Automatic underwater fish species classification with limited data using few-shot learning

Sébastien Villon, Corina Iovan, Morgan Mangeas, Thomas Claverie, David Mouillot, Sébastien Villéger, Laurent Vigliola

Automatic underwater fish species classification with limited data using few-shot learning

Sébastien Villon[a,*], Corina Iovan[a], Morgan Mangeas[a], Thomas Claverie[b,c], David Mouillot[c,d], Sébastien Villéger[c], Laurent Vigliola[a]

[a] ENTROPIE, IRD, University of New-Caledonia, University of La Reunion, CNRS, Ifremer, Labex Corail, Noumea, New-Caledonia, France

[b] CUFR Mayotte, France

[c] MARBEC, University of Montpellier, CNRS, IRD, Ifremer, Montpellier, France

[d] Institut Universitaire de France, Paris, France

[*] Corresponding author: sebastien.villon@ird.fr

Abstract:

Underwater cameras are widely used to monitor marine biodiversity, and the trend is increasing due to the availability of cheap action cameras. The main bottleneck of video methods now resides in the manual processing of images, a time-consuming task requiring trained experts. Recently, several solutions based on Deep Learning (DL) have been proposed to automatically process underwater videos. The main limitation of such algorithms is that they require thousands of annotated images in order to learn to discriminate classes (here species). This limitation implies two issues: 1) the annotation of hundreds of common species requires a lot of efforts 2) many species are too rare to gather enough data to train a classic DL algorithm. Here, we propose to explore how few-shot learning (FSL), an emerging research field, could overcome DL limitations. Few-shot learning is based on the principle of training a Deep Learning algorithm on "how to learn a new classification problem with only few images". In our case-study, we assess the robustness of FSL to discriminate 20 coral reef fish species with a range of training databases from 1 image per class to 30 images per class, and compare FSL to a classic DL approach with thousands of images per class. We found that FSL outperform classic DL approach in situations where annotated images are limited, yet still providing good classification accuracy.

Keywords: few-shot learning, Deep learning, video, marine biodiversity

Introduction

The world's ecosystems have entered an era of anthropogenic defaunation where human activities have triggered global decline in animal abundance, species range contraction and a new wave of species extinction [1]. This global change is threatening ecosystem services worldwide hence the stability of our food systems, economies, and health. Defaunation is more advanced in terrestrial and freshwater ecosystems than in the marine environment where it started centuries later. However, the pace of defaunation is accelerating in oceans mostly due to the advent of industrial fishing since a century ago [2]. Given this context of global changes rapidly affecting fish communities, it is imperative to monitor fish biodiversity over time, on a large scale and using non-destructive methods.

Fish biodiversity surveys in the marine environment are typically performed by divers. Although dive visual censuses provide a great deal of information on some shallow habitats, there are many limitations. First, divers are limited by depth and can hardly perform long dives to count fish below 30 m, ignoring mesophotic habitats and deeper ecosystems. Second, divers are limited by time and generally focus their 2-4 dives per day in the most speciose hard-substrate habitats, and ignore less rich and often immense adjacent soft-bottom habitats. Third, dive surveys provide data at a slow rate so that the compilation of global fish biodiversity database takes decades of efforts by multiple teams of highly skilled taxonomic divers (e.g. [3], [4]). This is a major restriction to the necessary temporal monitoring of global marine ecosystems, although a few time series exist in some countries[1] [5].

---

[1] AIMS, Long-Term Monitoring Program: Visual Census Fish Data (Great Barrier Reef) https://apps.aims.gov.au/metadata/view/5be0b340-4ade-11dc-8f56-00008a07204e

Underwater videos (UV) are increasingly used [6] to overcome the limitations of diver-based surveys to quickly collect large amounts of data. For instance, more than 15,000 video stations were deployed in 58 countries in just three years for the first global assessment of the conservation status of reef sharks [7]. Furthermore, underwater video surveys can be performed in many habitats, with some example in shallow reefs [8], sandy lagoons [9], deep sea [10], and even in the pelagic ecosystem [11]. Deploying underwater video stations does not require expert taxonomists and is now quite inexpensive with the improvement of cheap action cameras since a few years. The bottleneck to analyse this data now resides in the manual processing of the videos. Indeed, manually extracting fish biodiversity and abundance data from raw videos requires unsustainable workload by highly trained taxonomic experts. Although this annotation work can be improved through citizen science [12]–[14], such time-consuming and expensive task cannot match the increasing size of datasets, up to 20,000 hours of videos for global surveys [7] and the necessary monitoring of global oceans over time.

As the demand for automatic methods to analyze underwater videos is rising, the latest generation of deep learning algorithms (DL), and in particular convolutional neural networks (CNNs) are increasingly used for species identification [14]–[17] and fish detection [18]–[21]. However, these algorithms require a large dataset of annotated images (thumbnails hereafter) in order to train a robust model, able to provide satisfying results. Therefore, this method still requires collecting an important image dataset manually annotated by experts. This is especially problematic in highly diverse faunas such as coral reef fish that encompass nearly 6500 species worldwide [22]. Furthermore, a universal pattern in species distribution, including fish communities, is that both rare and common species are found in every community, with the fraction of rare species more important in rich ecosystems, such as coral reefs [23], [24]. It is therefore almost impossible to gather enough thumbnails of rare species to efficiently train a deep neural network in a “classic” way, which requires thousands of images per species [25]–[27]

There are two ways to tackle this problem of lack of data. The first one consists of directly addressing the data itself, through data augmentation [28]–[30]. The second option is to change the classification algorithm. Few-shots learning (FSL) algorithms [31], [32] are designed to compute a classification task (query, noted *Q*) with only a few thumbnails to train (Support Sets, noted *SS*), and it has been increasingly studied since 2017 [33]. Few-shots learning methods are divided into three main approaches. Metric-based methods are embedding both queries (*Q*) and support sets (*Ss*), before assigning to the query a class, according to distances computed between *Q* and *Ss* ([34]–[36]). The second approach consists of 1) training a model on a large database, and 2) adapt this model to a new task with few examples, while not forgetting the concepts learned previously [37], [38]. Finally, optimization-based methods are designed to adapt quickly to new tasks, hence able to learn a classification task with few examples [33], [39], [40]. Optimization-based algorithms showed promising results in deep learning few-shot classification [33], [41], [42]. Such methods propose to pre-train (or “meta-train”) a model with existing databases (e.g. MiniImageNet [43] , Ominglot [44]) on different tasks so it can adapt easily to a new one. For object identification, a task is defined by the classes the model has to discriminate. Once this model, called “meta-model” has been trained, it can then be tuned to operate on a new task with a very limited dataset, usually only 1-5 thumbnails per class.

In this study we propose to compare the efficiency of optimization-based few-shot learning and standard large dataset deep-learning methods to identify coral reef fish species on images. More specifically, we aim to determine how well a classic deep learning architecture trained with thousands of images and the benefit of data augmentation (hereafter DL) and FSL algorithms perform in situations where training thumbnail dataset is large or limited. To achieve this, we first trained a classic DL architecture built for image classification [45] on a large dataset of 69,169 thumbnails, and on a more limited dataset of 6,320 thumbnails for 20 coral reef fish species. Then, we trained a few-shots, optimization-based learning algorithm [39] on the exact same training datasets while varying the number of shots from 1 to 30. Finally, we compared the capacity of DL and FSL models to correctly identify species on an independent thumbnail dataset, and modelled the asymptotic relationship between classification accuracy and the number of thumbnails in the training datasets for both classic DL and FSL algorithms

Material and methods

Thumbnail datasets

We used three fish thumbnail datasets (*T0*, *T1*, and *T2*) extracted from 175 underwater videos recorded on reefs around Mayotte Island (Western Indian Ocean) using GoPro Hero 3+ and GoPro hero 4+ cameras with a resolution of 1920x1080 pixels. A thumbnail is defined as an image containing a single labelled fish belonging to one of the 20 most common fish species in the videos, and representing a broad range of sizes, colors, body orientations, and background (Supp. Fig. 1, Supp. Fig. 2).

T0 is composed of 69,169 thumbnails extracted from 130 videos, with a range of 1,134 to 7,345 thumbnails per species (Table *1*). *T1* is composed of 6,320 thumbnails extracted from 20 videos with 40-1,436 images per species whereas *T2* is composed of 13,232 thumbnails extracted from 25 videos with 55-3,896 images per species. Thumbnails size originally ranged from 55x55 pixels to 500x450 pixels, but were resized to 84x84 pixels before being processed through FS and DL algorithms.

The datasets *T1* and *T2* correspond to two real scenarii where videos were recorded during two trips in the field of a week each.

The three thumbnails datasets are fully independent, as they were extracted from videos recorded at different sites, with different conditions (weather, lighting, depth, time of the day, seascape) and on different days.

To train our DL architecture, we applied data augmentation to *T0* and *T1*. For each natural thumbnails in *T0* and *T1*, we created 9 thumbnails through contrast augmentation or diminution, and horizontal flip. We then obtained augmented datasets composed of 691,690 (AT0) and 63,200

(AT1) images respectively Supp. Table 1. Further details on thumbnail datasets and data augmentation are given in [46].

Table 1: Number of natural thumbnails extracted from the videos to build our three datasets

Family	Species	Training dataset *T0*	Training dataset *T1*	Test dataset *T2*
Acanthuridae	*Acanthurus leucosternon*	3,259	235	491
Acanthuridae	*Acanthurus lineatus*	1,008	114	864
Acanthuridae	*Naso brevirostris*	1,134	539	1932
Acanthuridae	*Naso elegans*	7,345	1,435	3,896
Acanthuridae	*Zebrasoma scopas*	4,970	48	579
Chaetodontidae	*Chaetodon auriga*	2,134	737	502
Chaetodontidae	*Chaetodon guttatissimus*	1,182	221	68
Chaetodontidae	*Chaetodon trifascialis*	5,234	41	630
Chaetodontidae	*Chaetodon trifasciatus*	4,421	71	82
Labridae	*Gomphosus caeruleus*	3,131	57	173
Labridae	*Halichoeres*	3,192	40	287

	hortulanus			
Labridae	*Thalassoma hardwicke*	4,951	181	275
Lethrinidae	*Monotaxis grandoculis*	3,893	797	1,422
Monacanthidae	*Oxymonacanthus longirostris*	2,553	54	55
Pomacentridae	*Abudefduf vaigiensis*	5,124	376	216
Pomacentridae	*Amblyglyphidodon indicus*	1,188	636	1,310
Pomacentridae	*Chromis opercularis*	1,525	81	93
Pomacentridae	*Chromis ternatensis*	3,640	300	156
Pomacentridae	*Pomacentrus sulfureus*	5,409	270	142
Zanclidae	*Zanclus cornutus*	3,876	86	59
TOTAL		69,169	6,320	13,232

Experimental design

To compare classic deep-learning and few-shot algorithms in situations of large or small thumbnail datasets, we led five experiments using datasets T0, T1, T2, AT0 and AT1 described in Supp. Table 1 :

1) We trained a classic DL algorithms architecture with our biggest dataset *AT0* as a baseline for the DL accuracy;
2) We trained the same DL architecture with the same hyper-parameters (e.g. model architecture and training process) but on a much more limited dataset (*AT1*). Hyper-parameters are the parameters defining the architecture (number of layers, number and size of convolutions, connections between layers) and the training process of a Deep Model (learning rate, neurone activation, back-propagation compotation).;
3) We trained the same DL architecture with limited datasets obtained by subsampling T0 to 250 and 500 images per class (here after "species" when we are referring to our experiments), corresponding to 2500 and 5000 thumbnails in *AT0*;
4) We pre-trained a FSL architecture on the 64 training classes of MiniImageNet (Supp. Fig. 3) and used T0 to build support sets (*SS*) with 1, 5, 15 and 30 thumbnails for each fish species;
5) We pre-trained the same FSL architecture on MiniImageNet and used the more limited T1 dataset to build support sets with 1, 5, 15 and 30 images per species.

We used ResNet 100 [45] as our classic deep-learning algorithm. Resnet is a convolutional neural network (CNN), a DL architecture which is able to both extract features from images and classify these images thanks to those features [47]. In order for a CNN to build an image classification model, the architecture is fed a large dataset, composed of pairs of labels and images. Using this dataset, the algorithms change their inner parameters in order to minimize the classification error, through a process called back-propagation. The ResNet architecture achieved the best results on ImageNet Large Scale Visual Recognition Competition (ILSVRC [43]) in 2015, considered the most challenging image classification competition. It is still one of the best classification algorithms, while being easy to use and implement.

For the few-shot implementation, we used the Reptile algorithm [39]. Few-shot learning algorithms are specific DL algorithms, whose goal is to be able to fit a model with very few training images. The Reptile algorithm is based on the well-known MAML architecture [33], and more precisely on the first-order version of MAML [48]. The Reptile algorithm is based on the division of the training dataset into a number of tasks T_i, a task being a learning problem. Through repetitively changing the task during the first training phase (known as meta-training), this algorithm produces a quick learner, i.e. a learner than can quickly adapt to a new task with a small number of examples.

Here, the few-shot algorithms were tested on a classic *n-ways k-shots* procedure, *n* being the number of classes per support set, and *k* the number of images per class in the support set. For instance, a 5-ways 1-shots consists of training 5 classes with supports sets composed of 1 image per classes (e.g. species). We set n=5 [34], [36], [40], [42], [49] and allowed *k* to vary between 1 shot and 30 shots for both experiments 4 and 5. We did not use data augmentation for FSL experiments for several reasons. First, the goal of FSL is to adapt quickly with a very limited number of images. Second, to have similar settings for method comparison. There were no data-augmentation in the original paper, so we reproduced that. It also allowed us to compare our results with those obtained on benchmarks. Third, the reason behind the use of raw data instead of augmented data in few-shot learning paper is that with very few training samples and few conditions, the risk of overfitting by using the same image modified multiple times is far greater than in classic approaches with important datasets with many conditions.

Model comparison

All the DL and FSL models were tested on the independent *T2* dataset.

First, we compared the results of experiments 1 and 2 in order to estimate the decrease in performance of a classic ResNet DL architecture when trained on a large dataset *AT0* (i.e. between 11,340 and 73,450 images per species after data augmentation, with an average of 3458 natural thumbnails per species) or trained on a more limited dataset *AT1* (i.e. between 400 and 14,360 images per species after data augmentation, with an average of 315 natural thumbnails per species).

Second, we compared the results of experiments 1 and 4 in order to evaluate if the ResNet architecture outperforms the Reptile architecture in a real-case situation where thumbnail dataset is large (*T0 and AT0*).

Finally, we compared the results of experiments 2 and 5 to determine whether and to which extent a Reptile model performs better than a ResNet model in a real-case situation where thumbnail dataset is limited (*T1 and AT1*).

In order to better evaluate the performance of ResNet and Reptile algorithms, we also modelled the relationship between model accuracy and the number of thumbnails used to train the models. To achieve this, we fitted the following asymptotic function to the results of experiments 1, 3 and 4 (obtained through training DL and FSL architectures on datasets of various size obtained from AT0 and T0):

$$Accuracy = Accuracy_{\infty} \cdot (1 - exp(-R \cdot N_{image})) \quad \text{(eq.1)}$$

where $Accuracy_{\infty}$ is the asymptotic model accuracy when the number of thumbnails N_{image} is infinite, and R is the rate at which the asymptote is reached.

Equation 1 was fitted by non-linear mixed-effect modelling (NLME [50]) using species as a random effect .This method is widely used for fitting asymptotic processes. It allows estimating and comparing asymptotic accuracies of both FSL and DL algorithms, and the number of image to reach these asymptotic accuracies. The number of images required to reach the asymptotic accuracy was calculated as the number of images corresponding to an accuracy of 0.99 times the asymptotic value, meaning the asymptote was reached within 1%.

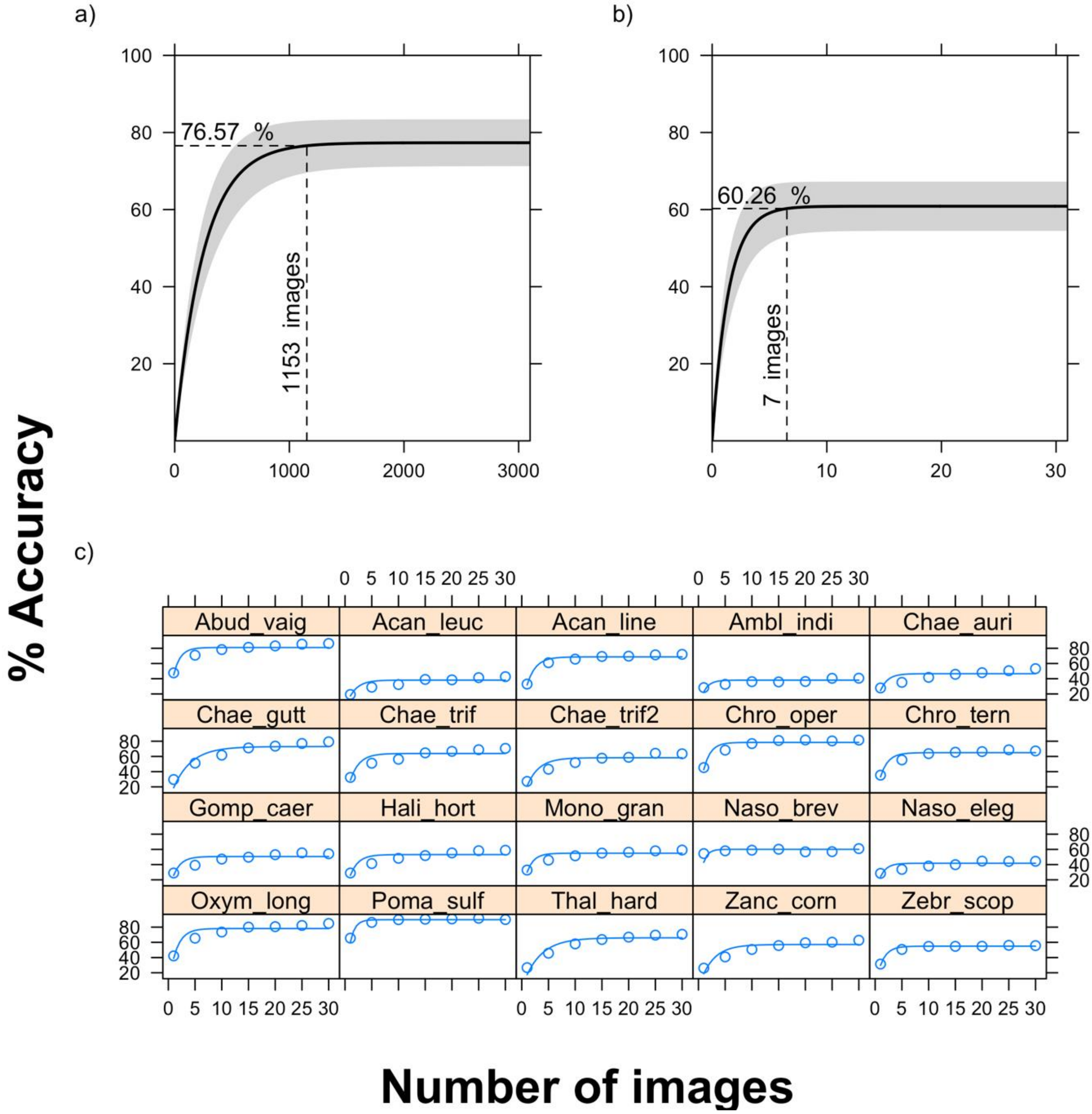
Results

The deep ResNet model trained on the large *AT0* dataset (3458 natural thumbnails in average per species) during the first experiment obtained a mean accuracy (i.e. percentage of correct classification) of 78.00% (standard deviation (SD) of 15.16%) on *T2* test-dataset (Table 2). With this model, accuracy varied among species between 54.14% (*Naso brevirostris*) and 99.07% (*Abudefduf vagiensis*). The same ResNet DL model trained on smaller *AT1* (315 natural thumbnails in average per species) during the second experiment showed highly degraded performance with a mean accuracy of only 42.21% (SD=24.95%). Among species variation ranged with this model from only 3.49% (*Chaetodon trifascialis*) to 85.86% (*Chaetodon auriga*).

The few-shot Reptile architecture trained on limited T1 dataset during our fifth experiment obtained a mean accuracy of 32.04% for the 1-shot learning (SD= 12.70%) and 51.77% mean accuracy for the 30-shots learning (SD= 18.96%) (Table 2,). In this scenario of limited *T1* training dataset, the few-shot Reptile algorithm nearly equalled the ResNet DL model with only 5 shots (41.47% accuracy for 5-shots learning on T1 vs 42.21% for DL on A*T1*), and performed better beyond 10 shots (45.92%

of accuracy on *T2* with 10-shots learning). A pairwise proportion test showed a p-value <0.0001, assessing that FSL was significantly better than DL in this scenario beyond 10 shots (Supp. Table *3*) accuracy of Reptile models had a standard deviation from 12.70% with one-shot learning, to 18.96% with 30-shots learning, indicating important variation in accuracy among species. However, this standard deviation was smaller than that of the ResNet algorithm trained on the same *AT1* limited dataset (24.95%).

Figure 1: Relationship between the number of natural thumbnails per species for training and the accuracy of deep-learning and few-shot learning models. Non-linear mixed effects asymptotic model fit for (a) DL architecture at the fixed-effect level, and for FSL architecture at (b) the fixed-effect level and (c) the random-effect level. Grey areas represent 95% CI in fixed-effects estimates. Dotted lines represent the NLME estimate of the number of images per species required to reach the 99% asymptote value. We obtained similar magnitude with the 95% asymptote value, reached with 750 images for DL and 5 with FSL.



The same few-shot Reptile architecture trained on subsets of T0 during the fourth experiment obtained even better results than when trained on T1, with a mean accuracy on T2 of 34.57% for 1-shot, 50.23% for 5-shots, and up to 64.92% for 30-shots (Table 2).

Mixed-effects modelling (NLME) of T0 and AT0 experimental data showed a clear pattern of asymptotic increase of accuracy with the number of natural thumbnails for both Resnet and Reptile architectures (Figure 1).

NLME models included significant species random effect for both DL and FSL (Log-likelihood tests, P<0.0001).

The fixed-effect asymptotic value of accuracy was higher for ResNet model ($Accuracy_{\infty}$ = 77.34%, 95% CI: 71.26-83.41%) than for Reptile model ($Accuracy_{\infty}$ =60.87%, 95% CI: 54.48–67.26%), illustrating higher classification power of ResNet over Reptile when large numbers of thumbnails are available. However, the slope of the asymptotic model was two-orders of magnitude higher for Reptile (0.707, 95% CI: 0.559-0.854) than for ResNet architecture (0.0040, 95% CI: 0.0032-0.0048), illustrating the high capacity of Reptile FSL algorithm to learn from only a few images. NLME modelling further showed that average asymptotic accuracy was reached with only 7 natural thumbnails per species for Reptile architecture, compared to 1153 natural thumbnails per species for ResNet, confirming the strong power of Reptile method in situation of limited thumbnail training dataset. However, model random effects showed that some variation existed among species. For Reptile architecture, asymptotic accuracy values ranged from 38.09% (*Amblyglyphidodon indicus*) to 89.78% (*Pomacentrus sulfureus*), and was reached with 4 to 16 training images per species. For DL architecture, species asymptotes varied from 62.72% (*Monotaxis grandoculis*) to

96.81% (*Abudefduf vaigiensis*), and could be reached with 786-1776 thumbnails per species (*Supp. Table 4*).

Table 2: Accuracy of our ResNet deep-learning (DL) and Reptile few-shots learning (FSL) models trained on T0 or T1 thumbnails datasets for different number of shots. Accuracy is the % correct classification of models on T2 test dataset. DL models were trained from T0 and T1 after data augmentation (AT0 and AT1).

	DL		FSL					
	T0	T1	T1			T0		
Image per species (on average)	3458	315	1 shot	5 shots	30 shots	1 shot	5 shots	30 shots
Abudefduf vaigiensis	99.07	69.91	16.08	11.39	11.38	47.67	70.9	86.35
Acanthurus leucosternon	86.15	44.67	25.51	30.71	38.74	19.23	28.8	42.66
Acanthurus lineatus	59.72	20.37	39.86	56.04	72.50	32.93	61.01	72.02
Amblyglyphidodon indicus	58.78	60.78	25.75	26.74	32.86	28.26	32.55	40.64
Chaetodon auriga	87.05	85.86	18.16	25.68	36.56	27.8	35.18	53.20
Chaetodon guttatissimus	85.50	44.12	33.58	44.21	58.26	29.61	51.18	79.29
Chaetodon trifascialis	90.00	3.49	29.02	25.48	28.44	27.14	43.17	63.51
Chaetodon trifasciatus	87.80	28.05	38.73	50.63	66.72	32.41	51.07	70.63

Chromis opercularis	61.29	9.68	44.01	61.81	62.94	45.34	68.28	81.50
Chromis ternatensis	59.61	55.77	18.91	24.94	35.07	35.4	55.44	67.22
Gomphosus caeruleus	75.72	20.81	26.01	38.99	58.74	28.96	39.16	54.22
Halichoeres hortulanus	82.93	17.07	31.82	44.81	57.01	28.94	41.35	58.87
Monotaxis grandoculis	57.10	53.37	32.03	41.52	50.64	32.8	45.85	59.13
Naso brevirostris	54.14	68.60	47.06	54.26	64.66	54.47	58.08	61.00
Naso elegans	93.24	79.43	34.54	43.11	52.17	28.47	33.71	44.36
Oxymonacanthus longirostris	96.43	14.54	39.29	53.48	66.26	42.15	65.44	84.86
Pomacentrus sulfureus	90.14	61.97	70.90	88.18	93.93	65.53	86.21	90.00
Thalassoma hardwicke	90.90	51.64	25.64	44.4	67.96	26.72	45.7	70.60
Zanclus cornutus	81.36	40.68	18.33	31.28	44.44	26.08	40.82	62.72
Zebrasoma scopas	63.04	13.30	25.56	31.81	36.05	31.42	50.69	55.70
MEAN	78.00	42.21	32.04	41.47	51.77	34.57	50.23	64.92
SD	15.16	24.95	12.70	16.93	18.96	11.14	14.75	14.55

Discussion

Our experiments demonstrated that few-shot learning methods based on Reptile architecture can be effectively used to drastically reduce the number of annotated images for underwater fish identification. Accuracy levels obtained with few-shot learning algorithm trained with only five training images are close to those of a standard Deep Learning architecture such as ResNet trained with 400-14350 images per species. Further, FSL architecture trained with 10 images outperformed a ResNet 100 architecture trained with at least 400 images per species. This is a very promising result in situations where many species need to be identified from models trained with a few images, a typical characteristic in marine biodiversity applications.

However, the important standard deviation among the different trained species (18.96 SD on 30-shots) showed that few-shot algorithms may not be robust enough to discriminate among similar species showing only subtle differences. Nevertheless, in our 2nd experiment, our ResNet model achieved an accuracy under 40% for all the species with fewer training images than 1140 (after data augmentation, i.e. 114 natural images), and only 7 species were identified with an accuracy greater than 45%. These species were represented with a range of 2700-14350 images during the training phase. We also show better results with the model trained on T0 than the model trained on T1. As expected, increasing the number of images per shot rely on better performances as well as increasing the per species images variability. However, in real conditions, few-shot learning is to be used in a context where very few images per classes are at disposal. Therefore, the dataset T1 corresponded more to a real use case scenario.

Thus, there is a trade-off to make between accuracy and robustness on one hand, and the cost of video annotation by experts on the other.

Modeling the accuracy of neural networks using NLME allowed to understand the number of images per species required for the Few-shot and Deep architecture to reach 99% of their maximum potential accuracy. In our case study, there was a 150-fold factor between the average number of images required for a Deep Learning architecture (1153 images) and for a Few-shot architecture (7) to reach asymptotic accuracy. However, it is important to note that these numbers could vary according to the number and complexity of classes fed to the deep classifier.

In this work we used a Reptile FSL architecture. As the field of few-shot learning is quickly improving, new methods are proposed at a fast rate. While Reptile obtained a mean accuracy of 61.98% on the MiniImageNet dataset (the most used benchmark for few-shot learning methods) through a 5-shots learning, [51] recently achieved 80.51% of accuracy on the same dataset. Although further studies are required, we can reasonably assume that the improvements of FSL algorithms will further expand the possible use of few-shot learning for real-life use cases.

Applied to marine and coral reef ecology, such methods requiring few examples to fit a model on an identification task could be used for studies on species rarely seen on screen. A key characteristic of highly diverse ecosystems is that they are composed of few very common species and a large proportion of less-common and rare species. Hence, the important effort required to build databases with a sufficient number of images of all these rare species is the main bottleneck preventing the use of Deep Learning on a large number of species. The improvement of few-shot learning algorithms offers promises to build efficient identification models to automatically process images and videos to localise and identify rare fish species. Such models could then be paired with more classic deep architectures, more efficient to identify abundant species with the leverage of important datasets.

Acknowledgements:

This study was funded by the French National Research Agency project ANR 18-CE02-0016 SEAMOUNTS.

References

[1] R. Dirzo, H. S. Young, M. Galetti, G. Ceballos, N. J. B. Isaac, and B. Collen, "Defaunation in the Anthropocene," *Science (80-.).*, vol. 345, no. 6195, pp. 401–406, 2014.

[2] H. S. Young, D. J. Mccauley, M. Galetti, and R. Dirzo, "Patterns , Causes , and Consequences of Anthropocene Defaunation," *Annu. Rev. Ecol. Evol. Syst.*, no. August, pp. 333–358, 2016.

[3] J. E. Cinner *et al.*, "Meeting fisheries, ecosystem function, and biodiversity goals in a human-dominated world," *Science (80-.).*, vol. 311, no. April, pp. 307–311, 2020.

[4] R. D. Stuart-smith *et al.*, "Integrating abundance and functional traits reveals new global hotspots of fish diversity," *Nature*, vol. 501, no. 7468, pp. 539–542, 2013.

[5] A. Heenan *et al.*, "Long-term monitoring of coral reef fish assemblages in the Western central pacific," *Sci. Data*, vol. 4, pp. 1–12, 2017.

[6] S. K. Whitmarsh, P. G. Fairweather, and C. Huveneers, "What is Big BRUVver up to ? Methods and uses of baited underwater video," *Rev. Fish Biol. Fish.*, vol. 27, no. 1, pp. 53–73, 2017.

[7] M. Aaron MacNeil *et al.*, "Global status and conservation potential of reef sharks," *Nature*, no. July 2019, 2020.

[8] J. Juhel, L. Vigliola, L. Wantiez, T. B. Letessie, J. J. Meeuwig, and D. Mouillot, "Isolation and no-entry marine reserves mitigate anthropogenic impacts on grey reef shark behavior," *Sci. Rep.*, vol. 9, no. November 2018, pp. 1–11, 2019.

[9] M. Cappo, G. De, and P. Speare, "Inter-reef vertebrate communities of the Great Barrier Reef Marine Park determined by baited remote underwater video stations," *Mar. Ecol. Prog. Ser.*, vol. 350, pp. 209–221, 2007.

[10] V. Zintzen, M. J. Anderson, C. D. Roberts, E. S. Harvey, and L. Andrew, "Effects of latitude and depth on the beta diversity of New Zealand fish communities," *Sci. Rep.*, vol. 7, no. July, pp. 1–10, 2017.

[11] Tom B Letessier *et al.*, "Remote reefs and seamounts are the last refuges for marine predators across the Indo- Pacific," *PLoS Biol.*, vol. 17, pp. 1–20, 2019.

[12] C. J. Torney *et al.*, "A comparison of deep learning and citizen science techniques for counting wildlife in aerial survey images," *Methods Ecol. Evol.*, vol. 10, no. October 2018, pp. 779–787, 2019.

[13] E. C. Mcclure *et al.*, "Artificial Intelligence Meets Citizen Science to Supercharge Ecological Monitoring," *Patterns*, vol. 1, no. 7, p. 100109, 2020.

[14] M. Willi *et al.*, "Identifying animal species in camera trap images using deep learning and citizen science," *Methods Ecol. Evol.*, vol. 10, no. 1, pp. 80–91, 2019.

[15] Z. Miao *et al.*, "Insights and approaches using deep learning to classify wildlife," *Sci. Rep.*, no. May, pp. 1–9, 2019.

[16] M. Lasseck, "Audio-based Bird Species Identification with Deep Convolutional Neural Networks Audio-based Bird Species Identification with Deep Convolutional Neural Networks," no. January, 2020.

[17] Y. Shiu *et al.*, "Deep neural networks for automated detection of marine mammal species," pp. 1–12, 2020.

[18] D. Rathi, S. Jain, and S. Indu, "Underwater Fish Species Classification using Convolutional Neural Network and Deep Learning. (arXiv:1805.10106v1 [cs.CV])," no. June, 2018.

[19] A. Salman *et al.*, "OCEANOGRAPHY : METHODS Fish species classification in unconstrained underwater environments based on deep learning," pp. 570–585, 2016.

[20] S. Villon *et al.*, "A Deep Learning algorithm for accurate and fast identification

of coral reef fishes in underwater videos," *PeerJ Prepr.*, vol. 6, p. e26818v1, 2018.

[21] H. Qin, X. Li, J. Liang, Y. Peng, and C. Zhang, "DeepFish: Accurate underwater live fish recognition with a deep architecture," *Neurocomputing*, vol. 187, pp. 49–58, 2016.

[22] P. Chabanet, S. R. Floeter, A. Friedlander, J. Mcpherson, and R. E. Myers, "Global Biogeography of Reef Fishes : A Hierarchical Quantitative Delineation of Regions," *PLoS One*, vol. 8, no. 12, 2013.

[23] A. P. Hercos, M. Sobansky, H. L. Queiroz, A. E. Magurran, and A. Andre, "Local and regional rarity in a diverse tropical fish assemblage," *Biol. Sci.*, vol. 280, pp. 81–101, 2013.

[24] G. E. Jones, M. J. Caley, and P. L. Munday, "Rarity in Coral Reef Fish Communities," *Coral reef fishes Dyn. Divers. a complex Ecosyst.*, pp. 88–101, 2002.

[25] L. Liu, T. Zhou, G. Long, J. Jiang, and C. Zhang, "Many-Class Few-Shot Learning on Multi-Granularity Class Hierarchy," *IEEE Trans. Knowl. Data Eng.*, pp. 1–14, 2020.

[26] A. Li, T. Luo, Z. Lu, T. Xiang, and L. Wang, "Large-Scale Few-Shot Learning : Knowledge Transfer With Class Hierarchy," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7212–7220.

[27] P. Zhuang, Y. Wang, and Y. Qiao, "WildFish : A Large Benchmark for Fish Recognition in the Wild," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, vol. 2, pp. 1301–1309.

[28] J. Wang and L. Perez, "The Effectiveness of Data Augmentation in Image Classification using Deep Learning," *arXiv Prepr. arXiv1712.04621.*, 2017.

[29] D. A. Van Dyk and X. Meng, "The Art of Data Augmentation," *J. Comput. Graph. Stat.*, vol. 8600, no. 2001, pp. 1–50, 2012.

[30] S. C. Wong, M. D. Mcdonnell, G. Adam, and S. Victor, "Understanding data augmentation for classification : when to warp ?," in *2016 international conference on digital image computing: techniques and applications (DICTA)*, 2016, pp. 1–6.

[31] L. Fei-fei, R. Fergus, S. Member, and P. Perona, "One-Shot Learning of Object Categories," *EEE Trans. pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 594–611, 2006.

[32] M. Fink, "Object Classification from a Single Example Utilizing Class Relevance Metrics," in *Advances in neural information processing systems*, 2005, pp. 449–456.

[33] C. Finn, P. Abbeel, and S. Levine, "Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks," *arXiv Prepr. arXiv1703.03400*, 2017.

[34] F. Sung, Y. Yang, and L. Zhang, "Learning to Compare : Relation Network for Few-Shot Learning Queen Mary University of London," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 1199–1208.

[35] Liu Yanbin *et al.*, "Learning to proagate labels: Transductive propagation network for few-shot learning," in *arXiv preprint arXiv:1805.10002*, 2019, pp. 1–14.

[36] J. Victor, Garcia Bruna, "FEW-SHOT LEARNING WITH GRAPH NEURAL NETWORKS," in *arXiv preprint arXiv:1711.04043, 2017.*, 2018, pp. 1–13.

[37] S. Gidaris, P. Paristech, N. Komodakis, and P. Paristech, "Dynamic Few-Shot Visual Learning without Forgetting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognitio*, 2018, pp. 4367–4375.

[38] B. Hariharan, R. Girshick, and F. Ai, "Low-shot Visual Recognition by Shrinking and Hallucinating Features," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3018–3027.

[39] A. Nichol and J. Schulman, "Reptile : a Scalable Metalearning Algorithm," *arXiv Prepr. arXiv1803.02999, 2018*, pp. 1–11, 2018.

[40] Q. Sun and Y. L. T. Chua, "Meta-Transfer Learning for Few-Shot Learning," *Conf. Comput. Vis. Pattern Recognit.*, pp. 403–412, 2018.

[41] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a Few Examples: A Survey on Few-Shot Learning arXiv : 1904 . 05046v2 [cs . LG] 13 May 2019," 2019.

[42] M. A. Jamal and H. Cloud, "Task Agnostic Meta-Learning for Few-Shot Learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[43] O. Russakovsky *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *Int. J. Comput. Vis.*, pp. 211–252, 2015.

[44] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum, "The Omniglot challenge : a 3-year progress report," *COBEHA*, vol. 29, pp. 97–104, 2019.

[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[46] S. Villon, D. Mouillot, M. Chaumont, and G. Subsol, "A new method to control error rates in automated species identification with deep learning algorithms," *Sci. Rep.*, vol. 10, pp. 1–13, 2020.

[47] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, pp. 436–444, 2015.

[48] A. Nichol, J. Achiam, and J. Schulman, "On First-Order Meta-Learning Algorithms," *arXiv*, pp. 1–15, 2018.

[49] Y. Wang, Q. Yao, and L. M. Ni, "Generalizing from a Few Examples : A Survey on Few-shot Generalizing from a Few Examples : A Survey on Few-shot," *ACM Comput. Surv.*, vol. 53, no. June, 2020.

[50] J. Pinheiro and D. Bates, *Mixed-effects models in S and S-PLUS*. 2006.

[51] H. Li, D. Eigen, S. Dodge, M. Zeiler, and X. Wang, "Finding Task-Relevant Features for Few-Shot Learning by Category Traversal," in *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, vol. 1.

Supplementary

			
Abudefduf vaigiensis	*Acanthurus leucosternon*	*Acanthurus lineatus*	*Amblyglyphidodon indicus*
			
Chaetodon auriga	*Chaetodon guttatissimus*	*Chaetodon trifascialis*	*Chaetodon trifasciatus*
			
Chromis opercularis	*Chromis ternatensis*	*Gomphosus caeruleus*	*Halichoeres hortulanus*

			
Monotaxis grandoculis	*Naso brevirostris*	*Naso elegans*	*Oxymonacanthus longirostris*
			
Pomacentrus sulfureus	*Thalassoma hardwicke*	*Zanclus cornutus*	*Zebrasoma scopas*

Supp. Fig. 1 : The 20 reef fish species considered in this study

581
582
583
584

585 *Supp. Fig. 2: Diversity of individuals of the same species and of their*
586 *environments.*

587

Supp. Fig. 3: Examples of classes' images in MiniImageNet

Supp. Table 1: Dataset usage during our experiments

Dataset name	**Building**	**Number of annotations**	**Usage**
T0	Human Annotation	69,169	Building of supports sets of 1,5,15 and 30 images for our fourth experiment
T1	Human Annotation	6,320	Building of supports sets or 1, 5, 15 and 30 images for our fifth experiment
T2	Human Annotation	13,232	testing dataset
AT0	Data augmentation applied on T0	691,690	DL training for our first experiment
AT1	Data augmentation applied on T1	63,200	DL training for our second experiment

Supp. Table 2: Mean accuracy obtained with FSL models trained with 1,5,10,15,20,25 and 30 images per species. All the images used for the supports set are from T1.

	Number of thumbnails in the Support set						
Species	1	5	10	15	20	25	30
Abudefduf vaigiensis	16.08	11.39	13.18	12.59	14.02	12.98	11.38
Acanthurus leucosternon	25.51	30.71	34.90	36.68	37.16	36.95	38.74
Acanthurus lineatus	39.86	56.04	63.52	66.21	70.02	70.93	72.50

Amblyglyphidodon indicus	25.75	26.74	28.06	28.80	30.24	32.20	32.86
Chaetodo auriga	18.16	25.68	30.67	32.63	33.78	35.42	36.56
Chaetodon guttatissimus	33.58	44.21	47.40	49.26	54.12	55.91	58.26
Chaetodon trifascialis	29.02	25.48	27.44	28.02	28.18	29.65	28.44
Chaetodon trifasciatus	38.73	50.63	53.47	60.17	63.19	64.38	66.72
Chromis opercularis	44.01	61.81	61.85	63.85	64.67	61.98	62.94
Chromis ternatensis	18.91	24.94	26.27	31.21	33.87	33.49	35.07
Gomphosus caeruleus	26.01	38.99	46.84	52.88	53.96	58.69	58.74
Halichoeres hortulanus	31.82	44.81	50.73	53.52	54.11	55.85	57.01
Monotaxis grandoculis	32.03	41.52	45.68	47.59	48.34	49.19	50.64
Naso brevirostris	47.06	54.26	61.01	59.46	59.30	62.40	64.66
Naso elegans	34.54	43.11	48.38	50.19	51.31	50.55	52.17
Oxymonacanthus longirostris	39.29	53.48	59.71	58.84	62.70	64.37	66.26
Pomacentrus sulfureus	70.90	88.18	90.92	93.08	92.67	94.13	93.93
Thalassoma hardwicke	25.64	44.40	57.80	62.03	64.97	67.86	67.96
Zanclus cornutus	18.33	31.28	37.27	41.71	41.70	42.95	44.44

Zebrasoma scopas	25.56	31.81	33.32	34.62	37.02	37.22	36.05
Mean	32.04	41.47	45.92	48.17	49.77	50.86	51.77
SD	12.70	16.93	17.69	18.00	18.10	18.52	18.96

Supp. Table 3: Probability values of the pairwise proportional test used to assess the significance of the difference between accuracies obtained through few-shot models with 5, 10, 15, 20, 25 and 30 shots and deep learning model trained on T1.

	DL	FS1	FS5	FS10	FS15	FS20	FS25
FS1	<2e-16						
FS5	0.28	<2e-16					
FS10	1.20E-08	<2e-16	3.80E-12				
FS15	<2e-16	<2e-16	<2e-16	0.0016			
FS20	<2e-16	<2e-16	<2e-16	4.10E-09	0.0379		
FS25	<2e-16	<2e-16	<2e-16	1.30E-14	8.90E-05	0.2362	
FS30	<2e-16	<2e-16	<2e-16	<2e-16	3.90E-08	0.0059	0.28

Supp. Table 4: Value of the asymptotic accuracy predicted by the NLME models, and number of natural images required for both Deep Learning architecture and Few-shot Learning architecture to reach 99% of this asymptote.

	Deep Learning		Few-Shot Learning	
Species	Number of images required to reach 99% of the asymptote.	Accuracy asymptote value	Number of images required to reach 99% of the asymptote.	Accuracy asymptote value
Naso brevirostris	1,506.47	67.43	5.04	38.10
Monotaxis grandoculis	1,769.71	62.72	7.59	38.14
Amblyglyphidodon indicus	1,766.55	62.89	5.34	41.76
Chromis ternatensis	1,492.65	67.75	6.16	46.55
Acanthurus	1,425.04	69.30	6.42	50.59

lineatus				
Zebrasoma scopas	1,776.17	62.74	5.92	54.94
Gomphosus caeruleus	1,341.80	71.51	6.90	53.15
Chromis opercularis	1,112.28	78.67	5.85	54.95
Zanclus cornutus	1,069.22	80.57	3.66	60.08
Halichoeres hortulanus	1,250.29	74.18	11.60	57.14
Acanthurus leucosternon	1,184.97	76.31	10.21	58.28
Chaetodon guttatissimus	868.64	90.93	6.46	64.95
Chaetodon auriga	1,053.06	81.31	14.97	65.94
Chaetodon trifasciatus	1,076.99	80.34	7.70	63.88
Chaetodon trifascialis	1,087.93	79.94	7.42	68.66
Pomacentrus sulfureus	960.99	85.63	15.41	72.98
Thalassoma hardwicke	994.27	84.00	5.86	78.42
Naso elegans	1,074.08	80.48	6.66	78.13
Oxymonacanthus longirostris	833.75	93.29	5.68	80.97
Abudefduf vaigiensis	785.95	96.82	4.12	89.78
Mean	1221.54	77.34	7.45	60.87