UNIVERSITAT DE BARCELONA

FUNDAMENTAL PRINCIPLES OF DATA SCIENCE MASTER'S THESIS

Deep Learning to Count Fish in Sonar Images

Author:
Penny TARLING

Supervisor:
Sergio ESCALERA
Mauricio CANTOR
Albert CLAPÉS

A thesis submitted in partial fulfillment of the requirements for the degree of MSc in Fundamental Principles of Data Science

in the

Facultat de Matemàtiques i Informàtica

January 18, 2021

UNIVERSITAT DE BARCELONA

Abstract

Facultat de Matemàtiques i Informàtica

MSc

Deep Learning to Count Fish in Sonar Images

by Penny TARLING

Counting fish in underwater imagery is a time-consuming task but gives invaluable information to biologists, conservation practitioners, and fishery managers. Deep learning can be deployed to automate this process. We demonstrate its effectiveness in the context of a rare cooperative foraging system between wild Lahille's bottlenose dolphins (*Tursiops truncatus gephyreus*) and artisanal net-casting fishers who forage together to catch migrating mullet fish (*Mugil liza*). The benefits in terms of foraging success accrued by interacting fishers and dolphins remains unclear, mostly because the murky waters complicate the estimation of mullet availability. Given that data from commercial fisheries indicate a rapid decline in the regional mullet stock, and that population monitoring indicate that the frequency at which dolphins and fishers interact has also been decreasing, it is imperative to understand the foraging benefits to both predators before this unique socio-ecological system collapses. In using underwater sonar imagery, we overcame the low water visibility when recording mullet schools. However, the resolution of these images is inherently lower than those of an underwater camera, making the task of training a machine learning model to estimate fish abundance more challenging. Thus, beyond the biological and conservation relevance for this traditional fishing practice, automatically and accurately estimating fish density in low-resolution sonar imagery comes with its own technical challenges and methodological merits.

Here we trained a convolutional neural network (CNN) with a new dataset of 500 annotated underwater sonar images to directly regress a sample image to a corresponding density map, which is then integrated to give a count estimate of the number of mullet. This technique is widely adopted in other counting tasks but has rarely been used in wildlife counting. One reason being due to the severe lack of labelled data. Inspired by works in crowd-counting, we address this challenge, with a multi-task network which learns to simultaneously rank unlabelled pairs of sample images according to number of mullet in a self-supervised task, and regresses a labelled sample to produce an estimated fish count. To account for the substantial noise in our images and the difficulties in counting fish when there are many occlusions and overlaps between individuals, we incorporate aleatoric uncertainty regularization into our approach. This both improves the accuracy in the model's predictions as well as giving the user an estimated "uncertainty" score of a given sample. Experimental results show that deep learning is effective for counting fish in sonar images, and the techniques we adopt improve the accuracy in our model predictions as well as other comparable state-of-the-art approaches: In samples containing between 0-438 mullet, our network predicted the count with a mean absolute error of 6.48, a decrease in the mean absolute error by 4.61 from our base model.

Acknowledgements

I would like to thank all my supervisors for their incredible support, effort and expertise in helping me complete this work.

To Sergio for embracing this project and guiding me to explore novel methods, and Mauricio for enabling us to work on his exciting ongoing research. As well as sharing his extensive data, Mauricio's guidance in ultimately ensuring the practical use of this project's results was invaluable.

I am also indebted to Albert for his technical support and patience with running deep models on the University of Barcelona's GPUs, and guidance in bringing the project together.

Finally, I would also like to thank all the Professors involved in the Master's Programme (Fundamental Principles of Data Science) at the University of Barcelona for their stimulating and informative courses. Without this teaching, I would not have acquired the skills needed or inspiration for this project.

Contents

Abstract **i**

Acknowledgements **ii**

1 Introduction **1**

2 Related Work **6**
2.1 Crowd Counting 6
2.1.1 Self-supervised learning 6
2.1.2 Uncertainty 7
2.2 Counting fish and other animals 7

3 Dataset **9**
3.1 Data collection in the wild and sonar technology 9
3.2 Labelled dataset 10
3.3 Unlabelled dataset 12

4 Method **13**
4.1 Base model 13
4.2 Regularising: data augmentation 14
4.3 Regularizing: self-supervised task 15
4.4 Regularizing the loss term: aleatoric uncertainty 16

5 Results **18**
5.1 Experimental setup 18
5.1.1 Comparison with state of the art: balance regularization 18
5.1.2 Summary of methods trialed 19
5.2 Evaluation metrics 20
5.3 Results of ablation study 21
5.3.1 Base model and with labelled data augmentation 21
5.3.2 Multi-task (supervised + self-supervised task) 21
5.3.3 Regularizing the loss term 22
Balance regularization 22
Aleatoric uncertainty regularization 23
Combining balance and uncertainty regularization. 26

6 Conclusion **29**

Bibliography **32**

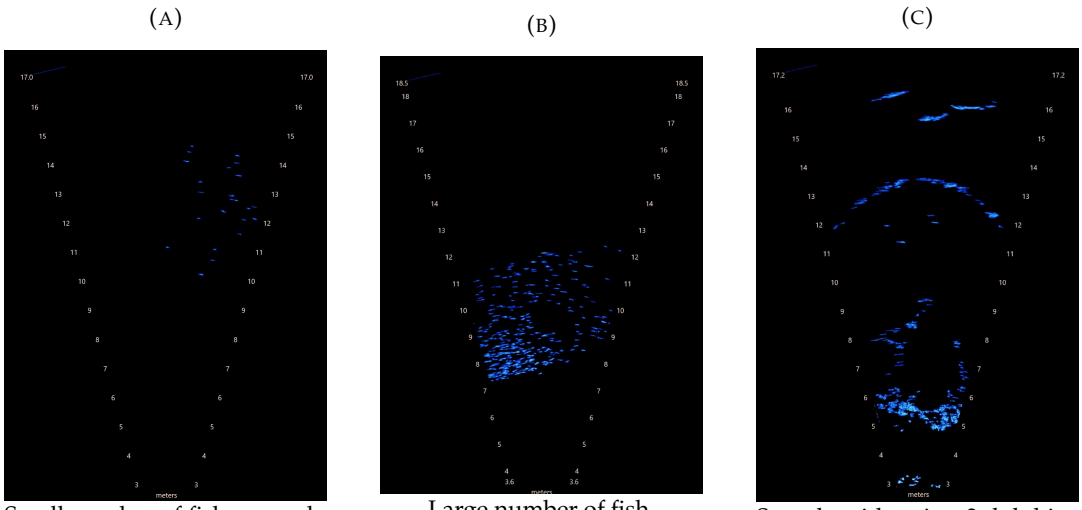
Chapter 1

Introduction

Recent years have seen a surge in interest in computer vision, and specifically deep learning, to gain a high-level understanding of images. Analysing natural images with computer vision expedite a number of important tasks, including image classification (Pham et al., 2021), facial recognition (Yan et al., 2019), object detection (Ren et al., 2015) and instance segmentation (He et al., 2017). Since these are common, time-consuming and labour intensive tasks, they have drawn much cross-disciplinary research attention.

Object *counting* is another example with its wide ranging applications: In Biology, technological advances in non-invasive sampling techniques has led an increasing demand for analytical tools that can automatically count natural 'objects' in an ever-growing volume of image and video data. Microbiologists, for example, can now rely on computer vision tools to count cells in microscopic images (Falk et al., 2019). There has also been progress towards conservation ecologists and agricultural managers being able to automatically count crops, plants and trees in aerial or satellite images (Ammar and Koubaa, 2020), and population ecologists being able to automatically quantify livestock (Xu et al., 2020). Often these problems have been addressed by object detection and image segmentation, whereby "regions of interest" are located to identify different objects. In some contexts, for example cell counting, where cells are overlapping and clumped together, these methods are not as effective. Xie, Noble, and Zisserman, 2018 overcame this by directly regressing images of cells to a cell spatial density map. The ability to automatically process a large volume of natural images, using relatively cheap and non-invasive approaches, is an enormous benefit of using computer vision to tackle old problems in Biology.

To address these problems, one can look to other well-developed applications of computer vision research, such as vehicle counting for traffic management (Oñoro-Rubio and López-Sastre, 2016) and crowd counting for security and surveillance (Gao et al., 2020). The latter in particular has been the focus of extensive research and where we have drawn much inspiration from. The common approach in crowd counting now is with the use of deep convolutional neural networks (CNNs) which directly regress an image to its corresponding density map (e.g. Cao et al., 2018; Liu, Weijer, and Bagdanov, 2018) . This density map can then be integrated to produce a final count estimate of the number of people and has the advantageous of only needing point annotations in labelling, instead of bounding boxes. In population and conservation biology, such techniques for crowd counting should be particularly useful for processing natural images with thousands of "objects" of interest, such as plants and animals. Providing a rapid and accurate estimate of individuals of different species in large volumes of images is of invaluable importance for



(A) Small number of fish, sparsely distributed

(B) Large number of fish, swimming compactly

(C) Sample with noise: 3 dolphins and fishing net

FIGURE 1.1: **Underwater sonar images from our dataset:** here we see a range of scenarios which our model needs to be adaptable to

improving biological monitoring and inform decision-making in biological conservation (Lamba et al., 2019).

While there has been recent advances in the use of deep learning for biological monitoring in terrestrial habitats (e.g. Norouzzadeh et al., 2018), comparatively less effort has been given to marine and freshwater systems. Underwater imagery analysis can expand and facilitate biological monitoring of underwater ecosystems and fisheries stock assessments. Despite a growing interest in using deep learning for this, the main focus of research has been on the task of species detection and classification (e.g. Moniruzzaman et al., 2017; Salman et al., 2019), rather than counting fish and estimating its abundance directly. In addition, this research effort typically uses images generated by single lens reflex cameras and therefore is restricted to habitats with good underwater visibility. Sonar imaging systems have become an increasingly common tool for capturing underwater images where visibility is a constraint, such as at night and in deep and very turbid waters (e.g. Boswell, Wilson, and Cowan Jr., 2008; Lankowicz et al., 2020). This is because sonar technology does not rely on light, but instead use sound energy to generate real-time digital underwater images from the returning echoes. However, the resolution of these images is inherently lower than that of underwater cameras; therefore, the task of counting aquatic animals in sonar images comes with additional challenges. It can be difficult, even for the human eye, to distinguish between "objects" which are captured without details in the sonar images (Figure 1.1). An automated counter needs to be able to accurately count the number of target species in images of very few numbers to very high, as this can vary dramatically. Furthermore, labelled, annotated data is severely limited and costly to acquire from the needed human input.

This thesis aims to solve the task of estimating fish abundance in turbid environments in a unique context (Figure 1.2): during the traditional fishing between artisanal net-casting fishers and wild dolphins targeting migrating mullet schools, in southern Brazil (e.g. Simões-Lopes, Fabián, and Menegheti, 1998; Peterson, Hanazaki, and Simões-Lopes, 2008; Cantor, Simões-Lopes, and Daura-Jorge, 2018). To do so,

we develop a deep learning model and provide a new dataset using images that have been selected and manually annotated from 105 hours of sonar video footage recorded in this unique natural setting (Cantor M, unpublished data). Beyond the technical merit and relevance in solving the task of processing low-resolution sonar-based underwater footage, there is also a practical and conservation relevance in the ability to automatically and accurately count mullet fish at the spatial scale that matters for the dolphin-fisher interaction.

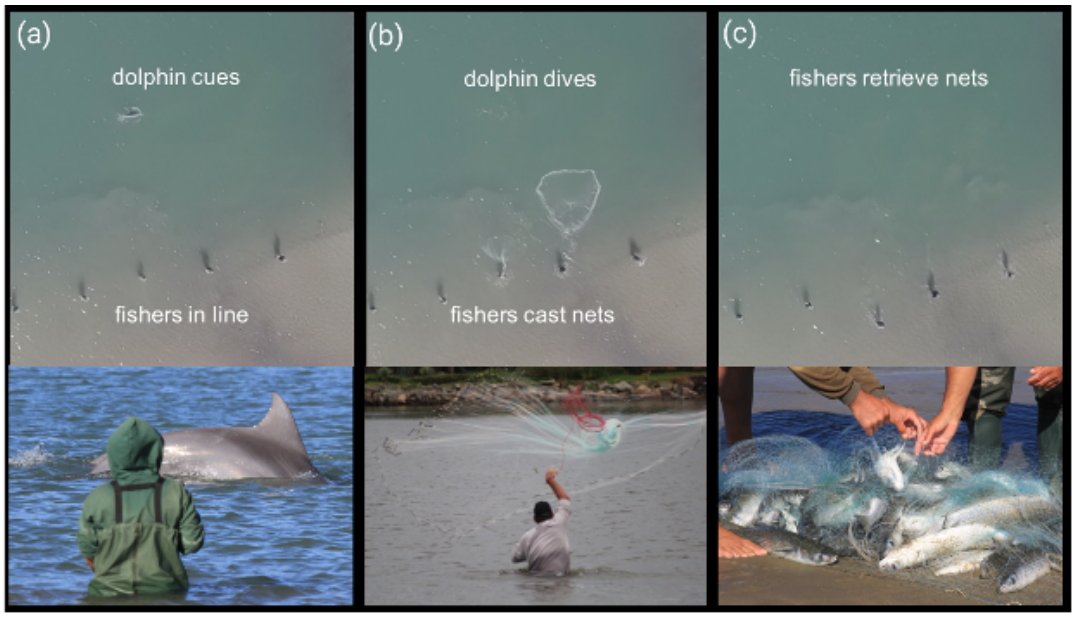


FIGURE 1.2: **The traditional cooperative foraging between wild dolphins and artisanal net-casting fishers targeting migrating mullet.** For over a century, Lahille's bottlenose dolphins have been seen herding mullet schools towards the edge of the estuarine canal in Laguna, southern Brazil, where artisanal fishers wait in shallow waters. Since the estuarine waters are murky, fishers cannot track the mullet schools but they can track the dolphins' behaviour. (a) Fishers wait in line at the edge of the canal for the dolphins' foraging cues (here, a sudden dive near the coast) which (b) fishers interpret as the right moment and place to cast their nets, presumably on top of (c) the passing mullet schools. Drone images by M. Cantor, A.M.S. Machado; Photographs by D.R. Farine, E.M. Ehrhardt, F.G. Daura-Jorge.

These fisher-dolphin foraging interactions are thought to represent one of the few remaining cases of human-wildlife cooperation. In certain estuaries in northern Argentina and southern Brazil, wild Lahille's bottlenose dolphins (*Tursiops truncatus gephyreus*) herd migrating mullet schools (*Mugil liza*) towards the coast where a line of artisanal fishers wait for stereotyped foraging behaviours by the dolphins, which they interpret as the right moment to cast their nets (Simões-Lopes, Fabián, and Menegheti, 1998). Although the traditional fishing practice between dolphins and fishers has been around for over a century (Simões-Lopes, 1991) and considered as mutually beneficial for both predators (e.g. Simões-Lopes, Fabián, and Menegheti, 1998; Daura-Jorge et al., 2012; Cantor, Simões-Lopes, and Daura-Jorge, 2018), the foraging benefits both predators accrued remains to be properly understood and quantified. The turbid waters complicate the estimation of the abundance of prey,

thus requiring a reliable method such as the sonar-based underwater imaging system for assessing the mullet schools in the estuarine waters with very low visibility in real-time.

To properly evaluate the speculative benefits of this interaction, the first steps are quantifying precisely both the (i) availability of mullet schools at the very local scale at which the dolphin-fisher interaction takes place; and (ii) the proportion of the mullet available that is caught by fishers and dolphins when interacting and when foraging independently (Cantor, In Prep.). Quantifying these benefits for interacting dolphins and fishers is crucial to determine whether their interactions are indeed mutual and, if so, to determine the minimum conditions of prey availability at which this traditional, century-old interaction can persist (Cantor, In Prep.) and remain resilient in face of the global trend of decline fisheries seen at local, regional and global scales (e.g. Worm, 2009; Pauly and Zeller, 2016; Hilborn, Amoroso, and Anderson, 2020). Given the real concern that the mullet stocks in southern Brazil are in decline (Sant'Ana et al., 2017) and causing a decline in the frequency of dolphin-fisher interactions, evaluating whether these changes can collapse this unique socio-ecological system becomes imperative (Cantor, In Prep.).

The underwater sonar images from this system pose further challenges for the automatic counting of mullet. For instance, these images can vary substantially from containing zero to several hundred mullets; dolphins and casting fishing nets can also be seen, making it difficult to distinguish between the target fish and these other "noisy objects". Here, for the first time, we combine self-supervised learning and an uncertainty regularization to count fish numbers in sonar images. Our results show an improvement in accuracy from our base model, with the implementation of these methods:

We have thus addressed the constraint of limited available labelled data by incorporating a self-supervised task into our framework and leveraging unlabelled data. Inspired by Liu, Weijer, and Bagdanov, 2018 work on crowd counting, we have generated pairs of images, where one image in the pair is a sub-section of the parent image. Thus it is known that the parent image must contain a greater or equal number of fish to its respective comparative image. It is then possible to train a Siamese network to rank these images according to fish abundance. This task is used to improve the training of the traditional supervised task of counting fish in labelled image data. We then trained these two tasks together in a multi-task network which simultaneously learns to rank the unlabelled data and count the number of fish in labelled data.

Particularly important in this context is the issue of noise in data as well as the difficulties for a human to accurately count the number of fish in densely populated images. Figure 1.1(B) shows an example image with a dense school of fish. The fish are seen simply as blue blobs with no detailed features, making it difficult to decipher the number present when they swim close together or even overlap. This will likely lead to inaccuracies, even inconsistencies, when manually annotating the data. To date, to the best of our knowledge, the handful of deep learning models applied to directly counting fish, output a point estimation only. Inspired by the recent work of Oh, Olsen, and Ramamurthy, 2020 who adopted the approach in crowd counting, we have trained our network to simultaneously produce an uncertainty measure alongside each point prediction. Not only do our results show this leads to

a higher level of accuracy, but it also provides the user with an understanding of the uncertainty surrounding a given result. The user is then able to examine these samples further or treat them with caution. When considering the application of these methods for aquatic monitoring and conservation, or fisheries and aquaculture management, over or underestimating populations could lead to adverse consequences such as biased decisions. Therefore, a greater understanding of how much a prediction can be depended on is crucial.

To the best of our knowledge, only one other study has applied deep learning to count fish in the same type of sonar images (i.e. multi-beam sonar cameras that can record moving objects): Liu et al., 2018 incorporated a regularizing technique to increase the weight of less common samples in their dataset, i.e. those with high numbers of fish. This in turn increased the overall accuracy of the network's predictions, particularly for this subgroup of samples. We experimented with this regularizing approach in our methodology, testing it with and without the self-supervised task and uncertainty regularization for comparison. Our results show that our novel ways to solving this task, improves upon this state-of-the-art approach.

In summary the main contributions of our work are:

- A new labelled dataset of 500 images taken in a natural environment along with >1million additional unlabelled images from 105hours of video
- A self-supervised task leveraging unlabelled data, to improve the supervised counting task, to the application of fish counting
- Uncertainty estimating applied to fish counting, shown also to also improve accuracy in count predictions
- An improvement in results from the existing state-of-the-art work applying deep learning to counting fish in comparable sonar images

Chapter 2

Related Work

Here we review the related studies to our work. We have taken much inspiration from research using computer vision to count people in images of crowded scenes. Thus, in what follows we explain how crowd counting has evolved in recent years and the techniques that we found especially appropriate for our task. Next, we review examples of studies on counting animals, and in particular fish, however this line of research is far less explored than that of crowd counting.

2.1 Crowd Counting

Machine learning for crowd counting firstly evolved from traditional handcrafted and detection based methods (e.g. Leibe, Seemann, and Schiele, 2005), to regression based methods, which relied on manual feature extraction (e.g. Chen et al., 2012). For a more detailed review, we refer our reader to Loy et al., 2013. Finally, as with other computer vision tasks, in recent years deep CNNs have outperformed these earlier approaches. One reason is the severe limitation with former methods in the lack of spatial and local feature information used in the training process, crucial in crowd scene analysis. Density based methods are now the common approach for this reason, with CNNs directly regressing an image to its corresponding density map.

Wang et al., 2015 was one of the first papers to introduce deep CNNs to crowd counting, with a focus on learning to ignore non-person noisy objects in images. From this seminal work, research has progressed to overcome other challenges posed by images of crowd scenes by improving the scale and context awareness of models. Zhang et al., 2016 built a multi-column network to adapt to the variation of scale present within and between images. Cao et al., 2018 advanced the computational efficiency of this approach with the use of an Inception (Szegedy et al., 2014) like architecture. Subsequently, Liu, Salzmann, and Fua, 2019 improved performance and generalisation by adaptively learning the scale of contextual information of patches within images. For a more comprehensive summary of CNN based methods, we refer the reader to Gao et al., 2020.

2.1.1 Self-supervised learning

Despite the increasing popularity, there are still a limited number of labelled crowd scene datasets. It is very time consuming to meticulously annotate an image with the location of each person, particularly in dense scenes with thousands of people. As a result, datasets tend to contain only several hundred images. To access more data in training, Liu, Weijer, and Bagdanov, 2018 leveraged unlabelled data from Google searches of crowd scenes and simultaneously trained a self-supervised task

to improve the training of the supervised crowd counting task. Pairs of unlabelled data were generated by taking a crop of one image and training a network to rank the 2 images in terms of number of people present. The crop must contain the same or fewer number of people compared to the original and so this can be carried out as a self-supervised task without the need for any additional labelling. The authors found that training these tasks in a multi-task network produced more accurate results than first training the model on unlabelled data and then fine tuning with the smaller labelled dataset, as is commonly seen when self-supervised tasks are combined with supervised tasks.

2.1.2 Uncertainty

A limitation with all CNN methods discussed so far is that their output is solely a point estimation, without any understanding of how certain this prediction is. The need to model uncertainty in computer vision was highlighted by the work of Kendall and Gal, 2017 and these ideas were incorporated into the task of crowd counting in Oh, Olsen, and Ramamurthy, 2020. Uncertainty in predictions can arise from both:

- Epistemic uncertainty: This is uncertainty in the model due to lack of knowledge. One approach to quantify this is by placing a prior over the model's parameters and determining how much these parameters will vary depending on training data. High epistemic uncertainty would result in a broad posterior distribution over the parameters. Epistemic uncertainty can theoretically be explained away with infinite data.

- Aleatoric uncertainty: This is uncertainty in the data and arises from genuine observational noise. No matter how much data we have, there will always be a degree of uncertainty in the prediction of noisy images. To quantify this in regression problems, the noise parameter also needs to be learned alongside the model weights.

In the task of crowd counting, where labelled data is limited, genuine noise and occlusions occur and there are likely inaccuracies in manually annotated images in dense scenes, it is important to have an understanding of uncertainty. This was the motivation behind the work of Oh, Olsen, and Ramamurthy, 2020 and for our application of it to fish counting. We have focused solely on heteroscedastic aleatoric uncertainty (the assumption that observational noise varies with the input data) but propose exploring epistemic uncertainty as further work.

2.2 Counting fish and other animals

Estimating fish abundance has traditionally relied either on intrusive field methods, indirect methods such as estimates of catch per unit effort from fisheries, or genetic or observational methods, such as tissue sampling and underwater surveys (e.g. Pope, Lochmann, and Young, 2010). Among the more recent non-invasive methods are the use of multimedia data, such as underwater imagery. However, a major bottleneck faced by researchers working with multimedia sampling is processing large volumes of data, which quickly become laborious and time-consuming to process manually. There have been an increasing number of research papers thus using machine learning to handle this task. Formerly, these involved hand crafted techniques

such as blob detection (Toh, Ng, and Liew, 2009) or the manual extraction of features, such as edge detection (Fabic et al., 2013), to then be used in regression techniques. These methods are however limited in a wild setting and/or in deep water, where occlusions are common and fish will naturally overlap each other in images. Deep learning can overcome these limitations. Zhang et al., 2020 deployed a hybrid CNN based on a multi-column CNN and dilated CNN to count farmed Atlantic salmon fish. This is somewhat different to our task, in that natural images were used and so fish present more distinct features. Data was also collected in an enclosed mariculture net cage so it is unlikely other noisy objects were present.

One key challenge in fish counting is the large variation in numbers between images. This is made more challenging by likely imbalances in image data generated in a wild setting. To count fish from sonar images, Liu et al., 2018 incorporated a regularizing term to address this issue and to overcome the commonly seen result in crowd counting works where numbers are underestimated and overestimated in dense and sparse crowds respectively. There is also a distinct lack of available labelled datasets to use in training. Schneider and Zhuang, 2020 used a variety of techniques to augment thousands of "side scan" sonar images from a small starting dataset to count the number of fish and dolphins, up to 34 and 3 respectively per image. Note, side scan sonars are different from the sonar camera used to collect data for this study: side scan sonars only image static non-moving targets but the camera can be attached to a moving object.

There are a few examples where deep learning methods have been used to directly count populations of other animals, although still surprisingly few given the benefit this could bring to biological research and conservation. The difficulty in obtaining labelled data is likely a limiting factor. Arteta, Lempitsky, and Zisserman, 2016 overcame this by citizen-science approaches, that is using online volunteers to annotate thousands of images of a penguin population in Antarctica. Penguins were counted with a multi-task network which predicted foreground-background segmentation to aid the direct counting task. Predictive uncertainty could also be modelled from the annotation discrepancies within an image as a consequence of multiple labellers.

Thus, there is still huge scope for further exploratory work in this area to address the constraint of available data and to deal with images captured in the wild being potentially challenging to work with. (For example, due to image quality, images being of low resolution, there being significant variation between images and the likely presence of occlusions and noise.) We have sought to make advancements in overcoming these challenges through publishing a new labelled dataset and with our multi-task network which incorporates both an auxiliary self-supervised learning task. This trains on more abundant unlabelled data, and gives an estimation of uncertainty to help identify particularly noisy data samples where predictions may be less reliable.

Chapter 3

Dataset

How we outline how the data were collected in the wild, and explain the preprocessing steps to acquire our labelled dataset along with some important descriptive statistics of it. Finally, we briefly summarise the extraction of unlabelled data, a far more straight forward procedure.

3.1 Data collection in the wild and sonar technology

Sonar-based underwater videos were recorded to quantify the availability of mullet schools (*Mugil liza*; figure 3.1(e)) during the cooperative foraging interactions between Lahille's bottlenose dolphins (*Tursiops truncatus gephyeus*) and artisanal net-casting fishers (Cantor, M. unpublished data). The videos were recorded in Laguna, southern Brazil (figure 3.1(a)), at the main dolphin-fisher interaction site, the Tesoura beach (figure 3.1(b)), a 100-m length beach at the inlet canal connecting the Laguna lagoon system to the Atlantic Ocean (e.g. Cantor, Simões-Lopes, and Daura-Jorge, 2018). The interaction site was sampled during 18 days in May-June 2018, from 09:00 to 17:00, during the peak of the mullet reproductive migration (e.g. Lemos et al., 2016), resulting in a total of 105h of underwater footage.

Since the water transparency at the lagoon canal is very low (from 0.3 to 1.5m visibility; Secchi disk, collected in situ), mullet schools were recorded by deploying an Adaptive Resolution Imaging Sonar, ARIS 3000 (Sound Metrics Corp, WA, USA; figure 3.1(c)). Such sonar imaging systems are efficient for generating real-time underwater images in conditions of zero visibility, and so it can be used to detect aquatic biota at night, or in deep, turbid and/or murky waters (e.g. Boswell, Wilson, and Cowan Jr., 2008; Lankowicz et al., 2020). The system uses 128 sound beams to emit pulses from frequencies between 1.8 MHz and 3 MHz and convert returning echoes into digital, bird's-eye view images (figure 3.1(d), figure 3.2). The sonar model and omitting frequencies were chosen specifically so they did not interfere with the dolphins, lower frequencies would overlap with their audible range. We deployed the sonar at the edge of the canal, along the line of artisanal fishers aim towards the inlet channel (figure 3.1(c,d)). Videos were recorded at 3 FPS, and the images ranged between 3-7 and 3-20 meters from the line of fishers toward the canal (figure 3.1(d)), depending on the emission frequency used. A summary of the data collection details can be found in Table 3.1

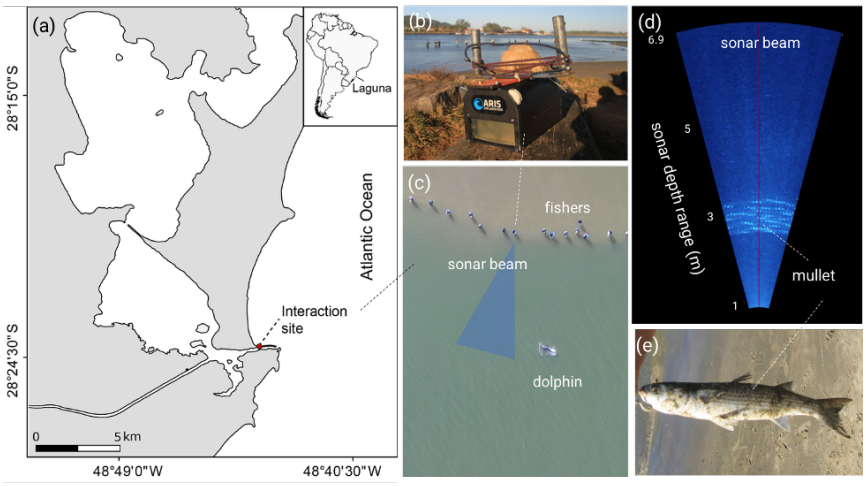


FIGURE 3.1: **Field sampling of mullet schools during the foraging interactions between wild dolphins and artisanal net-casting fishers.** (a) Study site in the canal that connects the lagoon complex adjacent to Laguna, southern Brazil, to the Atlantic Ocean. (b) The sonar camera, Adaptive Resolution Imaging System (ARIS 3000), used to generate real-time underwater videos of the passing mullet schools at the spatial scale (c) relevant for the interacting dolphins and fishers (6-20m). (d) Still from a sonar video depicting the bird's-eye view of a passing mullet school in front of the line of fishers. (e) Example of a typical mullet fish caught by the fishers. Sonar images by M. Cantor; Drone images by A.M.S. Machado; Photographs by D.R. Farine.

3.2 Labelled dataset

The ARIS video files obtained equate to over 1 million still frame images. For our study, 500 images were selected for labelling. These were carefully chosen to include a wide range of possible sample types, from low to high fish counts and from minimal to substantial noise. Because our labelled dataset is small and we wanted to maximise the chance of our model adapting to these varying observations, our dataset was not a representative sample of the field data collect. For example, we estimate substantial noise is present in $<$ 5% of images collected in the wild, but >30% of samples out of the 500 annotated, contained substantial noise. We estimate $\sim$25% of wild images contain no objects at all, but they make up just a handful of our labelled dataset as we expect this will be enough for training. Conversely, a higher proportion of the labelled dataset contain images with large, dense schools of fish compared to the true distribution, as we believe these images will require more data samples for training to achieve accurate predictions. Figure 3.4 shows the distribution of fish abundance in our labelled dataset. By training our model on the most challenging images, we can expect it to be more adapted to these. Importantly, however, without compromising on its ability to estimate abundance in straight forward images, i.e. images with limited noise and / or with sparsely disbursed and low numbers of mullet: fewer images of this type should be required for training. Testing our model with a challenging representation of the wild data will also give us more confidence to the extent it can be applied to biological research. Table 3.2 summarises the entire data collected and the annotated subset.

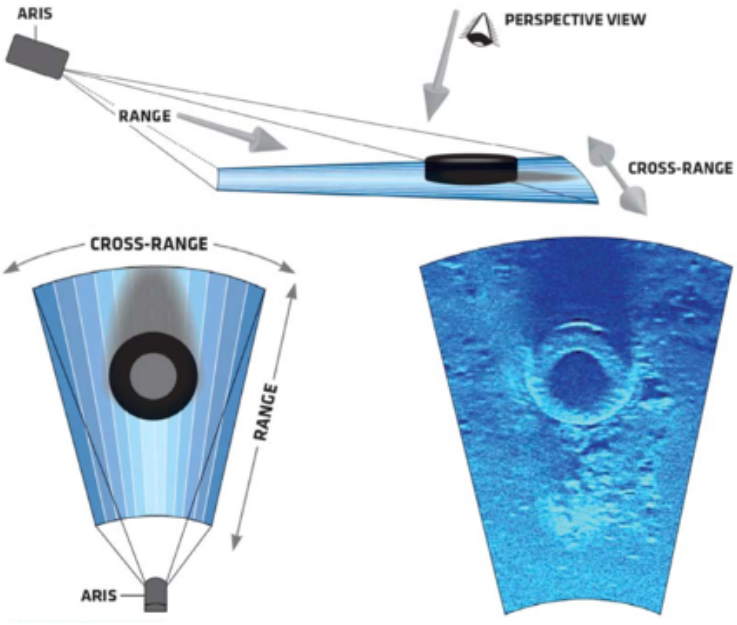


FIGURE 3.2: **Schematic of the digital images generated by the Adaptive Resolution Imaging System.** The sonar imaging system uses 128 beams to emit sound pulses and to convert their returning echoes into a digital image. Note that the sonar projects a wedge-shaped volume of acoustic energy and that the perspective is a bird's-eye view of the object (in this example a tire). These images were reproduced from the ARIScope Software User Guide v2.6.3 (©SoundMetrics Coorp)..

Date	Location	Sonar Camera	Sound Frequency	Water Visibility	Range
21/05/18 - 07/06/18	Laguna, Brazil	ARIS 3000	1.8MHz - 3MHz	0.3m - 1.5m	3m - 20m

TABLE 3.1: **Details on data sampling in the wild**

Total			Annotated Images		
Hours Recorded	Frame Rate	No. Images	No. Images	Mean (Fish)	Range (Fish)
105	3 FPS	>1million	500	41	0 - 438

TABLE 3.2: **Summary of the whole dataset and descriptive statistics on the labelled subset.**

As the videos taken were filmed at different ranges, the 500 images were cropped to show a geographical area of 4x9m^2 and all the same distance from the camera, thus biological population sampling will be comparable between input samples. The images were then all resized to the average using bilinear interpolation to 320 x 576 pixels (576 x 320 model input size). These images were annotated using the Visual Geometry Group Image Annotator (*VIA*). A point annotation was used the mark the coordinates of a fish (as close to the centre as possible) and a bounding box

was drawn around any noise (Figure 3.3). The point annotations of fish can then be used to derive corresponding ground truth density maps (Chapter 4). The bounding boxes to label noise were used for subsequent data augmentation, explained in the following Chapter 4 and Table 4.1.

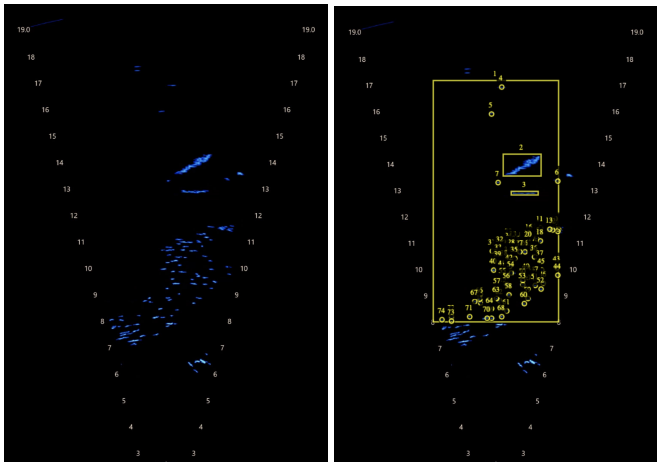


FIGURE 3.3: **Example of a sample and its corresponding annotations.** The large bounding box marks where the image will be cropped so all input samples represent a consistent size of geographical area and at a consistent distance from the camera. The smaller bounding boxes mark where noise is present. Each point marks the location of a fish.

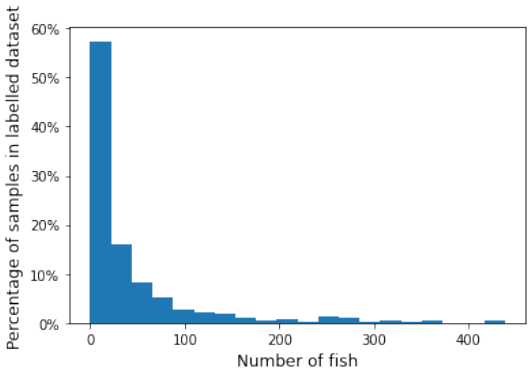


FIGURE 3.4: **Distribution of labelled dataset by number of fish.** This subset of data is skewed towards samples with low numbers of fish. This distribution is even more exaggerated in the complete dataset.

3.3 Unlabelled dataset

Unlabelled data can be extracted simply by choosing videos at random, cropping these to the same geographical area as above and resizing to 320 x 576 pixels. Subsequent crops within each sample image were then generated following the algorithm explained in Chapter 4 and Table 4.2.

Chapter 4

Method

The steps in our final methodology were firstly to augment the labelled dataset to increase training samples from 500 to 5,672. One branch of the network is thus trained on a supervised task, regressing the images to corresponding density maps. We then built a multi-task network to include the Siamese architecture to simultaneously train the self-supervised task with our unlabelled data. Finally, we added an additional layer to our output so our model learns to predict the noise variance within each labelled sample. This entire framework is shown in Figure 4.1. All code for this can be found in our GitHub repository: https://github.com/ptarling/Deep-Learning-to-Count-Fish-in-Sonar-Images

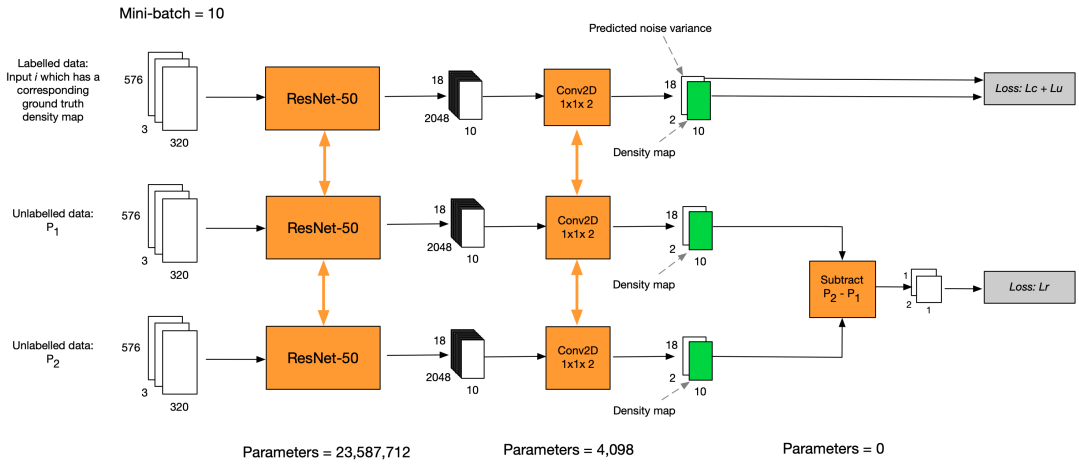


FIGURE 4.1: **Pipeline of our final network.** The multi-task network is trained end-to-end to simultaneously regress labelled images to corresponding density maps and rank the unlabelled images in order of fish abundance. All parameters are shared (represented by the orange arrows) thus incorporating the self-supervised task adds no parameters to the base model. The inclusion of an additional layer in our output to estimate noise variance only adds a further 2k parameters, equivalent to 0.01% of the total number.

4.1 Base model

The input to our network is a whole image (cropped and standardized in size to represent a consistent geographical area as explained in Chapter 3. As is common practice in crowd counting studies, we train our model to regress to a corresponding density map. Ground truth density maps were generated from original images using

a Gaussian Kernel with standard deviation 1, as seen in many related works (e.g. Cao et al., 2018; Liu, Weijer, and Bagdanov, 2018). Density maps can thus be simply integrated to give an estimated fish count. We can then train our base model directly on fish count using a simple L_1 absolute loss function which we refer to as L_c to distinguish the "count" loss hereinafter. For a single image i:

$$L_{c_i} = |c_i - \hat{c}_i| \tag{4.1}$$

where c_i is the count from the ground truth density map and $\hat{c}_i$ is the predicted count from the predicted density map.

The backbone of the base model is ResNet-50 (He et al., 2015). A deep architecture, but with identity short cut connections to improve gradient flow, and is advantageous in that weights can be initialized with those from training on the ImageNet dataset (*ImageNet*). Using pre-trained features in neural networks has been proven to improve results (Liu, Weijer, and Bagdanov, 2018). We remove the fully connected layer (FC-1000) and in its place a 1 x 1 convolutional layer is added to produce the corresponding 2D density map, 1 layer in depth (without uncertainty regularization).

4.2 Regularising: data augmentation

Starting from a relatively small dataset for training deep networks of 350 labelled images, we used a variety of techniques to multiply this. Following the work of (Schneider and Zhuang, 2020) we augmented 5,322 images believing this would be sufficient for the model to distinguish between dolphin and mullet. We propose experimenting with a greater number of synthetic images as further work. From initial exploratory experiments, we knew particular challenges to overcome in model performance were distinguishing between noise and fish and counting fish accurately in dense images (where they are seen swimming compactly and there is a high degree of overlap). Thus data augmentation was mainly generated from original densely populated images and images containing noise. Because the sonar camera used here captures objects from a "birds-eye view" (Figure 3.2), fish in images will be seen as the same relative size regardless of distance (in length) from the camera. (Depth of swimming will cause differences in scale but it does not result in great variations here.) Scale-awareness is therefore not a key factor to consider when training the network. Hence, it does not make sense to enlarge or reduce the scale of images so all crops are placed on a new blank background, similar to Schneider and Zhuang, 2020, or superimposed onto a different image. Table 4.1 shows an example of one augmentation algorithm in greater detail and some high level techniques used are listed here:

- Random crops then random placement on blank background
- Random crops of dolphin(s), which were then randomly translated and superimposed on to other non-noisy images
- Random small rotations between -20 and 20 degrees - rotations greater than this is not a realistic representation
- Horizontal flips

TABLE 4.1: **Example algorithm for data augmentation**

Algorithm 1: Superimpose dolphins onto non-noisy images	
Input:	Image with dolphin(s), number of augmented images per original
Step 1:	Crop to bounding box(s) of dolphin (this can be outside image bounding box)
Step 2:	Randomly choose new left location(s) of dolphin: from *le_bb* to *ri_bb - width of dolphin bounding box*
Step 3:	Randomly choose new upper location(s) of dolphin: from up_bb to *up_bb* + $\frac{3}{4}$*(height of image) - height of dolphin bounding box*. Dolphins not usually seen nearer to camera
Step 4:	**for** number of augmented images per original: i) Randomly choose non-noisy image iii) Superimpose dolphin(s) on to image iv) Get corresponding coordinates for fish still visible in frame **end for**
Output:	New image sample containing noise (dolphin(s))

Where "up_bb", "lo_bb", "le_bb", "ri_bb" are the upper, lower, left and right edge locations of original image bounding box.

- Random crops of fishing net(s), which were then randomly vertically translated (slightly and randomly superimposed with other non-noisy images

4.3 Regularizing: self-supervised task

To leverage the vast number of unlabelled image samples in our dataset, and compensate for the relatively small labelled training dataset, we followed the work of (Liu, Weijer, and Bagdanov, 2018) to incorporate a self-supervised task to build a multi-task network. The multi-task network simultaneously learns to rank unlabelled pairs which are generated according to alogrithm 2 (Table 4.2) as well as estimate the count of fish (Figure 4.1). Even with limited unlabelled data (which is not a constraint here), it is possible to generate 3+2+1 = 6 pairs of new data samples from a single image following this general approach (the full size image can pair with the 3 different smaller crops, 75% crop can pair with 50% and 25% and 50% can pair with 25%). The inputs to the network is now 1 image from the labelled dataset and a pair of images from the unlabelled dataset. Weights are shared between all branches, and no additional learnable layers are implemented, thus the number of parameters trained is the same as for the base model. We therefore know that any improvement in results is not due to more complexity in the model. The model is trained end to end so the dataset of unlabelled pairs increases the overall size of the training dataset with the aim of improving accuracy and generalisation of the supervised counting task.

TABLE 4.2: **Algorithm for generating unlabelled pairs of images**

Algorithm 2: Generate random pairs of unlabelled data	
Input:	Unlabelled image, size of crop k
Step 1:	**for** k in {0.25,0.5,0.75}: i) Crop image at: upper: *up_bb* + k(*image height)* left: *le_bb* lower: *lo_bb* right: *ri_bb* - k(*imagewidth*). ii) Choose random new location of crop: upper: from *up_bb* to *lo_bb - height of crop*: left: from *le_bb* to *ri_bb - width of crop* iii) Place crop on new blank background **end for**
Step 2:	Randomly choose original image, or 3/4 original image, and pair with random choice of smaller crop
Output:	Pairs of images, the larger being 100% or 75% of original image and smaller crop being 75%, 50% or 25% of original. (NB. if larger pair is 75%, this does not pair with itself).

Where "up_bb", "lo_bb", "le_bb", "ri_bb" are the upper, lower, left and right edge locations of original image bounding box.

The self-supervised task has a Siamese architecture with a global average pooling layer added to each branch to produce a single count estimate of each image in the pair. A further layer is added which subtracts the count estimate, $\hat{c}$ of image 1 in the pair, P_1 from image 2, P_2. This task can then be trained with a standard pairwise ranking hinge loss, applicable to a Siamese architecture. For a given pair i:

$$L_{r_i} = \max(0, \hat{c}(P_{2_i}) - \hat{c}(P_{1_i}) + \epsilon) \tag{4.2}$$

where ϵ is the margin of error, set to zero here. It is known from the cropping and ordering within pairs (Table 4.2), that $c(P_2) \leq c(P_1)$ and thus if the model predicts this order correctly the loss value for this pair will be 0. Otherwise the difference of the two will be added to the total loss: The greater the difference is, the greater the increase in loss. This way the model can learn the correct order within a pair according to number of fish (Liu, Weijer, and Bagdanov, 2018). It is not necessary to know the exact count of either image, hence enabling the self-supervised task.

4.4 Regularizing the loss term: aleatoric uncertainty

As we know there are varying levels of noise within our dataset, as well as there being a challenge of labelling images accurately and consistently, particularly in those displaying a high number of fish, there is obvious value in having a greater understanding of the relative uncertainty arising from each data sample. In addition, by reducing the weight of noisy samples, training should improve on non-noisy samples and overall lead to higher accuracy in predictions achieved.

Inspired by Kendall and Gal, 2017 and Oh, Olsen, and Ramamurthy, 2020 we add an additional output to our network so it will predict the noise variance, σ^2, as well as a count estimate for each image, i. We adjust our loss function with an "aleatoric uncertainty regularizer" so there is now a trade off between two components: the adjusted L_c/σ^2 and the predicted noise variance:

$$L_{c_i} + L_{u_i} = \frac{|c_i - \hat{c}_i|}{\sigma_i^2} + \log \sigma_i^2 \tag{4.3}$$

During training, the model learns to increase the value of σ^2 when the difference between c and $\hat{c}$ is large to decrease its contribution to the overall loss, but minimise the value of σ^2 when the difference is small. This way, the model is able to learn to ignore noisy samples, weakening their impact on training. The L_u component is added to the overall loss term so the model is penalised for increasing σ^2 to prevent it from simply learning to make σ^2 large for all samples.

Note the actual model output is $\log \sigma^2$ for greater numerical stability (Kendall and Gal, 2017) (to avoid dividing by zero). Also, as variance should be positive, this ensures the model cannot learn to make the predicted noise output negative to drive down loss: we multiply the L_c component by $e^{-\text{"predicted noise"}}$ which would result in a large multiple if the predicted noise variance was negative.

We add an additional output to our network. This has the same dimensions as the density map output and can be integrated for the desired scaler estimation of noise variance. An additional $\sim$ 2k parameters are trained for the network to learn this parallel prediction, but this number is negligible compared to the size of the base model. No additional parameters are trained for the regression to a density map. Thus we can be confident that any improvement in estimating fish count is likely due to the uncertainty regularisation rather than increased complexity in the model.

Chapter 5

Results

In this section we first give details of our experimental setup, including the incorporation of "balance regularization" (Liu et al., 2018) from the comparable state of the art methodology for fish counting in sonar images. We then give an overview of each of the 9 methods applied in our ablation study, followed by an explanation our experimental results. All code can be found in: https://github.com/ptarling/Deep-Learning-to-Count-Fish-in-Sonar-Images

5.1 Experimental setup

The 500 image dataset was split into 350 training images, 70 validation and 80 test. We made sure that the distribution of data in these sets was reasonably consistent to minimise bias in results. The error in predictions of the validation set was calculated at the end of each epoch throughout training and the results were used to save the optimal model weights. After all experiments were completed, we ran our models on the test data to analyse performance.

Mini-batch sizes of 10 images were used, or 10 images and 10 pairs of images for the multi-task networks (Figure 4.1). Input size of samples was 576 x 320 x 3 and thus memory capacity needed to be adhered to with small mini-batch sizes. Everything was programmed in Python and built with Keras and Tensorflow 2.2, which allows for custom training loops. As we were combining a Siamese architecture, with our base model, plus generating sample weights within each batch in some cases and / or adding a predicted variance output, we needed control over all stages of building and compiling our models. All loss terms were also custom built. We trained models where weights were initialised from the beginning (ImageNet weights, He et al., 2015; *ImageNet*) for 300 epochs and further experiments for up to an additional 200 epochs (as weights were initialised from training previously - stated in detail below). The Adam Optimizer (Kingma and Ba, 2014) was used for minimizing the loss term, with a learning rate of 0.0001. For base models, which were initialised from the start (with ImageNet weights, He et al., 2015; *ImageNet*), we lowered the learning rate to 0.00001 after 200 epochs. Final experiments required approximately 10 days of continuous training across 2 GPUs.

5.1.1 Comparison with state of the art: balance regularization

Like Liu et al., 2018, our dataset is imbalanced (Figure 3.4), a common occurrence with data collected in the wild. There are a far greater number of images containing less than 50 fish compared to those with several hundred. We therefore adjust our loss function with a sample balancing regularizer based on the approach of Liu et al.,

2018, which weights samples during online compilation of batches. We then directly compare and combine our method with the use of this regularizer. We group all images into 3 classes according to number of fish, c:

- Class 1: 75% of images, $c < 50$
- Class 2: 18% of images, $50 \leq c < 150$
- Class 3: 7% of images, $c \geq 150$

The following "information entropy-based" balance regularizer is then applied so that the weight of a sample is negatively correlated with the number of samples of its same class in the batch:

$$L_{ieb_i} = -\log\left(\frac{Nclass}{Nbatch}\right)|c_i - \hat{c}_i| \tag{5.1}$$

where $Nclass$ = number of images in the class of image i class and $Nbatch$ = number of images in batch

5.1.2 Summary of methods trialed

The following ablation studies were performed with corresponding loss function, L, for image (and pair, P) i. c_i is the actual fish count in image i, and $\hat{c}_i$ is the predicted count:

(i) **Base model**:

$$L_{c_i} = |c_i - \hat{c}_i|$$

(ii) **Base model with balancing regularizer:**

$$L_{c_i} + \beta L_{ieb_i} = |c_i - \hat{c}_i| - \beta \log\left(\frac{Nclass}{Nbatch}\right)|c_i - \hat{c}_i|$$

We found β = 0.1 worked best for our study, anything higher than this failed to train. Note this is different from Liu et al., 2018 who used $\beta = 1$ but they worked with patches of up to 8 fish meaning much smaller absolute differences in dense inputs. In contrast, the larger absolute differences in our patches can be over 50x greater than the smaller absolute differences. This high absolute difference is also seen in less common images which the regularizer adds more weight to. Because of this, $\beta > 0.1$ resulted in the regularizing term significantly dominating the overall loss function in certain mini-batches and the model was unable to learn.

(iii) **Base model with uncertainty regularizer:**

$$L_{c_i} + L_{u_i} = \frac{|c_i - \hat{c}_i|}{\sigma_i^2} + \log \sigma_i^2$$

(iv) **Base model with balancing + uncertainty regularizer:**

$$L_{c_i} + \beta L_{ieb_i} + L_{u_i} = \frac{|c_i - \hat{c}_i|}{\sigma_i^2} + \beta L_{ieb_i} + L_{u_i}$$

(v) **Base model with augmented labelled data:**

$$L_{c_i} = |c_i - \hat{c}_i|$$

This is the same network and loss function as the base model **i** but training on augmented data as well.

(vi) **Multi-task:**

$$L_{c_i} + \alpha L_{r_i} = |c_i - \hat{c}_i| + \alpha \max(0, \hat{c}(P_{2_i}) - \hat{c}(P_{1_i}))$$

For this study, we kept the hyperparameter $\alpha = 1$.

(vii) **Multi-task with balancing regularizer:**

$$L_{c_i} + \alpha L_{r_i} + \beta L_{ieb_i} = |c_i - \hat{c}_i| + \alpha \max(0, \hat{c}(P_{2_i}) - \hat{c}(P_{1_i})) + \beta L_{ieb_i}$$

(viii) **Multi-task with uncertainty regularizer:**

$$L_{c_i} + \alpha L_{r_i} + L_{u_i} = \frac{|c_i - \hat{c}_i|}{\sigma_i^2} + \alpha \max(0, \hat{c}(P_{2_i}) - \hat{c}(P_{1_i})) + L_{u_i}$$

(ix) **Multi-task with balancing + uncertainty regularizer:**

$$L_{c_i} + \alpha L_{r_i} + \beta L_{ieb_i} + L_{u_i} = |c_i - \hat{c}_i| + \alpha \max(0, \hat{c}(P_{2_i}) - \hat{c}(P_{1_i})) + \beta L_{ieb_i} + L_{u_i}$$

Experiments **i**, **ii**, **iii** and **iv** are initiated with ImageNet weights for the ResNet-50 architecture (He et al., 2015) and weights from additional layers and nodes with Xavier initialisation (Glorot and Bengio, 2010). Experiments **v** and **vi** are initiated with weights from **i**, **vii** from **ii** and **viii** and **ix** from **iii**. This way any progress seen from each ablation study can be considered fairly.

For each model, experiments were run 3 times. Thus, for methods **i**, **ii**, **iii** and **iv**, the weights were initialised from the start on 3 independent trials and these 3 independent sets of weights used to initialise the other methods, which were in turn run 3 times. An average of these 3 trial runs was then taken. 5.1.

5.2 Evaluation metrics

Following common practice in crowd counting papers, and also adopted in fish counting papers that have been discussed above, we evaluate our results using the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) where:

$$MAE = \frac{1}{N} \sum_{i}^{n} |c_i - \hat{c}_i| \tag{5.2}$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i}^{n} (c_i - \hat{c}_i)^2} \tag{5.3}$$

where N is the number of test samples, c_i is the actual count of fish in sample i and $\hat{c}_i$ is the predicted count in sample i.

TABLE 5.1: **Results on test set of each individual trial run and the average result across these**

Method		Average		Experiment 1		Experiment 2		Experiment 3	
		MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
i) Base	400	11.09	23.88	10.65	22.73	11.88	25.44	10.74	23.47
ii) + L_{ieb}	400	10.27	22.01	10.28	21.91	11.52	26.99	9.00	17.13
iii) + L_u	400	8.89	20.24	8.79	19.16	8.37	20.45	9.50	21.10
iv) + L_{ieb} + L_u	400	11.27	25.22	11.09	23.03	11.02	25.20	11.69	27.42
v) + aug. data	5,672	7.88	17.20	7.94	16.48	8.17	17.35	7.53	17.78
vi) Multi-task	5,672 + 11,344	7.05	**14.27**	7.18	15.39	6.54	13.88	7.42	13.53
vii) + L_{ieb}	5,672 + 11,344	7.87	16.67	9.91	21.81	6.16	13.13	7.53	15.07
viii) + L_u	5,672 + 11,344	**6.48**	14.81	6.26	13.66	6.31	15.12	6.88	15.65
ix) + L_{ieb} + L_u	5,672 + 11,344	7.25	16.99	7.64	19.26	6.44	16.02	7.67	15.67

5.3 Results of ablation study

Results of the experiments are shown in Table 5.1. Figure 5.1 shows the results of each method split by loose categories, to understand the effect of each ablation study on different types of samples, and in particular when analysing the impact of the targeted regularizers. Categories are shown in line with the groups chosen for assigning weights according to the L_{ieb} regularizer. Except the most common group of $c < 50$ fish / sample has been broken down further to show results for samples with less than 25 fish separately. The reason being that these images will be mostly sparsely distributed and relatively easy to count. Beyond 25 fish, more occlusions and overlaps will occur between individuals. Samples which contain large elements of noise, usually either dolphins or fishing nets, are also shown as a separate group to see how each model and associated loss term handles this particular challenge. For each group the MAE for a given sample has been divided by the average ground truth fish count for that group, plotted on the y-axis, so the error scores are somewhat normalised and can be compared between groups.

5.3.1 Base model and with labelled data augmentation

The base model, **i** achieved a MAE of 11.09 MAE and 23.88 RMSE averaged across the 3 runs. Each incremental study building on the base model, improved these scores. **v** which uses the same model architecture and loss function, but where training data has been augmented to increase the number of samples from 350 to nearly 6,000, improves the test MAE score to 7.88 and RMSE score to 17.20. This equates to a respective 3.17 and 6.68 reduction from the base model **i**, a 28% decrease in both.

5.3.2 Multi-task (supervised + self-supervised task)

As method **v** showed that training with synthetic data as well, notably improved performance, we train all our multi-task networks with this larger dataset and compare our results with **v**. The pure multi-task network, **vi**, which adds the pair-wise ranking hinge loss to the loss term and trains on unlabelled data as well as the labelled samples, reduces the MAE score to 7.05, by a further 11% and RMSE to **14.27**,

by a further 17% from **v**. It actually achieved the lowest RMSE score out of all approaches taken, showing it is the least susceptible to extreme values. This supports our hypothesis that effectively tripling the size of the training set by leveraging unlabelled data, improves the overall accuracy of the model on unseen data. Figure 5.1 allows for a greater understanding of what is driving this improvement: the last two columns show **vi** better predicts samples with high numbers of fish and samples with noise compared with **v**. This is likely because by adding unlabelled data, we increase the number of samples that fall within these two challenging categories, which will disproportionately require more training data. High number fish samples also make up the smallest percentage of the labelled training data, adding to the likelihood that models **i-v**, will not be as well adapted to this category. These samples are also the most time-consuming to annotate. Thus, adding more images like this through unlabelled data allows for an efficient way access these samples to help increase the accuracy of the corresponding predictions. A common outcome of trained neural networks' predictions, is a regression to the mean of training data. The mean of our training data is 41 fish. Thus the multi-task's ability to predict images of extreme high values compared to the single-network, is also a positive sign that it is more robust to not simply regressing to the mean.

5.3.3 Regularizing the loss term

Balance regularization

Incorporating the balancing regularizing term, L_{ieb} (Liu et al., 2018), in our base model, and with the smaller original labelled dataset, **ii**, improves the MAE and RMSE score to 10.27, by 7%, and to 22.01, by 8% respectively from the base model **i**. But it did not actually improve upon the results of the multi-task network **vi**, when added to the loss term here, **vii**. We would expect the L_{ieb} to improve results of the less common, more densely populated images as it increases the weighting of these samples. This was the case when comparing **ii** to the base model **i**, but not when comparing **vii** to the other multi-task networks, **vi, viii, ix** (Figure 5.1). In general, examining the average difference in count versus prediction, all methods suffer from under predicting the fish count in samples where noise is not present (but over predicting where noise is present). Perhaps this is due to the fact that, because a high proportion of our samples contain noise, the model trains to interpret some patterns as noise that are actually fish and in general under predicts the fish count. A common outcome of deep learning models, is a regression to the mean so the generated prediction is equally right or wrong for all samples; this could also be seen here with underestimating for non-noise and overestimating for noise. When looking at the raw average differences instead of absolute difference, it is apparent however that both **ii** and **vii** methods, which include the balancing regularizer, on average over predict the fish count for the less common class of relatively dense images, where $50 \leq c < 150$ fish. The regularizer therefore has had the expected effect on these samples in driving predictions up and overcoming the commonly seen outcome in crowd counting papers of under prediction in dense scenes. It has however over predicted these samples too much, resulting in still a higher overall MAE than comparative methods. (Figure 5.1 - Base + L_{ieb} and Multi-task + L_{ieb} show highest "normalised" MAE versus comparative models for $50 \leq c < 150$ fish). This prediction behaviour was not seen in very dense images, $c \geq 150$, where **vii** generally under predicted the fish count. This also resulted in the highest MAE compared to

other multi-task networks, **vi, viii, ix** (Figure 5.1). One plausible explanation could be overfitting these types of samples during training by adding too much weight to them. Hence, not performing so well on unseen data.

It is also evident that **ii** and **vii** are relatively less able to predict images with substantial noise. Figure 5.1 shows the "normalised" MAE score for **ii** was notably greater than the base model **i** and the base model with uncertainty regularization **iii**, and that of **vii** was greater than the pure multi-task network **vi** or multi-task with uncertainty regularisation **viii**. This can likely be explained because noisy images tend to contain a below average number of fish. So the balancing regularizer is actually also reducing the relative weight of noisy images in training, resulting in less accurate predictions on noisy test data.

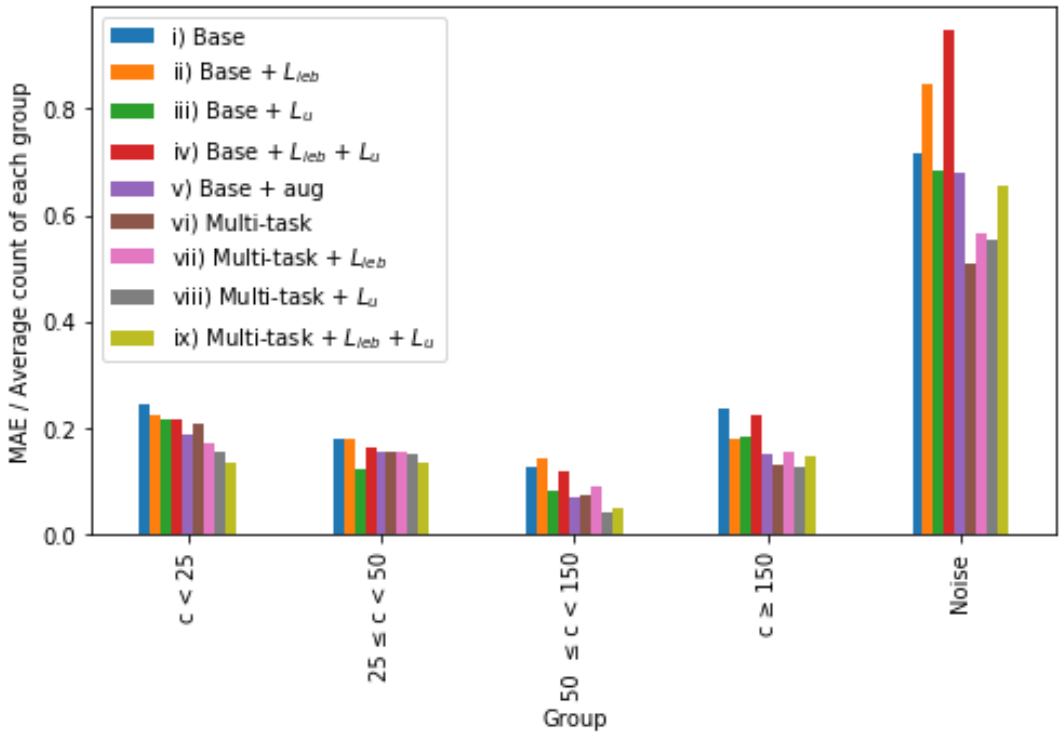


FIGURE 5.1: **Error analysis for samples grouped into categories depending on number of fish or noise present.** The MAE for a sample has been divided by the average actual count within each group so results are somewhat normalised and can be compared between groups. The percentage of samples that fall within each group are: $c < 25$: 34% , $25 \leq c < 50$: 10%, $50 \leq c < 150$: 14%, $c \geq 150$: 9%, Noise: 34%. The reason c < 50 (our first class for balancing regularization) is altogether lower than 75% as stated in 5.1.1, is because many of these samples have been put into the "Noise" category for this analysis.

Aleatoric uncertainty regularization

Adapting the loss term to include the L_u uncertainty in methods **iii** and **viii** regularisation improved both the comparative base model, **i** and the multi-task network

vi. In fact, the multi-task network with uncertainty regularization **viii** achieved the lowest MAE error score of **6.48** out of all the approaches tested, a 0.57, or 8%, improvement on the second best score from the pure multi-task network, **vi** and a 4.61, or 42% improvement from our base model **i**. Figure 5.1 suggests how the regularizer has affected training by showing how well adapted the model is to different sample categories. Interestingly the multi-task + L_u achieves the best results (lowest "normalised" MAE score on the bar plot) for more densely populated fish images with limited noise, where images contain more than 50 fish. As for **vi**, but to an even greater extent, this is a positive indication that it is not simply regressing predictions to the mean. But it works comparatively less well when noise is present (the last column shows it produced a "normalised" MAE score higher than the pure multi-task network, **vi** . This is as expected because essentially adding this regularizing term, allows the model to ignore noisy images in training so it learns to predict non-noisy images more accurately, but in turn there will be a trade off in its ability to handle noisy images.

The second benefit of the this approach should be a reliable predicted noise variance. Figure 5.2 includes scatter plots for the best performing three out of the four methods where the loss term has been modified with uncertainty regularization, **iii**, **viii** and **ix** to see if there is a correlation between each model's two outputs: predicted noise variance (σ^2, outputted as $log\sigma^2$) and predicted count (from which we calculate the absolute error). The left scatter plots, show there is a moderate positive correlation between the absolute error score and predicted variance score of a sample. For the two multi-task networks, the correlation statistic r, reaches 0.68 and 0.73, without and with the balancing regularizer respectively ($p < 0.001$, one-tailed test, in all cases). The histogram plots show variance predictions are heavily skewed towards lower values. This is also seen in the scatter plots where a high number of samples are clustered in the bottom (left) corners. Over 80% of samples have a predicted noise variance (σ^2) of < 3 for any of these three methods. In practice the user can then choose to treat sample results with high relative variance scores with caution, investigate further or even ignore altogether. In turn this should lead to better accuracy in interpreted findings.

Three test sample images along with their corresponding ground truth and predicted density maps are shown in the first two columns (A & B) of Figure 5.3 so results can be compared locally. The density maps can be interpreted like typical heat maps, where areas of red indicate dense regions of mullet. Prediction outputs are from the base model, **i**, (C) and multi-task with uncertainty regularization, **vii**, the best performing model from our experiments for comparison (D). The first test image contains no noise, but has a relatively high number of mullet, 74, swimming compactly in places resulting in occlusions and overlaps, and thus making it difficult to distinguish between individuals. The base model performs reasonably well, over predicting by 7 (91% accuracy) but we can see the improvement in accuracy with **vii** which gives a perfect count prediction (100% accuracy). The predicted noise variance σ^2 is 2.96, which is meaningful. It is < 3, so falls outside the top 20% of high predicted variance score samples, which is good as the prediction is relatively accurate and we would not wish to disregard it, but it is still relatively high suggesting there is some sample noise (in this case occlusions) making it more difficult to count.

In the case of the middle image, mullet are more sparsely distributed so should

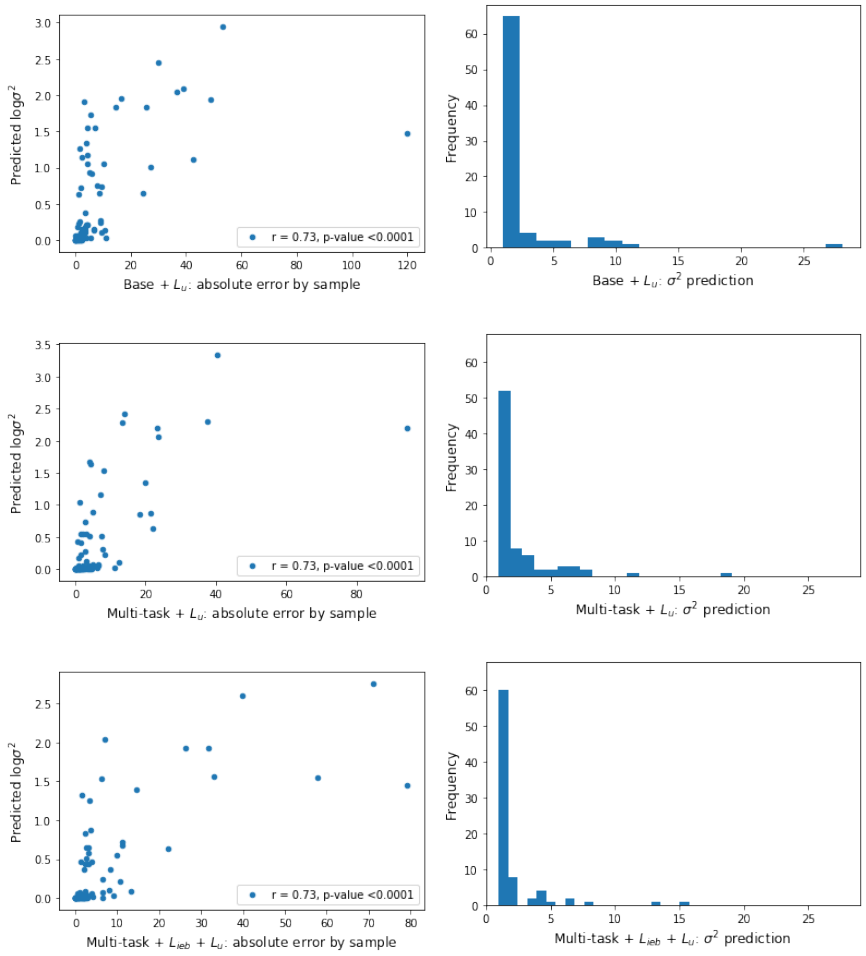


FIGURE 5.2: Left: **Scatter plots to show relationship between predicted noise variance and absolute error score, for methods iii, viii and ix.** Right:**Histogram presenting corresponding distributions of predicted noise variance scores.**

be easier to count. Indeed model **vii** (D) gives a near perfect prediction (96% accuracy) and we can see how well the ground truth and predicted density map coincide. There is a random blue blob in the bottom left corner. This is could be a faint blob visible to the machine but missed by the human eye. Arguably the automated count could be more reliable here! Notably, the predicted noise variance is just 0.08, a useful score given the lack of noise and accuracy in prediction. The base model, (C) under predicted the count by 9 (63% accuracy). Given by the lack of red where fish are seen swimming more compactly, it seems it did not detect overlaps between fish instead counting individuals together.

The bottom test image contains substantial noise (fishing net). Both models clearly identify parts of the fishing net as noise as we cannot see red patches in the density map corresponding to the areas in the test sample with most prominent nets. Furthermore, if the models assumed all blue patterns seen here in the test image were fish, we would expect the predicted count to be be much higher. Both models predict the count to be 14 and it appears to be counting the blue objects inside the fishing net as fish. This is even inconclusive for a human labeller; they could either be fish or splashes as a result of the fishing net. Anything seen within fishing nets was consistently manually labelled as noise as it was too difficult to distinguish and thus we expect some discrepancies from model predictions and our ground truths. Even though the multi-task + L_u performs slightly worse than the base model here (taking the predictions with a higher degree of precision), it simultaneously gives a high predicted noise variance of 4.66, alerting the user that there is likely noise and potential inaccuracies in this estimation.

Combining balance and uncertainty regularization.

When modifying the loss term in our base model to include both the balancing and uncertainty regularizer, **iv**, we do not see an improvement in results than using just one of these regularizers, **ii** or **iii**, or even from the base model itself, **i**. MAE was 11.27 and RMSE 25.22, both higher than the results of the base model, 11.09 and 23.88 respectively. This outcome was seen again from our experiments with a multi-task network: we do not see an improvement in performance from multi-task with both loss function regularizing terms, **ix**, compared with multi-task with only uncertainty regularization, (Figure 5.1). 7.25 MAE compared to **6.48**. In fact, overall predictions are less accurate than when even compared to the pure multi-task model **vi** which gave an MAE score of 7.05. It is therefore apparent that these two regularizers do not compliment each other and the uncertainty regularizer beats the balancing regularizer in performance.

This pattern is particularly highlighted in the results for samples containing substantial noise (Figure 5.1). The far right column shows **iv** and **ix** produced the highest ("normalised") MAE score out of all the single-task and multi-task networks respectively. As explained above, this is likely because both regularizers have the effect of reducing the weighting of noisy images in training and thus when used together, the performance on this type of data in testing is worse. We do however see the best comparative results for more sparsely populated test samples. This is a somewhat surprising result as both regularizers are counter-acting each other in their behaviour in terms of relatively weighting these samples. One explanation may be that images with small numbers of fish, sparsely distributed, should be straight forward

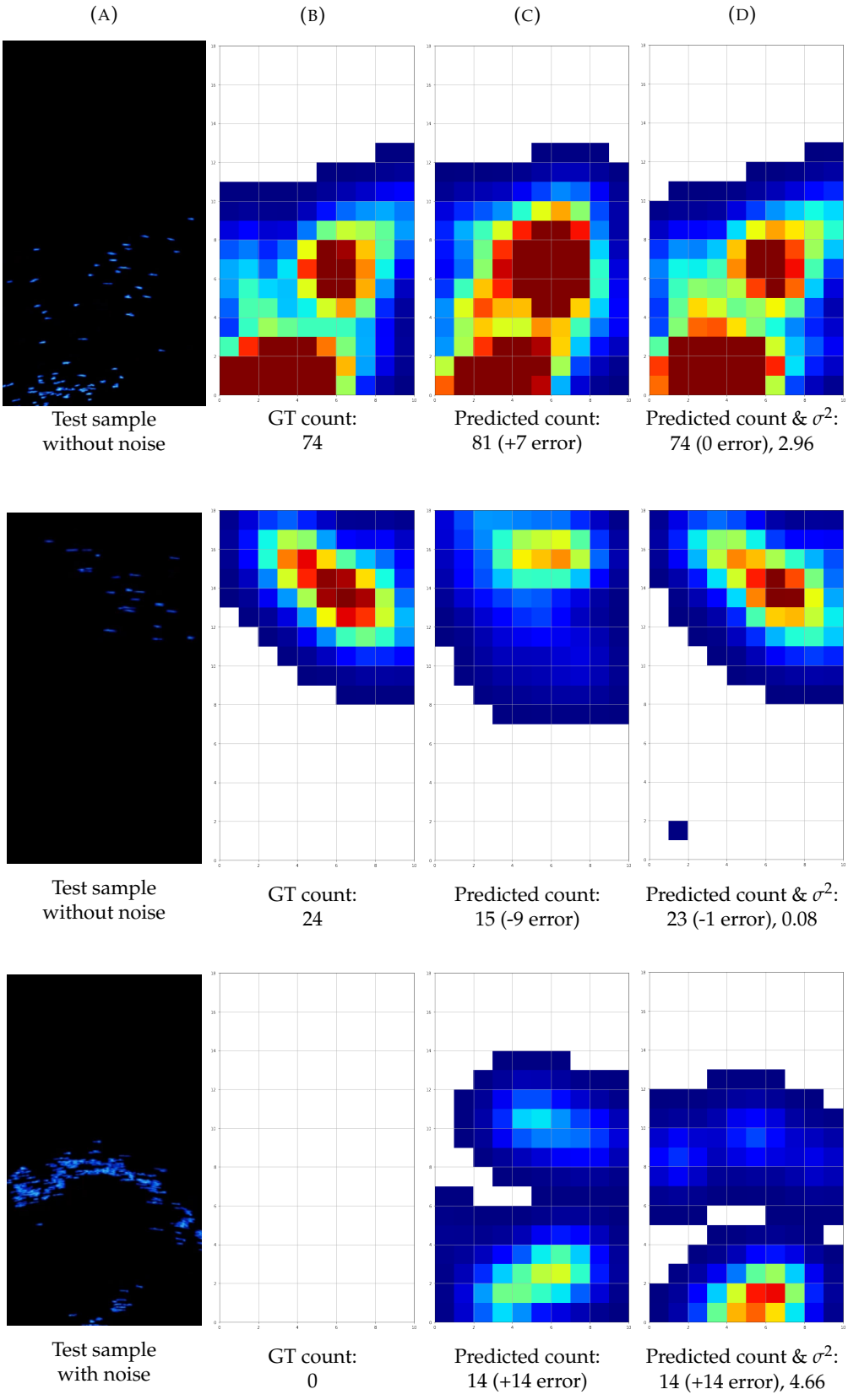


FIGURE 5.3: From right to left: (A) example images with (B) corresponding ground truth (GT) and predicted density maps for (C) base model, **i**, and (D) Multi-task modified with uncertainty regularization, **viii**, the best performing model in our experiments. The density maps can be interpreted as a typical heat map where areas of red indicate dense regions of mullet

for the network to count accurately and require less training. Thus even though the balancing regularizer decreases the weights of these samples, it also compounds the effect of the uncertainty regularizer of decreasing the weight of noisy samples. In turn, this could lead to an harmonious intersection of smaller weighting given to low fish number samples in training, but an even more pronounced reduction to the impact of noise, that leads to the best performance on unseen data for this class.

Chapter 6

Conclusion

In this thesis, for the first time, we combined a self-supervised task with uncertainty regularisation in a deep learning model to count mullet fish (*Mugil liza*) in underwater sonar images. The key challenges with this type of problem are the lack of data (particularly annotated data) available, the low resolution of sonar images and noise present in data collected in the wild. We address these by 1) introducing a new large dataset, with a subset of 500 labelled images, 2) leveraging unlabelled data in a self-supervised task and 3) incorporating a measure of uncertainty to not only improve training but gain a greater understanding of the noise present in samples.

We show that deep learning is an effective way to count fish in images which are low resolution, present many occlusions and overlaps between individuals and can contain significant levels of noise. Nine different methods with varying datasets, model architectures and loss functions were trialed 3 times. We see an improvement in predicted results on our test data with each novel technique (in the context of fish counting) implemented: from innovative ways to augment the annotated data, to building a multi-task network to simultaneously train the self-supervised task, to finally combining this with uncertainty regularisation. From this, we were able to obtain an average MAE score of just 6.48, which was a 42% improvement on our comparative base deep network. Note, this is the result for a biased sample, deliberately chosen to contain a significantly higher proportion of noisy samples and those with dense schools of fish. This enabled us to test the model with the most challenging data but we are confident the average accuracy in predictions will be substantially greater with a true sample representation of the data.

We put forward two models for practical use depending on the specific use case: The multi-task network with uncertainty regularisation which leads to the lowest MAE averaged over all samples and gives an approximation of sample noise. And the multi-task network without uncertainty regularisation which is more accurate in predicting fish count when noise is present and is more robust to extreme counts. Both these models were able to estimate the number of mullet present to a sufficient level of accuracy for biological research, in samples ranging widely in the number and spatial distribution of fish.

Our approach can be compared to that used by the state of the art in deep learning for fish counting in moving sonar imaging (Liu et al., 2018) (we have only found one other paper attempting this with our type of sonar images). Our methodology builds upon this by adding a Siamese network to a regression model in a multi-task network. We incorporated the balancing regularization proposed in their paper with both our base model and multi-task network but we found in both cases that our approach of using uncertainty regularization was more effective here.

From labelling a dataset such as this, we understand the challenges it presents. It is difficult even for the biological experts to determine fish numbers in very dense images and at times distinguish between noise and fish. To this end, there will very likely be inaccuracies, bias and even inconsistencies in the labelling which will have affected the training capacity of the model and lead to discrepancies between predictions and ground truths. For this particular real-life application, we propose therefore incorporating a classifier in the multi-task network. From examining predicted density maps we believe the model will be good at detecting categorically whether noise is present or not, e.g. if a fishing net is present in a sample, even if it cannot accurately detect it at every pixel within a sample. As this type of noise is apparent in a very small proportion of images in the whole dataset, these sample results could be discarded from population estimation altogether, if the classifier detects this type of noise. For wider applications, similar classifiers can be built depending on the particular challenges at hand.

Building upon our derived network, future research can explore a more sophisticated back-end. With the ResNet architecture as the front end, our input samples have been be reduced by 2^5 and so we have likely lost valuable information through this. Particularly as counting problems should make use of spatial relationships and local patterns, training a model to learn to regress back to the original size of the input through learnable upsampling could be beneficial to achieving more accurate results. Thus this could be a fruitful avenue for further work. We propose exploring a U-Net type architecture, with a decoder backend, first proposed in medical imaging research (Ronneberger, Fischer, and Brox, 2015), or a dilated convolutional backend as used in Oh, Olsen, and Ramamurthy, 2020 and Zhang et al., 2020.

From our results, we know that training on unlabelled data with the incorporation of the pair-wise hinge loss boosts performance. There is potential for further improvement by experimenting with, and fine tuning the hyperparameter, α, the multiple of the hinge loss that this is added to the overall loss term. Given the benefit we found of incorporating aleatoric uncertainty into training and prediction, suggested next steps could be to broaden this with the inclusion of epistemic uncertainty. Particularly as we know there must be a degree of model uncertainty given our relatively small labelled dataset. Furthermore, measuring epistemic uncertainty will allow for a confidence interval range around each point prediction, resulting in a more comprehensive interpretation of population estimations.

Another interesting line of research could be with the experimentation of regularizing the loss term to account for accuracy in results in terms of percentage error rather than simply absolute error. In training, the model could be penalised more for higher absolute errors on lower ground truth counts of fish, to help reduce the percentage error in predictions. This metric is not usually incorporated or discussed in crowd counting papers or even in comparable fish counting papers (Liu et al., 2018) but it could be a useful metric to measure from a biologist's point of view.

Even though we have applied our approach to a specific, unique, cooperative foraging system between Lahille's bottlenose dolphins (*Tursiops truncatus gephyreus*) and artisanal fishers in Laguna Brazil, the effective techniques we have applied can

be used to analyse wide ranging underwater imaging. This can be hugely valuable in being able to automatically access underwater populations in a cheap, efficient and non-intrusive way support conservation efforts worldwide as well as wide-ranging biological research (Lamba et al., 2019).

Bibliography

Ammar, Adel and Anis Koubaa (May 2020). "Deep-Learning-based Automated Palm Tree Counting and Geolocation in Large Farms from Aerial Geotagged Images". In: *arXiv e-prints*, arXiv:2005.05269, arXiv:2005.05269. arXiv: 2005.05269 [cs.CV].

Arteta, C., V. Lempitsky, and A. Zisserman (2016). "Counting in the Wild." In: *ECCV*. URL: https://doi.org/10.1007/978-3-319-46478-7_30.

Boswell, Kevin M., Matthew P. Wilson, and James H. Cowan Jr. (2008). "A Semiautomated Approach to Estimating Fish Size, Abundance, and Behavior from Dual-Frequency Identification Sonar (DIDSON) Data". In: *North American Journal of Fisheries Management* 28.3, pp. 799–807. DOI: https://doi.org/10.1577/M07-116.1. URL: https://afspubs.onlinelibrary.wiley.com/doi/abs/10.1577/M07-116.1.

Cantor M., Daura-Jorge FG Farine DR (In Prep.). *Foraging synchrony drives ecological and cultural resilience of human-dolphin mutualism*.

Cantor, Mauricio, Paulo C. Simões-Lopes, and Fábio G. Daura-Jorge (2018). "Spatial consequences for dolphins specialized in foraging with fishermen". In: *Animal Behaviour* 139, pp. 19 –27. ISSN: 0003-3472. DOI: https://doi.org/10.1016/j.anbehav.2018.03.002. URL: http://www.sciencedirect.com/science/article/pii/S0003347218300812.

Cao, Xinkun et al. (2018). "Scale Aggregation Network for Accurate and Efficient Crowd Counting". In: *ECCV*, 734–750.

Chen, Ke et al. (Jan. 2012). "Feature Mining for Localised Crowd Counting". In: *British Machine Vision Conference*. DOI: 10.5244/C.26.21.

Daura-Jorge, F. G. et al. (2012). "The structure of a bottlenose dolphin society is coupled to a unique foraging cooperation with artisanal fishermen". In: *Biology Letters* 8.5, pp. 702–705. DOI: 10.1098/rsbl.2012.0174. URL: https://royalsocietypublishing.org/doi/abs/10.1098/rsbl.2012.0174.

Fabic, J. N. et al. (2013). "Fish population estimation and species classification from underwater video sequences using blob counting and shape analysis". In: *IEEE*, pp. 1–6. DOI: 10.1109/UT.2013.6519876.

Falk, Thorsten et al. (2019). "U-Net: deep learning for cell counting, detection, and morphometry". In: *Nature Methods* 16, pp. 67–70.

Gao, Guangshuai et al. (2020). "CNN-based Density Estimation and Crowd Counting: A Survey". In: *ArXiv* abs/2003.12783.

Glorot, Xavier and Y. Bengio (Jan. 2010). "Understanding the difficulty of training deep feedforward neural networks". In: *Journal of Machine Learning Research - Proceedings Track* 9, pp. 249–256.

He, Kaiming et al. (2015). "Deep Residual Learning for Image Recognition". In: *CoRR* abs/1512.03385. arXiv: 1512.03385. URL: http://arxiv.org/abs/1512.03385.

He, Kaiming et al. (2017). "Mask R-CNN". In: *CoRR* abs/1703.06870. URL: http://arxiv.org/abs/1703.06870.

Hilborn, Ray, Ricardo Oscar Amoroso, and Christopher M. Anderson (2020). "Effective fisheries management instrumental in improving fish stock status". In:

Proceedings of the National Academy of Sciences 117.4, pp. 2218–2224. ISSN: 0027-8424. DOI: 10.1073/pnas.1909726116. URL: https://www.pnas.org/content/117/4/2218.

ImageNet. URL: http://www.image-net.org/.

Kendall, Alex and Yarin Gal (2017). "What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?" In: *CoRR* abs/1703.04977. URL: http://arxiv.org/abs/1703.04977.

Kingma, Diederik and Jimmy Ba (Dec. 2014). "Adam: A Method for Stochastic Optimization". In: *International Conference on Learning Representations*.

Lamba, Aakash et al. (2019). "Deep learning for environmental conservation". In: *Current Biology* 29.19, R977 –R982. ISSN: 0960-9822. DOI: https://doi.org/10.1016/j.cub.2019.08.016. URL: http://www.sciencedirect.com/science/article/pii/S0960982219310322.

Lankowicz, Katelynn M. et al. (2020). "Sonar imaging surveys fill data gaps in forage fish populations in shallow estuarine tributaries". In: *Fisheries Research* 226, p. 105520. ISSN: 0165-7836. DOI: https://doi.org/10.1016/j.fishres.2020.105520. URL: http://www.sciencedirect.com/science/article/pii/S0165783620300370.

Leibe, B., E. Seemann, and B. Schiele (2005). "Pedestrian detection in crowded scenes". In: 1, 878–885 vol. 1. DOI: 10.1109/CVPR.2005.272.

Lemos, Valéria et al. (May 2016). "Tracking the southern Brazilian schools of Mugil liza during reproductive migration using VMS of purse seiners". In: *Latin American Journal of Aquatic Research* 44, pp. 238–246. DOI: 10.3856/vol44-issue2-fulltext-5.

Liu, L. et al. (2018). "Counting Fish in Sonar Images". In: *2018 25th IEEE International Conference on Image Processing (ICIP)*, pp. 3189–3193. DOI: 10.1109/ICIP.2018.8451154.

Liu, W., M. Salzmann, and P. Fua (2019). "Context-Aware Crowd Counting". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liu, Xialei, Joost Weijer, and Andrew Bagdanov (June 2018). "Leveraging Unlabeled Data for Crowd Counting by Learning to Rank". In: pp. 7661–7669. DOI: 10.1109/CVPR.2018.00799.

Loy, Chen Change et al. (Oct. 2013). "Crowd Counting and Profiling: Methodology and Evaluation". In: vol. 11. DOI: 10.1007/978-1-4614-8483-7_14.

Moniruzzaman, Md. et al. (2017). "Deep Learning on Underwater Marine Object Detection: A Survey". In: *Advanced Concepts for Intelligent Vision Systems*. Ed. by Jacques Blanc-Talon et al., pp. 150–160.

Norouzzadeh, Mohammad Sadegh et al. (2018). "Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning". In: *Proceedings of the National Academy of Sciences* 115.25, E5716–E5725. ISSN: 0027-8424. DOI: 10.1073/pnas.1719367115. URL: https://www.pnas.org/content/115/25/E5716.

Oh, Min-hwan, Peder Olsen, and Karthikeyan Ramamurthy (Apr. 2020). "Crowd Counting with Decomposed Uncertainty". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34, pp. 11799–11806. DOI: 10.1609/aaai.v34i07.6852.

Oñoro-Rubio, Daniel and R. López-Sastre (2016). "Towards Perspective-Free Object Counting with Deep Learning". In: *ECCV*.

Pauly, Daniel and Dirk Zeller (2016). "Catch reconstructions reveal that global marine fisheries catches are higher than reported and declining". In: *Nature Communications* 7, p. 10244.

Peterson, Débora, Natalia Hanazaki, and Paulo César Simões-Lopes (2008). "Natural resource appropriation in cooperative artisanal fishing between fishermen and dolphins (Tursiops truncatus) in Laguna, Brazil". In: *Ocean Coastal Management* 51.6, pp. 469 –475. ISSN: 0964-5691. DOI: https://doi.org/10.1016/j.ocecoaman.2008.04.003. URL: http://www.sciencedirect.com/science/article/pii/S0964569108000458.

Pham, Hieu et al. (2021). *Meta Pseudo Labels*. arXiv: 2003.10580 [cs.LG].

Pope, Kevin, Steve Lochmann, and Michael Young (Jan. 2010). "Methods for Assessing Fish Populations". In:

Ren, Shaoqing et al. (2015). "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks". In: *CoRR* abs/1506.01497. URL: http://arxiv.org/abs/1506.01497.

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox (2015). "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *CoRR* abs/1505.04597. URL: http://arxiv.org/abs/1505.04597.

Salman, Ahmad et al. (Feb. 2019). In:

Sant'Ana, Rodrigo et al. (2017). "Bayesian state-space models with multiple CPUE data: the case of a mullet fishery". In: *Scientia Marina* 81.3, 361–370. DOI: 10.3989/scimar.04461.11A. URL: http://scientiamarina.revistas.csic.es/index.php/scientiamarina/article/view/1727.

Schneider, Stefan and Alex Zhuang (2020). *Counting Fish and Dolphins in Sonar Images Using Deep Learning*.

Simões-Lopes, Paulo C. (1991). "Interaction of coastal population of Tursiops truncatus (Cetacea, delphinidae) with the mullet artisanal fisheries in southern Brazil". In: *Biotemas* 4, pp. 83–94.

Simões-Lopes, Paulo C., Marta E. Fabián, and JoÃ O. Menegheti (1998). "Dolphin interactions with the mullet artisanal fishing on Southern Brazil: a qualitative and quantitative approach". en. In: *Revista Brasileira de Zoologia* 15, pp. 709 –726. ISSN: 0101-8175. URL: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0101-81751998000300016&nrm=iso.

Szegedy, Christian et al. (2014). "Going Deeper with Convolutions". In: *CoRR* abs/1409.4842. arXiv: 1409.4842. URL: http://arxiv.org/abs/1409.4842.

Toh, Y. H., T. M. Ng, and B. K. Liew (2009). "Automated Fish Counting Using Image Processing". In: *IEEE*, pp. 1–5. DOI: 10.1109/CISE.2009.5365104.

VIA. URL: https://www.robots.ox.ac.uk/~vgg/software/via/.

Wang, Chuan et al. (2015). "Deep People Counting in Extremely Dense Crowds". In: *Proceedings of the 23rd ACM international conference on Multimedia*.

Worm B., Hilborn R. Baum J. K. Branch T. A. Collie J. S. Costello C (2009). "Rebuilding global fisheries". In: *Science* 325, pp. 578–585.

Xie, Weidi, J. Alison Noble, and Andrew Zisserman (2018). "Microscopy cell counting and detection with fully convolutional regression networks". In: *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* 6.3, pp. 283–292. DOI: 10.1080/21681163.2016.1149104. URL: https://doi.org/10.1080/21681163.2016.1149104.

Xu, Beibei et al. (2020). "Automated cattle counting using Mask R-CNN in quadcopter vision system". In: *Computers and Electronics in Agriculture* 171, p. 105300. ISSN: 0168-1699. DOI: https://doi.org/10.1016/j.compag.2020.105300. URL: http://www.sciencedirect.com/science/article/pii/S0168169919320149.

Yan, Mengjia et al. (2019). "VarGFaceNet: An Efficient Variable Group Convolutional Neural Network for Lightweight Face Recognition". In: *CVPR*. arXiv: 1910.04985 [cs.CV].

Zhang, S. et al. (2020). "Automatic Fish Population Counting by Machine Vision and a Hybrid Deep Neural Network Model". In: *Animals* 10.2. URL: https://www.mdpi.com/2076-2615/10/2/364.

Zhang, Y. et al. (2016). "Single-Image Crowd Counting via Multi-Column Convolutional Neural Network". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.