

Expertise Guided Saliency Prediction in an Underwater Context

Jason Summers

903702

Submitted to Swansea University in partial fulfilment
of the requirements for the Degree of Master of Science



Swansea University
Prifysgol Abertawe


Department of Computer Science

Swansea University

30th September 2020

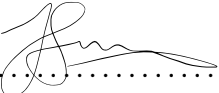
Declaration

This work has not been previously accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed  (candidate)
Date 29/09/2022

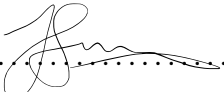
Statement 1

This work is the result of my own independent study/investigations, except where otherwise stated. Other sources are clearly acknowledged by giving explicit references. I understand that failure to do this amounts to plagiarism and will be considered grounds for failure of this work and the degree examination as a whole.

Signed  (candidate)
Date 29/09/2022

Statement 2

I hereby give my consent for my work, if accepted, to be archived and available for reference use, and for the title and summary to be made available to outside organisations.

Signed  (candidate)
Date 29/09/2022

Abstract

Offshore wind farm (OWF) inspections are subject to many challenges both physical and contextual, footage is often noisy and full of naturally salient objects with little to no relevance to the inspection. One of the key challenges of these inspections is effectively finding multiple/similar frames in the footage that show defects. This is because conventional embedding methods might focus too heavily on irrelevant but salient objects. In this research I conduct a pilot user study to capture the eye movements of experts and non experts (in the area of OWF inspections) while viewing underwater footage. This data was then used to train a model to generate salient masks mimicking expert and non expert gaze patterns. The model successfully replicated gaze locations and expert polarity based on the data from only two experts and five non experts. The model showed evidence of identifying relevant and irrelevant regions but more data will be needed to reliably judge the semantic relevancy of objects, hence this work serves as the foundation for a future study.

Acknowledgements

Firstly, I would like to thank my supervisory team Professor Mark Jones, Dr Su Yang and Dr Martin Porcheron for their guidance and counsel during my project. I would also like to thank my external/industry supervisor Catherine Seale and Vaarst as a whole for their fantastic support. Finally I would like to thank my family and fellow CDT researchers at Swansea for helping me get to where I am now.

Contents

1	Introduction	1
1.1	Offshore Assets	1
1.2	The Value of Expertise	3
1.3	Research Questions:	3
1.4	Collaborations	4
1.5	Human Centred Design and Responsible Innovation	4
2	Literature Review	7
2.1	Machine Learning	7
2.2	Discovery Phase	8
2.3	Summary	15
3	Methodology	16
3.1	User Study	16
3.2	Final Data	19
3.3	Model Design and Training	22
4	Results	25
4.1	Training and Validation Loss	25
4.2	Enhancing	26
4.3	Final Saliency Maps	26
5	Conclusion and Discussion	28
5.1	Future Work	28
	Bibliography	29

Appendices	33
A Implementation of Important Algorithms and Functions	34

Chapter 1

Introduction

1.1 Offshore Assets

In recent years there has been a greater worldwide recognition of climate change and the threat that it poses [1] and with this there has been a greater interest in developing effective renewable energy sources. One such renewable energy source is wind farms which are commonly set up both in windy land areas and offshore zones. These offshore wind farms (OWF) utilise the large amounts of open ocean across the world and feed back this energy to the mainland. Looking at the UK alone, OWF made up 50% of all renewable power generation in 2014 with numerous OWF projects in development and increases in subsidies from the government [2]. This interest shows no signs of stopping with developments and OWF power capacity being predicted to increase faster than in the last two decades [3]. Indeed the global OWF capacity is projected to increase fifteen-fold to 2040, becoming a \$1 trillion industry over the next two decades [4].

OWFs, and other offshore structures and assets, are in greater need of inspection and monitoring compared to structures on land. This is because structures in an underwater environment may be subject to biological growth, turbulent currents, corrosion and physical degradation, all of which may contribute to an increase in fatigue [5]. Check-ups and maintenance for these structures presents many challenges along the same thread, due to the depth and length of the inspections, many structure surveys are conducted with remotely operated underwater vehicles (ROUV or ROV) or autonomous underwater vehicles (AUV). These vehicles are used to record hours of footage using a range of sensors, particularly optical. This footage can then be analysed to identify where there is damage in the structure. The engineers responsible for inspecting structural conditions can find themselves having to scroll through hours of footage to find the few instances of highly

important events (e.g. damage) that may be present. The large volume of data that is captured requires the appropriate tools, machine learning (ML) is adept at handling big data to learn or replicate patterns [6] and is therefore the perfect tool for this job. However, the conditions for how this data was captured are highly specific and have many characteristics that set it apart from many other areas of study in ML, these characteristics present many unique challenges and problems.

Problems with Underwater Media

Although the learning capacity of ML is great, this can sometimes be a double edged sword. If provided flawed data, a model might learn to predict or replicate based on these flaws. It is therefore important for both a models robustness and trustworthiness that any and all problems present in the raw data are discussed and understood.

Challenge 1 - Occlusion and Illumination: Inconsistencies in visibility can arise from the varying depths, capture angles, sedimentation and properties of the water which can greatly affect a models ability to classify elements of an underwater image [7]. Objects such as bubbles, debris and biological life can often block regions of the footage, potentially causing the inspector to miss key parts of a structure. Light degradation can also be an issue, the Lambert-Beer empirical law states that decay in the intensity of light depends on the properties of the medium through which the light travels, so the water itself can alter the colour and illumination of objects [8, 9]. For every 10 meters of depth underwater, the light available to us is halved, so from the light available at the surface, just 50% of the light will be available at 10 meters down, and only 25% 20 meters down [10]. Visibility can also vary with weather, storms can create turbulent water, creating a more complex medium that amplifies the issues above. Light degradation and absorption are not constant for all wavelengths of light, longer wavelengths like red and orange are more easily absorbed in contrast to shorter wavelengths like blue and green. Light from the surface is therefore altered with a filtering the deeper you go down, meaning the footage often has a blue-green tint.

Challenge 2 - Distractions: As mentioned, footage for OWFs are often hours long but only contain a few instances of interest. One approach to finding these instances is to group frames in the footage based on similarity. This is a task that can be performed with ML processes such as image embedding which may capture and learn semantic elements of an image. In this regard the previously mentioned objects (such as bubbles, fish, etc.) do more than block segments of the image. These objects are naturally highly salient with respect to the rest of the image and

may have the capacity to skew the similarity of an images representation in the embedding space towards other images that also show these objects. For inspections these objects may not relate to any defects and thus can harm search efforts.

Challenge 3 - Non-stable Video Capture: Turbulent water causes the ROV to have unpredictable movement and video capture and can lead to a number of issues such as those mentioned in challenge 1, beyond these it can also hinder the temporal value that video could provide. The stability of footage is an important factor in accessing the trajectory of objects in the footage [11] which in turn, can affect the object detection capabilities of ML models [12, 13]. This is unfortunate as structural defects are likely to be generally stable in contrast to fish, bubbles, etc. and this added complexity makes it harder to utilise the temporal component to categorise irrelevant objects based on movement.

1.2 The Value of Expertise

Of course there remains the question; how can we define relevance for a ML model? Inspections of underwater assets require an understanding of how damage can occur in such an environment, what elements of the footage are indicative of damage, and knowledge of the assets themselves. Incorporating this knowledge into a model is vital to combating the issues behind challenge 2, but is this something that can be easily captured as data? With these inspections being primarily visual, one way to capture knowledge may be to track the way an expert observes the footage and to observe what objects or regions they focus on.

1.3 Research Questions:

For this research I set out to answer a number of questions:

Question 1: What are the technical and user based issues behind the current survey methods in industry?

Question 2: Is it possible to translate human experience into data?

Question 3: Can a model learn to identify relevant and non relevant areas within media in a technical domain?

1.4 Collaborations

This project was done in collaboration with a stakeholder called Vaarst, a Bristol based company focused on automating offshore asset inspection. By attending biweekly meetings with my industry supervisors I was able to gain valuable ideas and directions to ensure that my research remained connected to the key issues in the field. Furthermore it enabled me to talk directly with engineers, surveyors and ML researchers working with the current inspection system, which provided me with feedback and knowledge on the system's issues.

1.5 Human Centred Design and Responsible Innovation

This project was also done in partnership with the Engineering and Physical Sciences Research Council (EPSRC) who are dedicated to conducting responsible research and innovation (RRI) in fields that improve our interaction with technology and with the world around us.

1.5.1 RRI

With the fast growth of AI and ML it is more important than ever to consider the potential consequences and risks behind our research. The EPSRC and parent organisation UK Research and Innovation (UKRI) define responsible innovation as following the following principles: [14]

- Responsible innovation is a process that seeks to promote creativity and opportunities for science and innovation that are socially desirable and undertaken in the public interest.
- Responsible innovation acknowledges that innovation can raise questions and dilemmas, is often ambiguous in terms of purposes and motivations and unpredictable in terms of impacts, beneficial or otherwise.
- Responsible innovation creates spaces and processes to explore these aspects of innovation in an open, inclusive and timely way.

1.5.2 Human Centred Design

Working directly with an industrial partner means this project must be done with an understanding of how any produced algorithms or techniques will be implemented in such an industry. It is important that these techniques do not generally endanger jobs. Hence my projects design was based around improving the current tools used by professionals, making their jobs easier and

allowing the industry to grow. The models produced in this project were also designed around the way current professionals in the subject field operate. By focusing on and capturing how experts conduct their job, any implementations may work more harmoniously with these experts.

Transparency:

These improvements will however affect decisions in the real world if implemented in the future. Thus these improvements should have, when possible, an element that informs us how a decision was altered. In the case for this project, a core component of the design is the use of salient masks. These masks can easily form heat maps indicating where in an image the model considers to lie the most important objects. This is not only designed to aid future models in comparing the contextual similarity of images, it is also designed for us to judge how it compared the images. As a result of this explainability, we may be able to build our trust in these models.

1.5.3 Sustainability

Another avenue of impact to consider is how this research affects the greater world. The UN have long promoted a series of sustainable development goals [15], this project contributes to several.



Goal 7 - Affordable and Clean Energy.

This project has the potential to enable more effective and less time consuming inspections of OWFs, subsequently the cost for developing and maintaining such clean energy sources may decrease. With more development and research in this area, the industry will be able to grow, bringing us closer to a point where we do not have to rely on fossil fuels.



Goal 9 - Build resilient infrastructure, promote inclusive and sustainable industrialization and foster innovation.

These methods need not be kept for one company or nation but contribute to the worldwide industry of underwater inspection. A large collection of the world have a stake in offshore enterprises, be it in energy, conservation or research and by being open about our research every community can benefit and grow.



Goal 13 - Take urgent action to combat climate change and its impacts.

This goal works hand in hand with goal 7, by encouraging the growth of the sustainable energy industry, we can begin to prevent further damage to our planet. Of course this alone is not a solution to climate change, but it means we can tackle one major source of the problem.



Goal 14 - Conserve and sustainably use the oceans, seas and marine resources for sustainable development.

As mentioned in goal 9, some underwater enterprises are conservation efforts, and projects like this one, which focus on relevant salient prediction, can be used to increase the accuracy of image classifiers or labellers. This could lead to better documentation of fish populations, or the conditions of the underwater landscape.

Chapter 2

Literature Review

In order to tackle these issues we will first need a basic understanding of the modern machine learning landscape, then it will be possible to define key regions of research that can tackle these issues. This literature review will consist of two main sections that mirror this reasoning. First I will cover the basic and then relevant ideas in the field of machine learning. Secondly there will be a discovery phase that covers papers that were systematically found using relevant search terms.

2.1 Machine Learning

We start with first defining the field/technology of machine learning, broadly speaking, machine learning is simply statistically refined functions capable of predicting or replicating patterns. There are many types of machine learning but all share a reliance on data. This statistical basis has made machine learning the obvious tool for many problems in a plethora of fields based on seemingly unpredictable behaviours and can therefore be found in many industries. This versatility has been one of the major reasons why machine learning quickly became one of the fastest growing technical fields over the last twenty years [16]. Another reason for this growth is that machine learning studies algorithms that have the ability to improve, that is to say they learn, with experience [17, 16]. It is no coincidence that this mirrors human nature, as you will see it is a habit of researchers in this field to copy the success of nature. Finally there is the more straightforward point of technology development, the increased availability of data and the general increase in power and affordability of computer components [18] has meant that we are able to do faster and deeper experimentation.

2.1.1 Types of Data; Supervised and Unsupervised Learning

The types of data available for a problem may vary, in some cases it is easy to obtain *labels* that help describe the data to the algorithm, in other cases it is not possible to obtain such details. This will change the way one approaches the problem and there are naturally different methods to deal with different situations, each with their positives and negatives. The extent to which a model or algorithm is given labels to help it improve is referred to as its supervision. Some key areas on the spectrum of supervision are as follows:

- **(Fully) Supervised learning:** The model is provided labels for each element of the data.
- **Unsupervised learning:** The model is provided no labels.
- **Semi-supervised learning:** A small portion of the data has provided labels.
- **Reinforcement learning:** The model is only provided a numerical score for guidance.

As mentioned, the initial raw video data for this project was based on long inspections, these videos lack any form of object labelling and with the previously outlined issues, attempting to label using previously trained models may provide some issues. Referring back to the research question 2 on page 3, one of the aims of this research will be to translate experience into data. This experience can help define regions that are and are not relevant, and thus, it may be possible to label the data, not in a way that simply looks at object classification but rather in a way that reflects the way an expert would analyse the image.

2.1.2 Areas of Interest

2.2 Discovery Phase

The following papers were found using particular search terms inputted into Google Scholar, the first 20 papers were selected for each term. These search terms were selected after my initial research, and because they reflect many of the defining challenges behind the topic of this paper. From here I will define a keyword based filter to search the abstracts of each paper, this is to eliminate any papers that are likely to be irrelevant. One reason for this approach is to ensure that there is no bias when selecting literature, the degree to which a paper supports or opposes my initial ideas has no bearing on its selection. The order the paper is reviewed in is arbitrary. Each section is referred to by the search term that it relates to. Although I will consider and review all

papers that result from this filtering process in order to reduce bias, there will be some inherent bias in the way Google Scholar presents these results, such as the influence of a papers popularity.

2.2.1 Underwater Data Labelling

To start with we will observe some of the current methods with dealing with underwater data, specifically I wish to address the problem of detecting regions of interest in visual data. Hence, the 20 papers were filtered on the basis that the abstracts contained any of the keywords **detect**, **visual**, **mask** but did not contain the phrases **3D** or **navigation**. Some of the papers included in the search related to 3D data, namely research for underwater navigation and point cloud data which does not strongly relate to the topic of this paper, hence their removal.

Bhattacharai et al. [19] conduct research into the detection of bubbles to help the mapping of underwater CO₂ seepage points from the seafloor with good results. They use the You Only Look Once (YOLO) framework, a realtime object detector [20], to train on a custom labelled dataset. In many ways this research has links to my own, the starting data was video footage subject to the same visual handicaps. Bhattacharai et al. chose to split the video into frames to be treated as separate images. The team then manually labelled the bubbles for a selection of images for training. The results of the study were promising, the detector was able to accurately detect the majority of bubbles in the foreground after training on a small dataset of 35 images. A second dataset of 50 images, including 15 segmented images, was then used for training and resulted in a greater number of background bubbles being detected. Bhattacharai et al. provide great detail on the workings of the YOLO framework and its application in an underwater setting. They demonstrate the benefits of preprocessing, showing how even a relatively small dataset can get excellent results. However the degree to which this model can generalise is likely low due to the amount of data used.

Konovalov et al. [21] trained a binary classifier to detect the presence of fish in a given image. They did so with a significant dataset of images that utilised both publicly available images and manually labelled images originating from video frames from a range of habitats. The detector itself was based off of the Xception architecture, a convolutional neural network (CNN) [22] with a modified output for binary classification. Multiple configurations were made, including one where the global max-pool was moved to be the last layer and the one-class dense layer was converted to a convolution layer. This configuration produced a heat map of values between 1 and 0, signifying areas where fish are present. This method also produced the best results of the configurations considered. Segmenting highly salient objects such as fish is a comparatively easier

task to identifying defects, which require experience and knowledge to recognise. Identifying these already salient objects is a small step to defining what is or is not relevant. Although this work focuses on identifying regions not of interest, in this case fish, the idea of using masks to indicate regions is an interesting and effective one.

Among the papers found were a number of studies on seagrass detection, two of which utilise different forms of the region-based CNN (or simply R-CNN). Moniruzzaman et al. [23] provide a deep outline of the Faster R-CNN architecture [24] and its use in an underwater context. They also create a dataset with the unfortunate inclusion of images taken in a laboratory context, it could be said that these images have unrepresentative lighting that could affect the models performance. Pamungkas et al [25] conduct a similar study except with a model based on mask R-CNN [26], an extension on Faster R-CNN. The main difference being an improvement in masking and object segmentation abilities. The training process is described in much detail including information on valuable preprocessing techniques. Finally Raine et al. [27] compare Resnet and Visual Geometry Group (VGG) models in classifying the species of seagrass. They address common issues with underwater datasets, eventually collecting their own dataset. By dividing the images into patches, they were able to efficiently categorise regions without the need for dense polygon, pixel or bounding box labels within each training image. Research that produces datasets on specific forms of biological life, like these papers, may not seem important for the case of OWF inspections initially and indeed they may not contribute much to my current research, but this may not be the case for future research. They offer us the chance to create a specialised dataset that could train a model to identify, and isolate, what might be classified as distracting.

Next are two studies focused on the detection of fish, similarly to the seagrass papers we can see the usage of both Faster and mask R-CNN methods. Conrady [28] conducts a thorough investigation into the population of a local species of fish. He captures, forms and analyses his own dataset, training a model to a validation of 80.35% for detecting and segmenting fish images. This paper serves mostly as a proof of concept for the application of the mask R-CNN in the underwater context. Ottaviani et al. [29] goes beyond to not only discuss variations in training methodology for Faster R-CNN, but also to highlight the phenomenon of concept drift, detailed as "...a drop of performance over the time, mainly caused by the dynamic variation of the acquisition conditions." [29].

Also found was Tarling et al. study on a density-based regression approach to count fish in low resolution sonar images [30]. This study conducts another "video to frame" style of preprocessing,

labelling a small portion to form both an unlabelled and labelled dataset. This allowed them to train a multi-task network, consisting of one supervised branch that learns to estimate fish count via a density mapping, and a separate set of parallel branches that ranks unlabelled images based on number of fish. In the context of this counting task, their model proves to be highly effective, and even generalisable, shown by its use and success when applied to DeepFish dataset [31]. Although this paper shows Resnets applicability to underwater images, this counting task has little connection to the problems for my paper.

2.2.2 Attention Mechanism Based Saliency

Next to explore is the literature for general model architectures around detecting saliency of objects. Included in the search is the term 'attention mechanism', this is in reference to an important area of research in ML that contain models such as transformers, which are designed to focus on important parts of an input [32]. More specifically, these models contain an exchange of a sequence of hidden variables or states between encoders and decoders, these hidden states correspond to entries in the input sequence and thus allow the output to be influenced most by specific parts of the input sequence [33]. This mechanism's use in predicting important regions of an image make it an ideal tool for my problem. In order to effectively filter the search results, my filtering will include terms that relate to eye tracking, distinctions between relevant and non-relevant saliency, and exclude papers focused on text and facial data. Thus the 20 papers were filtered on the basis that the abstracts contained **video**, **fixation**, '**releval**' (to include relevant and relevance) and did not contain **text** or **face**.

First we have a pair of publications from Wang et al., focused on examining and building on the field of video saliency. The first of the two, published in 2018, lays some needed foundations for the field by introducing a new dataset named DFK1K [34], consisting of a thousand video sequences of varying subjects, contexts and background complexity. This variation sets this dataset apart from similar studies, it contains many manually added labels describing the movement of the camera or subject and the type of subject (human, animal, object, etc.). Data on the saliency of the content of the video was captured using a thorough eye tracking study. This study consisted of 17 participants who were shown videos in the dataset and tracked using a Senso Motoric Instruments (SMI) RED 250 tracker. This was followed by a new model design using an attention based CNN-LSTM (long short term memory) architecture to map a saliency mask based on the frame of a video. The incorporation of temporal elements made this model a strong candidate for this project, and the design for the eye tracking study is somewhat a paragon but the extent to which I can replicate

this type of study is questionable. This is for two reasons, limitations in time, and high volume and noisiness of data. Wang et al.'s follow up publication in 2019 under IEEE updates this work, finalising the proposed model, known as Attentive CNN-LSTM Network (ACLNet) [35]. This concludes the development of an important benchmark for video saliency detection, one that benefits the design of the eye tracking experiment in this project.

Next in the results is a paper by Dahou et al. [36] who developed a multi-stream saliency model for use in an unorthodox context. The main structure comprises of both an encoder-decoder system and system focused on each face of a cube representing the 360° data. Positive results in comparison to the state-of-the-art models at the time shows the value in an encoder-decoder design. Although they highlight the efficacy and relevance for attention based approaches to saliency prediction, the significant differences in subject footage limit the applicability of the main components of this research. Liu et al. [37] show an alternative encoder-decoder architecture that utilises tokens to help not only identify salient objects but also to generate boundaries. The immediate issue with this research is that although the latent representation and tokens generated add to a system that is shown to be effective, the datasets the model was tested on typically had centred subjects and backgrounds with little noise. Furthermore the predictions being made simply consider the saliency of the entire object and make no distinction in saliency between a persons face and their shoes for instance, the entire person is labelled as salient. This method may do well at identifying the distracting objects of underwater footage but they ultimately ignore the subtleties of complex scenes. The previous work of Wang et al. with saliency based on eye tracking serve to better approximate a humans definition of what is salient in an image and possibly by extension, what is important.

Yan et al. [38] provide a VGG based model to predict visual saliency, included is a semantic perception subnetwork. This subnetwork sets this model apart from other saliency predictors, by adjusting channel features, the model is able to incorporate the importance of an object into the saliency prediction. This importance is verified by examinations of datasets with rich semantic subjects. Further innovations are made through re-parameterisation, helping to increase the efficacy and robustness of the final product. The inclusion of all the subsystems is shown to improve the predictions during validation. This paper provides a fresh outlook that is highly suitable to this project provided that a similar importance system can be defined with regards to experts in the field. Next is a paper by Wang et al. [39] who deliver a straightforward proposal of a saliency prediction model that utilises an improved version of the feature pyramid attention (FPA) design.

This project is notable for the ways in which it experiments with feature extraction without raising the computational cost, improving previous methods by reducing training times.

Finally, we come to an important paper by Lou et al. [40] who present a new architecture that incorporates vision transformers into a CNN encoder-decoder system to capture contextual information and predict visual saliency. Specifically the model uses an adjustable CNN encoder backbone linked with multiple transformers. Each transformer is fed a different set of feature maps extracted from the CNN encoder, these are then fused together in the decoding stage to produce a saliency map. The result is a model that identifies perceptually relevant components of an image. This model does exceptionally well when evaluated on public datasets and by combining many desirable factors in respect to this projects aims, it becomes an ideal candidate architecture to replicate human inspection.

2.2.3 Eye Tracking Saliency Video

The final search term is designed to apply the knowledge from the previous terms and gain some insight for carrying out an effective eye tracking study. To limit the results to those that concern data similar to mine, I have eliminated papers who's abstracts contain the words **3D** (or similar) and **360**. I also removed two papers that have already been covered by the previous term. Finally I have limited the results to include those who's abstracts contain **map**, **frame** or **expert**. This is to, again, close in on the core challenges of this research.

First is a paper by Coutrot and Guyader [41] who present a fusion method, combining eye tracking data with mapping for dynamic and static saliency to form a master saliency map. Data is examined, plotted and analysed at the frame level to observe the temporal evolution of the feature weights, this was done for 50 frames (2 seconds) per video. The result is a number of observations on the behaviour of participants, including the finding that these feature weights dramatically vary as a function of time and of the semantic visual category of the video. Specifically they highlight, with support from other research, the changes in exploration strategies when observing a video, the phenomenon of centre bias at the start of stimuli and the impact of an objects movement on its saliency. This study is very comprehensive and covers the intriguing subject of temporally impacted saliency, one that may prove useful for classifying relevant saliency for OWF footage. Chevet et al. [42] also present a fusion method but not to the same usefulness as Coutro and Guyader. As the title of their paper suggest, this is not a method that directly uses eye tracking data as guidance, the

objective is instead to replicate human attention. Perhaps some aspect of relevancy can be defined by irrelevancy, and therefore by defining what is distracting by mimicking human attention, I may be able to tackle some of the issues I have outlined.

Next is an impressive paper by Tsiami et al. [43] who look to predict the saliency of an object in a video with an enriching process. This enrichment is done utilising the audio from the video, as well as eye tracking data. This new layer yields a model that largely out performs the other state-of-the-art models. The inclusion of audio allow the model to affix relevant information to the objects in the footage making the predicted saliency more contextually relevant. The major downside with regards to my own research is that not only is audio not included in the I data obtained, but audio underwater would suffer comparable distortions to the visual element of underwater footage.

Lyudvichenko et al. [44] utilise eye tracking data from one observer to produce high quality saliency maps that mimic multiple observers. The objective is to use saliency maps to prioritise bit allocation for important/salient regions. They describe a transformation process incorporating a Gaussian blurring filter to better model visual attention. They conclude that many modern visual-attention models can be improved with such transformations. Although the application is unrelated to the problems behind this project, these transformations look to be an easy way to boost the accuracy and robustness of a visual-attention based model.

Next Podder et al. [45] look into the behaviours of participants while viewing various videos, a Tobii eye tracker is used to detect pupil size, gaze location and blinking data. The videos were also analysed by a group of experts to judge the important contextual regions of objects and a graph based visual saliency model is used to define additional salient points. These metrics and classifications are analysed and plotted to assess the behaviour of the participants. This opens up a new avenue in how we look at an experts interaction to stimuli, not only might their eye movement reflect their knowledge, changes in blinking or pupil size might indicate interest or excitement. Unfortunately no strong evidence was found in the case of this study for pupil size matching any particular category of stimuli. The researchers highlight some limitations with the study, including a lack in number and diversity of participants. In a future study it would be valuable to consider these metrics.

Finally we have two publications from Jinag et al. [46]. Both focus on the development of a novel video saliency prediction method named DeepVS as well as a new database. As part of this method, the team create a series of sub architectures to extract and utilise features to predict inter-frame

saliency. The architectures are designed to combine dynamic and temporal features which has significant value with respect to this project. One of the major contributions of this research is the large-scale eye-tracking database of videos (LEDOV), a database that consists of 32 subjects' fixations on 538 videos. A range of considerations and observations are discussed including centre bias and the increased saliency of moving objects. The dynamic saliency components of this research may be highly useful to the problems behind OWF footage.

2.3 Summary

In conclusion it is evident that eye tracking is a valuable source for defining a custom measure of saliency. One notable method for doing this is by generating a saliency mask using an encoder-decoder system. There are a number of studies that highlight the value of temporal data in combination with eye tracking data to detect dynamic saliency, but given the time limitations of this project, I will be focusing on the area based saliency in an image. With this in mind the TranSalNet becomes an obvious model choice due to its feature extracting capabilities. The question now becomes; What observation behaviours can be distinguished between experts and non experts?

Chapter 3

Methodology

3.1 User Study

Having already procured suitable footage I planned a user study to capture the training and experience of OWF surveying experts. The eye tracking had to be done in person, giving me the chance to talk directly to experts in the industry as well as conducting my experiment. Including industry experts in these processes ensures that they remain a core focus of my research, it is important to incorporate human-centred design philosophies [47] to see to it that RRI is enforced. This section will outline the planning and execution of this study. Unfortunately, given the time available for this project, I was not able to first trial my methods and secondly obtain a large number of participants. It would be more fitting to therefore regard this as a pilot study for future studies.

3.1.1 Expert Engagement Visit

In order to conduct this study I travelled to the offices of the collaborating company Vaarst to see first hand the work that goes into the technology used in underwater surveying. I spent the day engaging with members of the offshore structural maintenance industry and was able to ask them about their experiences during their work. This included ML engineers, data scientist and surveyors, the latter being considered the experts for this study. They informed me on how they conduct their surveying, saying that although they did not waste time watching the entire video in realtime, they often manually scrolled through the video to find each defect, which was still time consuming and fatiguing. For this pilot study, the interactions between experts and the data has been limited to just eye tracking but these comments show that this may not be reflective of an experts day-to-day experience. Although it may come with its complications,

3. Methodology

a future study of similar design could include a second interaction via video scrolling. Such a study could not only yield data on important regions in an image but it could also help to identify what regions of a video are in most need of inspection.

3.1.2 Survey Footage

The data procured was a selection of videos provided by the collaborating company Vaarst. Included was a longer video of a routine wind turbine inspection and a shorter surveying of the HMS J6, a sunken submarine off the coast of Seahouses. Both videos share some similarities in format but show very different styles of capture. Both are high definition videos showing structures surrounded by fish and covered in marine growth, both videos also have non-ideal lighting. The main differences are to do with how much they vary in subject focus, perspective, lighting, stability and other occlusive factors.



Figure 3.1: Three frames from the wind turbine inspection footage, highlighting the variation in content.

The wind turbine inspection is captured from a constantly moving ROV which is in contrast to the stable video capture for the submarine footage. This stability unfortunately leads to the submarine video having not much variety and subsequently being more prone to fatigue.



Figure 3.2: Two frames from the HMS J6 footage.

3. Methodology

For this reason, I chose to prioritise the footage for the wind turbine footage during the main experiment. I chose 15 minutes of the wind turbine inspection that represented a wide variety of visual features and 2 minutes for the submarine footage. This combined with an eye tracking calibration period brings the total time of the experiment to around 18 minutes.

3.1.3 Experimental Design

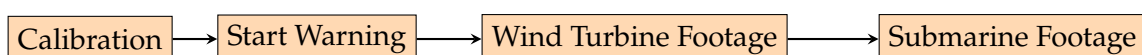
Hardware and Setup:

For the study I used a Tobii Pro X3-120 Eye Tracker connected to a Windows laptop running Tobii Pro Lab. The eye tracker has a small form factor and can easily be placed via a tripod in front of a laptop displaying media. The tracker and laptop stayed in the same position for the whole experiment. There are similar eye trackers in the form of glasses but they have less of a focus on the subjects gaze towards displayed media and more of a focus on the gaze towards their surroundings. This increases the noise of the data and makes it harder to relate the registered gaze to the displayed media, hence the static eye tracker was chosen.



Figure 3.3: The experimental setup

The tracker is able to register the eye positions of participants within a certain range, outside of this range the tracker may register only one or likely no eyes. During calibration the participants can see a representation of this range and were asked to limit their head movement after finding a good spot within the range. The calibration stage consisted of the users fixating on dots in a specific order before the experiment in order to estimate the geometric characteristics of a subject's eyes. This serves as a basis for a fully customised gaze point calculation. After the calibration period a brief text warning that the study is about to begin is shown, then the footage is played for the participants in the order:



Participants:

As mentioned, the source of experts for this study was Vaarst but unfortunately due to the limited time for this project, the number of participants able to be involved was limited to two experts and five non-experts. These non-experts consisted of ML researchers or personnel in HR or finance. Their work is adjacent to the industry and did not lead to them having any particular expertise in performing inspections.

Ethics:

This study posed no ethical concerns, no notable deception or potentially harmful actions, either to the participants or environment were done. All participant details have been removed or anonymised and the figures 3.3 were included only after receiving consent. Although the participants spent up to twenty minutes looking at a screen, possibly leading to eye strain and discomfort, they were informed before that they should stop the experiment should they feel uncomfortable or strained.

Task Description:

Both experts and non experts were informed of a task, this task being to inspect footage of underwater structures for defects. Both groups were also informed of the hypothesis that non-experts may be susceptible to distractions. This was done informally throughout the day when meeting the staff. It may be possible that participant bias exists for both groups either with the experts ignoring, and non-experts purposely focusing on biological life. It may be that the results have been affected by some ambiguity where the task was concerned. The purpose of this pilot study was to trial these methods and uncover the potential problems. In a future study it will be important to standardise the task for both groups and announce it in a formal manner to reduce the likelihood of modified behaviour.

3.2 Final Data

In this section I will discuss the data collected from the user study and the processing it went through before it was applied to the model. One large portion of the processing was to adjust the data to work with the original video data itself and modifications were also done to the videos in order to make them more manageable.

3.2.1 Data Pipeline

Cleaning:

Tobii Pro Lab produced a large dataset covering over 90 metrics including timings, coordinate data for both eyes, eye movement type and even pupil size. Eye fixation coordinates and timings were given in multiple forms with different reference points, such as coordinates in pixels or in centimetres. Much of the data has limited use for the task of mask generation and many columns produced from the software were empty or repetitive and could be easily cleaned. After cleaning, the dataset consisted of a recording timestamp in μs , participant name and expertise level, and the pixel x, y coordinates of the participants in the display area coordinate system (DACS). This coordinate system was chosen because of the unit being pixels, which allowed for an easy transition between gaze coordinates and image pixels for generating an appropriate mask. There is the unfortunate factor that this may not exactly reflect the aspect ratio of the media in other cases, however the wind turbine footage had a near standard aspect ratio, being 1.81.

Asynchronous Recording and Frame Standardisation:

One immediate issue that became clear was that with the data being recorded frequently, and in microseconds, it was unlikely that every reading would be in perfect sync with the frames of the video. This meant it was necessary to group the readings into the frame they should belong to using the recording time. Because there were more readings than frames, there are multiple readings per participant per frame. This raises a question about whether it is safe to take the extra readings as extra data, or could this capture unwanted rapid eye movement rather than fixation. The solution for the next version of this study will be to evaluate a reading based on the eye movement type alongside other metrics in order to determine when exactly a participant is focusing on an object. For this study this additional data was all used as I hypothesise that clusters of points can provide much of the detail on the level of fixation from a participant. Because the video is $25fps$ the equation for what frame a timing belongs to is as follows:

$$F = \lceil t/l \rceil \quad (3.1)$$

Where F is the frame a timing belongs to, t is that timing in microseconds and l is the frame length in microseconds. The total number of frames for the wind turbine video is 22119 but the temporal aspect of the footage means a random sampler is needed to avoid using similar images in training.

Multi-Class Saliency and Mask Generation:

Before the mask can be generated, an issue must be addressed, the data produced includes both coordinates for experts and non experts. It is important to express during training that the expert gaze predictions should be distinct from the non experts, to conform with the output of the TranSalNet these labels were placed into the same mask. This means four cases needed to be represented in total, the case for expert, non expert, neutral (neither expert or non expert) and finally, the cases for both participant types gaze. Having the hypothesis of non experts being more easily distracted for the main task by irrelevant objects, it should make sense that their gaze data should be labelled as being of minimal importance, and as being opposite to experts. What the label for the unlikely combination of expert levels should be is somewhat more debatable, it is both possible that the experts have been distracted by irrelevant objects or the non experts, being aware of the experiment focus, are focusing on the structure. Similarly, the neutral label is not obviously defined but may likely be best suited as being exactly in-between the two expert level labels. Both these variables may require experimentation in the future to fine tune the final design. The final raw labels chosen for this project were 1 for expert, 0 for non expert, 0.4 for neutral and 0.6 for the combined label.

With pixel level coordinate data available, mask generation becomes a simple task of assigning the required label to the appropriate location of a new array or tensor that mirrors the video frame. The immediate issue with this however is this results in an image with neutral labels being significantly greater in number compared to other labels. To remedy this issue, a Gaussian filter was incorporated into the data loader, followed by a boost in contrast.

Video Editing:

The wind turbine inspection data was chosen as the main subject due to its more orthodox aspect ratio and longer length, meaning there is more data available for training. The video was however much too large to be imported as is, it was therefore downsized to a size of 392 by 216 pixels for easier processing. This was done using the open source software FFmpeg. These new dimensions are exactly 20% of the original dimensions and were picked for yielding a video with a manageable size. The transforms within the model mean that other and possibly larger resolutions can be easily tested in future studies.

3.3 Model Design and Training

In this section I will discuss how the data collected was used to train a pre-established architecture, TranSalNet [40]. This architecture was chosen because of both its similarity in terms of objectives and because it is able to identify perceptually relevant characteristics of the input. Some adjustments are made to best fit the data available to the models design and the loss function was altered. These adjustments and the difference in context and training mean that the model used is not directly comparable to the original, and an investigation on the ramifications of these changes is beyond what is possible in this project due to time restrictions.

3.3.1 Dataloaders

The video footage is imported using the torchvision's `read_video`, a video reading API that is able to prepare the video as tensors for use in PyTorch. The masks, to also be referred to as labels, are loaded in as tensors, having been generated and pickled from the data pipeline. Separate torchvision transforms are defined for both the images and labels, both resize the input to 288 by 384 to match the data from the original paper. The previously mentioned Gaussian blur is applied to mask here during application of the transforms. The final parameters for the blur are a kernel size of 81 by 81 and a sigma randomly selected from the range (5, 7). These parameters were selected as they gave the desirable level of blurring, and may again be a factor to experiment with in the future.

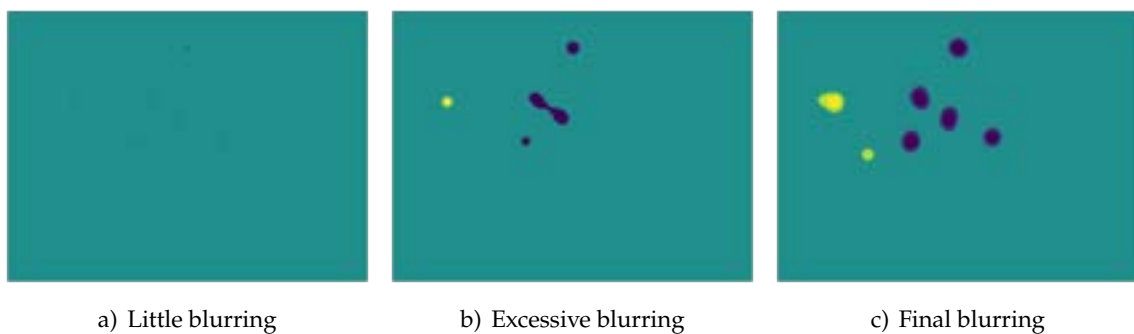


Figure 3.4: A range of blurring settings, yellow and purple signify gaze location for experts and non experts respectively on frame 10000.

The variation to the sigma forces the model to learn to deal with less desirable or less obvious labelling, potentially making for a more robust model. A boost to the contrast is also performed to prevent the fading of labels due to blurring. Finally both images and masks are forced to

3. Methodology

convert to tensors. From here a typical PyTorch data loader is created to fetch, and load to GPU, the required pair of data elements.

3.3.2 Model Design

As mentioned this project will be utilising the TranSalNet, a complex model comprised of both an encoding segment, itself made of a CNN backbone and a set of transformer encoders, and a decoding segment. Three feature maps are extracted from the CNN encoder, each with different spatial sizes. These feature maps are fed into an assigned transformer encoder to enhance the long-range and contextual information. In order to maintain positional information a positional embedding is created before the transformer stage. The CNN decoder fuses these long-range context-enhanced feature maps to help to implement pixel-level classification for saliency map prediction.

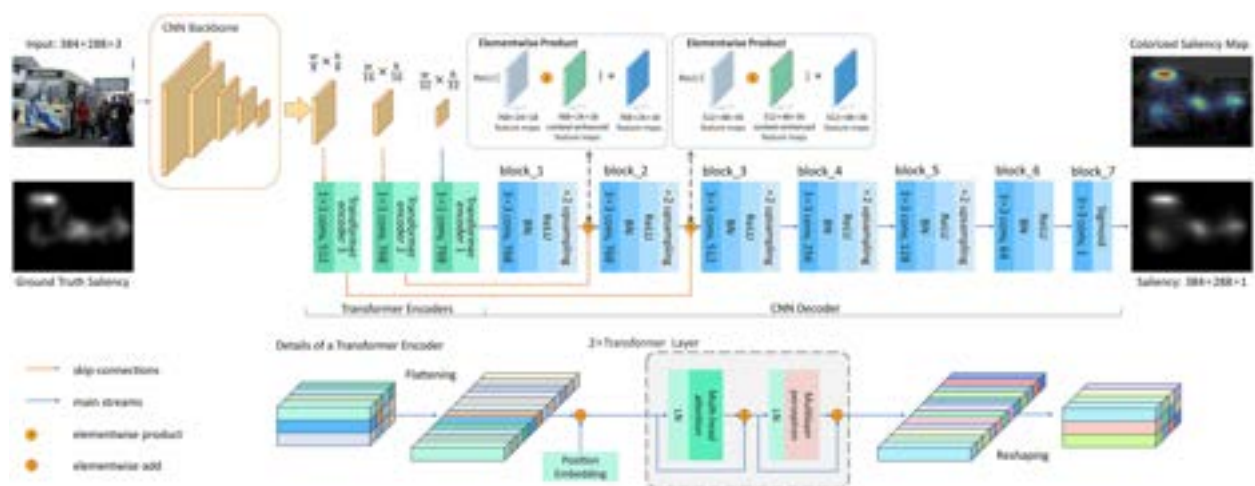


Figure 3.5: The design of the TranSalNet [40].

The core design of the model is unchanged for this project as the dataloader's transforms allow for varying input and training masks sizes to seamlessly work. What is open to change is the CNN backbone, the source code allows for easy integration of a number of variations of both ResNet and DenseNet. In the original paper, Lou et al. found DenseNet-161 to be more effective than ResNet-50 [40] but due to the increased model size and training times of the DenseNet CNN, ResNet-50 was chosen. The ResNet component was given pretrained weights, trained on Imagenet. It should be noted again that this implementation is done utilising data from a pilot study and thus serves to be more an experiment than a final model.

3.3.3 Custom Loss Function and Training

The loss function for this model has a notable distinction from the original paper; In order to comply with the data available, the function has become one of ground truth saliency map and the predicted map, and not one of any fixation map. Excluding this, the function is the same [40]. It is a linear combination of the Kullback-Leiber divergence (KLD), linear Correlation Coefficient (CC), and Similarity (SIM). Let \mathbf{y}^s and $\hat{\mathbf{y}}$ be the ground truth saliency map and predicted saliency map respectively, and i indicates the i th pixel of \mathbf{y}^s and $\hat{\mathbf{y}}$, our loss function is defined as:

$$L(\mathbf{y}^s, \hat{\mathbf{y}}) = \lambda_1 L_{\text{KLD}}(\mathbf{y}^s, \hat{\mathbf{y}}) + \lambda_2 L_{\text{CC}}(\mathbf{y}^s, \hat{\mathbf{y}}) + \lambda_3 L_{\text{SIM}}(\mathbf{y}^s, \hat{\mathbf{y}}), \quad (3.2)$$

where λ_1 , λ_2 and λ_3 are the weights of each metric, and

$$L_{\text{KLD}}(\mathbf{y}^s, \hat{\mathbf{y}}) = \sum_i y_i^s \log\left(\epsilon + \frac{y_i^s}{\epsilon + \hat{y}_i}\right), \quad (3.3)$$

where ϵ is a regularisation constant and set to 2.2204×10^{-16} ;

$$L_{\text{CC}}(\mathbf{y}^s, \hat{\mathbf{y}}) = \frac{\text{cov}(\mathbf{y}^s, \hat{\mathbf{y}})}{\sigma(\mathbf{y}^s)\sigma(\hat{\mathbf{y}})}, \quad (3.4)$$

where $\text{cov}(\cdot)$ is the covariance and $\sigma(\cdot)$ is standard deviation;

$$L_{\text{SIM}}(\mathbf{y}^s, \hat{\mathbf{y}}) = \sum_i \min(y_i^s, \hat{y}_i). \quad (3.5)$$

The weights and constants for the loss function are inherited from the original paper. Training was done using a windows machine utilising a Nvidia 3090 GPU running Python 3.10. The main dependencies include PyTorch 1.12, numpy 1.22.3, OpenCV 4.6 and CUDA 11.6. The training loop was similar to the loss functions in that the only difference from the original paper is the modifications done to remove the dependency on fixation plotting. This included removing a sub loss function, the Normalized Scanpath Saliency (NSS). The data was split into training and validation datasets, the split was done by Sci-kits learn's `train_test_split` function, with a 66.7% and 33.3% split for training and validation data respectively.

Chapter 4

Results

4.1 Training and Validation Loss

The model was trained with a limit of 20 epochs but was stopped early after 16 epochs. This early stop was a result of an algorithm in the original paper designed to prevent overfitting. Training took five hours and was re-run several times, all to similar results.

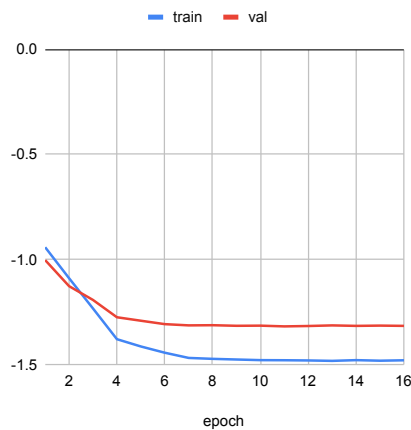


Figure 4.1: The training and validation loss over the full 16 epochs.

There is a noticeable characteristic of this loss progression, it starts in the negatives and continues to drop, plateauing at around -1.5 for the training and -1.3 for the validation. This unorthodox progression (starting so low, below zero and continuing) is likely an artefact of the modifications made to the loss function of the original paper. The removal of the NSS sub loss function means that there is one less sub function that aims to be maximised, hence the reduced loss.

4.2 Enhancing

The maps produced largely resemble the ground truth with regards to location of regions but not in terms of clarity and definition, the generated maps look washed out/faded. To solve this issue I applied a boost in contrast similar to the one done in the dataloader. The final contrast factor used by the `adjust_contrast` torchvision transform was seven but like the dataloader, this value was chosen purely for aesthetic reasons. While for this project this does not greatly affect the result, it only boosts clarity, if these maps are used in future algorithms, all variables like these should be tested.

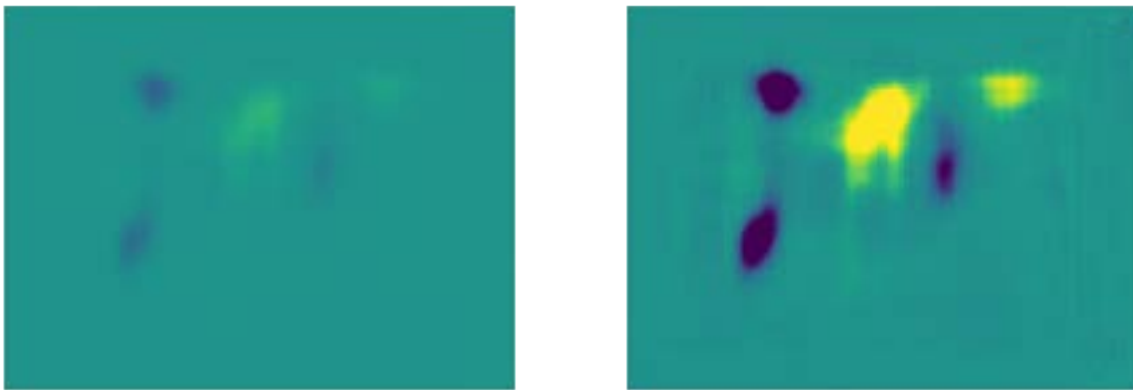


Figure 4.2: Before and after increasing contrast for generated masks, yellow represents predicted expert gaze, purple for non experts.

4.3 Final Saliency Maps

After enhancing the images, well defined regions are visible, in the majority of cases these regions match the location and expert polarity of the ground truth. On top of this, in many cases the number of regions reflects the number and spread of participants gaze values. In this regard the model does very well but in terms of assigning relevancy to a specific object in the image, the effectiveness of this initial model is questionable.

4. Results

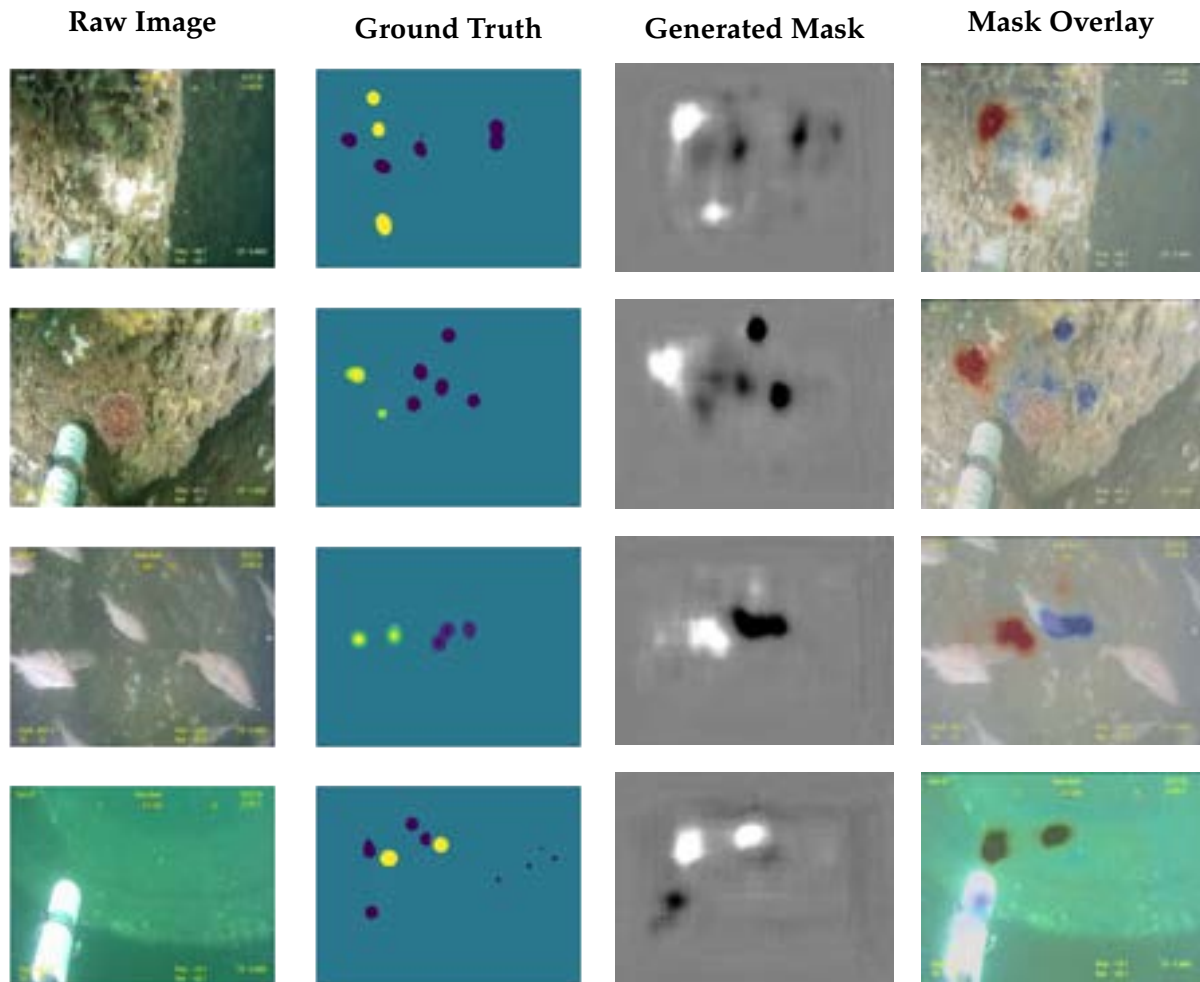


Figure 4.3: A selection of frames and their corresponding masks, for the overlay red represents expert gaze and blue non experts.

As visible from these examples, there is evidence of the model predicting non expert gaze on objects that are thought to be irrelevant, such as fish and the ROV probing instrument. There is also examples of non expert gaze prediction in spaces away from the structures, such as regions of the murky background. Experts are typically looking for whatever they can to assess the integrity of the structure, often this will be areas that are exposed due to reduced marine growth. In the second row we can not only see a captured example of this behaviour but also evidence of this behaviour being replicated by the model. The true validity of these patterns can only be determined by conducting a second user study, such as a questionnaire, on the experts.

Chapter 5

Conclusion and Discussion

This methodology shows great promise for generating saliency maps based on the eye tracking data, distinct areas can be identified and in many cases the placement aligns with the theory. There are however a number of variables that need to be fine tuned before this model can be effectively used in future research. One primary area to improve upon is the user study itself, the length of the videos opened up the possibility for eye fatigue, shorter videos with breaks in-between will likely yield more reliable data as well as creating a more comfortable participant experience. More participants and a better way at recognising stacked gaze data may improve the irrelevant object classification capabilities of the model.

5.1 Future Work

Throughout this work there have been many instance where my methodology could improve, not only the ones mentioned above but also in model implementation. I plan to continue this research by more attentively designing a second larger user study to capture data, and create a more robust pipeline to utilise this data. The next step for this work, excluding the more complete study, is to implement these masks into an image embedding model. Should an something such as an autoencoder be trained to embed images whilst considering expert guided saliency, further dimensionality reduction methods could be used to cluster images based on similarity. By involving saliency masks I should be able to reduce the impact of distracting features on image comparisons. Succeeding in this, a tool can be created to aid these inspectors in conducting more efficient and effective inspections.

Bibliography

- [1] IPCC, *Climate Change 2022: Mitigation of Climate Change. Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, P. Shukla, J. Skea, R. Slade, A. A. Khourdajie, R. van Diemen, D. McCollum, M. Pathak, S. Some, P. Vyas, R. Fradera, M. Belkacemi, A. Hasija, G. Lisboa, S. Luz, and J. Malley, Eds. Cambridge, UK and New York, NY, USA: Cambridge University Press, 2022.
- [2] P. Higgins and A. Foley, "The evolution of offshore wind power in the united kingdom," *Renewable and sustainable energy reviews*, vol. 37, pp. 599–612, 2014.
- [3] M. Bilgili and H. Alphan, "Global growth in offshore wind turbine technology," *Clean Technologies and Environmental Policy*, pp. 1–13, 2022.
- [4] "Offshore wind outlook 2019," <https://www.iea.org/reports/offshore-wind-outlook-2019>, accessed: 2022-8-23.
- [5] A. Mourão, J. A. Correia, B. V. Ávila, C. C. de Oliveira, T. Ferradosa, H. Carvalho, J. M. Castro, and A. M. De Jesus, "A fatigue damage evaluation using local damage parameters for an offshore structure," in *Proceedings of the Institution of Civil Engineers-Maritime Engineering*, vol. 173, no. 2. Thomas Telford Ltd, 2020, pp. 43–57.
- [6] A. Cam, M. Chui, and B. Hall, "Global ai survey: Ai proves its worth, but few scale impact," 2019.
- [7] A. Jalal, A. Salman, A. Mian, M. Shortis, and F. Shafait, "Fish detection and species classification in underwater environments using deep learning with temporal information," *Ecological Informatics*, vol. 57, p. 101088, 2020.

- [8] S. Raveendran, M. D. Patil, and G. K. Birajdar, "Underwater image enhancement: a comprehensive review, recent trends, challenges and applications," *Artificial Intelligence Review*, vol. 54, no. 7, pp. 5413–5467, 2021.
- [9] R. Schettini and S. Corchs, "Underwater image processing: state of the art of restoration and image enhancement methods," *EURASIP journal on advances in signal processing*, vol. 2010, pp. 1–14, 2010.
- [10] W. Zhang, L. Dong, X. Pan, P. Zou, L. Qin, and W. Xu, "A survey of restoration and enhancement for underwater images," *IEEE Access*, vol. 7, pp. 182 259–182 279, 2019.
- [11] S. S. Shadrin, O. O. Varlamov, and A. M. Ivanov, "Experimental autonomous road vehicle with logical artificial intelligence," *Journal of advanced transportation*, vol. 2017, 2017.
- [12] P. Shruthi and R. Resmi, "Path planning for autonomous car," in *2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)*, vol. 1. IEEE, 2019, pp. 1387–1390.
- [13] A. Ess, B. Leibe, K. Schindler, and L. Van Gool, "Moving obstacle detection in highly dynamic scenes," in *2009 IEEE International Conference on Robotics and Automation*. IEEE, 2009, pp. 56–63.
- [14] "Framework for responsible innovation," UKRI. [Online]. Available: <https://www.ukri.org/about-us/epsrc/our-policies-and-standards/framework-for-responsible-innovation/>
- [15] United Nations General Assembly. (2015) The 17 goals - sustainable development goals. [Online]. Available: <https://sdgs.un.org/goals>
- [16] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015.
- [17] T. M. Mitchell and T. M. Mitchell, *Machine learning*. McGraw-hill New York, 1997, vol. 1, no. 9.
- [18] D. E. Liddle, "The wider impact of moore's law," *IEEE Solid-State Circuits Society Newsletter*, vol. 11, no. 3, pp. 28–30, 2006.
- [19] P. Bhattarai, S. Krupiński, V. Unnithan, F. Maurelli, N. Secciani, M. Franchi, L. Zacchini, and A. Ridolfi, "A deep learning approach for underwater bubble detection," in *OCEANS 2021: San Diego – Porto*, pp. 1–5, ISSN: 0197-7385.

- [20] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.
- [21] D. A. Konovalov, A. Saleh, M. Bradley, M. Sankupellay, S. Marini, and M. Sheaves, "Underwater fish detection with weak multi-domain supervision," in *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, ISSN: 2161-4407.
- [22] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.
- [23] M. Moniruzzaman, S. M. S. Islam, P. Lavery, and M. Bennamoun, "Faster r-CNN based deep learning for seagrass detection from underwater digital images," in *2019 Digital Image Computing: Techniques and Applications (DICTA)*, pp. 1–7.
- [24] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.
- [25] S. A. Pamungkas, I. Jaya, and M. Iqbal, "Segmentation of enhalus acoroides seagrass from underwater images using the mask r-CNN method," vol. 944, no. 1, p. 012010, publisher: IOP Publishing. [Online]. Available: <https://doi.org/10.1088/1755-1315/944/1/012010>
- [26] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.
- [27] S. Raine, R. Marchant, P. Moghadam, F. Maire, B. Kettle, and B. Kusy, "Multi-species seagrass detection and classification from underwater images," in *2020 Digital Image Computing: Techniques and Applications (DICTA)*, pp. 1–8.
- [28] C. Conrady, "Automated detection and classification of red roman in unconstrained underwater environments using mask r-CNN," accepted: 2022-02-18T04:49:35Z Publisher: Faculty of Science. [Online]. Available: <https://open.uct.ac.za/handle/11427/35704>
- [29] E. Ottaviani, M. Francescangeli, N. Gjerci, J. d. Río Fernandez, J. Aguzzi, and S. Marini, "Assessing the image concept drift at the OBSEA coastal underwater cabled observatory," vol. 9, pp. 1–13, accepted: 2022-04-12T10:46:29Z Publisher: Frontiers Media. [Online]. Available: <https://upcommons.upc.edu/handle/2117/365750>

- [30] P. Tarling, M. Cantor, A. Clapés, and S. Escalera, “Deep learning with self-supervision and uncertainty regularization to count fish in underwater images,” vol. 17, no. 5, p. e0267759, publisher: Public Library of Science. [Online]. Available: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0267759>
- [31] A. Saleh, I. H. Laradji, D. A. Konovalov, M. Bradley, D. Vazquez, and M. Sheaves, “A realistic fish-habitat dataset to evaluate algorithms for underwater visual analysis,” 2020.
- [32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.
- [33] C. Raffel, M.-T. Luong, P. J. Liu, R. J. Weiss, and D. Eck, “Online and linear-time attention by enforcing monotonic alignments,” in *International Conference on Machine Learning*. PMLR, 2017, pp. 2837–2846.
- [34] W. Wang, J. Shen, F. Guo, M.-M. Cheng, and A. Borji, “Revisiting video saliency: A large-scale benchmark and a new model,” in *Proceedings of the IEEE Conference on computer vision and pattern recognition*, 2018, pp. 4894–4903.
- [35] W. Wang, J. Shen, J. Xie, M.-M. Cheng, H. Ling, and A. Borji, “Revisiting video saliency prediction in the deep learning era,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 1, pp. 220–237, 2019.
- [36] Y. Dahou, M. Tliba, K. McGuinness, and N. O’Connor, “Atsal: An attention based architecture for saliency prediction in 360° videos,” in *International Conference on Pattern Recognition*. Springer, 2021, pp. 305–320.
- [37] N. Liu, N. Zhang, K. Wan, L. Shao, and J. Han, “Visual saliency transformer,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 4722–4732.
- [38] F. Yan, Z. Wang, S. Qi, and R. Xiao, “A saliency prediction model based on re-parameterization and channel attention mechanism,” *Electronics*, vol. 11, no. 8, p. 1180, 2022.
- [39] Y. Wang and M. Zhu, “Saliency prediction based on lightweight attention mechanism,” in *Journal of Physics: Conference Series*, vol. 1486, no. 7. IOP Publishing, 2020, p. 072066.
- [40] J. Lou, H. Lin, D. Marshall, D. Saupe, and H. Liu, “Transalnet: Towards perceptually relevant visual saliency prediction,” *Neurocomputing*, vol. 494, pp. 455–467, 2022.

- [41] A. Coutrot and N. Guyader, "Learning a time-dependent master saliency map from eye-tracking data in videos," *arXiv preprint arXiv:1702.00714*, 2017.
- [42] C. Chamaret, J.-C. Chevet, and O. Le Meur, "Spatio-temporal combination of saliency maps and eye-tracking assessment of different strategies," in *2010 Ieee International Conference on Image Processing*. IEEE, 2010, pp. 1077–1080.
- [43] A. Tsiami, P. Koutras, and P. Maragos, "Stavis: Spatio-temporal audiovisual saliency network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4766–4776.
- [44] V. Lyudvichenko, M. Erofeev, Y. Gitman, and D. Vatolin, "A semiautomatic saliency model and its application to video compression," in *2017 13th IEEE International Conference on Intelligent Computer Communication and Processing (ICCP)*. IEEE, 2017, pp. 403–410.
- [45] P. K. Podder, M. Paul, T. Debnath, and M. Murshed, "An analysis of human engagement behaviour using descriptors from human feedback, eye tracking, and saliency modelling," in *2015 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*. IEEE, 2015, pp. 1–8.
- [46] L. Jiang, M. Xu, T. Liu, M. Qiao, and Z. Wang, "Deepvs: A deep learning based video saliency prediction approach," in *Proceedings of the european conference on computer vision (eccv)*, 2018, pp. 602–617.
- [47] B. Shneiderman, "Human-centered artificial intelligence: Three fresh ideas," *AIS Transactions on Human-Computer Interaction*, vol. 12, no. 3, pp. 109–124, 2020.

Appendix A

Implementation of Important Algorithms and Functions

```
1 def standard_time_series_ppt(df, participant):
2     pdf = get_ppt_df(df, participant)
3     pdf = pdf.fillna(method='ffill',axis=0)
4     pdf = pdf.fillna(method='bfill',axis=0)
5     baseline = pdf['Recording timestamp [ s ]'].tolist()[0]
6     pdf['Recording timestamp std [ s ]'] = pdf['Recording timestamp [ s ]'] - baseline
7     return(pdf)
8
9 def standard_time_series_stitch(df):
10    df_list = []
11    ppts = df['Participant name'].unique()
12    for ppt in ppts:
13        df_to_stitch = standard_time_series_ppt(df, ppt)
14        df_list.append(df_to_stitch)
15    stitched_gaze_df = pd.concat(df_list)
16    return(stitched_gaze_df)
```

Listing A.1: An algorithm to standardise the timings for each participant.

A. Implementation of Important Algorithms and Functions

```
1 frame_length_muS = 1/vid_fps * 10**6
2
3 def round_down(x):
4     if(x == 0):
5         return int(1)
6     else:
7         return(int(math.ceil(x / float(frame_length_muS))))
8
9 df['frame'] = df['Recording timestamp std [ s ]'].apply(round_down).astype(int)
```

Listing A.2: An algorithm to identify and label the correct frame for each reading.

```
1 def get_coords(level, frame):
2     try:
3         temp = level_frame_group.get_group((level,frame))
4         return list(zip(temp['Gaze point X [DACs px]'],temp['Gaze point Y [DACs px]']))
5     except:
6         return [(0,0)]
7
8 def return_label(level, x, y, current):
9     if level == 'expert':
10        return 1
11    elif(level == 'non expert' and current == 1 ):
12        return interest_value
13    else:
14        return 0
15
16
17 def label_both_frame_grid():
18     frames_labels = []
19     for frame in range(1, frame_count + 1):
20         frame_grid = torch.full([height, width], neutral, dtype = torch.float16)
21         for level in ['expert','non expert']:
22             coords = get_coords(level, frame)
23             for xy in coords:
24                 x = xy[0]
25                 y = xy[1]
26                 current_value = frame_grid[y][x]
27                 frame_grid[y][x] = return_label(level, x, y, current_value)
28         frames_labels.append(frame_grid)
29     frames_labels = torch.stack(frames_labels, axis=0)
30     return frames_labels
```

Listing A.3: An algorithm to create a mask based on coordinate values of participants gaze.