.org <http://ceur-ws.org> (September 22-25 2020)
22. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C., Berg, A.: SSD: Single shot multibox detector. In: Proceedings of the European Conference on Computer Vision, Springer (2016) 21–27
23. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: Proceedings of the International Conference on Computer Vision, IEEE (2017) 2961–2969
24. Chollet, F.e.: Keras. https://keras.io (2015)
25. Everingham, M., Van Gool, L., Williams, C., Winn, J., Zisserman, A.: The pascal visual object classes challenge 2007 (voc2007) results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html (2007)
26. Arazo, E., Ortego, D., Albert, P., O'Connor, N., McGuinness, K.: Pseudo-labeling and confirmation bias in deep semi-supervised learning. arXiv preprint arXiv:1908.02983 (2019)
27. Hermans, J., Spanakis, G., Mö ckel, R.: Accumulated gradient normalization. arXiv preprint arXiv:1710.02368 (2017)
28. Lin, T., Goyal, P., Girshick, R., He, K., Dollar, P.: Focal loss for dense object detection. In: Proceedings of the International Conference on Computer Vision, https://doi.org/10.1109/iccv.2017.324, http://dx.doi.org/10.1109/ICCV.2017.324 (October 2017)
29. Ghani, A., Isa, N.: Underwater image quality enhancement through composition of dual-intensity images and Rayleigh-stretching. SpringerPlus **3**(1) (2014) 757
30. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
31. Lin, T., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. http://arxiv.org/abs/1612.03144 (2016)
32. Lin, T., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.: Microsoft COCO: Common objects in context. In: Proceedings of the European Conference on Computer Vision, Springer (2014) 740–755