Underwater Archaeological Object Detection using Photogrammetric Fusion

Ethan Zammit

Supervisor: Dr. Dylan Seychell

Co-supervisor: Prof. Ing. Carl James Debono

June 2024

Submitted in partial fulfilment of the requirements for the degree of M.Sc. ICT in Artificial Intelligence (Hons.).

L-Università ta' Malta
Faculty of Information & Communication Technology

Acknowledgements

First and foremost, I would like to express my gratitude to my supervisor, Dr. Dylan Seychell, and my co-supervisor, Prof. Ing. Carl James Debono. Their guidance, expertise, and constant reassurance have played a pivotal role in the success of this study. I would like to extend this appreciation to Prof. Timmy Gambin, Mr. John Wood, and the rest of the Department of Classics and Archaeology for their knowledge sharing, data provision, and guidance in archaeological and photogrammetric concepts. This also applies to individuals involved in the annotation sessions, especially Mr. Pablo Morando, for their time and effort in making this study possible.
Moreover, I would like to thank my family, my girlfriend, closest friends, and colleagues for their unyielding trust, support, and encouragement, which have motivated me to strive to be the best version of myself.
Finally, I would also like to thank MDIA for their trust and for funding my academic journey through the Pathfinder scheme.

Abstract

The sea holds many secrets, yet the surveillance of underwater scenery still demands complex operations and sharp observation. Underwater conditions pose several challenges, including blurriness, degradation, and light distortion, all proving to be detrimental to detection performance. These challenging conditions have attracted researchers' attention, leading to consistent improvements in detection accuracy. Building on these advancements, our work focuses on maritime archaeology sites, specifically detecting artefacts from the Tower Wreck in the Xlendi Underwater Archaeological Park, some even dating back to 300 BCE. This was done through the compilation of a multi-class dataset for underwater amphora detection, labelled by field experts. The final dataset covers $625m^2$ and consists of 864 images. Based on this dataset, various object detection architectures, including single-shot detectors (YOLO), transformer-based models (DETR), and two-stage detectors (Faster R-CNN), were evaluated. Transformer and two-stage models were found to underperform, possibly linked to their increased data requirements. On the other hand, YOLOv7-tiny achieved the best overall performance with a mAP_{50} of 86.14%. Further analysis was performed by using depth maps generated using the photogrammetric model, and saliency estimation techniques. Various fusion methods were then compared, which mixed these maps into the original imagery. This modification provided marginal improvements over base models, with YOLOv7-tiny mAP_{50} increasing to 86.34%. Finally, photogrammetric techniques were used to project 2D detections to 3D coordinates on the orthomodel. From these coordinates, several methods of processing were compared to best display these in an aggregated orthomosaic visualisation. These findings highlight the potential for bridging the gap between maritime archaeology and computer vision, paving the way for more efficient and accurate underwater archaeological surveys.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Problem Definition

It is estimated there are over 3 million shipwrecks at the bottom of the world's oceans, waiting to be discovered (Grzadziel, 2022), often revealing fundamental knowledge about previous civilisations, such as shipbuilding techniques, trade routes, sea battles, and transportation of material (Kızıldağ, 2022). Each of these discoveries could completely change our understanding of previous lives, yet with less than 1% actually being explored (Bennett, 2016) it highlights the necessity for better tools for the exploration, monitoring and preservation of maritime artefacts.

Such tools for exploration may be significantly enhanced using artificial intelligence (Paraskevas *et al.*, 2023). Neural Image enhancement techniques can improve underwater images, making artefacts easier to identify (Ertan *et al.*, 2024). Neural 3D reconstruction is also an emerging field, which can create detailed models of sites from multiple images (Mildenhall *et al.*, 2020). Additionally, AI-based depth estimation techniques can provide accurate measurements of underwater sites. Moreover, advanced object detection algorithms can automatically identify and classify artefacts from images and video footage, reducing manual effort and increasing accuracy (Paraskevas *et al.*, 2023; Yang *et al.*, 2023).

Despite the massive leaps in traditional object detection, Underwater Object Detection (UOD) remains one of the most challenging tasks in computer vision (Xu *et al.*, 2023). These complications are multifaceted, one of the major issues being the lack of sufficient data. Obtaining underwater images is more difficult than obtaining images in the atmosphere, making it difficult for data-driven deep-learning models to achieve satisfactory results (Xu *et al.*, 2023). Moreover, the quality of underwater images is often poor, due to issues such as non-uniform lighting, blur, and colour distortion, which often leads to loss of colour and texture information (Espinosa *et al.*, 2023; Fu *et al.*, 2023). Furthermore, due to the varying landscape of the seabed, some artefacts may be partially buried or in dense rocky areas, making it difficult even for trained eyes to identify them.

Considering these challenges, many researchers and individuals have opted to tackle

this issue, where most modern solutions revolve around the adoption of mainstream object detection models to new datasets (Espinosa *et al.*, 2023). Our research is targeted towards a general improvement in the value of vision techniques for archaeological object detection, proposing a solution across several fronts. The lack of available data is addressed through the proposal of a new archaeological object detection dataset surrounding the Tower Wreck, in Xlendi, Gozo. This problem is further addressed through the deep integration with photogrammetric techniques, which enrich the images with rich positional information. The difficult nature of the imagery was additionally targeted through the introduction of saliency and depth estimation techniques as auxiliary features to object detection methods, aiming to saturate the usage of the available data. Furthermore, the object detection and archaeological domains are closely bridged through a bidirectional link between 2D-pixel positions and 3D geographic coordinates, which further promotes the explainability and real-world impact of our results, including estimating geographic coordinates for each artefact detected and an aggregated visualisation tool for the entire work area of our data.

1.2 Motivation

The exploration, monitoring and preservation of cultural artefacts has traditionally been done manually, often requiring hours of research, preparation and orchestrated diving operations (Fayaz *et al.*, 2022). Apart from the dangers posed to the divers, there is also the requirement for skilled individuals to interpret such observations (Beijbom *et al.*, 2012).

In recent years, advancements in the computer vision field have paved the way for autonomous underwater vehicles (AUVs) and remotely operated vehicles (ROVs) to facilitate underwater data collection tasks (Beijbom *et al.*, 2012; Fu *et al.*, 2023; Mallet and Pelletier, 2014). Nevertheless, the increase in accessibility for such data further necessitates the betterment of tools for automated analysis and interpretation, which still lack the accuracy and features required (Fayaz *et al.*, 2022).

Despite the rampant research attention to UOD, most studies focus on marine organism detection, rather than cultural objects, which as discussed, poses its unique set of difficulties. The handful of studies that particularly target cultural object detection are often limited to the abstract scenarios of 2D detection, which puts a limit on the interpretability and tangible benefits for archaeologists and personnel who need them the most. This discrepancy between the advancements in AI and the growing requirements in archaeology and other dependant areas calls for an infrastructure for deeper integration of domains. Such integration would allow achievements in object detection techniques

to be more easily transferred to the direct improvement of tools for archaeological interpretation and enhancement to capabilities of readily available tools.

1.3 Aim and Objectives

Given the prevailing importance of applying advancements in object detection techniques to applied domains such as archaeology, we found this to be an interesting problem to tackle and set out to help push the state-of-the-art in developing a practical solution. The main aim of this research is to study the applicability and improvement of object detection techniques for underwater archaeological object detection. Summarised by the research question: *How effectively can object detection techniques detect and classify archaeological artefacts in underwater environments?* To achieve our main goal and answer this question, a number of objectives were determined:

O1. Compile a Maritime Archaeology dataset of artefacts in various landscapes, properly labelled and classified by professionals.

O2. Evaluate and compare the performance of various state-of-the-art object detection techniques on underwater archaeological objects.

O3. Explore the performance impact of supplementing object detection with auxiliary visual saliency and photogrammetric depth maps.

O4. Increase the interpretability of detections using automated localisation techniques for geotagging artefacts and projecting labels to an aggregated orthomosaic of the dataset.

1.4 Proposed Solution and Contributions

This study builds on the knowledge gained from similar research, proposing a set of solutions to the outlined objectives. The primary contributions of our work are as follows:

- **The compilation of a new underwater archaeological object detection dataset.** Through a collaboration with the Department of Classics and Archaeology, which provided a set of images and a photogrammetric model, a new object detection dataset was compiled and annotated. Starting from the type of imagery, the area selected encapsulates varying seabed landscapes, rocky outcrops and varying types of artefacts, ensuring that the real-world distribution of artefacts is properly represented. Annotation sessions were then held, where field experts annotated the 864

images with bounding box labels. Another pass was then organised where the artefacts were individually identified into classes according to their level of preservation. This contribution is especially valuable given the rarity of high-quality underwater imagery, combined with the immense depth of the archaeological site, and further considering the joint expertise required for the proper annotation of data.

- **The fine-tuning of several state-of-the-art object detection techniques.**
  This study focused on exhaustively comparing and analysing the performance of several SOTA object detection methods, which were fine-tuned to underwater archaeology, and carefully evaluated. Namely, the architectures Faster-RCNN (Ren *et al.*, 2015), YOLOv5 (Jocher, 2020), YOLOv7 (Wang *et al.*, 2023), YOLOv8 (Jocher *et al.*, 2023), YOLOv9 (Wang *et al.*, 2024), DETR (Carion *et al.*, 2020), Deformable DETR (Zhu *et al.*, 2021) and RT-DETR (Zhao *et al.*, 2023) were compared. These methods were found among the most prevailing in UOD and computer vision literature. Furthermore, the automated classification of artefacts based on their level of preservation, to our knowledge, was yet unexplored through computer vision techniques, laying important groundwork for further research.

- **The analysis of saliency and depth as auxiliary Information on our area.**
  In light of maximising the fruitfulness of the available data, the analysis of enrichment through visual saliency and depth estimation techniques was studied. Depth maps were estimated by leveraging the Digital Elevation Model (DEM), estimated from photogrammetry, which was a unique approach to fusion with 3D models. On the other hand, the visual saliency estimation techniques chosen were found to be commonly used within the domain, which included Itti (Itti *et al.*, 1998), Deepgaze (Patacchiola and Cangelosi, 2017) and InSPyReNet (Kim *et al.*, 2022). These saliency estimation techniques were first individually evaluated in salient object detection, gaining insight into their potential contribution to the main object detection techniques. Then, several merging techniques were studied to combine this information with the imagery. Despite salient object detection not being a new field in vision, the use of saliency as auxiliary information was also not commonly studied, exposing a potential path of improving detection in difficult and noisy scenarios.

- **Increasing explainability of detections through photogrammetry.**
  Location information provides essential context to many practices, especially in applications requiring careful interpretation of scenes (Wilson *et al.*, 2023). Current object detection techniques lack this information, as being limited to pixel positions, the absolute location of detected items may not be as interpretable. For this reason,

the deep integration between photogrammetric and object detection techniques is proposed, boosting the real-world value obtained from mainstream object detection technique advancements, to tangible benefits in archaeological studies and other domains. This ideology is further highlighted through similar works by Gené-Mola *et al.* (2020) and Al-anni and Drap (2024), which were able to efficiently leverage photogrammetric fusion techniques to gain relative positional information from individual images. The first stage involved automated geotagging of each detection produced by the vision techniques, obtaining absolute 3D world coordinates, and allowing for more digestible results. Furthermore, the geographic bounding boxes were aggregated onto a geographically aligned orthomosaic, presenting a positional and contextual visualisation tool.

Through these advancements, the study pushes the envelope on the benefits harnessed from object detection technique research, whilst simultaneously bridging these advancements to the archaeological domain and others alike. Apart from the research and experimentation itself, a contribution is additionally posed towards an end-to-end detection system, which from data already available, such a system can be used by researchers to obtain accurately annotated and classified labels with geographic reference and aggregated visualisation tools.

1.5 Document Structure

The rest of this document is partitioned into five primary chapters. In Chapter 2, a background of grounding concepts and techniques is presented, building the necessary context and concepts for our research domain. This chapter starts with an overview of what maritime archaeology entails, the typical manual techniques followed, and the importance of the archaeological field on society. Afterwards, the working area of our data is put into perspective, presenting the type of artefacts, the scale, and the difficulty of the task. This further includes the main concepts behind photogrammetric techniques, and the mathematical relation between images, estimated camera positions, detected objects and the world positions.

Chapter 3 delves into the surrounding literature, which develops the required concepts to better understand what the state-of-the-art techniques involve, starting from dissecting the modern object detector into its main building blocks. Then, a comparison between two primary backbone paradigms is presented, namely convolution-based techniques and transformer-based techniques, which have their own applications. Additionally, the ideation and current work on visual saliency estimation and depth are presented,

along with similar work leveraging these as additional features. Finally, localisation techniques for object detection were discussed, especially methods using photogrammetric fusion.

Chapter 4 presents the methodology of this research, first focusing on the general approach taken for our solution, and then honing into each individual component, which builds up to achieving our aim and objectives. This starts by explaining the rigorous process of dataset annotation and compilation, followed by the experimentation framework adopted to shortlist promising object detector techniques. Furthermore, the saliency and depth map augmentation techniques are presented, and followed by the handcrafted techniques required for proper projection and suppression of detections to the orthomodel and orthomosaic. Finally, the evaluation techniques are presented and explained, displaying the metrics used to evaluate our experiments and primary objectives.

Chapter 5 deals with presenting the results obtained from our experiments, where each objective is individually addressed, and the relevant results are tabulated, analysed and discussed, aiming to interpret the results obtained. A staged approach was taken, where after each stage the most promising techniques were kept for further evaluation, converging to an optimal configuration. This is finalised by detailed discussion and analysis of the top-model results, aiming to better understand real-world performance.

Chapter 6 concludes this study, revisiting each objective and summarising what key takeaways were obtained from interpreting our experiment results. The potential for future work is then addressed, displaying several paths one could develop from our work.

2 Background

Photogrammetry, computer vision, and maritime archaeology are typically distinct concepts encompassing a plethora of topics, studies and research within their own domains. This study converges these disciplines, making it imperative to grasp the nuances inherent in each. While the focus lies on computer vision, this chapter establishes some of the key knowledge to understand the rest of our work, starting with the elementary concepts of maritime archaeology and putting the dataset into perspective. Then the primary processes behind photogrammetry are discussed, both in terms of philosophy and mathematical theory. Furthermore, an overview of image enhancement techniques is then given as this is a prominent research topic in the underwater object detection domain.

2.1 Maritime Archaeology

This section gives an overview of the particular application chosen for this study. It starts by scraping the surface of the archaeological concepts required, and the target objects. This is then led to discussing the nature of the data we are working with, its historical significance and a peek into its challenging nature. Finally, the relevance of photogrammetric documentation of archaeological sites is discussed, in relation to the type of data made available to computer vision methods.

2.1.1 What is an Amphora ?

Amphorae are clay jars specifically designed for maritime transport (Gambin, 2012), an example of which may be found in Figure 2.1. Their archaeological significance however lies in their uniqueness to periods and the almost fingerprintable characteristics which are attributed to their origins. In the case of the Xlendi Tower wreck, three major amphorae were identified, being the Maltese Punic Amphora, and two variants of the Flat-Bottomed Amphora.

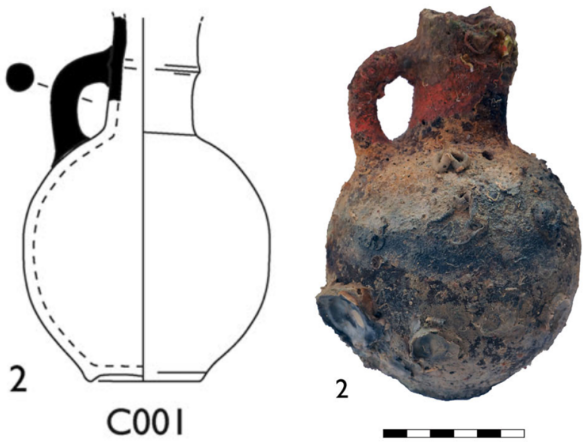


Figure 2.1: A Phoenician round-mouthed jug depiction adapted from (Gambin *et al.*, 2021).

The Maltese punic amphora is linked to the 5^{th} century BCE, which was locally produced on the Maltese islands, following a typical Phoenician ovoid shape (Heritage Malta, 2023). The flat-bottomed amphorae are linked to the periods between 410-300BCE and have often been found in neighbouring sites such as in Sicily. They were originally characterised by red-painted lines, yet these do not preserve well in underwater environments. It is these unique and traceable characteristics of typologies which make amphorae so important, in that apart from being a chronological indicator, they also serve as an insight into trade patterns, hinting to even what types of items were being traded and with whom (Anastasi *et al.*, 2021).

2.1.2 Xlendi and The Tower Wreck

The harbour at Xlendi Bay was known as a centre of commercial activity in the past, even connecting to other hubs across the central Mediterranean. This was further confirmed by several foreign-led studies that documented and retrieved a large number of cultural relics in the 1960s. The majority of such relics may be found at the Gozo Museum of Archaeology (Heritage Malta, 2023).

The discovery was initially made in 1993 when a deep-water submarine investigation outside Xlendi Bay discovered a dense dispersion of amphorae at 105 metres of depth. Due to its proximity to the 17th-century Xlendi coastal guard tower, this deep-water location is today known as the Xlendi Tower Wreck. In order to better comprehend the area's archaeological potential, remote sensing studies were conducted between 2006 and 2008 with the goal of, among other things, relocating the deep-water site that was

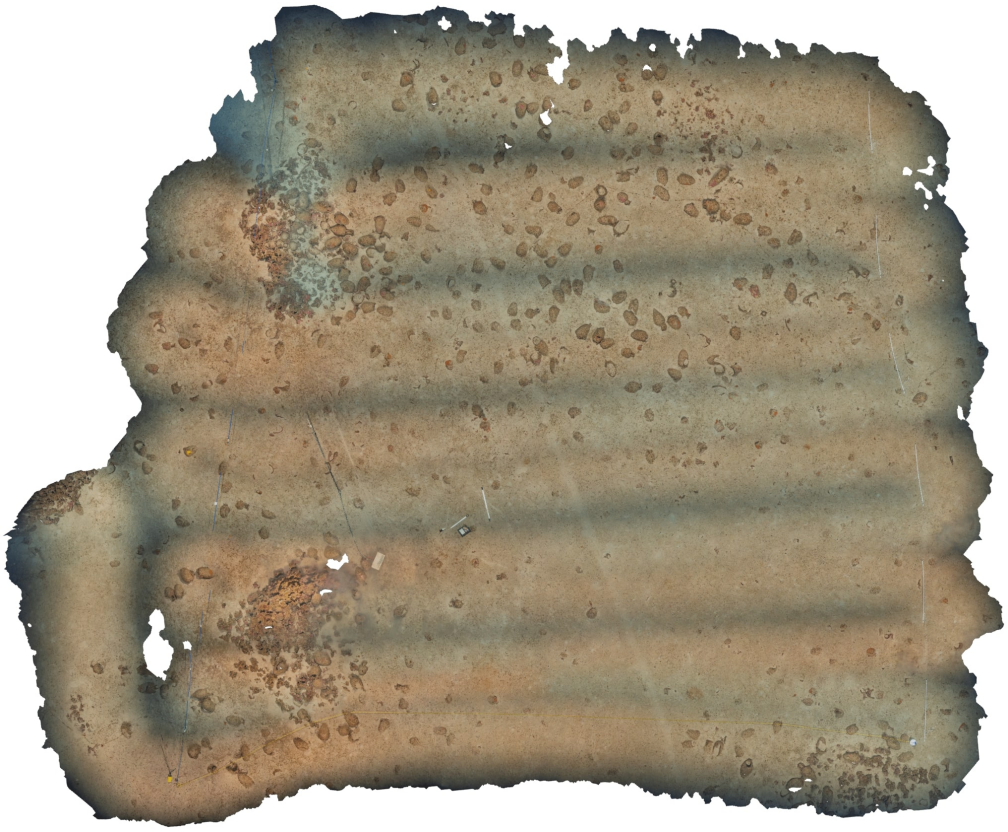


Figure 2.2: Orthomosaic of the entire 625 m^2 area.

found in 1993. The site's extent was measured, and it was discovered to be made up of a densely packed region of about 67,000m^2 (Heritage Malta, 2023). According to preliminary research, the artefacts were found to belong to the 3^{rd} century BCE, during Malta's Punic period. Further investigation is being carried out to determine the site's potential link to the shipwreck. An extensive initiative was commenced in 2021, with the aim of documenting the Tower Wreck site (Heritage Malta, 2023).

2.1.3 Tower Wreck Data and Photogrammetric Documentation

The exploration and documentation of the Tower Wreck has long been worked on by Heritage Malta (2023), where the enormous area is being targeted in chunks of 25m x 25m within a single dive operation. This divide-and-conquer method makes iterative documentation possible, where the area can be scanned in separate operations, and then aggregated into one single large representation through photogrammetry.

Photogrammetry is a method of creating a virtual 3D record with millimetric accuracy from images and videos, which is being used to document the site (Gambin *et al.*, 2023).

For this process, a series of overlapping images need to be captured all over the site, typically also emphasising lighting, and utmost precision in calibration setup. Exploration of maritime archaeology sites brings invaluable insights, especially in the context of deep-sea sites, where the artefacts and the seabed are often better preserved than shallower areas (Drap *et al.*, 2015). Unfortunately, it is a double-edged sword, as this often tends to be the most tedious aspect of such archaeological projects, where most archaeologists and marine biologists may not easily access the sites for analysis, especially since oxygen diving beyond 50m is prohibited and requires considerably more training, orchestration and skill (Drap *et al.*, 2015). Moreover, nautical archaeological sites have ever-increasing threats, such as newer techniques of trawling, an industrial fishing method that destroys the entire surface of objects on valuable archaeological sites (Drap *et al.*, 2015). Even then, if not for external factors causing damage, archaeological excavations themselves are often destructive (Drap *et al.*, 2013). All these active threats to such valuable sites prompt the necessity for proper documentation of artefacts. For this reason, photogrammetry is often introduced in such underwater archaeological operations, enabling an efficient method for comprehensive documentation, without impacting any of the objects themselves (Costa *et al.*, 2016; Drap *et al.*, 2015; Jones and Church, 2020; Yamafune *et al.*, 2017). This further provides all the associated benefits of digitisation, allowing experts and even the general public to access these remote areas in comprehensive detail (Gambin *et al.*, 2021).

Figure 2.2 displays an orthomosaic of the entire $625m^2$ (25m x 25m) area we will be working with for this project, showing the scale and nature of the area overall. One may also observe the route taken by the diver, starting from the bottom left. The diver then proceeded to go over the area horizontally, slowly moving upwards. Notably, this means that apart from overlaps in the direction the diver is going, there are also overlaps when the diver passes through the same area again. An example of the nonlinear underwater light may be observed in the horizontal darker lines, which were areas that were generally further away from the lighting equipment, hence quickly suffered low light and averaged a lower light level overall. Around the centre of the image, one may also observe 2 scale bars, which are used as reference scales for distance calibration in photogrammetric software. There is also a colour palette, which offers a grid of known colours, which are used to attempt to reverse the discolouration experienced due to the underwater light absorption. In terms of artefacts, one may see different densities of deposits around the area, where locations such as the top left contain densely packed areas of amphorae, which are sitting beside a rocky area. On the other hand, areas such as the centre-right have very little amphorae deposits.

Conveniently, these high-quality, relatively well-lit raw images that are used to make

photogrammetric models may be used as a basis for creating an object detection dataset, which enables the automated detection, classification and even localisation of relevant artefacts within imagery.

2.2 The Photogrammetric Process

The process of building and establishing a photogrammetric model is truly elegant, especially in terms of the efficient use of time and photographic resources (Gambin *et al.*, 2023). One of the preliminary stages of photogrammetry deals with data collection, establishing several requirements for image collection. The first requirement involves the capturing of a degree of overlap between images, and these images must also cover the entire area to be studied. Although this may not seem as large of a task, reaching the archaeological site itself is often a tedious process, and the diver's times may be restricted. In fact, for a Phoenician shipwreck in Xlendi discussed by Gambin *et al.* (2023), divers were limited to just 12 minutes on the site, making time extremely valuable, and further increasing the need for efficient methods of recording.

2.2.1 Image Matching Process

A particularly interesting part of the process lies is the second stage of data processing, where techniques such as the Scale-Invariant Feature Transform (SIFT) (Lowe, 1999) algorithm are used for homologous point estimation (Drap *et al.*, 2013). Looking into how the SIFT algorithm works step by step, the first task is to construct a scale space to ensure that features are scale-independent. This is done by creating a series of blurred images at different scales, from which the Difference of Gaussians (DoG) is computed to highlight significant changes in intensity. The formula for DoG at any given pixel position (x, y) and scale σ is given by Formula (2.1) (Lingua *et al.*, 2009).

$$DoG(x, y, \sigma) = (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y) \tag{2.1}$$

where, k is the scale factor, and $I(x, y)$ represents the same pixel value on the input image. Moreover, $G(x, y, \sigma)$ is the Gaussian kernel at pixel (x, y) and scale σ, defined by Formula (2.2) (Lingua *et al.*, 2009).

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{2.2}$$

Given the size of imagery, this formula is only used to calculate the local gradient directions for each key point, which are considered from local extrema in the DoG, whereas

points with low contrast or poor localisation are eliminated by comparing them with their neighbours. Next, the orientation is assigned to each key point to ensure they are rotation invariant. This is done by calculating the local gradient directions for each key point. Finally, the descriptor is created for each of these key points, which summarises the local features in vector form as obtained from the gradient magnitudes and orientations. Key points and their descriptors from different images can be compared for feature matching, typically using a similarity measure such as Euclidean distance.

Notably, the SIFT algorithm is considerably slow, due to the hefty matching that needs to be carried out around each key point, such a process may be unsatisfactory for larger projects. An alternative to the HOG technique is the Features from Accelerated Segment Test (FAST) algorithm, which, efficiently detects key points by employing a corner detection strategy based on the intensity variations around a pixel. However, it is a heuristic, hence sacrifices some of the robustness to noise provided by HOG and in turn SIFT (Edward and Drummond, 2006).

2.2.2 Structure from Motion and Dense reconstruction

Traditional photogrammetric methods were typically based on the basis of human stereo vision, where two separate viewpoints with a known distance are used to perceive depth, through a form of disparity estimation (Iglhaut *et al.*, 2019). However, depth may also be calculated if either the object or the observer is in motion, in which case a snapshot may be used from both angles and given known parameters (Iglhaut *et al.*, 2019). This forms the basis of Structure from Motion (SfM) and stereo matching algorithms classified as Multi-View Stereo (MVS) techniques.

SfM techniques ingest the initial key points established by image-matching techniques, serving as the next iteration. The main process behind this is known as the Bundle adjustment, dealing with camera pose estimation. Mathematically, bundle adjustment formulates an optimisation problem where the goal is to find the set of parameters that minimises the sum of squared differences between the observed image points and their corresponding projections. After this process, a set of parameters representing the camera poses as well as camera intrinsic parameters such as focal length and distortion coefficients are established. Additionally, the SfM process generates an improved iteration of the 3D model, known as the sparse cloud, which is richer than the tie points (Iglhaut *et al.*, 2019).

The sparse cloud contains some light depth and texture information, yet is still far from capturing detailed meshes of the surface. The process of dense reconstruction based on MVS begins with the selection of reference images, also known as seed points. These

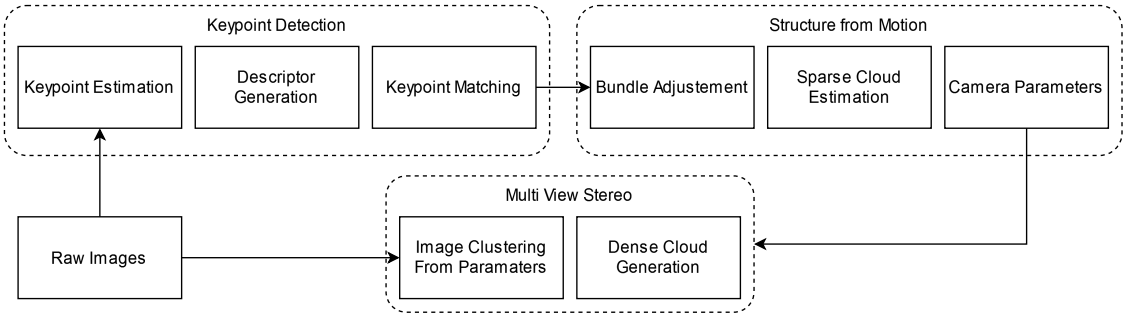


Figure 2.3: SfM-MVS end-to-end process starting from raw images to dense reconstruction.

seed points are derived from the sparse tie points that were previously generated by the SfM algorithm (Ma *et al.*, 2023). Once the seed points are established, the next step involves estimating the depth and normal vector of the pixels that are mapped by these seed points. This estimation is based on the intrinsic and extrinsic parameters of the camera.

Following this, small square patches are created around the seed points and projected onto the other images. The pixel depth and normal vector are then optimised by minimising the reprojection error. In order to ensure the accuracy of the reconstruction, noisy points are removed using the photometric consistency method of normalised cross-correlation, which in essence is used to measure the similarity between pixel intensity values in different images. Subsequently, the depth and normal values of the seed points are assigned to the adjacent pixels. This process is repeated until all pixels have been searched. Finally, depth maps are reconstructed based on the pixel depth and normal vector. The depth pixels are projected into 3D space to obtain the 3D point clouds, completing the dense reconstruction process.

This forms the basis of the process known as SfM-MVS, illustrated in Figure 2.3. SfM-MVS describes the entire pipeline starting from raw images, running a keypoint matching technique such as SIFT, then forming a sparse point cloud and estimating camera parameters using SfM, and finally computing a dense cloud of the area through MVS.

2.2.3 Agisoft Metashape

As one may quickly come to realise, the SfM-MVS methodology requires a considerable amount of separate processes and forms to be quite a complex system of operations. The complexity of such systems is especially determinantal in the case of rapid progress checks, such as those regularly required in daily diving operations, which help to ensure alignment, coverage and even plan future dives (Gambin *et al.*, 2018).

For this reason, off-the-shelf software is often utilised to facilitate the process, making it available for the general public, popular examples of which include Autodesk ReCap Photo and Agisoft Metashape (Jones and Church, 2020). These tools propose to expedite the end-to-end SfM-MVS process, starting from images, providing tools for alignment, and calibration, matching images, calculating dense point clouds, orthophotos and providing interactable visualisation tools.

Furthermore, Metahsape offers a Python SDK, allowing access to vast functionality, including the projection of points from 2D to 3D, marker point creation, conversion from model points to geographic, camera parameter access, orthomosaic generation and exporting of depth maps amongst more, allowing for deep integration between the utilities offered by the software and external software pipelines such as data enrichment for object detection (AgiSoft LLC., 2023).

2.2.4 Emergence of Neural 3D mapping

In recent years, 3D reconstruction has been seeing increased research attention through the exploration of neural approaches. As opposed to the numerical and algorithmic approaches we have discussed, newer techniques strive to create more visually sound scenes whilst utilising relatively fewer resources, both in terms of imagery required, and computation resources associated (Kerbl *et al.*, 2023; Mildenhall *et al.*, 2020).

Neural Radiance Fields (NeRF) (Mildenhall *et al.*, 2020) were proposed as a novel technique for 3D scene reconstruction. These require images and respective camera positions, which may either be obtained through rich data collection or estimated via methods like SfM. Different to traditional photogrammetric techniques, the NeRF optimisation process involves training a deep neural network to map 5D input coordinates, representing spatial location and viewing direction, into a volume density and view-dependent emitted radiance. By learning this mapping, NeRF can reconstruct the scene in a neural representation, which it can then use to generate new views. A radiance field, refers to a function that describes how light propagates through a 3D scene, it essentially captures the brightness of light rays at different points in the scene and from various viewing directions as described from camera poses. Albeit requiring camera poses, NeRF is still an unsupervised learning process, as its point is to learn to estimate the intricate details and lighting effects of the individual scene at hand. Once the model is trained, new views can be generated on the fly, making it suitable for immersive experiences where users can navigate through and interact with 3D scenes dynamically. Additionally, NeRFs' ability to capture fine details and realistic lighting effects enables the creation of visually stunning virtual environments, which goes beyond the point of traditional photogrammetry. That

being said, it's important to recognise that NeRF may not be the optimal choice for tasks requiring precise measurements, such as archaeological scenarios. While NeRF excels at visual fidelity, achieving the level of accuracy of traditional, well-done photogrammetric approaches can be challenging due to the inherent limitations of the technique.

Gaussian splatting (Kerbl *et al.*, 2023) offers a compelling alternative for visualising 3D scenes, particularly in applications where preserving fine details and achieving smooth rendering are paramount. This technique involves projecting 3D points onto a 2D image plane and then rendering them using Gaussian kernels to create smooth, visually appealing visualisations. One of the key advantages of Gaussian splatting is its ability to produce high-quality images with realistic depth cues and shading effects. However, it's important to note that Gaussian splatting shares many of the disadvantages to NeRF, as it is primarily a visual technique and lacks the ability to generate a comprehensive 3D model of the scene.

As discussed, the current choice behind visualisation techniques stems from a balance between measurable accuracy to visual quality, which depends on the application at hand. Variations on these mentioned techniques aim to strike a balance between, and newer and better techniques are emerging. This movement towards better 3D environment estimation further calls for techniques which are able to leverage this positional information, both for improving detection accuracy, but also for enriching outputs with positional context.

2.3 Object Localisation Within Imagery

Object localisation refers to the process of gaining contextual location information for detected objects. Especially within remote areas with limited access, the benefit of knowing where a detection fits into the bigger picture poses several benefits for the interpretability of results. This may be approached as a projection problem, where we want to establish and access the existing relation between the camera position and the object being targeted by the image.

To grasp the inner workings of such a process, it's best to first familiarise ourselves with a pin-hole camera model, which gives idealistic insight into how a real-world object is represented as an image, and the projections involved to translate in-between (Bianco *et al.*, 2013; Hartley and Zisserman, 2004). Depicted in Figure 2.4, we may note the three main components within this, the first being the object at point w, which has coordinates $[X_w, Y_w, Z_w]$ in the world coordinate system. The second component is the optical centre O_c, which is the origin of the camera coordinate system, with coordinates $[X_c, Y_c, Z_c]$.

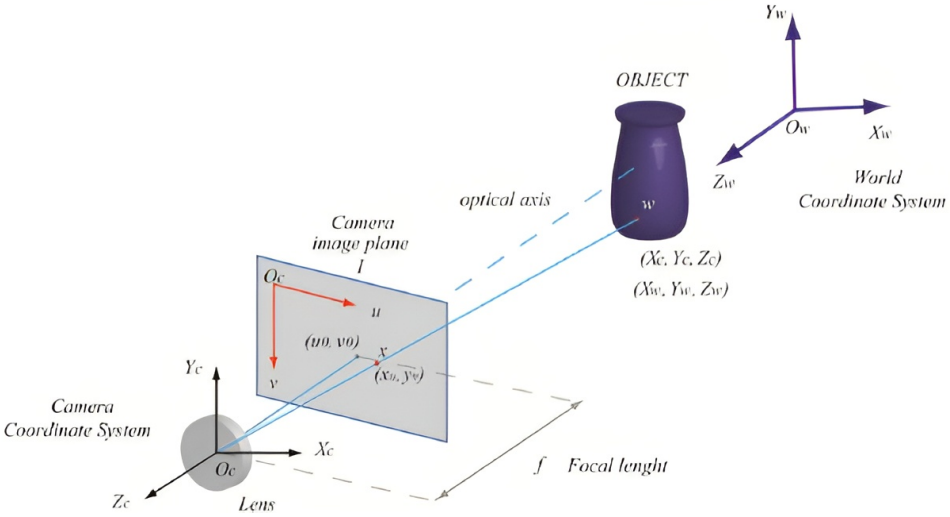


Figure 2.4: The pin-hole camera model by Bianco *et al.* (2013).

The third component is the image point x, which is the projection of w onto the image plane I, with coordinates $x = [x_u, y_v]$. The image plane is parallel to the X_cY_c plane and located at a distance f from the optical centre, where f is the focal length of the camera. The line passing through the optical centre and perpendicular to the image plane is called the optical axis. The pin-hole camera model assumes that the rays of light from the object point to the image point pass through the optical centre, forming a perspective projection. This projection can be mathematically described by (Bianco *et al.*, 2013):

$$x = \frac{f}{Z_c} \begin{bmatrix} X_c \\ Y_c \end{bmatrix} \tag{2.3}$$

This equation relates the coordinates of the object point in the camera coordinate system to the coordinates of the image point in the image plane. However, to obtain the coordinates of the object point in the camera coordinate system, we need to apply a rigid-body transformation that converts the coordinates from the world coordinate system to the camera coordinate system. This transformation can be expressed by:

$$X_c = RX_w + T \tag{2.4}$$

where R is a rotation matrix and T is a translation vector.

These two equations (2.3, 2.4) form the basis of the geometric model of the camera, enabling the mathematical projection from one of the coordinate systems to another. One such projection operation may be going from a particular position on the 2D image plane and projecting the orthomodel. This starts by casting a ray from the optical centre O_c, which passes through the pixel position x. When this ray is followed, this represents a

line of potential 3D points which would be represented as this pixel. Without additional information, we wouldn't be able to determine where this point on the cast ray is located, hence the exact location of the original object point along this ray. We also would not have access to R and T, which represent the camera's position. This is where photogrammetry comes into play, from which we may obtain the camera pose and depth information. The camera pose allows us to understand where the camera's coordinate system is based, whereas the depth information is what allows us to select the correct point along this ray and find the exact 3D coordinates (X, Y, Z) of the object point (Bianco *et al.*, 2013).

In reality, modern wrappers around these techniques are heavily abstracted. The usage of facilitators such as the Agisoft Metashape SDK (AgiSoft LLC., 2023) provides ray-casting features, where we can start from the origin of the camera (which is pose estimated), cast a ray towards the focal plane at the pixel position (u, v) and then using the surface of the estimated orthomodel, we can find the point intersection with the point cloud, giving us the estimated location of where the point would be in the photogrammetric environment. Through geographical alignments, usually done through the use of Ground Control Points (GCPs), being in modelled space with known geographic locations, we can then convert photogrammetric coordinates to world geographic coordinates. Such a process could potentially be leveraged in combination with object detection techniques, where the bounding box vertices are projected to form 3D object detection through 2D techniques.

2.4 Image Enhancement Techniques

Subaqueous imagery lacks quality in several ways, where adverse effects such as distortion, spectrum absorption, blurriness and discolouration are commonly experienced in deep underwater scenes (Espinosa *et al.*, 2023). These effects often tend to hinder common object detection models, especially with small and clustered targets commonly targeted in UOD. In light of this, there have been several researched techniques aiming to partially reverse some of these effects (Hummel, 1977; Lu *et al.*, 2016; Pizer *et al.*, 1987; Xu *et al.*, 2023; Zuiderveld, 1994). Nonetheless, most techniques focus on getting more normalised images, in essence, ones which look closer to on-land. Although these are usually perceived by humans as better looking, this does not necessarily translate to better feature extraction from vision models, as useful texture and colour information may still be sacrificed in the process (Xu *et al.*, 2023).

Classical image enhancement methods, deeply rooted in fundamental photometric principles, offer valuable tools for improving image quality. These techniques, while not

always perfect, are often fast and efficient, capable of achieving satisfactory results for many applications, without requiring lengthy training or inference processes.

A photography concept commonly leveraged is white balance correction, which is a process used to adjust the colour temperature of an image to ensure that whites appear white under different lighting conditions (Bainbridge and Gardner, 2016). A common approach to white balance correction involves using a colour card as a reference (Bainbridge and Gardner, 2016). Although this method may seem somewhat naive, it can effectively reverse colour casts by identifying reference colours and adjusting the overall colour balance accordingly. However, it's important to note that the efficacy of this approach may be limited by factors such as non-uniform lighting and refraction in underwater imagery, as well as variations in colour temperature across the image (Lu *et al.*, 2016). When artificial light is used, some methods (Lu *et al.*, 2016) aim to address this by considering the camera distance from the surface, yet complications arise when this light is mixed with ambient light, needing to consider more variables, and being more susceptible to noise.

Other popular techniques make use of the concept of image histograms, which represent the distribution of pixel intensities in an image. An image enhancement technique that follows this ideology is Histogram Equalization (HE) (Hummel, 1977). HE utilises the histogram to enhance contrast by redistributing pixel intensities across the entire range. However, it has the drawback of equalising the global histogram, which can lead to over-enhancement in regions with low contrast. To address the limitations of HE, Adaptive Histogram Equalization (AHE) (Pizer *et al.*, 1987) was introduced. AHE performs histogram equalisation locally in different regions of the image, thereby preserving local contrast better than global methods. This local adaptation helps to avoid over-enhancement in low-contrast regions. However, AHE can lead to amplified noise and artefacts due to its aggressive contrast enhancement. In an effort to minimise these effects, Contrast Limited Adaptive Histogram Equalization (CLAHE) (Zuiderveld, 1994) was introduced. CLAHE restricts the contrast enhancement in each local region to prevent over-amplification of noise and artefacts. By limiting the contrast enhancement, CLAHE aims to maintain a natural appearance in the enhanced image. Nonetheless, even CLAHE may still over-amplify noise and cause over-saturation, particularly in regions with high local contrast or sharp transitions. These limitations have spurred the development of more advanced and context-aware image enhancement techniques in recent years, such as through neural approaches.

The transition towards neural approaches has achieved remarkable results, where Convolutional Neural Networks (CNN), transformer, or even Generative Adversarial Network (GAN) based techniques are often leveraged to improve the quality and look of images (Akkaynak and Treibitz, 2019; Espinosa *et al.*, 2023; Islam *et al.*, 2020b). Put plainly,

these methods work by generating new images that are similar enough to the original, yet aiming to have the degrading features reduced (Islam *et al.*, 2020b). These are typically trained on a dataset where images have already been enhanced through intensive or laborious methods, such as using colour restoration, giving targets that the model should aim to reach. There are also instances where the clear images are taken on land, and the images are degraded retroactively. GANs typically have more restorative power than classical methods, as they excel in generating new information which fits the images. However, they are also computationally more expensive than traditional methods (Islam *et al.*, 2020b). These methods are also very sensitive to the quality of the training data, where if the training data is not representative of the image domain to be enhanced, their performance can suffer. Other neural approaches use CNN-based architectures, such as proposed by Espinosa *et al.* (2023). Their architecture is based on the U-net encoder-decoder structure, which can even reach real-time speeds (40fps), making such techniques not only effective but also fast.

All in all, there are many approaches to image enhancement, and a decision is to be made based on the data available, speed, quality and whether these should be targeted towards visual effects or towards better learnability for object detection techniques.

2.5 Summary

This chapter serves as a pool of relevant information which one requires to better understand the purpose, application, restrictions and methodology of our study. Initially, the nature of maritime archaeology, our working area and the details of the diving operations are discussed, giving the context behind the origin of the data whilst establishing a target audience of where such advancements will impact most.

Secondly, the photogrammetric process is discussed, outlining its strengths and its particular relevance to archaeological site documentation. Furthermore, the major processes of the photogrammetric process were analysed, including Image matching techniques like SIFT, the concept of Structure from Motion, point clouds, the estimated camera parameters and the Multi-View Stereo for dense cloud generation. Moreover, the emergence of neural scene reconstruction is explored, which provides a promising future for ease of obtaining 3D scene information, making such advancements even more impactful.

Thirdly, the relevance of photogrammetry for object detection is targeted through the use of localisation techniques, including the bidirectional mathematic link that may be established between images and the world. This link was a key point within our work,

where we studied how to best use this to increase both the accuracy and value of object detection predictions.

Finally, two primary approaches to image enhancement were outlined, being classical and neural techniques. Classical methods such as White-Balance optimisation or CLAHE are typically based on photographic concepts. On the other hand, GAN-based methods generate new images that are similar to the original, and are especially useful for severely degraded images such as underwater images, yet have considerable computational costs.

3 Literature Review

The analysis of significant literature forms the cornerstone of any academic research, paving the way for innovative expansion and refinement of established concepts. The first section of this chapter outlines the major building blocks of modern object detection architecture, which over the past few years has seen several changes in definition. Next, sections 2 and 3 focus on notable takes on the main parts of detectors, including the newest developments between convolution and transformer-based approaches. Section 4 highlights how these advancements in generic object detection models are translated to underwater object detection, given its unique set of limitations. Section 5 follows this by addressing similar work employing enrichment techniques through the use of depth and saliency. Finally, Section 6 delves into similar studies leveraging photogrammetric techniques, especially in underwater environments, serving as case studies for specific approaches.

3.1 Dissecting the modern Object Detector

The general architecture of the object detector has seen several shifts over the past years, and keeping up with the latest advancements is becoming increasingly difficult. This section serves as a snapshot of common concepts within most modern architectures, which share similar functionality and purpose across techniques. Particularly, a modern detector typically has three main parts, namely a backbone, a neck and one or more detection heads. The rest of this section delves into the main philosophy of each part.

3.1.1 Preprocessing and Augmentation

The data preprocessing step is crucial in preparing imagery for training, aiming to maximise the amount of information that can be extracted from the readily available data. A modern method of application is the Albumentations package (Buslaev *et al.*, 2020), which offers several augmentation techniques. These include items such as random cropping,

flipping, CLAHE for colour normalisation (see Section 2.4), shuffling in RGB channels, saturation shuffles, and randomised contrast. Each of these techniques alters the image representation in different ways, increasing the diversity of the training data. Buslaev *et al.* (2020) further emphasise how image augmentations improve the generalisation capabilities and overall robustness of deep learning models. This is achieved through presenting the model with a wider variety of training examples, in turn allowing the model to learn how to circumvent difficult conditions and reduce the risk of overfitting.

3.1.2 Backbone

The backbone is an essential aspect of modern object detectors, greatly impacting both the resultant speed and accuracy of the general detection architecture (Zou *et al.*, 2023). This part is typically the initial phase of the model, responsible for the ingestion of the data as the feature extraction phase. The purpose of the backbone is to summarise and condense the image's most relevant features, to be consumed in later stages of the model. For this reason, this is often represented as having a funnel shape, which in the case of CNNs, is characterised by sets of convolutions and max pooling layers, which quickly decrease the resolution of the image, forming a semantically rich feature map (Sharma and Mir, 2020).

Since AlexNet (Krizhevsky *et al.*, 2012) was proposed, being one of the first backbones as we know them today, backbones have seen increasing widths and depths, aiming to maximise retained information (Liang *et al.*, 2022). Popular examples include VGG, ResNet, DenseNet and ResNeXt, which have progressively built out the main attributes and features common across modern implementations. Moreover, there has been a shift towards more efficient backbone architectures, including MobileNet and EfficientNet, which have been specifically designed towards stripping inefficient components, further being suitable for real-time and low-resource applications (Howard *et al.*, 2017; Tan and Le, 2019).

3.1.3 Neck

With the rich information captured within the backbone, this stage aims to make the best use of these extracted features. Typically the challenge lies in striking a balance between semantic and localisation information, both of which are fruitful for object detection tasks.

Semantic features refer to the high-level, abstract features that provide a vast understanding of the image content. These features are typically extracted from the deeper layers of the backbone, which capture complex relational patterns and structures within

the image. On the other hand, localisation features refer to the low-level, spatial features that provide precise location information about the objects in the image. These features are typically extracted from the shallower layers, which retain more of the original spatial resolution of the image. The neck is designed to fuse these two types of features to create a rich, multi-scale feature representation, which allows the network to effectively detect objects of various sizes and at different scales.

3.1.4 Detection Head

The large amount of information extracted from the previous stages of the model requires complex understanding and interpretation, which occurs in the detection head. Not only does this involve classifying the type of object but also determining its position and size within the image. Variations in the detection head can often see substantial changes in the model behaviour, hence often seeing researchers' attention in better interpreting features. This could be broadly classified into two main categories, being anchor-free and anchor-based models (Liu *et al.*, 2020; More and Bhosale, 2023).

Anchors refer to predefined bounding boxes that are used as starting points for predicting the location and size of objects in the image (Liu *et al.*, 2020). Anchor-based models adjust these anchors to match the objects in the image, which can help stabilise the learning process and improve the model's ability to detect objects of various sizes and aspect ratios. Anchor-based may be further categorised into two main detectors, namely single-stage and two-stage detectors (More and Bhosale, 2023).

Two-stage methods follow two primary stages in the detection phase, where the first stage involves a region proposal network, which generates a set of bounding boxes that potentially contain objects. The second stage involves selecting the most promising boxes and refining their positions. A classic example of this approach is the R-CNN (Girshick *et al.*, 2013) and its successors, Fast R-CNN (Girshick, 2015) and Faster R-CNN (Ren *et al.*, 2015). As discussed by More and Bhosale (2023), two-stage techniques tend to be associated with slower performance due to the more complicated detection stage both in predicting a large number of potential regions and in filtering them out using techniques such as Non-Maximum Suppression (NMS). Additionally, these have often been associated with a decreased ability to consider global context, yet have typically achieved higher levels of accuracy. This trade-off between accuracy and computational efficiency has led to a shift towards single-stage techniques.

Single-stage detectors, as the name suggests, aim to detect objects in one single phase, directly generating the output rather than intermediary proposals. One of the first models to propose this approach was the Single Shot MultiBox Detector (SSD) (Dragomir

et al., 2016), which was built upon VGG16. Instead of using a region proposal step, SSD assigns scores to predictions at runtime, eliminating the need for a separate proposal step. At the time of its introduction, SSD surpassed the performance of Faster R-CNN while being up to three times faster, addressing the computational inefficiency of two-stage techniques. In recent implementations, single-shot methods have even achieved similar accuracy to two-stage techniques whilst being both simpler and typically faster, making them ideal for most modern applications.

That being said, anchor-free models have also been gaining traction, where unlike anchor-based models, which use predefined bounding boxes as starting points for predicting the location and size of objects in the image, anchor-free models predict the location and size of objects directly from the feature maps (Liu *et al.*, 2020). Anchor-free models also have two main approaches, being keypoint detection and dense prediction. Keypoint detection methods focus on detecting specific points of interest in the image, such as the corners or the centre of the object, to define the bounding box. Examples include CornerNet, which regressively predicts the top left and bottom right points of an object, and CenterNet which additionally predicts the centre point of the object. These methods essentially, base the detection on a specific point relative to the estimated object position (Liu *et al.*, 2020). On the other hand, dense prediction methods follow a similar ideology to semantic segmentation, where pixel-level predictions are made, producing a heatmap around the detected objects, which is then used to estimate a bounding box around it (Liu *et al.*, 2020). These anchor-free techniques aim to avoid the complicated computation of anchor generation and matching, and they are more flexible and adaptable to different datasets and scenarios (Liu *et al.*, 2020).

3.2 Convolutional Object Detection Techniques

This section delves through some of the essential convolution-based object detection techniques. Convolution has been extensively used throughout many concepts in the vision domain, be it enhancement, image classification and object detection amongst others. The widespread use has also led to them becoming heavily optimised towards efficiency and accuracy.

3.2.1 Common Convolutional Concepts

This subsection delves into some specific object detection building blocks which have gained relevance and popularity within the domain. These are typically infused into object detection techniques in different architectures, which serve very specific purposes,

as will be outlined in each. Such concepts become essential knowledge when trying to understand and especially compare detection models, necessitating understanding of the intended purpose and resultant effect of each layer.

3.2.1.1 Skip Connections

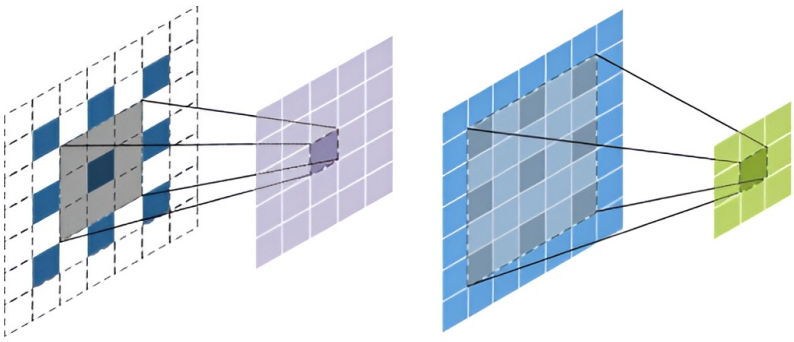
Skip or residual connections have become ubiquitous in many AI domains, particularly due to their efficacy and little overhead introduced. These are connections which feed information directly from earlier parts of the network to deeper layers, aimed at counteracting the vanishing/exploding gradient problem experienced in deep neural networks (He *et al.*, 2015). Although it may seem counter-intuitive, as information is able to skip parts of the model, this is only applied in the backpropagation stage. Such models typically achieve better performance, as in later stages of the model the information may be too far off from the original image, making backpropagation less effective. For this reason, skip connections help the model refer back to earlier states, increasing overall information propagation throughout the entirety of the network. This technique was popularised by ResNet (He *et al.*, 2015), where the whole idea behind the architecture was the introduction of these connections, which achieved remarkable improvements over predecessors, highlighting the possible tangible improvements.

3.2.1.2 Deconvolution

Deconvolutions have seen considerable adoption in the vision domain, where the top-level definition of the operation may be simply described as doing the opposite of typical convolutions (Fu *et al.*, 2017; Qin *et al.*, 2019).

In essence, while typical convolutions involve aggregating information within a defined area of an image, deconvolution operates by starting with a single value and expanding it into a larger result patch through convolution operations. This process is accomplished by introducing padding around the pixels of the source image, effectively creating a gap. Consequently, during the convolution process, the resulting patches become larger than the original image. For a visual representation, refer to Figure 3.1(a), where the padding applied to the original source image and the resulting patch size are clearly depicted.

A practical technique utilising deconvolution techniques was presented by Fu *et al.* (2017). In this instance, it was used to form an asymmetrical hourglass figure similar to an encoder-decoder structure. The initial layers of the architecture follow the typical feature extraction conical shape, where the image is summarised into smaller and smaller layers. However, at the very end, the authors opted for several layers of deconvolution which



(a) Deconvolution operation

(b) Dilated convolution operation

Figure 3.1: Deconvolution and Dilated convolution operations as presented by Qin *et al.* (2019).

re-enlarge the image from the latent space, allowing for finer localisation, especially on small objects. This has also seen adaptation by Chen *et al.* (2022) for underwater scenes, where low-lighting and low-quality scenes may benefit from the deconvolution process for added contrast and detail.

3.2.1.3 Dilated Convolutions

Dilated convolutions aim to increase the receptive field of convolutions, or more plainly their range, through the introduction of gaps within the kernels. As opposed to the contiguous area typically used in normal convolution operations, dilated convolutions can capture information from a larger area, whilst retaining the same computational complexity (Chen *et al.*, 2022; Qin *et al.*, 2019).

Qin *et al.* (2019) leverage dilated convolutions as a replacement for max pooling. The authors claim that due to its nature, max pooling suffers from information loss, which negatively affects the localisation accuracy. For this reason, dilated convolutions were introduced as a learnable alternative, which achieve a similar effect. This is depicted in Figure 3.1(b). Notably, this contrasts with deconvolutions as this dilates the kernel itself rather than the underlying data, in essence skipping pixels altogether, meaning a wider portion of the original area is captured in any given instance.

3.2.1.4 Deformable Convolutions

With the benefits aforementioned in dilated convolutions, this approach takes it a step further. Instead of expanding linearly outward with some predefined gaps, deformable convolutions introduce learnable 2-dimensional offsets within operations, which are cal-

culated at runtime. This allows the model to focus on different parts within the convolution area, which tends to give more liberty to the models.

The equation for deformable convolutions is given by Yang *et al.* (2023):

$$Y(p_0) = \sum_{p_n \in R} w(p_n) * X(p_0 + p_n + \Delta p_n) \tag{3.1}$$

where the output of the feature map $Y(p_0)$ is defined in terms of the input feature map X, the convolution kernel R, the position p_n being the n^{th} point in the convolution kernel, the weight $w(p_n)$ associated with p_n, the position p_0 being the current position on the output feature map, and the two-dimensional offset Δp_n, being the offset for the sampling point in the deformable convolution. It is exactly this final learnable offset which defines these types of operations, where the model can learn to hone into specific parts of the patch, sort of emulating a similar effect to that achieved by attention modules. Notably, this offset value is often not a whole number, where interpolation is required to calculate the feature value at the theoretical point.

3.2.1.5 Feature Pyramid Networks

Emerging from the issue of deeper backbone techniques, Lin *et al.* (2017) proposed Feature Pyramid Networks (FPNs), typically used as a neck stage of the model, targeting better propagation of semantic and localisation data between low-level and high-level feature maps.

As discussed in Section 3.1, modern backbones are deep structures, which aim to extract valuable features from images. Shallower layers have a smaller receptive field, which is due to the limited size of convolution kernels, yet still large size of the features. In essence, the kernel is proportionally smaller to the image size. It is also for this reason, that shallower layers have better localisation information, as each operation is more granular to the relative size of the image. As the features are reduced, the kernel size becomes proportionally larger, meaning a single convolution has access to a more relative area of the original image. As expected, at these deeper parts, the localisation information is not as rich, yet more global patterns can be targeted for semantic information.

It is this exact discrepancy that FPNs aim to normalise, proposing a method of allowing localisation information to deeper layers within the network. Illustrated in Figure 3.2, the features extracted from all layers are then fused into a mirror of the first stage, where separate prediction heads then branch out of each scale, targeting different sizes.

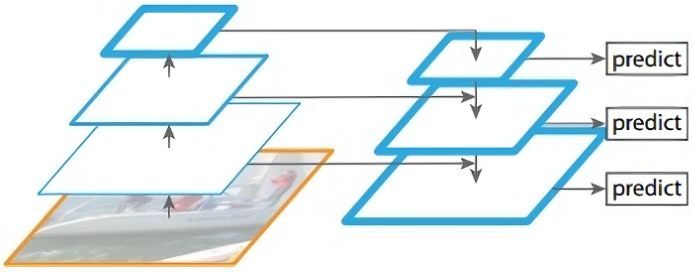


Figure 3.2: Feature Pyramid Network (Lin *et al.*, 2017).

3.2.2 Convolutional Model architectures

For the past decade, deep learning-based computer vision techniques have been dominated by CNNs (Feng and Tang, 2023). These models excel at extracting local features, such as lines or corners, and then building their way up to recognise larger structures. During their reign in the vision domain, CNNs have been optimised for efficiency, achieving real-time speeds and impressive accuracy.

3.2.2.1 YOLO Family

One of the most successful convolutional model families has been the You Only Look Once (YOLO) suite (Redmon *et al.*, 2015). The history of YOLO models is marked by constant innovation, with different research groups contributing advancements and competing on various generations of architecture. Nonetheless, these models have reached a point of high accuracy and efficiency, seemingly approaching the saturation point for traditional CNN-based object detection. The defining characteristic of YOLO models is their single-shot nature. Unlike region-proposal based models (Girshick *et al.*, 2013), YOLO performs the entire object detection process in a single forward pass through the network, making it significantly faster.

Among the chain of YOLO advancements, YOLOv5 (Jocher, 2020) has seen widespread adoption within the computer vision community. Building upon previous versions, YOLOv5 introduced residual connections and an FPN, making it considerably more versatile across several scales, and increasing its accuracy. YOLOv5 also had the added benefit of being written in PyTorch library in Python rather than C on the Darknet library used in earlier versions. Although differences in speed may be of concern, these are quickly outweighed by the much larger community and support seen by PyTorch developers, whilst still achieving better performance than YOLOv4.

A few iterations later, YOLOv7 was proposed by Wang *et al.* (2022), focused on further enhancing real-time detection accuracy without compromising inference times. The

authors claim to achieve this by significantly reducing model parameters, and effectively improving speed and accuracy simultaneously. They introduce the term "Bag of Freebies" to describe these general improvements that solely impact training costs and not inference costs, such as modifications to the model structure and loss functions. Comparing the base model to YOLOR, the previous state-of-the-art model, the authors report a 43% reduction in parameters with a minimal increase in Average Precision (AP) of approximately 0.4%.

Later, YOLOv8 was introduced (Jocher *et al.*, 2023), which once more proposed several benefits over previous generations. A major improvement in YOLOv8 is its anchor-free nature, where it directly predicts the object centroid, leading to faster NMS, which is a time-consuming step during inference. Additionally, the model utilises online image augmentation through mosaics, which helps the training process recognise challenging situations by introducing new object orientations, occlusions, and mixed placements within the training images. Aside from purely architectural advancements, the authors have placed considerable emphasis on making the model intuitive to train and adapt to custom datasets, aiming for wider user adoption.

One of the most recent advancements in the YOLO family comes from YOLOv9 (Wang *et al.*, 2024), announced by the same authors of YOLOv7. In this iteration, the authors propose several novel techniques, including Programmable Gradient Information (PGI) and Generalized Efficient Layer Aggregation Network (GELAN). PGI is designed to combat the concept of data dilution, a common issue in deep learning models where information tends to get diluted as it passes through layers. This works by introducing an auxiliary reversible branch, a secondary training path that is designed to generate reliable gradients that supply the main branch with information during backpropagation. Notably, for inference these auxiliary branches can be completely removed from the model, improving efficiency, size and speed. On the other hand, GELAN is a novel network design that optimises the structure of the model for efficient information flow. This technique is based on ELAN, being a design strategy for deep learning networks. The primary motivation behind ELAN is to manage the shortest and longest gradient paths, thereby enabling deeper networks to learn and converge more efficiently. By analysing the gradient path, the weights of different layers can learn more diverse features, resulting in better predictions and reducing latency. As an extension, GELAN can support any computational block, rather than specific ones, increasing the applicability across various methods. Additionally, apart from the considerations of ELAN, GELAN also considers the computational complexity, accuracy, number of parameters, and inference speed, allowing users to better tailor the model towards the intended inference device and maximising parameter utilisation (Wang *et al.*, 2024).

3.2.2.2 Efficiency Oriented Techniques

In parallel to the mentioned YOLO advancements, pushes towards more efficient utilisation of resources were made, which respect the limitations of low-power devices and edge computing. EfficientNet is a class of convolutional neural networks proposed by Tan and Le (2019) at Google in May 2019, achieving high accuracy with orders of magnitude fewer parameters than other methods. The model emphasises thinning out the redundant parts of the network, making it more computationally efficient. EfficientNet's main innovation lies in its ability to balance accuracy and efficiency by carefully scaling the model's depth, width, and resolution at compound scales. Apart from reducing potentially redundant parameters within these, this method further reduces the manual hyperparameter tuning required to train the network. Despite the reduction in parameters, the model achieved state-of-the-art performance at the time (Tan and Le, 2019). Being targeted at image classification, Efficientnet can also be used as a backbone towards object detection methods, retaining the primary benefits and advancements.

EfficientNet was further improved upon by EfficientDet (Tan *et al.*, 2020), which combines the EfficientNet backbones with a bi-directional feature pyramid network (BiFPN), which adds weights to feature importance, whilst enhancing feature fusion within the network. Additionally, this technique is targeted towards being an end-to-end object detection architecture, rather than being solely a backbone. The authors were able to achieve better efficiency than prior models, across several applications. For instance, EfficientDetD7 achieved 52.2 AP on the COCO dataset with just 52M parameters, being more than 4 times smaller than previous detectors with similar AP. Moreover, compared to previous detectors with similar results, the models are up to 4.1x faster on GPU, further suggesting their efficiency on typical hardware resource-constrained devices.

An even later addition to the efficiency-focused family was EfficientNetV2 (Tan and Le, 2021), which specifically aims at improving upon training efficiency. Through the adoption of a training-aware Neural Architecture Search (NAS) and the previous compound scaling techniques, these models are able to train much faster than SOTA techniques, whilst still being smaller. Once again being targeted towards classification tasks, EfficientNetV2 achieves 87.3% top-1 accuracy on ImageNet, which the authors claimed to have even outperformed similar vision transformer techniques, whilst training more than 5x faster using the same resources.

3.3 Vision Transformers

Transformers have dominated the Natural Language Processing (NLP) space for quite some time, where bigger and more efficient models are still frequently emerging, showing no signs of saturation (Dosovitskiy *et al.*, 2021). These models use self-attention mechanisms to learn relationships between all the elements in an image, rather than just the relationships between nearby pixels like in CNNs.

The introduction of transformers for computer vision occurred quite recently when Dosovitskiy *et al.* (2021) released their paper titled "An image is worth 16x16 words". The authors achieved the first major breakthrough in adopting transformers to the vision domain, creating the so-called Vision Transformer (ViT). The major setback that was hindering earlier transformer adoption to the vision domain, was the sheer amount of pixel values that images contain. Whilst NLP typically deals with hundreds of words, CV often deals with thousands of individual pixels in modern image resolutions. The authors were able to circumvent this issue by grouping pixels into 16×16 patches, and considering these as learnable, hence the paper's name. When considering that computing global self-attention is a quadratic operation, this optimisation becomes even more noteworthy, single-handedly making the adaptation feasible. An illustration of this process may also be seen in Figure 3.3, showing the initial image, patch splitting, positional encoding, the transformer-encoder and the classification $[CLS]$ token. These parts make up the main building blocks of a modern vision transformer, which we will elaborate on in the following subsections.

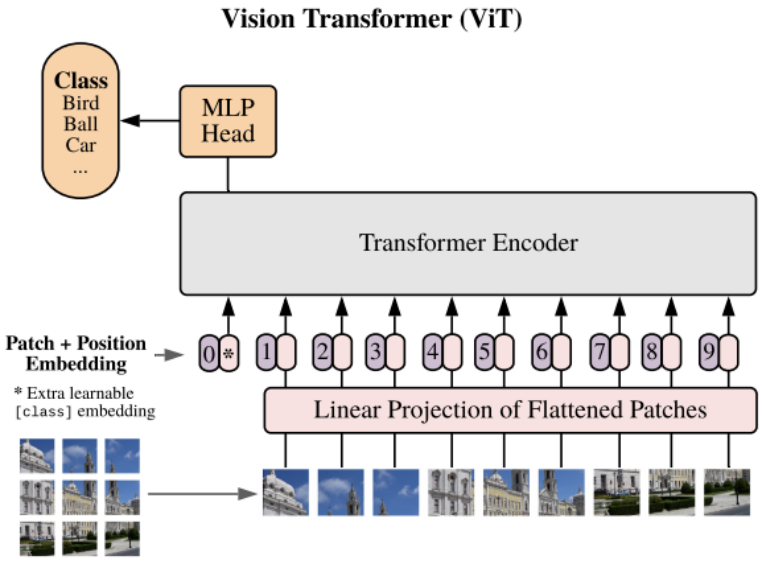


Figure 3.3: ViT Architecture (Dosovitskiy *et al.*, 2021).

3.3.1 Common Vision Transformer concepts

Vision transformers pose a new paradigm in many domains, with radical shifts from traditional techniques. Given their prevalence and potential impact on the domain, it is imperative to understand their purest strengths and scalability, whilst keeping in mind the associated limitations. This subsection outlines the major parts of a vision transformer, delving into the most important concepts and processes.

3.3.1.1 Multi-Head Self-Attention Modules

Attention is the concept of weighing a specific area of an image more than the rest, based on some estimated relevance (Cheng *et al.*, 2016). Self-attention stems from NLP where the meaning of a particular word is based on the context established by other words within the sentence (Vaswani *et al.*, 2017). In the context of vision, pixels are the information, hence in this case it deals with the interrelation between some specific pixel and all other pixels within the image. This is also where the expensive nature of the traditional transformer comes from, where this self-attention operation scales quadratically to the input size (Liu *et al.*, 2021c).

These attention operations typically occur in one or more attention heads, which have the role of focusing on a specific concept. Each attention head has three main components, namely the $query$, key, and $value$ (Vaswani *et al.*, 2017). This is associated with the way retrieval systems work, by searching for some $query$, comparing the matches with some key, and returning some relevant results representing the $value$. The comparison between the $query$ and the key can use any vector similarity comparison, such as cosine distance. The query and key objects are multiplied, in the same way it is for cosine similarity, forming the attention filter, a learnable $n \times n$ matrix representing the attention between each item and all others. Using this attention filter as a mask, this is multiplied with the $value$, giving us the filtered value, only highlighting the most important features.

The multi-head aspect of the self-attention aspect comes from the fact that the model often has multiple of these sets of attention filters in each module, all associated with learning different relational concepts. For example the original ViT paper used 12-16 heads, and calculated the self-attention between pixels within the same 16×16 patch (Dosovitskiy *et al.*, 2021).

3.3.1.2 Inductive Bias

In vision, CNNs rely on spatial hierarchies and local correlations, where although this likely makes training easier, due to the hand-holding local limitation of CNNs, it also limits the

maximum information absorption, and often lacks global context. Conversely, transformers tend to exhibit the opposite of this property, where the low inductive bias allows more global pattern recognition, yet the positional relation of patches needs to be learnt from scratch. Understanding this distinction helps in leveraging the strengths of each model for the optimal solution, depending on the requirements at hand.

3.3.1.3 Tokenization and Positional Embedding

An important stage in the transformation of input images occurs during tokenization. As illustrated in Figure 3.3, input images are initially grouped into isolated arrays of pixels, referred to as linear embeddings (Dosovitskiy *et al.*, 2021). In the context of vision, an embedding refers to a continuous representation of an image patch that captures important visual information and spatial relationships (Dosovitskiy *et al.*, 2021). This representation allows the model to understand and interpret visual content by encoding the image into a vector space where similar images are closer together and dissimilar images are farther apart.

Since these embeddings are practically linear arrays of patches, this initial representation lacks inherent spatial relationships, which is an important aspect of most vision tasks. To address this, positional embedding mechanisms are employed to encode positional information into the input sequence, allowing the model to discern spatial relationships between tokens. A common example of position embedding involves incorporating sine and cosine functions of different frequencies to represent spatial coordinates relative to each token's position within the input sequence (Vaswani *et al.*, 2017).

By integrating tokenization and positional embedding stages, vision transformers equip themselves to handle spatial relationships and global context effectively, enabling them to excel in various computer vision tasks. This approach reflects a departure from the traditional CNN architecture, emphasising the importance of adaptability and flexibility in capturing complex visual patterns and relationships.

3.3.1.4 Transformer Encoder

The transformer encoder is responsible for capturing the inter-dependencies within image patches, representing the backbone part of a detector. This is achieved through the use of self-attention mechanisms on the input patches, allowing the model to attend to different parts of the image and capture both local and global relationships. This results in a set of enriched patch embeddings that preserve important visual information and spatial relationships. In turn, these may be used for facilitating downstream tasks such as image classification, object detection, and semantic segmentation.

In certain applications, such as ViT, an encoder is found to be sufficient. Here, the transformer's role is primarily to enrich vector representations of image patches, enabling operations like attention to be performed effectively. Other techniques may require both an encoder and a decoder, depending on the complexity of the expected output, such as sequence generation or autoregressive prediction.

3.3.1.5 Latent Space and Interpretation

The concept of a latent space refers to the abstract, high-dimensional mathematical representation the most relevant parts of the image as ingested by the encoder. One might struggle to understand how the process of summarising an image into a compressed, richer version strictly relates to other downstream tasks. However, this summarisation is crucial as it condenses complex image information into a format that can be effectively analysed and directly used for decision-making.

In classification tasks such as ViT, the $[CLS]$ token, first conceptualised in NLP with the BERT model (Devlin *et al.*, 2018), plays a crucial role. Since in self-attention, all tokens are compared to all others, the value of the class token will be dependent on all other values in the embedding. This effectively serves as a summarised output of the model, whereby using it for evaluation, the model will further learn to prioritise its value based on relevant other parts of the text. This same concept is used in image classification tasks, where the content of the embeddings may be summarised by this single class token, serving as the output.

However, achieving relevant output is not always as simple, and in other tasks such as object detection, a single label does not suffice. Due to the added detail, more interpretation is required to get additional information such as bounding box positions, and an arbitrary number of predictions, which adds further dimensionality on top of classification. A method of doing this is implemented by Fang *et al.* (2021), where 100 randomly initialised $[DET]$ tokens are used as placeholders for potential results. These are a step up from the single class token used in ViT, yet in other cases, decoders are required to better extract information.

3.3.1.6 Transformer Decoder

The transformer decoder, is responsible for interpreting the low-level features of the latent space, such as those generated by an encoder. This could either be used to transform these into formats easier to work with or could also be used for full-scale generative tasks.

In NLP, the transformer decoder plays a pivotal role in sequence-to-sequence tasks, transforming the feature-rich embeddings into coherent and contextually appropriate

output sequences. This same ideology is followed in vision, where for example DEtection TRansformers (DETR) (Carion *et al.*, 2020), utilise the decoder to interpret encoded features from the latent space, transforming them into sequences of proposed class labels and positional information. In such cases, this decoder component acts as part of the head of the detector, synthesising learned representations into actionable predictions.

In contrast to encoder-only approaches, decoder-only architectures have been adopted in scenarios where a full contextual understanding of the input is unnecessary. Notable examples in NLP include GPT-3 (Brown *et al.*, 2020) and GPT-4 (OpenAI *et al.*, 2024), designed to generate coherent text sequences based solely on the preceding context. Similarly, in computer vision, this translates to generative applications where pixel generation relies solely on preceding ones.

3.3.2 Vision Transformer Based Object Detection Architectures

One of the first applications of transformer to vision was indeed ViT (Dosovitskiy *et al.*, 2021), where the transformer encoder was used as the main feature extraction phase of the model.

The evolutions based on ViT continued, and one of the most influential milestones since the original ViT was the SWIN transformer, being chosen as the best paper at ICCV 2021 (Liu *et al.*, 2021c). This adaptation uses hierarchical, shifted windows to extract image features at different levels, whilst considerably increasing the efficiency and performance over ViT. This technique starts similar to ViT transformers, where it splits the image into a set of patches. These patch sizes are set to 4x4, which means a significantly higher amount of patches are used. These also have some amount of channels denoted as C, hence forming a resultant size of $\frac{width}{4} \times \frac{height}{4} \times C$ (Liu *et al.*, 2021c). These patches are then further grouped to form windows. These windows go through SWIN Transformer Blocks, which are the main contribution of the technique. SWIN blocks are used in pairs, where the first one carries out Window Multi-head Self Attention (W-MSA), and the second one uses a Shifted W-MSA (SW-MSA). Since ViT computed the self-attention globally, meaning each patch between all others, large images were still time-consuming (Sun *et al.*, 2022). In SWIN, the attention is computed between patches of the same window, which serves a major role in increasing efficiency. This covers most required relations, as related patches are likely to be adjacent. Moreover, when the windows are shifted, attention may be computed between previously non-adjacent patches, hence still enabling global relations with much simpler computation. After each pair of blocks, a process called patch merging is performed, where groups of 4 patches are grouped together and added to the channels. Hence taking the first operation, it converts from $\frac{width}{4} \times \frac{height}{4} \times C$ to

$\frac{width}{8} \times \frac{height}{8} \times 2C$ (Liu *et al.*, 2021c). As one may note, these hierarchical feature maps may be easily integrated with existing deep learning techniques, such as with Feature Pyramid networks described in Section 3.2.1.5 (Liu *et al.*, 2022b).

This methodology was further improved upon in Swin Transformer V2 by Liu *et al.* (2022b), which introduced several notable enhancements. Firstly, a novel normalisation technique called Residual-Post-Norm was proposed, which transitions the main components of the Layer Normalisation layer towards the end of each residual unit. This normalisation technique was found to produce milder activation across the network layers, in turn increasing stability during model training. Secondly, the authors introduced a log-spaced continuous position bias, a modification to the position embedding used in the ViT Transformer. This was found to be beneficial, especially when compared to absolute position embedding, by allowing the model to effectively transfer weights from low-resolution pre-trained techniques to higher-resolution applications. Lastly, the paper utilised SimMIM, a self-supervised pre-training method, which effectively bootstraps the model's initial conditions. This technique reduces the need for vast labelled images, making the model more data-efficient, which is particularly useful in scenarios where labelled data is scarce or expensive to obtain. ViT and the associated techniques however were limited to image classification tasks, and by themselves could not produce the localisation results required by object detection.

In parallel to the development of ViT, DETR (Carion *et al.*, 2020) was proposed as a solution specifically targeted towards detection rather than just classification. This model first starts with a convolutional summariser, which essentially decreases the width and height of the image, but increases the amount of channels. These are then fed into the transformer encoder and decoder stages which generate object queries. Object queries may be seen as potential areas of objects within images, which should remind us of region proposals outlined in two-stage convolutional modules. Instead of the $[CLS]$ token such as in the case of ViT, DETR uses a feed-forward network to interpret each object query into common detections (Carion *et al.*, 2020).

Notably, the loss calculation during training is not as straightforward, as the model may output results in different orders than the annotations, and it may also make incorrect classifications or provide inaccurate bounding box locations. To address this challenge, the authors developed a bipartite matching solution. Essentially, each object query consists of a class and a bounding box. The class may represent one of the possible classes in the dataset or a special "no class" token. This leaves us with two sets to match, both of size n, being the object queries and the annotations. If either class is the "no class" token, no loss can be calculated, and it is skipped. If there is a non-matching class but a matching

bounding box, it likely indicates a misclassification, resulting in a loss. Similarly, if there is a mismatch in bounding box locations, it indicates a localisation error, leading to the corresponding loss. However, determining which two items to compare is challenging, as the model does not follow any specific output order. This is where the Hungarian algorithm (Kuhn, 1955) comes into play, finding a one-to-one combination of matches that minimises the loss.

Albeit achieving successful results using an innovative architecture, DETR follows an unideal complex process, which has also seen difficulty in convergence and poor accuracy for smaller objects (Zhu *et al.*, 2021). This issue was later addressed by Deformable DETR (Zhu *et al.*, 2021), which adopts a similar ideology to deformable convolutions, as discussed in Section 3.2.1.4. In this instance, attention is not distributed across the entire image but instead has positional weighting, focusing only on a small set of key sampling points around the main reference point. This approach results in increased efficiency and faster convergence.

Other variations of DETR include DETR with Improved DeNoising Anchor Boxes (DINO) (Zhang *et al.*, 2022), which achieves higher efficiency and object detection performance compared to the original DETR. DINO introduces several improvements, including contrastive denoising training, which enhances the effectiveness of anchor boxes, and mixed query selection to improve anchor box performance. Furthermore, DINO utilises a look-forward-twice scheme, which predicts bounding boxes iteratively, improving the accuracy of box predictions.

Another significant advancement is Real-Time DETR (RT-DETR) (Zhao *et al.*, 2023), which addresses various size constraints associated with classic DETR and its associated techniques. This is done through the implementation of a more efficient transformer architecture and a novel multi-scale feature fusion strategy. The efficient transformer reduces the computational complexity by introducing a lightweight self-attention mechanism, while the multi-scale feature fusion strategy enhances the model's ability to handle objects of varying sizes. Furthermore, RT-DETR incorporates a dynamic inference mechanism, which allows it to adaptively adjust the number of transformer layers based on the complexity of the input, thereby achieving a balance between accuracy and computational efficiency. These improvements not only make RT-DETR more suitable for real-time applications but also demonstrate its potential in pushing the boundaries of object detection in terms of both speed and performance.

3.3.3 Attention-Based Hybrid Techniques

With the remarkable success of transformers in large-scale datasets, it is important to note that their widespread adoption is still predominantly seen in scenarios with a substantial number of training images (Liu *et al.*, 2023, 2021b; Pan *et al.*, 2021). Moreover, due to their resource-intensive nature, training transformer models can be a challenging task, particularly when limited resources are available. As a result, CNNs continue to be a popular choice for lower-resource implementations, often achieving superior performance and efficiency (Liu *et al.*, 2021b)

Inspired by the promising capabilities of transformers and the strengths of CNNs, researchers have turned to exploring hybrid approaches that combine the key factors of both technologies. Particularly, aiming to leverage the attention mechanisms and global context understanding of transformers while retaining the spatial information processing and hierarchical feature extraction capabilities of CNNs, including its efficiency, speed and ease of training.

An example of a hybrid technique is the ACMix Block (Pan *et al.*, 2022) , which effectively combines the strengths of attention and convolution techniques. The self-attention mechanism, known as the Global Attention Mechanism (GAM), effectively captures both channel and spatial aspects of the features, enhancing the significance of cross-dimensional interactions. On the other hand, convolutional layers excel at local feature extraction and are computationally efficient.

3.4 Underwater Object Detection

Underwater object detection is not a new application by any means, with sonar theories dating back to the 1490s, when Leonardo Da Vinci wrote “If you cause your ship to stop and place the head of a long tube in the water and place the outer extremity to your ear, you will hear ships at a great distance from you” (Urick, 1975). This environment of sonar was further developed in the two world wars, where the detection of submarines and landmines called for innovation in underwater detection techniques (Malumbres *et al.*, 2009; Spampinato *et al.*, 2008).

The use of vision was then quickly adopted, through the accessibility of close-range data obtained through AUVs (Foresti and Gentili, 2000). Later on, deep learning saw several applications to UOD, for instance, Marc *et al.* (2016) used a system of HOG + Support Vector Machines for Coral Reef Fish detection. During a similar time, Li *et al.* (2016) applied Fast R-CNN on fish species detection in underwater imagery. This series of events led to a fast-paced subdomain out of object detection emerging, being UOD.

In modern studies, the focus is often on adapting breakthroughs in land object detection to the underwater environment, often with added resilience to noise, and the various difficulties of underwater imagery (Guan *et al.*, 2023).

3.4.1 Underwater Object Detection Datasets

The modern community of UOD mainly stems from the Underwater Robot Picking Contest (URPC), a competition originally planned to incentivise the development of software for marine organism-picking AUVs. The datasets provided for the competition were however found extremely useful for Underwater Object Detection purposes, with classifications for organisms such as starfish, sea urchins, and holothurians amongst others. Competitors would then compete by innovating around existing object detection techniques, to get better detection results on the datasets. As expected, this springboarded the entire domain, where a substantial and commonly accepted dataset was present, and techniques were already being developed.

Datasets such as DUO (Liu *et al.*, 2021a) combine and reannotate the different URPC datasets published every year to form a bigger, publicly available organism dataset. Another popular UOD instance is the Brackish dataset (Pedersen *et al.*, 2019), claiming to be the first publicly available European underwater marine organism detection dataset, based in Limfjorden, a brackish strait in Denmark.

3.4.2 State-of-the-art Underwater Target Detection Techniques

This subsection delves into object detection techniques specifically tailored towards underwater scenery. The main motivation comes from the difficulties posed by subaqueous imagery, where many researchers focused on minimising the detrimental visual effects in underwater scenery (Chen *et al.*, 2020a; Guo *et al.*, 2020; Islam *et al.*, 2020b).

3.4.2.1 YOLOv7-ACMix

In their work, Liu *et al.* (2023) proposed an improved version of the YOLOv7 (Wang *et al.*, 2023) network, referred to as YOLOv7-AC. The primary motivation behind this was to address the limitations of conventional underwater target detection methods in their ability to handle global relations and noisy environments.

A major modification is the implementation of ResNet-ACMix blocks, which is comprised of an attention layer, sandwiched between 2 convolution layers, with a skip connection propagating information over all 3 items. As described in 3.2.1.1, residual or

skip connections are designed to avoid feature information loss and reduce computation. Whereas the hybrid between convolution and attention aims at finding a sweet spot between speed and accuracy as discussed in Section 3.3.3.

Additionally, the authors propose a modified E-ELAN module, called AC-E-ELAN. This block replaces the 3x3 convolution block in the E-ELAN structure with an ACmixBlock module. This modification, along with the incorporation of skip connections and 1x1 convolution architecture between ACmixBlock modules, improves feature extraction and network inference speed. Furthermore, YOLOv7-AC incorporates a Global Attention Mechanism in the backbone and head parts of the model to better leverage the benefits offered by transformers.

Experimental results show that the improved YOLOv7 network outperforms the original YOLOv7 model and other SOTA underwater target detection methods. The proposed network achieved a Mean Average Precision (mAP) value of 89.6% and 97.4% on the URPC dataset and Brackish dataset, respectively, and demonstrated higher frames per second compared to the original YOLOv7 model.

3.4.2.2 SWIN Backbone

Further dealing with the direct problem of underwater object detection techniques through computer vision, Liu *et al.* (2022a) adapted the SWIN transformer backbone specifically for UOD applications. This backbone was integrated within the Faster R-CNN architecture, being a two-stage approach. Additionally, the authors use a path aggregation network to combine shallow and rich features, which allows location information and overall information to be better fused and utilised. Moreover, the authors also add Online Hard Example Mining, which automatically selects harder examples for training to boost performance in such areas. Overall, the authors were able to exceed SOTA accuracy, where the technique was even able to detect originally unlabelled objects. The authors also note the limited speed of transformers, and for their future, the authors will target the use of single-stage detection models for speed optimisations, rather than two-stage.

Another successful implementation was proposed by Lei *et al.* (2022), where they used the SWIN backbone on the YOLOv5 Model. Furthermore, the authors focused on improving the multi-scale feature fusion of networks, which the authors claim allows the neural models to make better use of extracted features from the backbone as we discussed in Section 3.1.3. The authors were successful in exceeding general target detection model performances on the UPRC dataset. In their conclusions, the authors further advocate the importance of feature enhancement techniques as they may help deal with the complexities of underwater images. In terms of drawbacks, the authors found that misclassification

and false detections were still present, especially in complex scenery. Moreover, although satisfying real-time requirements, the model was still relatively large and took 45 hours to train.

3.4.2.3 SWIPEnet

Swipenet was proposed as a novel neural network, targeting small object detection in noisy underwater environments by Chen *et al.* (2022). It is of particular interest due to its classical approach towards architecture, being based on a heavily modified version of VGG16, now an older architecture, yet still seeing impressive results. The first modification was the usage of dilated convolutions, which we know increase the receptive field, and are able to capture more context around smaller objects. This has been covered in more depth in Subsection 3.2.1.3. Secondly, a custom loss function was proposed, named sample-weighted loss, which assigns a difficulty weighting to images. Chen *et al.* (2022) claim that for their small underwater object detection case, harder examples are likely noisier, and hence decrease the learning weighting to focus training on clearer examples. Finally, the authors propose Curriculum Multi-Class Adaboost, which is an innovative training paradigm, centred around the separation of noisy elements into their own detectors. This starts off by training a clean model, without any influence of noise, and then train further detectors to handle diverse noisy data, combining them into a robust ensemble.

3.4.3 Archaeological Object Detection

This section reviews related literature on archaeological object detection, addressing similar challenges in underwater environments and the complexities posed by archaeological artefacts. By examining these studies, we gain insights into the methodologies and approaches used to overcome these specific difficulties.

A similar system to our requirements was presented by Paraskevas *et al.* (2023), in which they compare several sizes of YOLOv8 detecting underwater pottery from a shipwreck off Modi Island, Greece. The authors found that the *small* variant of YOLOv8 performed best with a 75.5% mAP. Larger variants of the model were found to overfit, especially with limited data sizes. The technique was further found suitable for real-time detection and can be applied to several scenarios such as live detection for AUVs.

In another study, Yang *et al.* (2023) present a custom feature extraction network, named MDLA-DCN, specifically designed for AUV exploration. Leveraging deformable convolutions and the BAM attention module, the model adeptly extracts and optimises feature information from complex underwater scenes, achieving 92.8% mAP versus the

YOLOv7 baseline which achieved 89.9% mAP. The authors base these results on an original dataset covering underwater cultural artefacts, which is not publicly available.

Al-anni and Drap (2024) recently presented a similar system for 3D segmentation of underwater archaeological artefacts, particularly for outlining amphorae in the vicinity of a Phoenician shipwreck Xlendi, Gozo, Malta. The authors use YOLOv4 as the object detection technique, trained on thousands of automatically annotated imagery. The authors claim around 85-87% single-class performance, yet the metric used was not specified. Additionally, despite having a photogrammetric model and an orthophoto of the area, the detections were still carried out on the raw imagery. Notably, their presented work was heavily focused on the archaeological implications of techniques, and laid important groundwork for future work, especially for a more detailed analysis surrounding the vision techniques applied. There are also limitations associated with automated annotation, as it has the potential to leave difficult cases out. Although not specified, from the figure presented, the authors also focused on whole, or more visible pieces within the imagery rather than shards and buried artefacts, which is another potential area of exploration.

Separately, in tackling the general infrastructure of compiling a dataset, and deploying a recognition tool, the *Archaeological Automatic Interpretation and Documentation of cEramics* (ArchAIDE) project (Anichini and Gattiglia, 2018), develops an end-to-end system for the automated recognition and document of imagery from a smartphone. The process involves a series of steps, the first one being preprocessing, where a pipeline of processes for quality enhancement and noise reduction is applied. This includes classic augmentation techniques such as image resizing, and Colour Correction (CC), all helping to maximise the potential to recognise objects. Next, computer vision algorithms extract relevant visual features from the images, which include shape descriptors, colour histograms, texture patterns, and other characteristics that capture the distinct visual properties of archaeological ceramics. The extracted features are then used as input to clustering algorithms. These algorithms are trained on a comprehensive database of labelled archaeological ceramics, where each sample is associated with its corresponding class or category. Once the classification is performed, the results are interpreted and documented. This can include information such as the type of ceramic, its cultural origin, its historical period, and any associated metadata. The documentation provides valuable insights for archaeological analysis and research. By utilising computer vision and machine learning, the ArchAIDE project automates the process of recognising and interpreting archaeological ceramics. It enables the efficient analysis of large quantities of ceramics, reduces manual labour, and enhances the robustness of the interpretations, all presented in an intuitive smartphone app, bringing ease of understanding to non-technical individuals.

3.5 Depth and Saliency Enhanced Detection

This section is dedicated to studies which focus on the improvement of object detection performance through the use of visual saliency estimation or depth maps. The literature surrounding the fusion of these techniques was not found to be as mainstream, calling for further research on how and where to leverage these auxiliary maps to improve detection.

3.5.1 Depth Enhancement

In the context of depth enhancement, the integration of depth information with traditional RGB images has shown significant promise in various applications. Using modern hardware such as Kinect or stereo camera setups, obtaining depth has been greatly facilitated, making such studies more feasible (Ophoff *et al.*, 2019). Apart from additional hardware requirements, photogrammetric 3D surface models may be leveraged, enabling the estimation of high accuracy depth maps.

Ophoff *et al.* (2019) set out to investigate whether adding depth information helps SOTA object detectors perform better and further investigated several methods of fusion. The authors deeply integrated the depth, where they concluded that adding the depth layers towards the mid-late layers provides the best fusion results. Based on this, they found that the fusion significantly improves detection results when compared to the traditional YOLOv2 model.

In another study, Schwarz *et al.* (2015) focused on retaining the transfer learning potential of pretrained CNNs by employing a multi-step encoding process. This approach preserved the three colour channels in the image while emphasising the main Region of Interest (ROI) based on distance-from-centre calculations. Their findings highlighted the efficacy of this method in enhancing object classification, pose-estimation and instance segmentation. In contrast, Xu *et al.* (2017) introduced the Horizontal disparity, Height above ground, and Angle with gravity (HHA) feature encoding structure, providing a 3D representation of depth features in an image. Given the abundance of associated features, the authors proposed a novel architecture with three main components. First, they employed a standard RGB detection model based on AlexNet. Second, they used a Geometrical Feature Depth-Specific Detection Network that operated solely on the HHA channels. Finally, these two networks were merged using a Correlated Detection Network, which aggregated information from both modalities. Each network had its own heads and Region Proposal Network (RPN) models, allowing for individual evaluation. The authors reported consistent improvements in overall detection accuracy, though performance varied across different dataset categories.

3.5.2 Saliency Estimation

Visual Saliency refers to a concept in neuroscience, which refers to regions that attract an observer's attention, tending to the question "Where to look?" (Islam *et al.*, 2020a; Reggiannini and Moroni, 2020). Computationally, this is defined as the degree of distinctiveness of an object or a region from its surroundings. Efforts to estimate this mathematically have proven to be an interesting problem (Itti *et al.*, 1998; Koch and Ullman, 1985), where in recent years, machine learning-based approaches have emerged. This category of techniques aims to estimate the most prominent areas within mages in an efficient and accurate manner.

In the context of object detection, saliency maps can be used to identify the regions of an image that are most likely to contain objects, which can guide the object detection model to focus its attention on said regions (Islam *et al.*, 2020a). The opposite also holds where saliency maps can be used to filter out regions of an image that are unlikely to contain objects, which can help to reduce false positives (Cane and Ferryman, 2016). Saliency maps may either be used as a preprocessing step or may also be integrated within the models as an additional input channel, which would guide the object detection model into distinct areas.

The concept of saliency maps was first conceived by Koch and Ullman (1985), in which they studied the usage of saliency maps using computational models of visual attention. A saliency map is a topographic representation of visual saliency across an image, where regions with high saliency are more likely to attract attention. The authors proposed that these saliency maps are generated through a series of neural computations in the visual system, reflecting the significance of different image features for guiding attention. This later inspired the bottom-up saliency method as proposed by Itti *et al.* (1998). This method was based on 42 underlying feature maps, such as colour intensity and orientation, all of which the authors attributed to affect object saliency.

After years of iterative buildup on neural saliency estimation, Deepgaze (Patacchiola and Cangelosi, 2017) proved to be a significant leap forward to visual saliency estimation. Unlike traditional methods, Deepgaze leverages deep neural networks and uniquely previous fixation history. By considering temporal context, this technique can predict where an observer is likely to fixate next.

A case of saliency map augmentation for object detection performance was applied by Katyal *et al.* (2018) which particularly employed visual saliency techniques for foggy conditions. The authors use the CovSal (Erdem and Erdem, 2013) saliency estimation, a technique which aims to identify the centre of attention within an image mimicking human attention by considering colour, texture and spatial distribution. The results of

detection from the saliency map were then merged with YOLO results through a hand-crafted aggregation technique, based on the likelihood of detections appearing in both the YOLO version and the Saliency version.

3.5.3 Salient Object Detection

In recent years, the field of Salient Object Detection (SOD) has witnessed significant growth, with a focus on developing techniques to accurately localise and segment primary objects within images. These are particularly interesting due to their zero-shot approaches, where the saliency estimation techniques are not specifically trained towards the domain at hand but are trained towards extracting the main parts of images across several domains without retraining.

Early methods primarily employed contour detection based on the saliency maps themselves, representing a foundational approach in the field. While effective in many cases, these methods encountered challenges in managing noise levels and achieving consistent performance across diverse datasets, whilst still being quite naive in their detection methods. Such a technique was proposed by Tian *et al.* (2008), where an unsupervised object detection technique was implemented for time-sequenced images, based on Itti saliency estimations, and a series of postprocessing techniques to extract the final set of predictions.

These techniques have come a long way, however, and modern techniques such as the InSPyReNet framework were proposed by Kim *et al.* (2022). These are specifically designed for efficient salient object detection in high-resolution images and saliency map outputs. InSPyReNet specifically, leverages an image pyramid architecture, allowing for rapid processing while maintaining accuracy. The framework compares two distinct backbones being Res2Net and the Swin Transformer, providing flexibility and performance optimisation.

An interesting extension to salient object detection is also Saliency Ranking, which possesses the additional capability of ordering regions based on their visual significance within an image. One such algorithm was proposed by Seychell and Debono (2018) in their SARA algorithm. This method works by employing a grid pattern on the image, which is then used to segment the image into multiple regions. Each region is analysed individually, with a saliency score calculated based on factors such as the entropy score of the saliency map, proximity to the centre of the image, and depth score. The regions are then ranked according to these scores, providing a hierarchy of salient regions. This method offers a more comprehensive understanding of an image by not only identifying the salient regions but also determining their order of importance. This makes saliency

ranking particularly useful for multi-object images and complex scene understanding.

These advancements represent the potential strengths to be harnessed through enhancing the accuracy, efficiency, and robustness of salient object detection techniques, paving the way for applications in diverse domains such as image understanding, visual content analysis, and human-computer interaction.

3.6 3D Object Localisation

Object localisation deals with the process of finding the location of a detected object within a bigger setting, in turn gaining an additional layer of interpretability and tangible utility to detected results. A common starting point is often an orthomosaic, which through the use of GCPs can be geographically calibrated. However, object detection techniques tend to perform poorly on massive images, not only due to the unfeasible memory requirements associated but also to the relatively small size of objects within an orthomosaic (Akyon *et al.*, 2022).

Božić-Štulić *et al.* (2018) approach this by splitting the orthomosaic into smaller chunks, performing the detection on the chunks, and then concatenating the results back once completed. For their application, the authors make use of a UAV to collect aerial imagery, which is then processed to an orthomosaic using several of the photogrammetric techniques discussed in Section 2.2. The authors found that the collection through a UAV poses several benefits to ease of operation, whereas manual equipment would take considerably more effort to deploy on such a large area. They did however note the limitations in both quality and the attention required to cover the area. The GCP density was not found to be adequate to establish geographic reconstruction of objects, which limits the potential understanding of the area as well. The main strength lies in the potential monitoring, where a good quality orthomosaic may even be created on the fly for search and rescue operations.

However, splitting the image arbitrarily into chunks could pose several variables, such as relevant objects being on the border, decreasing the potential accuracy. Slicing Aided Hyper Inference (SAHI) (Akyon *et al.*, 2022), proposes a generic solution to the problem, where instead of naively splitting a larger image, a sliding window approach is used. This approach involves systematically moving a fixed-size window across the image, processing each window individually. Dealing with the border problem, a ratio of overlap is introduced, where the results are then merged together with Intersection Over Union (IoU) Thresholding to remove any duplicates.

Nevertheless, the amount and quality of data available is still decreased when working

on an orthomosaic, which adds complications due to holes, distortion, and irregular lighting amongst many more. One also loses on the oblique nature of photos, which in some cases, such as needing to differentiate whether an item is buried or not can be a useful feature. For this reason, researchers may opt to do the object detection directly on the raw images, and then leverage photogrammetric techniques to project the information to 3D where necessary. One such example leveraging the camera pose estimation was presented by Zhao *et al.* (2022), which performed the object detection on the raw images of the dataset, and then, using the camera location and rotation information, the points are projected onto the orthomosaic, through a similar process described in Section 2.3. Such a technique enables the strengths in quality and quantity of data by training on the raw images, whilst still projecting to the the photogrammetric model for interpretibility and localisation purposes.

A similar approach was presented by Gené-Mola *et al.* (2020), which perform 3D instance segmentation for agricultural object detection. The authors leverage structure form motion techniques to build a photogrammetric model of their area, which allows for positional estimation of imagery. From there, Mask R-CNN is employed for 2D instance segmentation on cropped subsections of the entire image, which is then projected back to the orthomodel, obtaining 3D positional segmentation of apples. The authors are also presented with duplication errors due to overlapping imagery, which was handled through the use of IoU thresholding, meaning that consecutive detections which had an IoU overlap higher than 50% are unified, strengthening the angles of detection. The metrics present around 86% Average precision for single-class segmentation, which the authors claim to perform well in comparison to state-of-the-art 3D fruit location systems.

Furthermore, in a recent study by Al-anni and Drap (2024) a methodology for efficient 3D segmentation through the fusion of traditional object detection techniques and photogrammetric tracking is presented. The authors present their work on the Phoenician Shipwreck off the close of Xlendi, Gozo, which is within the vicinity of our working area yet has completely different morphology and conditions. The Phoenician Shipwreck spans 12x5 metres, making a dense patch of land with a visible contour of a shipwreck. The authors effectively employ a 2D-3D photogrammetric link to project several points. The link was used to project several points from object detection forming a sparse segmentation mesh. All in all, this paper lays fundamental groundwork for proving the diverse applicability of photogrammetric fusion, especially in maritime archaeology settings.

3.7 Literature Review Summary

This chapter summarises several aspects of surrounding literature, whilst serving as a checkpoint of the latest advancements in the targeted domains. The first domain is object detection, where the main components of a modern object detector are discussed. Based on this, SOTA techniques and their original contributions are analysed, establishing a direction of literature towards faster, and more accurate architectures. This further includes the latest advancements in CNNs and transformers, which have shaken up the domain direction. Next, this chapter outlines how advancements are transferred to underwater scenery, forming the downstream task of Underwater Object detection and the culture around it. Lastly, object localisation studies are discussed, which effectively leverage 3D photogrammetry to improve the interpretability of computer vision results.

4 Methodology

An experimentation framework is proposed which targets our main set of objectives. The techniques applied are based on findings from surrounding literature, which we apply and improve to better tackle our specific dataset. This chapter outlines the setup, individual methods and evaluation metrics chosen. This includes a high-level description of the chosen architectures and techniques, and how the pieces fall together in the architecture. Then, the steps taken during the data compilation stage are discussed, where a dataset suitable for object detection is assembled. Next, the initial object detection comparison is outlined, where several promising techniques described in Chapter 3 are applied to our domain. The best-performing subset of models was then used to analyse the effects of saliency and depth, including an exhaustive search over the different channel and merging configurations. Finally, the best-performing configurations are chosen for localisation experiments, where through photogrammetric projection, each detection is linked to geographic coordinates, and methods were explored on aggregating these into a single orthomosaic projection.

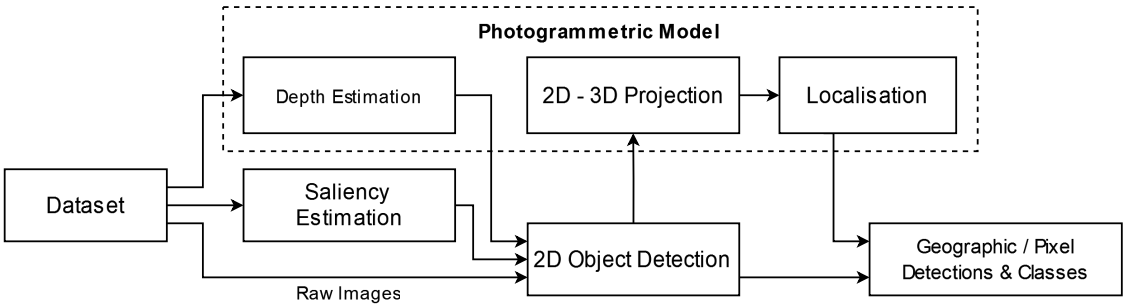


Figure 4.1: Top Level Architecture displaying the flow of data through the distinct processes. This highlights the bidirectional link to the photogrammetric model and the added functionality it grants.

4.1 System Architecture

An overview of the architecture is presented in Figure 4.1. Starting from the work of our first objective, a multi-class underwater object detection dataset forms the ground truth data to be utilised within the rest of the processes. These processes include comparing several object detection architectures, aiming to scientifically evaluate the most suitable model for our specific use cases, as targeted in our second objective. This was further extended in our third objective, where several fusion techniques for depth and saliency maps were explored, aiming to best integrate these with the original imagery, whilst minimising potential information loss. The overarching aim of photogrammetric fusion in our work is also clearly highlighted, displaying the bidirectional utility, first used to estimate depth maps from the original imagery. Following a similar ideology to the work presented by Gené-Mola *et al.* (2020) and Al-anni and Drap (2024), the photogrammetric link is again used to project the bounding boxes from pixel positions to 3D model coordinates. These results were finally utilised in a multi-step aggregation procedure, which displays the results in a single orthomosaic of the entire working area.

4.1.1 Experimental Setup

The majority of the tests were performed on the CCE Networks Lab Cluster at the University of Malta, which has the following configurations available:

$7\times$: 6 Cores $\times$ 2 threads - 20GiB RAM - RTX 3060Ti 8GiB GPU

$6\times$: 6 Cores $\times$ 2 threads - 20GiB RAM - RTX 3060 12GiB GPU

The availability of nodes heavily fluctuated according to system demand, yet typically multiple nodes were leveraged in order to run multiple experiments in parallel. The 8GB VRAM GPUs can only handle real-time techniques with low batch sizes, whilst the 12GB versions may be used for larger batch sizes, yet still not enough to run larger transformer techniques (Koot *et al.*, 2021). These being mid-range desktop GPUs also relates to real-world use cases for such techniques, where end users would not require enterprise-level hardware to achieve these results.

In terms of software, PyTorch[1] served as a primary tool for model development in this study, owing to its versatility and widespread adoption within the research community. This was further extended through the use of the Ultralytics[2] library, which easily allows the training of several state-of-the-art object detection methods and training setups such as augmentation, parameter tuning and evaluation. Furthermore, the PaddleDetec-

[1] https://pytorch.org/

[2] https://github.com/ultralytics/ultralytics

tion[3] library introduced a range of cutting-edge models, which further facilitated the experimentation process surrounding transformer-based techniques. For photogrammetric modelling, Agisoft Metashape and the respective SDK (AgiSoft LLC., 2023) were used, which proved to be particularly helpful for accessing 3D model information programmatically for custom exports and projection operations.

4.2 O1 - Dataset Compilation

The first objective, **O1**, deals with the compilation of a suitable object detection dataset, enabling the application of object detection models in our working area.

The wreck area covers a considerable area, estimated to span over 67,000m^2, with depths varying from around 100m to a maximum depth of approximately 110m below sea level. Such intricate imagery takes significant effort to interpret and manually annotate, meaning the first step was to identify a suitable subset of the area to be used as the dataset. Working on a pilot patch of land allows us to carry out a preliminary understanding of how the techniques perform. Focusing on the exploratory aspect, an area with diverse landscapes was chosen, covering areas with sparse objects, dense objects, and even rocky outcrops, all having varying difficulties and characteristics. This allows the model to be both trained and evaluated on a dataset that is representative of the bigger picture, aiming to get a realistic subset of how such a technique would perform in real settings. Image collection sessions typically cover 625m^2 per dive; hence it was decided to take an entire trip and use it as our set of images. This left us with a total of 864 images to work with as our working dataset.

The Agisoft photogrammetry software was used by the Classics and Archaeology department at the University of Malta, which produced an orthomodel of the area using the SfM process described in Section 2.2. This meant we could also leverage the Agisoft SDK to facilitate processes such as 2D-3D projection, depth map generation, and geographic operations, among others.

The next phase was the annotation process of the imagery. This is a crucial part of the work, as this turns simple photographs into annotated, usable information to train and evaluate models upon, effectively setting targets for supervised models to train towards. This was carried out in two primary stages. The first was dealing with the single-class annotation of anthropogenic objects within the work area. For this use case, six annotators with the proper domain knowledge participated in the annotation process, where the images were shuffled and equally split, forming six groups of 144 images each. A sec-

[3] https://github.com/PaddlePaddle/PaddleDetection

ondary pass was then carried out, where a field expert individually classified and refined the initial annotations, producing a multi-class archaeological object detection dataset.

The next phase dealt with the splitting of data into distinct sets for testing, training, and evaluation. A ratio of 80% training, 10% testing, and 10% validation was chosen, where 80% of the images were visible to the model for training, whilst keeping the rest hidden for metric calculation. Separately, a five-fold validation dataset was prepared, which splits the entire image set into five distinct chunks, where each fold is used as a test set exactly once. The images are shuffled before splitting, ensuring that diverse areas are captured in all sets. These choices were used as a domain standard, commonly adopted to ensure proper training, testing, and validation during the entire process.

4.2.1 Areas in the Dataset

There are several factors associated with each of the area types of data, which may give us insight and explanation as to how a detection model may perform on them. Sparser areas have easily distinguishable amphorae, as the lack of surrounding noise will likely enable the detector to confidently detect these. Artefacts in these areas also tend to be whole or buried, which may give an easier time for a detector to pick up on. Such an area may be seen near the bottom-middle of Figure 2.2. Denser areas without rocks, such as the top centre, may pose some marginally increased difficulty due to the tendency for overlap within detections. There may also be debris and other material that rests between artefacts, which may also be detrimental to performance. Finally, areas with rocky outcrops are most likely to be difficult to discern, which proves difficult even for human annotators. Such an area may be seen in the top left, where a rocky area with densely located artefacts may be noted. Not only do the rocks themselves tend to look similar to broken pieces of amphorae, but the densely packed number of amphorae may also interfere with the detection performance. Rocky outcrops also tend to act as shelters for other debris, ecology, and organisms, which all may be considered noise to the object detector.

4.2.2 Classes in the Dataset

The amphorae in the dataset were split into three main categories according to their level of preservation, namely whole, buried, and broken.

Whole amphorae are those which are visibly whole, with most of their structure visible and visibly intact. As a premonition, these tend to be very distinguishable. They typically have a larger size due to not being submerged and have well-defined shapes. These should

be expected to perform quite well, as their characteristics are very distinct from the other classes. Accurate detection of such artefacts is also considerably more important, as these are the most likely to be eligible for further analysis and even more detailed identification in downstream tasks.

Broken amphorae are those which are visibly broken. These need to have clear missing chunks, or even be shards themselves, to the point where the annotator can tell that it is not a whole artefact due to sharp irregular seams. This is quite a broad class, capturing anything from whole amphorae with a missing chunk to minuscule shards of pottery that are further buried in the sediment. This may be difficult for a detection model to consistently discern, whilst further being difficult due to the smaller nature of shards.

Buried artefacts are those which are buried in the sediment, where an assumption as to whether they are whole or broken cannot be made. Given some of the amphorae are mostly sedimented, with only a patch being visible, one cannot be sure whether this is broken underneath or is fully intact. Such artefacts may not always be easy to discern from whole ones, due to the round characteristics with no visible breaks, but may also be assimilated to broken pieces due to their smaller visible footprint, which may resemble broken shards of amphorae. This would especially be problematic if the detection was to be done on the orthomosaic itself, due to the uniform top-down perspective. However, the oblique nature of the raw imagery tends to better retain the depth information.

In general, the classes serve separate purposes, with defined rules for what makes up each one. However, given the difficulty in underwater scenery, it is often difficult even for domain experts to determine the properties. This is especially the case due to the ambiguous nature of buried and degraded artefacts, the non-ideal lighting conditions, and even the fuzzy definitions between something being completely whole and something being completely broken.

4.2.2.1 Distribution

In our case, we are dealing with considerably unbalanced class counts per instance. Knowing the distribution of data we are working with will allow us to better interpret the results later on, as this serves as a guide to which classes are expected to perform better due to increased class distributions or which classes contribute more to the average result. The numbers are presented in Table 4.1, where we may see that Whole has the most count, representing nearly half the entire distribution of the annotations. Given these are the easiest to identify and also amongst the most valuable to identify, having many of these is not necessarily a negative sign. These are closely followed by broken artefacts, which is also understandable. Due to shards being smaller yet numerous, it is unsurprising that

Table 4.1: Distribution of Class Annotations within primary splits.

Class	Training	Validation	Testing	**Total**
Whole	5,413	719	572	6,704
Buried	2,083	221	233	2,537
Broken	4,002	623	387	5,012
Total	11,498	1,563	1,192	14,253

a large quantity of these would be present. Finally, buried artefacts are the least common by a considerable margin, which will be interesting to see the detection performance on cases that don't necessarily fit into the other classes with confidence whilst being under-represented.

4.3 O2 - Initial Object Detector Comparison

Targeting the second objective **O2**, an exhaustive comparison was performed on several state-of-the-art techniques, commonly used within the object detection domain. The main purpose behind this analysis was to systematically evaluate how different architectures handle complications introduced in UOD, and the characteristics of our specific dataset conditions. To ensure a fair comparison, these were all trained and evaluated on the same dataset, and where possible, exterior factors were eliminated from the evaluation by keeping consistent configurations and experimental setups. This comparison targets three primary sets of detectors, namely, single-shot CNN detection techniques from the YOLO family, two-stage techniques based on Faster R-CNN and finally transformer techniques. This section first details the reasoning behind the architectures compared. Then, the several configurations for each model are outlined, which add up to a considerable search space. Finally, the evaluatory ranking procedure is outlined, which is used to reliably choose which configurations are kept for further analysis and which ones are considered sub-optimal.

4.3.1 Architectures Evaluated

Our analysis began with YOLOv5, a well-established model in the domain, serving as a basis for several other techniques and UOD-specific models (Jocher, 2020). Furthermore, we explore YOLOv7 (Wang *et al.*, 2023) as targeted by Yang *et al.* (2023) as their base model, which the authors found to perform well for underwater archaeological object detection. Additionally, YOLOv8 (Jocher *et al.*, 2023) was targeted as carried out by

Paraskevas *et al.* (2023), in which they found the smaller variant to perform better than larger ones due to their small dataset, a limitation we also share in our case. Additionally, we further evaluate the performance of YOLOv9 (Wang *et al.*, 2024), the cutting edge from the YOLO family of models, promising several advancements and performance on the COCO dataset. The comparison of a generation of techniques was chosen to provide insight into the relation between generic on-land advancements in CNN architectures and how these benefits transfer to underwater archaeology scenarios.

Apart from single-shot techniques, two variations of Faster R-CNN (Ren *et al.*, 2015) were compared, particularly ones using ResNet50, and SWIN-Tiny as backbones. The original Faster R-CNN backbone was based on VGG-16, which over the years has been improved upon time and time again. ReseNet50 was particularly influential on the development of Faster R-CNN as they introduced skip connections and residual blocks, enabling a deeper and more efficient feature extraction compared to VGG-16. On the other hand, SWIN-Tiny is a recent variant of the Swin Transformer architecture, which has garnered attention for its effectiveness in handling object detection tasks with fewer computational resources. By evaluating these variations of Faster R-CNN, we aim to assess the impact of backbone selection on the model's performance and efficiency.

Continuing the transition towards transformer models, our comparison extends to the DETR series, encompassing DETR (Carion *et al.*, 2020), Deformable DETR (Zhu *et al.*, 2021), and RT-DETR (Zhao *et al.*, 2023). Each modification of the architecture presents unique advancements and considerations. DETR is particularly influential as one of the earliest papers claiming to provide usable and accurate transformer performance for object detection tasks. Given the claims of Deformable DETR's improvements upon the original, this was also included, aiming to analyse the speed and convergence, which was deemed problematic with the original variant. Finally, given the large size of such models, especially compared with smaller single-shot CNN architectures, RT-DETR was implemented, aiming to analyse how the author's claims of exceeding YOLO performance and accuracy using transformers.

4.3.2 Configuration and Search Parameters

Each experiment performed may take several configurations which affect how the model trains and eventually performs. The first stage of configurations included searches over pretrained weights, colour correction and model sizes, all of which were deemed interesting points of exploration for the resultant performance. Additionally, several parameters were kept constant to minimise their effects on the variability of results.

These constant parameters included batch size, where this was kept at a value of 2.

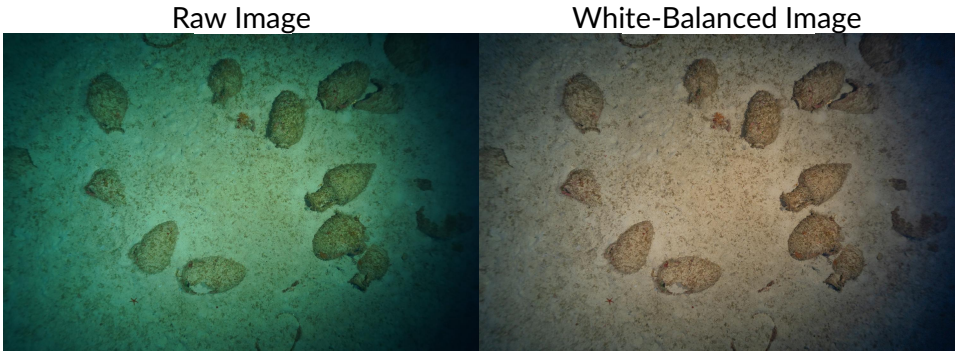


Figure 4.2: Raw and Colour Corrected Imagery Comparison, displaying the more normalised look achieved through white balancing.

This batch size was found to accommodate the larger models on the limited VRAM size of our GPUs, hence by keeping this constant across the board, variations between model configurations and GPUs could be minimised. Another consistent parameter was the image input resolution, which was kept at a constant, letter-boxed 1024×1024. This was chosen as a middle ground between retaining high image fidelity, whilst keeping modest memory requirements to fit all models, which grow exponentially by size. Moreover, epochs for YOLO models were kept at a constant 250 as discussed in similar literature. Faster R-CNN was set to 300 epochs, whilst transformer techniques, given their more data-hungry approach were trained at 500 epochs. Notably, early stopping was also enabled, meaning that after a set of consecutive stagnant results, the training was terminated.

In terms of searchable parameters, the pretrained configuration is a boolean value dictating whether to use transfer learning from the COCO pretrained version of the detection models and fine-tune our dataset, or whether to start from scratch using random weights. Pretrained models typically converge faster and may pose several benefits from being pretrained on a much vaster dataset, however, it may also cause the model to converge to local minima, missing important concepts which are specific to UOD.

Moreover, the colour correction parameter was also explored, which is a boolean value which determines whether to use a colour-corrected version of the dataset or not. This colour correction is minimal, being based on white balance correction aimed at reversing colour deviations on the colour card, described in Section 2.4. An example is presented in Figure 4.2 displaying an image before (left) and after colour correction (right). As may be seen, the image looks considerably more normalised, and most of the blue-green hue is counteracted. This is far from perfect, as the non-linear effects of underwater imagery mean the accuracy of the correction will vary by distance from the camera and lighting.

Yet given the typically equidistant image collection process utilised in our dataset this was not found to be problematic. This comparison aims to determine whether this balancing process provides any benefit to detectors, or whether it's only adapted towards a normalised look.

Finally, the different scaling sizes of YOLO and RT-DETR models were searched, which produce an internally different-sized model. YOLOv5 and YOLOv8 were tested on all available sizes, namely n, s, m, l, and x, which further unlocks an understanding of how predictive performance scales with the model sizes. YOLOv7 was tested on its tiny, normal and x scales, which are also the core sizes commonly used. YOLOv9, being a cutting-edge model, at the time of writing has not released the core models, so the larger variants C and E are being tested, which still serves as good grounds for comparison to other larger configurations. Moving on to transformer-based techniques, DETR For RT-DETR, the HGNetv2-l and HGNetv2-x sizes were tested, which are adapted for its faster, real-time approach. Finally, Faster R-CNN was tested with the Swin-Tiny and ResNet50 backbones, which allows for a varied comparison of how two-stage detector performance changes between transformer-based and convolution-based backbones.

A large selection of models was evaluated for this study, aiming to get firm results on how the performance of generic object detectors on mainstream benchmark land datasets translates to other downstream tasks. This step also serves as a shortlisting stage, where the least performant models are omitted from future experimentation, whilst promising ones are further analysed under K-Fold Cross-validation.

4.3.3 Cross-validation

The process of training an AI model has several factors of randomness which may alter its resultant performance. This is especially problematic in our case, where our comparison involves several architectures achieving similar scores, which require reliable performance ranking. For this reason, it's important to not rely on just holdout set results, and opt for a more rigorous evaluation, which involves testing separate parts of the dataset.

K-fold validation allows for the assessment of model stability and the impact of the training data's variability on the model's performance. It also mitigates the risk of overfitting, as the model must prove its effectiveness across multiple, varied training sets. Notably, this also implies training the same model k individual times, which drastically increases the number of total models to be trained, especially when combinations of other individual parameters are tested. For our use case, a k of 5 was chosen, which is standard in the domain, finding a balance between rigorous testing and the increase in training configurations.

With this in mind, this process was used as a confirmation tool, where the first set of the configurations was eliminated by evaluating the primary test set, whilst the highest-performing techniques were then evaluated using fold-based analyses, and averaged to get a more reliable metric. This choice was made with the aim of further reducing the total overhead of fold validation whilst using it when necessary.

4.4 O3 - Saliency and Depth Map Fusion

Relating to the third Objective **O3**, this research analyses the potential effects of integrating saliency and depth maps within our data as an enrichment step. This process supplements the detectors with auxiliary information, aiming to maximise performance. An overview of the experimentation framework for map fusion is displayed in Figure 4.3, which outlines the flow of data, starting from the methods for the map generation, the merging techniques to encode these within existing channels, and the resultant image sets. The primary parts of this process will be discussed in detail in the following sections.

4.4.1 Depth Estimation

Depth maps are visual representations that provide information about the distance of objects from the camera. These maps are essential in understanding the spatial structure of a scene, enabling a model to consider how far away different objects are. This is also useful in cases with uneven lighting, where harsh shadows may be hard to discern from objects using normal imagery, but depth maps tend to provide more accurate silhouettes (Ophoff *et al.*, 2019).

In our case, having monocular imagery, the depth map was calculated from the Digital Elevation Model (DEM), established during the stereo-matching phase of photogrammetry. The DEM provides an estimated surface level for the entire area of the model. Separately, during the bundle adjustment phase, the estimated camera positions are also refined, which can be used to emulate the original perspective of the cameras. Through the combination of these, a picture pointing at the DEM through the perspective of each estimated camera position is taken. This leaves us with a 1:1 matching depth map with the original images, which is used as an additional depth channel.

This depth map is especially useful, as it harnesses the utility of photogrammetry for the inputs of the detection algorithms, which up to our knowledge has not yet been explored. Furthermore, unlike saliency estimation, depth is completely new information, as

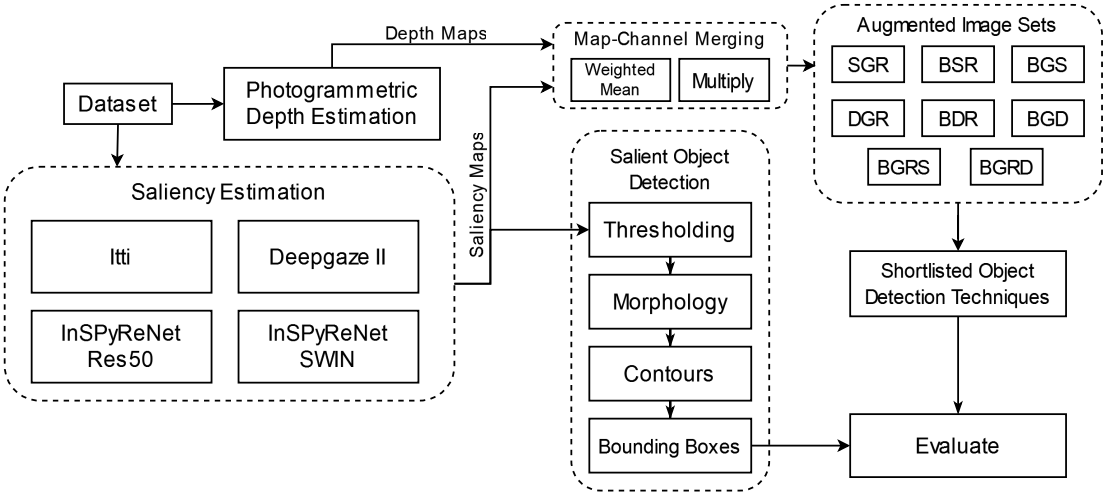


Figure 4.3: Saliency and Depth Enrichment Architecture. This displays the role of several estimation, fusion and processing techniques, displaying candidate configurations for comparison in dashed boxes.

it is based on the estimated structure in its entirety including other images, which would not be captured in any one single instance.

4.4.2 Visual Saliency Estimation Techniques

Visual saliency maps can effectively capture points of particular attention within an image, highlighting regions that are most likely to attract the human eye. As seen from the literature review, although salient object detection techniques are quite prevalent, it is not as common to fuse saliency maps with input data to be used by generic object detectors. The theory suggests that precomputed saliency maps may guide the detector technique to focus on specific parts of the image. This might also turn out to have an opposing effect, as the model might overly concentrate on salient aspects, potentially neglecting non-salient yet relevant items. For our purposes, four primary visual saliency techniques were selected and compared, being Itti (Itti *et al.*, 1998), Deepgaze (Patacchiola and Cangelosi, 2017), and two versions of InSPyReNet (Kim *et al.*, 2022).

Itti's technique (Itti *et al.*, 1998) employs a mathematical approach to saliency estimation, where it tries to calculate saliency based on modelled human preferences. These low-level features include intensity, colour, and orientation, providing a foundational perspective on visual attention. While this technique is less computationally intensive than running deep learning models, it may not capture complex visual patterns as effectively as modern methods. Nevertheless, it provides an interesting basis for comparison.

In contrast, Deepgaze (Patacchiola and Cangelosi, 2017) utilises deep learning to model visual saliency, aiming to get a more learnable approach to modelling. This approach typically offers improved performance over traditional methods, though it requires substantial computational resources, and was found not to have as high of a resolution. Another advanced method is InSPyReNet (Kim *et al.*, 2022), which is based on the Swin Transformer and Res2Net50 backbones. Since this technique is targeted towards SOD, the generated saliency maps are sharply focused on the primary objects of scenes. However, it also demands considerably higher processing power for inference than other methods, increasing the overhead of detection operations.

4.4.3 Salient Object Detection

Salient object detection deals with identifying important areas within a visual scene. This process is particularly useful due to its zero-shot nature, where new scenes do not require specific training, allowing for immediate application across diverse environments. As a preliminary test to determine whether the saliency results are relevant to object detection techniques, a baseline Salient Object Detection approach is adopted. The primary stages of this process are outlined below.

Saliency Map Generation: The process starts by generating saliency maps for each of the images in the test. This process is also repeated for each individual visual saliency estimation technique, aiming to evaluate how each one performs when used independently.

Thresholding: The next step involves applying thresholding to the saliency maps. This step aims to eliminate noisier, low-importance elements from the map, which could otherwise interfere with the detection process. Several threshold values are experimented with to find the optimal one for each technique. This leaves us with a saliency map that highlights regions with higher importance, potentially corresponding to objects of interest while reducing the impact of background noise.

Morphology Operators: Following thresholding, morphology operators, specifically erosion and dilation, are applied to the saliency maps. The purpose of these operations is to refine the segmented regions by removing smaller, unessential pieces and merging larger, more significant chunks. Erosion works by shrinking the boundaries of the foreground object, effectively removing small noise points, while dilation has the opposite effect, expanding the boundaries of the foreground regions to fill in gaps and connect disjointed parts. This results in a more consistent and connected representation of the salient objects, ensuring that the objects are more accurately represented.

Contour Extraction: After refining the saliency maps with morphology operations, contours are extracted from the processed maps. Contour extraction involves identifying

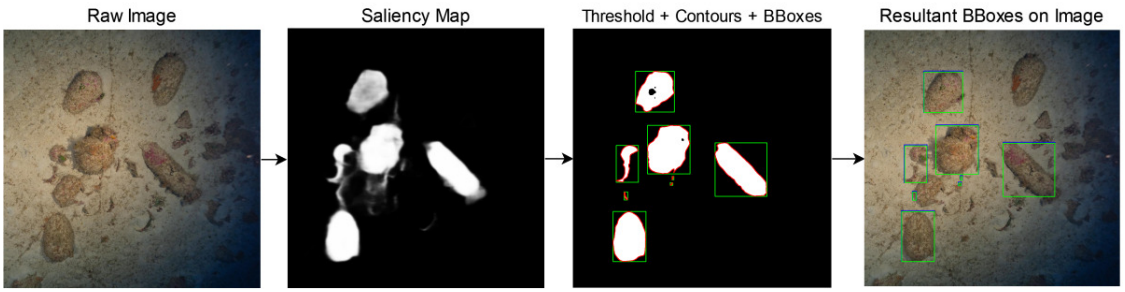


Figure 4.4: Classical Salient Object Detection pipeline, displaying the resultant image and bounding boxes over the different layers of processing.

the boundaries of the regions with high saliency detections, which are then quantised into distinct areas. This process works by tracing the edges of these regions, effectively creating a boundary outline around each detected salient object. The result is a set of contours that delineate the shape and extent of each high-saliency region.

Bounding Box Extraction: For the next step, bounding boxes are created around the extracted contours. This is achieved by calculating the minimum and maximum coordinates of each contour, which define the corners of the bounding box.

Box-Size Filtering: Once the bounding boxes are extracted, basic area thresholding is applied. This step is designed to discard bounding boxes that are unrealistically small or large, based on parameterised thresholds, searched and compared for each individual technique. The rationale behind this is that bounding boxes outside these thresholds are less likely to correspond to objects of interest. For example, very small bounding boxes might result from noise or minor details, while very large ones might encompass multiple objects or sometimes brightly lit ropes across the entire image. By filtering these out, we focus on bounding boxes that are more likely to represent actual objects, improving the accuracy and relevance of the detection results.

Evaluation: The final step involves the evaluation of the quality and utility of the bounding boxes derived from the saliency maps. This is carried out by comparing resultant bounding boxes to unclassified annotation bounding boxes, which serve as a reference. This comparison is used to determine whether the saliency maps alone can yield useful information for object detection, whilst highlighting areas for potential improvement in the detection pipeline.

Figure 4.4 illustrates this described SOD process. It starts from the raw image displaying several amphorae. Then, the second image displays the corresponding saliency map, generated using InSPyReNet. Next, the third image displays the map after thresholding, with red contours and green bounding boxes. The final image displays these generated bounding boxes on the original image.

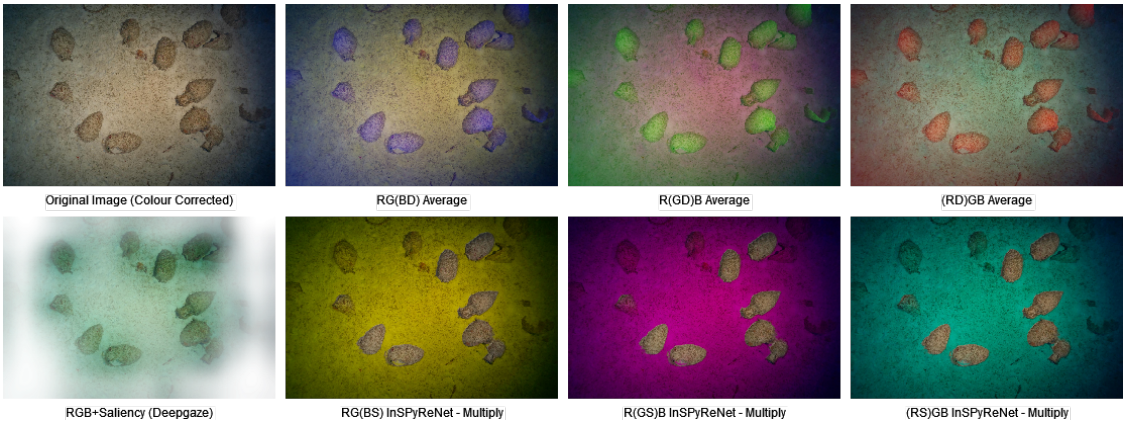


Figure 4.5: Resultant images after various channel fusion techniques. Differences may be seen between the intensity of the artefacts and the rest of the image.

4.4.4 Map-Channel Merging Techniques

With saliency and depth maps obtained, it becomes crucial to integrate this information effectively within the imagery and subsequently the object detection model. There are several approaches to tackling this integration, aiming to inject this information without losing excessive variation of the original images.

To explore the efficacy of map-channel merging methods, three main techniques were utilised. The first two include multiplication and weighted means as enrichment processes. By using depth or saliency as an enrichment step offers several benefits, particularly due to the additional exportability and training aids it unlocks. Since the model architecture itself is not modified, it can easily be switched out for newer or more relevant models, making the solution more generic. This advantage also means that, especially with smaller datasets, pretrained weights from other domains may be leveraged, potentially boosting performance and training convergence rates. This approach integrates the auxiliary information within the data itself through various handcrafted techniques, aiming to retain as much of the relevant original data as possible while injecting this additional information. However, this approach is not as dynamic in its integration of auxiliary maps, as the intensity of the maps is predefined by the initial fusion techniques, and underlying information may be sacrificed in the process. In order to retain as much of the original image details as possible, the enrichment operations are typically applied to only one of the channels at a time. This necessitates a comparison between the effects of targeting each individual channel, as different channels may contain separate details.

The third method involves modifications to the architectures to accept an additional fourth channel. This modification allows for the inclusion of all three original image chan-

nels (RGB) plus the additional saliency or depth information, providing the model with enriched data without sacrificing the original input. However, this also has its own disadvantages as will be discussed.

4.4.4.1 Channel Multiplication

An element-wise multiplication approach was explored. The idea behind this is that the map specifies areas the object detector should focus on. Hence, by multiplying the channel values with the map values, these regions are accentuated, making them more prominent while decreasing brightness in smaller map values. The effect can be seen in the second row of Figure 4.5, where the areas considered salient appear relatively bright and ordinary, whilst the surrounding areas sharply drop in channel intensity. Essentially, multiplication treats the map as a mask, fully trusting it to emphasise the significant parts of the image. However, this approach might lead to some information loss in the less emphasised areas. This can be mathematically modelled as:

$$\hat{C}(i,j) = C(i,j) \times M(i,j) \tag{4.1}$$

where C is the channel being manipulated and M represents the map.

4.4.4.2 Weighted Channel Mean

Additionally, an element-wise weighted mean approach was analysed. This technique combines the original channel values with the map values in a weighted manner, aiming to enhance the image data while retaining a portion of details from the original channels. This theory is rooted in image processing fusion techniques, where the range of the original channel is reduced and a newer range from the map is overlaid on top of it, essentially combining these. Although the model won't be able to discern whether this added intensity comes from the underlying image or the overlaid map, it should still have partial effects on the brightness of the channel, offering a less intrusive hinting technique. Mathematically, it is represented as:

$$\hat{C}(i,j) = (1-w) \cdot C(i,j) + w \cdot M(i,j) \tag{4.2}$$

where C represents the channel being processed, M represents the map being used, and w represents the weighting factor, which determines the intensity of the map's effect. This is seen in the first row of Figure 4.5, where the shallower areas have much brighter channels, whilst the surrounding areas are slightly faded.

4.4.4.3 4-Channel Models

To accommodate the additional information from the maps, the model architectures were customised to accept four channels instead of the original three. A custom dataloader was developed since typical vision libraries handle only three channels by default. Additionally, several of the default image augmentation processes depend on having three channels, such as *RandomHsv*, which needs to be omitted from 4-channel models, potentially affecting performance. Nevertheless, this modification allows for added channels, without sacrificing the original data. Being a fourth channel, this is often represented as opacity within visualisations, which is seen in Figure 4.5, where the map-intense portions of the image appear opaque, whilst the other areas appear fainter. Notably, the detector does not necessarily interpret this as opacity and still has access to all the channels in full detail and the additional map.

4.5 2D Model Comparison Techniques

This section outlines the model comparison process targeting objectives **O2** and **O3**. To evaluate the performance of object detection models, a scientific evaluation of how the predictions compare to ground truth data is required. Commonly adopted metrics, primarily Intersection over Union (IoU) and Mean Average Precision (mAP), are used for this purpose.

IoU measures the overlap between the predicted bounding box and the ground truth bounding box. It is calculated as the ratio of the intersection area to the union area of the predicted and ground truth boxes, with values ranging from 0 to 1, where a value of 1 indicates a perfect overlap. The formula for calculating IoU is described in Equation 4.3 Athira *et al.* (2021):

$$\text{IoU} = \frac{\text{Area of Intersection}}{\text{Area of Union}} \tag{4.3}$$

Mean Average Precision (mAP) is a crucial metric used to evaluate object detection models. This metric relies on the foundational metrics of precision and recall, which are classification metrics. To determine these metrics, each detection must be classified as either true or false, which is achieved using an IoU threshold. Typically, IoU thresholds range from 0.5 to 0.95, allowing us to classify detections as true positives if they exceed the threshold, or false otherwise.

Precision represents the proportion of true positives out of all predicted positives. It indicates how well a model predicts only the correct bounding boxes, avoiding false

positives. Precision is mathematically defined as:

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{FP} + \mathrm{TP}} \tag{4.4}$$

Recall represents the proportion of true positives out of all actual positives. It measures the model's ability to correctly identify all relevant instances. Recall is mathematically defined as:

$$\mathrm{R} = \frac{\mathrm{TP}}{\mathrm{FN} + \mathrm{TP}} \tag{4.5}$$

When varying the confidence thresholds, precision and recall typically oppose each other: high confidence indicates higher precision and low confidence typically indicates higher recall. This relationship is modelled using the Precision-Recall curve. The area under this curve represents the Average Precision (AP), which indicates the model's overall performance in terms of both accuracy and recall. The equation for calculating Average Precision is presented in Equation 4.6 (Wen *et al.*, 2023):

$$\mathrm{AP} = \int_0^1 P(R)\, dR \tag{4.6}$$

where P is the precision and R is the recall.

AP can only handle a single class and does not account for the performance of multi-class models. To address this, Mean Average Precision (mAP) is used, which averages the AP across all detected classes. The equation is presented in Equation 4.7 (Wen *et al.*, 2023):

$$\mathrm{mAP} = \frac{1}{C} \sum_{i=1}^{C} \mathrm{AP}_i \tag{4.7}$$

where C represents the total number of classes, and AP_i represents the Average Precision for class i.

These evaluation metrics provide standardised measurements to compare the performance of different object detection models, assessing both detection accuracy and localisation. In our case, mAP_{50} and $mAP_{50:95}$ are primarily used. mAP_{50} evaluates detection accuracy at a single IoU threshold of 0.5, indicating a model's ability to detect objects correctly without being overly strict about the exact overlap, making it suitable for various applications. On the other hand, $mAP_{50:95}$ provides average AP across IoU thresholds ranging from 0.5 to 0.95, ensuring that the model performs well in both detecting objects and accurately predicting their positions, which is valuable for applications requiring precise localisation.

By using these metrics, comprehensive insights into the model's strengths and areas for improvement can be gained, leading to better overall performance in real-world applications.

4.5.1 Gradient-Based Visualisation Tools

In assisting interpretability, gradient-based visualisation tools, such as Grad-CAM (Selvaraju *et al.*, 2017), provide valuable insights into the inner workings of AI models. Grad-CAM generates class-specific activation maps by using the gradients of the target class flowing into the final convolutional layer, helping us understand which parts of the input image are most influential for the model's predictions.

XGrad-CAM (Fu *et al.*, 2020) builds on this by incorporating an axiom-based approach, which improves the precision and interpretability of the activation maps. This method integrates a weighting mechanism that considers the importance of individual pixels in relation to gradient information, resulting in more precise and fine-grained visualisations. XGrad-CAM highlights the critical regions of the input that significantly influence the model's predictions, providing clearer insights into what features the model considers important. Given these potential insights, we will be using this tool for analysis, enabling us to generate relevant activation maps, and visually interpret the focus points of detectors.

4.6 O4 - Localisation Techniques

Localisation is a multi-step process that transforms bounding-box detections, which are pixel positions on images, into 3D orthomodel points and subsequently into world geographic coordinates. This technique requires an understanding of key trigonometric concepts, the camera model outlined in Section 2.3, and familiarity with the benefits and limitations of the photogrammetric process outlined in Section 2.2.

4.6.1 Overview

A detailed view of the entire localisation process is outlined in Figure 4.6. This starts by detailing the 2D object detection, which performs individualised localisation on pixel positions and classification. Then, these 2D points are projected onto the orthomodel, where the centre point is taken as the object geotagged and saved for documentation and further evaluation. Separately, the four corners of the bounding boxes are also projected, which undergo orientation correction to deal with the differing originating axis, clustering and suppression to minimise duplicate detections from multiple image sources and finally an optional normalisation step which turns the boxes from 4-pointed oriented boxes to 2-point horizontal boxes.

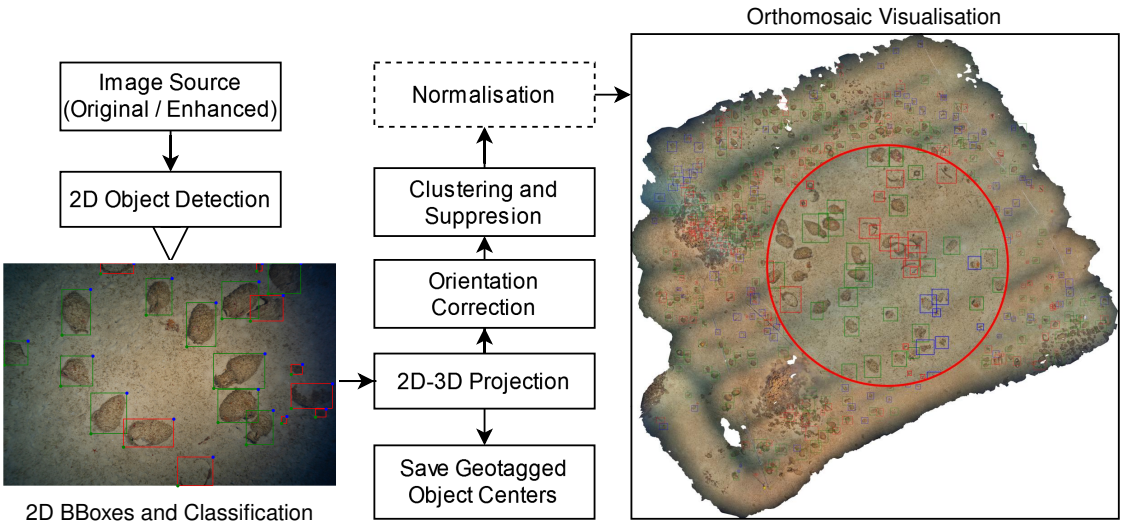


Figure 4.6: Localisation Process Architecture displaying the steps in projecting 2D predictions onto the orthomodel and visualisation on the orthomosaic.

4.6.2 2D-3D Projection

At a high level, it is crucial to understand how such a process is even possible, and how to leverage the bidirectional link between coordinate systems effectively. The ability to perform any projection relies on having both extrinsic and intrinsic camera parameters, which effectively means knowing, or estimating, the camera's position and orientation at the time the image was captured.

4.6.2.1 Estimated Camera Parameters

Intrinsic parameters include the focal length, principal point, and lens distortion coefficients of the camera. These parameters describe how the camera projects 3D points onto the 2D image plane and are usually obtained through a calibration process prior to image acquisition. On the other hand, Extrinsic parameters define the camera's position and orientation in the world coordinate system. These camera poses are determined during the photogrammetric stage, where the camera poses are jointly estimated through a process known as bundle adjustment. This involves optimising the parameters to minimise the reprojection error across all images in the dataset. This is achieved by triangulating points that are visible in multiple images to determine the camera's pose relative to these points.

4.6.2.2 Pixel to Model

Following the ideology presented in Figure 2.4, we know that if we have the camera pose, we have an estimate of where the camera was pointing when the image was taken. Given this positional information, we can cast a ray originating from the camera centre, towards the image plane, such that it passes through some pixel at point (u, v). This ray, when followed, will eventually intersect with the orthomodel, which gives us a 3D point equivalent to the pixel at (u, v). This repeated process allows us to effectively link from the local coordinates of each image, onto 3D coordinates in the model coordinate system, providing a shared space between all image detections. There are also cases where the ray never intersects with the model, where in cases such as the images at the very edge of the area see no overlap, and hence cannot be accurately included in the model. Any projections occurring at these points are omitted, as no accurate representation is found.

4.6.2.3 Model to Geographic Coordinates

Taking this step further, we may transform the 3D orthomodel coordinates into geographic coordinates (i.e., latitude, longitude, and altitude). This involves aligning the 3D model with the geographic coordinate system using Ground Control Points (GCPs) or other reference data that provide the necessary transformation parameters. GCPs are known geographic locations that can be identified in the images and the 3D model. These points are used to align the 3D model with the geographic coordinate system accurately. This alignment unlocks a conversion layer between internal positions within the orthomodel, and geographic locations, where a greater sense of scale and exportability are enabled. In our case, this process was standard procedure for the Classics and Archaeology department. Nevertheless, it is important to understand the process and associated nuances of aligning an orthomodel to the world.

4.6.2.4 Agisoft Wrapper

The Agisoft SDK streamlines these complex processes, offering direct integration with their software which facilitates the process of accessing the point cloud, casting rays, plane transformation and geographic coordinate casting. This standardises the approach to such tasks in the real world, ensuring consistency and reliability.

4.6.3 Orientation Correction

In the context of image processing, bounding boxes are typically defined in relation to the horizontal axis of the image, being the corners of extremities which capture the entire

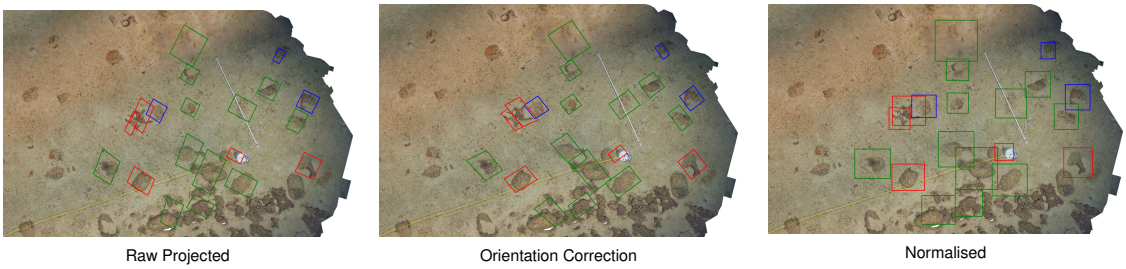


Figure 4.7: Projection Processing Pipeline featuring the orientation correction and normalisation results based. This highlights the individual effects of each step in cleaning detections on the orthomosaic.

ROI. Since cameras may be in any orientation when taking the image, when bounding boxes are projected from the image plane to the orthomosaic plane, the horizontal axis does not stay consistent, making the bounding boxes unusably distorted. To address this issue, it is necessary to establish the difference between the axes of the image and the orthomosaic and rotate the bounding boxes by the difference. These points are rotated about the centre of the bounding box, which is calculated using:

$$c_x = \frac{x_1 + x_2 + x_3 + x_4}{4} \qquad c_y = \frac{y_1 + y_2 + y_3 + y_4}{4} \tag{4.8}$$

where the x_n and y_n terms represent the corners of the corners of the bounding boxes, whilst c_x and c_y represent the calculated coordinates of the centre point.

This rotation difference may be approximated using the estimated extrinsic camera parameters. Extrinsic rotation parameters are typically defined in terms of $< \omega, \phi, \kappa >$, representing the angle of rotation around the X, Y and Z axis respectively. In our particular case, considering the camera is pointing downwards towards the seabed, and wanting the angle the image makes on the seabed, we are most interested in the κ value. In photogrammetry, this angle is typically represented as the counter-clockwise angle starting from vertical, thus, we need to convert this to the standard mathematic angle, which starts from the positive x-axis, being horizontal. This can be done by applying $\theta = \kappa + 90$, which accounts for the difference in the origin axis of the angle of rotation.

Knowing this centre point and the angle of rotation θ, we can find the new rotated points using:

$$\begin{aligned} x_i' &= \sin(\theta) \cdot (x_i - c_x) - \cos(\theta) \cdot (y_i - c_y) + c_x \\ y_i' &= \cos(\theta) \cdot (x_i - c_x) + \sin(\theta) \cdot (y_i - c_y) + c_y \end{aligned} \tag{4.9}$$

where x_i', y_i' are the new coordinates of the rotated points based on the raw projected coordinates x_i, y_i. The result of the orientation correction step may be seen in the first

two steps of Figure 4.7, where especially elongated boxes have the most visible rotation, which after correcting is better suited for the shape of the artefacts.

4.6.4 Bounding Box Perspective Normalisation

Due to the perspective distortion of each camera, a perfectly rectangular bounding box on the image plane is often shifted into an irregular quadrilateral on the orthomosaic plane.

The process of normalisation is fairly simple, where the minimum and maximum positions of each of the points are taken as bounds, and new bounding boxes are fitted to these in relation to the mosaic axis. This provides a normalised angle bounding box which approximately encapsulates the object without requiring excessive leeway. The result of this operation may be seen in the last image of Figure 4.7, where the orientations of the bounding box are normalised into horizontal ones.

That being said, not all operations benefit from normalisation, where this step does lose out on some localisation information. For this reason, if the tightest bounding boxes around objects are a priority, the oriented bounding boxes may provide better results. For our use case, a hybrid approach is taken, where stages such as clustering (described in Section 4.6.5) may benefit from considering polygonal operations for tighter IoU calculation. On the other hand, direct evaluation techniques such as mAP are compared between normalised boxes to minimise the bias of orientations.

4.6.5 Bounding Box Suppression and Clustering

As described in Section 2.2, the raw images do not necessarily cover unique areas, where for accurate photogrammetric reconstruction, a degree of overlap is standard. This does however imply that when detections are performed on raw images and projected to the common orthomosaic coordinates, the same area is likely to be seen in more than one image. This will cause several boxes for the same artefact, with variations according to perspective. Especially when there are dense areas of artefacts, this problem makes it difficult to discern whether close boxes are duplicate detections or truly overlapping detections. Such ambiguity is especially problematic for downstream tasks requiring accurate counting or location-based clustering, which would face considerable inaccuracy. There are several ways we attempted to counteract this issue, where we namely explore Non-Maximum Suppression (NMS), Weighted Boxes Fusion (WBF) and our own approach of Polygonal Area Locking (PAL). These are discussed in further detail in the following sections.

4.6.5.1 Non-Maximum Suppression

NMS is a concept where overlapping bounding boxes with lower confidence scores are suppressed in favour of more confident ones. This is typically used at the detection stage of object detection techniques, used to filter out noisier, low-confidence outputs surrounding objects. The full algorithm has been detailed in Algorithm 1 as described by Chu *et al.* (2020).

Algorithm 1: Traditional Non-Maximum Suppression

Input: Set of boxes $B = \{b_1, b_2, \ldots, b_n\}$, corresponding scores $S = \{s_1, s_2, \ldots, s_n\}$, and threshold N_t
Output: List of non-maximum suppressed boxes D and corresponding scores S
Initialise $D \leftarrow$ Empty list;
while $B \neq \emptyset$ **do**
  Select the box b_M with the maximum score s_M in S;
  Add b_M to D and remove it from B;
  for *each* $b_i \in B$ **do**
    if *iou*(b_M, b_i) $\geq N_t$ **then**
      Remove b_i from B and its score from S;
    end
  end
end
return D *and* S

By employing and comparing this within our scope, the aim is to achieve a similar effect, where we keep the maximum confidence detection, whilst suppressing duplicated detections of the same artefact.

4.6.5.2 Weighted boxes fusion

Whilst NMS is able to suppress overlapping boxes with non-maximum confidence scores, Weighted Boxes Fusion (WBF) (Solovyev *et al.*, 2021) is often proposed as an alternative, which aims to merge boxes from different sources, rather than completely removing them. This is especially promising in our case, where each repeated detection should realistically add confidence and value to the detection, rather than confusion.

Similar to NMS, WBF uses a scoring system. However, apart from considering only the confidence scores, WBF also takes into account the level of IoU for merging boxes. The fused box coordinates and confidence scores are calculated using the weighted average of the coordinates and confidence scores of all the boxes in a cluster.

By computing these weights for each box and normalising them based on their confidence scores, WBF assigns proportional importance to each detection, typically resulting

in a more accurate fusion of bounding boxes from multiple sources.

4.6.5.3 Polygonal Area Locking

Another take on a suppression technique was named Polygonal Area Locking (PAL), which is based on the enforcement of mutual exclusivity of image sources within an area. Essentially, this enforces that any one region within an orthophoto should be based on exactly one image source. This mentality is used to *lock* the bounds of a single image on the orthophoto and ignore further attempts to plot boxes in that region. This requires two primary operations, one being determining which region of the orthophoto the particular image is part of, and the other being to determine whether a new point is within an existing polygon or not.

In order to estimate the camera footprints, this is modelled as a convex hull problem. A convex hull is the smallest convex set of vertices that encloses all other vertices within it. The Quickhull algorithm (Barber *et al.*, 1996) was adopted in our case, where we start by finding the extremities of points by finding the minimum and maximum coordinates of bounding box corners. Then, by drawing an imaginary line between these two points, the furthest point from this line is found, which is added to the convex set. All points within this set are considered enclosed and not considered for the convex set, whilst the process is repeated until it connects all the outer points, forming a shape that wraps around all the detections.

The second task is to identify whether new points should be considered or skipped if they are within any of the polygons. For this task, the ray casting algorithm (Wegstein *et al.*, 1962) for determining whether a point is within a polygon or not can be used. This algorithm follows a very elegant principle, being that of casting a ray from the point in question, towards a consistent direction, such as horizontally. Then, the number of times that the ray intersects with the polygon is counted, where if the number is even, it means the point is outside of the polygon, as it means that it entered as many times as it exited. If it is an odd number, the point is inside, as it had to go out without ever going in.

These two solutions allow us to form the problem as a set of polygons around the orthomosaic which lock their own areas. A potential limitation of this occurs with ignoring edge items. In such cases where a bounding box partially intersects with some other polygon, it will be completely omitted. For this reason, a threshold of corners per bounding box may be set to counter this effect, for instance, if the threshold is set to 2, a particular bounding box may have 2 of its vertices in another polygon and still be allowed. Further limitations include the lack of a scoring system, hence the first detections of an area are preserved, whilst other images could have represented the area better.

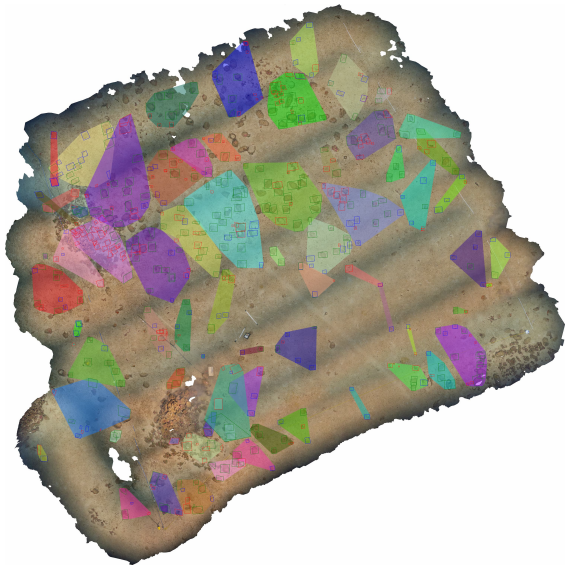


Figure 4.8: Generated Polygonal Area Locking regions, based on camera footprints, used to filter detections intersecting these.

The set of polygons generated from the test set was plotted in Figure 4.8, where each individual image is represented as its own polygon, encapsulating its set of bounding box points. Figure 5.5 displays the results of the PAL process with a threshold of 0, meaning it does not allow any overlap. This does seem to have the benefit of having very few duplicates retained, however, some of the clearer detections were also lost in the process, which might be overly strict. The full differences between thresholds were compared, aiming to evaluate the resulting differences.

4.6.6 Projection Evaluation

Evaluating the projection is challenging because ground truth data for the orthophoto is not readily available. While manual ground-truth annotations done directly on the orthophoto would provide the most accurate targets, this was not possible due to the difficulty and expertise required for such imagery. This limitation means that the exact positions of true target objects on the orthomosaic are unknown. However, the projection method is a deterministic mathematical process, hence the annotations themselves may be projected from pixel to geographic coordinates, and these may be used as the target boxes as an approximation.

Using these projected annotations, the mAP metric discussed in Section 4.5 can be employed to assess the level of retained accuracy after the projection step. Some deviation is expected, as the 2D boxes undergo projection, rotation, normalisation, and sup-

pression, each introducing potential inaccuracies. For this reason, it's important to monitor the extent of these errors to ensure that the resultant predictions on the orthomosaic are sufficiently close to the projected annotations.

Additionally, distance measurements between the centre points of boxes can be computed. By using IoU thresholds to pair annotations, we can calculate the geographical distance between pairs of overlapping boxes to interpret how box IoU translates to real-world distance. For this, the Haversine distance formula given in Equation 4.10 (Nassar *et al.*, 2020) is used:

$$d = 2R \cdot \arcsin\left(\left(\sin^2\left(\frac{O_{\text{lat}} - G_{\text{lat}}}{2}\right) + \cos(G_{\text{lat}}) \cdot \cos(O_{\text{lat}}) \cdot \sin^2\left(\frac{O_{\text{lng}} - G_{\text{lng}}}{2}\right)\right)^{0.5}\right) \tag{4.10}$$

where O_{lat}, O_{lng} represent the detection's predicted coordinates, G_{lat}, G_{lng} represent the coordinates estimated from ground truth object annotations and R represents the radius of the sphere, which in the case of Earth is taken as $6,371km$.

4.7 Methodology Summary

This chapter outlines the methodology employed to achieve the study's primary objectives. We began by compiling a multi-class underwater archaeological object detection dataset, comprised of 864 images and 3 classes. Based on this data, the comparison process of various state-of-the-art object detection models is described, including CNN-based YOLO variants, Faster R-CNN models, and transformer-based DETR series. In light of improving model performance, we further outline the methods explored for integrating depth and saliency maps within our inputs. The evaluation methods for 2D model predictions are also discussed, aiming to scientifically evaluate and rank resultant model performances across a variety of preprocessing and architecture differences. Primary results are reinforced using 5-fold cross-validation techniques. Additionally, localisation methods are developed to transform 2D detections into 3D geographic coordinates, incorporating orientation correction and bounding box suppression techniques like NMS, WBF, and our own approach named PAL. This is finalised by detailing the complexities in methods associated with projection accuracy deviation analysis.

5 Evaluation

This chapter presents the analyses and results obtained from the performed experiments. Utilising a multi-stage structure, these experiments are targeted towards exhausting the search spaces of each objective, finding the most suitable results for our use case, and retaining the most performant configurations. Through this repeated process of experimentation, we converge into the ideal set of configurations.

5.1 Evaluation Overview

The architecture encompasses three main stages, each aiming to address a specific objective, which answer to the overarching aim of this study. This has been visually presented in Figure 5.1, which details the primary stages and top-level flow of convergence towards the most applicable solution. Stage 1 (O2) encapsulates the base model training and configuration search, which is designed to determine which model types and architectures are most suitable for our use case. This stage focuses on identifying optimal configurations through systematic evaluation, which are then confirmed through K-Fold Cross-validation of top-performing models. Subsequently, Stage 2 (O3) employs saliency estimation techniques and photogrammetric depth estimation, to analyse how these may affect the object detection performance. These are firstly experimented with through a baseline of Salient Object Detection, which determines whether the maps on their own have noticeable localisation information. These components are then encoded within the image sets by searching over fusion configurations. These are compared to the base models by their respective resultant performance, to determine whether the overhead of adding these enrichment steps provides any noticeable changes, or otherwise. The final stage, Stage 3 (O4), transitions into photogrammetric 3D projection. This stage aims to analyse the differences between suppression techniques and estimate the retention of localisation accuracy after the detections are passed through the projection steps. This further presents the visualisation achieved through aggregated plotting on the orthomosaic, aiming for the interpretability of results.

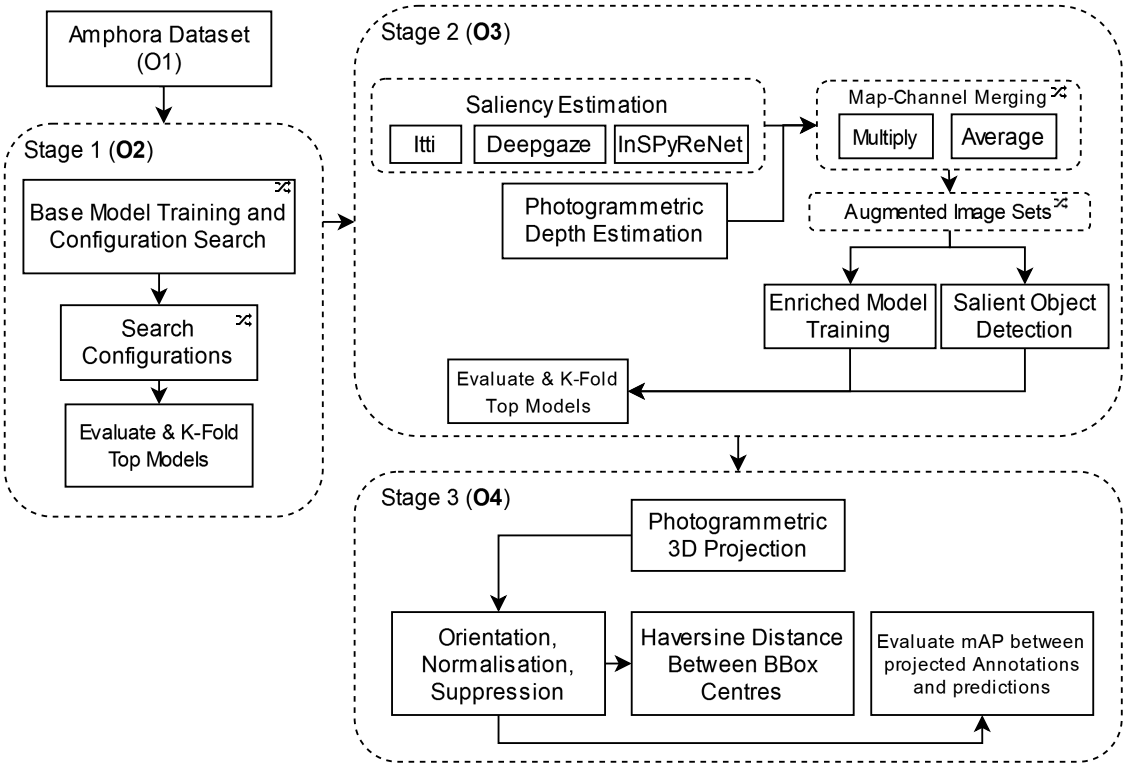


Figure 5.1: Evaluation architecture overview displaying the three distinct stages and pertaining comparisons.

5.2 Stage 1 (O2) - Base Model Comparison

Relating to Objective O2, a set of promising models was chosen for a detailed comparison and evaluation based on predictive performance. This stage is meant to survey how the unique qualities of the model architectures reflect in their ability to handle the conditions of our data. This also serves as a primary shortlisting of techniques, where unsatisfactory results may be omitted quickly from further experimentation, to systematically converge towards the best solution.

5.2.1 Architecture and Size Comparison

The very first experiment deals with the comparison of general model architectures on their accuracy, ease of convergence and how the accuracy scales by model size. These models have been run on pretrained weights from the COCO dataset and then fine-tuned on a colour-corrected version of our dataset. This is deemed to be the base model as the colour-corrected imagery is seen as more neutral and normalised, being closer to on-land images, while the pretrained weights are theorised to help the model converge faster, al-

Table 5.1: First stage evaluation of base models across multiple sizes.

Model	Size	Pretrained	CC	mAP_{50}	$mAP_{50:95}$
YOLOv9	c	Yes	Yes	87.56	53.88
YOLOv9	e	Yes	Yes	85.01	53.01
YOLOv8	n	Yes	Yes	85.02	52.88
YOLOv8	s	Yes	Yes	84.82	52.38
YOLOv8	m	Yes	Yes	85.99	53.24
YOLOv8	l	Yes	Yes	85.93	53.62
YOLOv8	x	Yes	Yes	85.57	54.02
YOLOv7	tiny	Yes	Yes	**87.65**	53.40
YOLOv7	normal	Yes	Yes	86.83	53.26
YOLOv7	x	Yes	Yes	86.83	52.87
YOLOv5	n	Yes	Yes	85.44	52.82
YOLOv5	s	Yes	Yes	86.52	53.78
YOLOv5	m	Yes	Yes	85.71	53.49
YOLOv5	l	Yes	Yes	87.25	**54.63**
YOLOv5	x	Yes	Yes	80.75	50.33
RT-DETR	l	Yes	Yes	83.53	52.92
RT-DETR	x	Yes	Yes	72.76	46.36
Deformable DETR	ResNet50	Yes	Yes	79.1	43.6
FRCNN	SWIN-tiny	Yes	Yes	83.9	49.6
FRCNN	ResNet50	Yes	Yes	81.7	50.2

beit not necessarily optimally. These experiments were performed on the primary holdout test to get an estimate of performance. The results obtained are presented in Table 5.1. Notably, the absolute best metrics are highlighted in bold, whilst the best configuration per primary architecture is underlined. Generally, the results obtained are good, in that a majority of the predictions are correctly positioned and classified, which especially given the difficulty of the domain shows a deep understanding of context and artefact preservation level. The results obtained are relatively close, with most architecture configurations achieving around 85% mAP_{50}, and $mAP_{50:95}$ above 50%.

Analysing the single-shot architectures, there were no clear distinctions between these, where YOLOv7 achieved the highest overall mAP_{50} of 87.65%. This was closely followed by YOLOv9 and YOLOv5. YOLOv5 also achieved the highest overall $mAP_{50:95}$ with 54.63%, which hints at its marginally better localisation. YOLOv8 has the lowest overall performance from the single-shot CNN results, especially in mAP_{50}. When analysing size configurations, these were mostly within the margin of error. The best mAP_{50} results

were achieved by YOLOv9c, YOLOv8m, YOLOv7-tiny size and YOLOv5l. It was noted that larger model sizes do not necessarily increase the overall performance. This falls in line with the findings of Paraskevas *et al.* (2023), which found that on smaller dataset sizes, larger model configurations often overfit. This also tells us that the bottleneck for performance is not limited by the model architecture itself, but more by the size and ambiguity present within the data.

Transformer architectures were found to typically underperform compared to single-shot techniques. Additionally, these techniques were found to be considerably slower to train, with most YOLO models converging within 200 epochs and approximately 4 hours, whilst DETR techniques required around 400 epochs, 10+ hours, and were still not reliably training. Additionally, DETR was found not to converge at all, even with rigorous parameter tuning and repeated experimentation, which confirms the discussions by Zhu *et al.* (2021) about its difficulty in converging. The highest performing method from this set of techniques was achieved by RT-DETR, achieving an 83.53% mAP_{50}. Although this was only marginally less than other techniques, the unreliable and long training time made it a difficult model to work with whilst providing no other benefits. For these reasons, DETR, RT-DETR and Deformable DETR are considered not appropriate for our data size and type, and will hence not be studied further to better focus on explaining more promising architectures.

When considering two-stage techniques, their performance was also found to be lower than single-shot techniques. This is likely linked to the models being more complex, hence requiring more data to properly utilise the architectures' inert benefits. The SWIN version achieved minimally better mAP_{50} and similar $mAP_{50:95}$ to the ResNet, whilst requiring roughly the same training time of approximately 4 hours. Although training was reliable, the consistently lower performance than other techniques meant it was not worth pursuing, hence these are not considered for future experimentation either.

In summary, the single-shot CNN techniques were found to be the most appropriate type of detector for our data, which will be the primary target of following experimentation. No clear distinctions between these are clearly visible, hence further investigation and ablation was required to better rank these.

5.2.2 Configuration Search Comparison

There were two primary configuration choices which required experimentation, being colour correction and pretrained weights. An ablation is a setup which keeps all other parameters consistent, apart from the one being evaluated, which will enable us to better analyse how performance scales without external variations.

Table 5.2: Colour Correction ablation results.

Model	Size	Raw		CC	
		mAP_{50}	$mAP_{50:95}$	mAP_{50}	$mAP_{50:95}$
YOLOv9	c	85.20	52.44	87.56	53.88
YOLOv9	e	85.25	53.29	85.01	53.01
YOLOv8	n	84.89	53.24	85.02	52.88
YOLOv8	s	85.72	53.80	84.82	52.38
YOLOv8	m	85.77	53.65	85.99	53.24
YOLOv8	l	85.28	53.25	85.93	53.62
YOLOv8	x	85.55	52.84	85.57	54.02
YOLOv7	tiny	86.51	53.44	**87.65**	53.40
YOLOv7	normal	86.15	53.05	86.83	53.26
YOLOv7	x	87.09	53.38	86.83	52.87
YOLOv5	n	84.80	52.63	85.44	52.82
YOLOv5	s	85.81	53.05	86.52	53.78
YOLOv5	m	86.51	53.74	85.71	53.49
YOLOv5	l	86.38	53.89	87.25	**54.63**
YOLOv5	x	85.21	52.98	80.75	50.33

5.2.2.1 Colour Correction Analysis

This experiment is based on the first experiment findings, where single-shot techniques are used to analyse the impact of colour correction on the resultant performance. It is known that underwater scenery poses many difficulties to visual clarity, caused by the challenging lighting conditions and the non-uniform propagation of light, especially affecting the red wavelength of channels. Through the application of white balancing, we aim to explore whether the normalised appearance affects the resultant detection accuracy. It's important to note that the pretrained/colour-corrected results are the same as the first experiment since the configurations are the same.

The results in Table 5.2 show that colour correction typically leads to superior performance. All models achieved their best mAP_{50} results with colour-corrected inputs, which potentially highlights their dependence on enhanced colour fidelity for optimal detection efficacy. For instance, YOLOv9c reached a mAP_{50} of 87.56% with colour correction, compared to 85.20% without it. Similarly, YOLOv5's best-performing configuration (large size) showed an $mAP_{50:95}$ of 54.63% with colour correction, higher than 53.89% achieved without it. Notably, although this does reliably affect the peak performances, averages sometimes vary, where YOLOv5x with colour correction is distinctly less performant than without it. However, this could also be a case of overfitting due to its large size.

Table 5.3: Pretrained weights ablation results.

Model	Size	No Pretraining		Pretrained	
		mAP_{50}	$mAP_{50:95}$	mAP_{50}	$mAP_{50:95}$
YOLOv9	c	85.50	53.51	87.56	53.88
YOLOv9	e	85.90	54.24	85.01	53.01
YOLOv8	n	84.96	52.78	85.02	52.88
YOLOv8	s	86.76	53.83	84.82	52.38
YOLOv8	m	86.65	53.77	85.99	53.24
YOLOv8	l	85.03	52.90	85.93	53.62
YOLOv8	x	84.31	52.59	85.57	54.02
YOLOv7	tiny	85.52	50.70	**87.65**	53.40
YOLOv7	normal	85.53	51.38	86.83	53.26
YOLOv7	x	85.44	51.86	86.83	52.87
YOLOv5	n	84.65	52.57	85.44	52.82
YOLOv5	s	84.18	52.50	86.52	53.78
YOLOv5	m	85.53	52.82	85.71	53.49
YOLOv5	l	83.95	51.87	87.25	**54.63**
YOLOv5	x	84.95	52.57	80.75	50.33

5.2.2.2 Pretrained Weights Analysis

The third experiment analyses whether pretrained weights on the COCO dataset prove beneficial to our performance or whether they hinder the models' ability to adapt to the starkly different underwater environments. These results are presented in Table 5.3.

The comparative analysis demonstrates that pretrained weights generally enhance model performance across various configurations and sizes. For instance, YOLOv9 shows an improvement in mAP_{50} from 85.50% without pretrained weights to 87.56% with pretraining, indicating that pretrained models adapt well even to challenging underwater conditions. This is also the case for other architectures, with typical gains particularly noted in YOLOv7 and YOLOv5. The YOLOv7 tiny variant reached a peak mAP_{50} of 87.65% with pretrained weights, a substantial increase compared to 85.52% without. YOLOv8 is a noticeable exception, where the small size configuration shows a decrease from 86.76% to 84.82% in mAP_{50} with pretrained weights. Such cases suggest that while pretrained weights generally provide higher peaks in performance and more robust training, it still does sometimes cause models to be marginally worse, which would need to be handled on a case-by-case basis.

These results are summarised in Figure 5.2, which displays the prominent differences between results, which typically result in positive performances. Whilst in some cases,

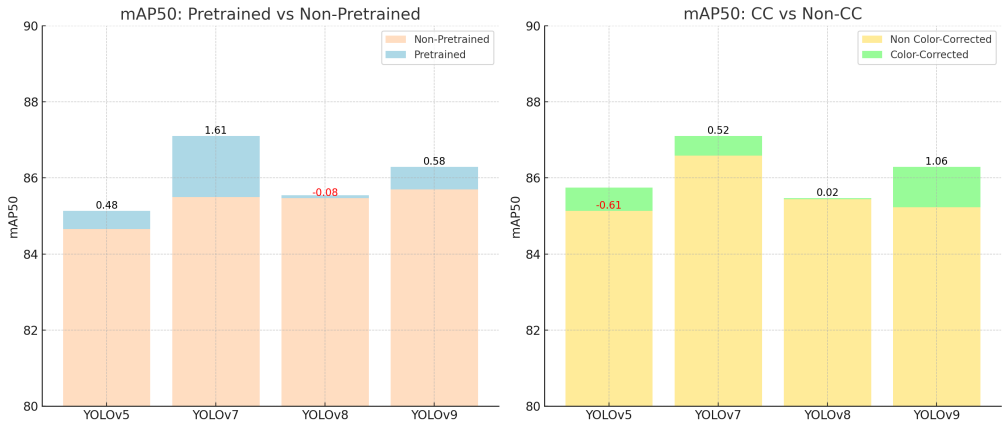


Figure 5.2: Average Differences per Model Between Configurations.

such as YOLOv5 CC, averages may be worse, the peaks are consistently higher, which will be our primary target. Given these results, we will be utilising the pretrained and colour-corrected versions for the rest of our experiments.

5.2.3 Detailed Fold-Based Evaluation

The top-performing architectures were further analysed under 5-fold validation to confirm the findings using more robust metrics, and again to reduce the margins of error. The averaged metrics per group of folds are presented in Table 5.4. Overall, the fold-based evaluation solidified the previous observations, offering more nuanced insights into model performance across different dataset distributions. The results from this detailed analysis showed that the models maintained consistent performance across different folds, confirming the reliability of the results. For example, YOLOv7 tiny maintained a high mAP_{50} of 86.14% and a decent average $mAP_{50:95}$ of 51.42%. Different object conditions revealed varying detection performances, with whole objects consistently achieving higher detection scores compared to broken and buried objects. This indicated the challenge posed by occlusions and object fragmentation in underwater environments. Notably, YOLOv5 medium size stood out with the highest $mAP_{50:95}$ scores for broken and buried objects, demonstrating its robustness in challenging conditions.

From these results, it was determined that YOLOv7-tiny provided the best mAP_{50} scores, and will hence be retained for further experimentation in later stages. Additionally, due to its highest $mAP_{50:95}$, YOLOv5m was also retained for further experimentation, aiming to analyse how the peak localisation is affected by the later stages. That being said, whilst these selected configurations are significantly better than some others, all the single-shot models were found to provide acceptable performance, and would likely

Table 5.4: 5-Fold Validation displaying mean performance results per configuration.

Model	Size	mAP_{50}	$mAP_{50:95}$			
			Avg	Whole	Broken	Buried
YOLOv9	c	84.99	52.04	67.97	43.32	44.84
YOLOv9	e	84.51	52.18	67.88	44.05	44.60
YOLOv8	n	84.92	51.86	67.78	43.75	44.05
YOLOv8	s	85.02	51.99	67.55	43.95	44.47
YOLOv8	m	85.45	52.57	68.11	44.68	44.91
YOLOv8	x	84.25	51.64	67.54	43.09	44.28
YOLOv7	tiny	**86.14**	51.42	66.68	43.60	44.04
YOLOv7	normal	86.10	51.76	66.78	44.12	44.42
YOLOv7	x	85.72	51.54	66.70	44.07	43.86
YOLOv5	n	84.88	51.97	67.55	44.43	43.92
YOLOv5	s	85.42	52.52	68.13	44.76	44.67
YOLOv5	m	85.38	**52.80**	68.27	44.98	45.16
YOLOv5	l	85.24	52.44	68.12	44.52	44.67

be satisfactory for many day-to-day operations.

5.3 Stage 2 (03) - Depth and Saliency Fusion Techniques

The second stage of evaluation focuses on the fusion of visual saliency and depth maps within the imagery, studying whether these have any noticeable effect on prediction performance. This is based on the theory that by injecting auxiliary information within images, object detectors may be able to leverage it to better differentiate between instances. As a baseline approach, SOD is performed, comparing several visual saliency estimation techniques on their standalone ability to localise artefacts. Subsequently, multiplication and average mean fusion techniques were compared over all 3 channels of the original images or added as an additional channel. Afterwards, the effect of the weight parameter w of the weighted mean is experimented with to further tune the potency. These experiments aim to explore many possibilities for detection techniques to pick up on the additional information without sacrificing the original channels.

Figure 5.3 displays the original colour-corrected image for a particular area and several resultant maps. The depth map is brighter in areas that are closer to the camera, whilst

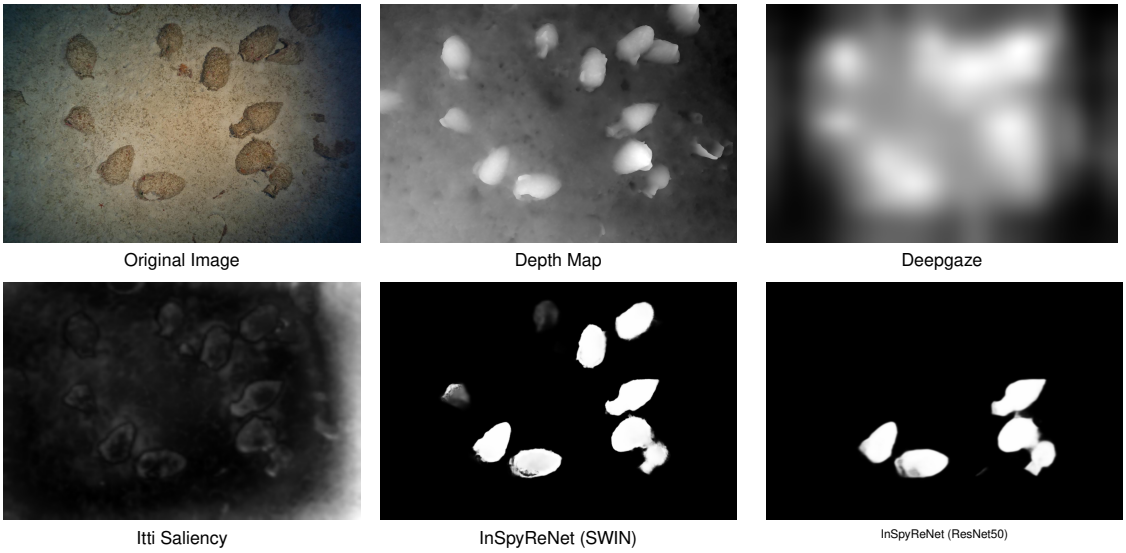


Figure 5.3: Resultant Saliency and Depth Map Comparison.

darker areas are further away. From this alone, the amphorae themselves can already be seen somewhat differently, where a detector may leverage this information to better localise artefacts.

We also display the results of several saliency detection techniques, where Deepgaze is seen as a smoothed-out low-resolution map, where we can see faint details corresponding to the original image, yet nothing articulated. Itti, on the other hand, has higher resolution, due to it being a mathematical approach. Most of the observed highlights are seen towards the edges of objects. It also seems to not properly handle gradient fading of colour, as it detects the vignetting as highly salient, relating back to its naive nature. Finally, InSPyReNet in both configurations gives a highly detailed saliency map with strong mask-like highlights of primary objects, yet completely omits surroundings. This also makes sense considering its primary object detection-oriented backbone, which is designed to discretise between detections and non-detections. When comparing the SWIN and the ResNet-based backbones, we may note that in this instance SWIN captures more of the artefacts, especially ones which are not as central or brightly lit which could be linked to the transformer's ability to look at the more global image rather than local features.

5.3.1 Salient Object Detection

Salient object detection serves as a baseline technique to better understand what information is brought forward solely using the saliency techniques. SOD has seen its fair share of studies, aiming towards a better zero-shot approach at detecting the most salient objects within imagery (Chen *et al.*, 2020b). In our instance it is being utilised as a baseline technique to other detection models, hence for simplicity, we take on a classical approach. This is based on the use of contours and a multi-stage preprocessing pipeline. The ranges tested for the preprocessing steps are presented below. Threshold values included None, 150, 175, and 200, which test from mid to low strictness in brightness. Erosion iterations tested were None, 5, 10, and 15. Similarly, the dilation iterations tested were None, 5, 10, and 15. For the minimum area, values ranged from 1,000 to 11,000, and for the maximum area, values ranged from 20,000 to 70,000 pixels. These variations aimed to estimate the impact on SOD accuracy. These results are presented in Table 5.5.

From the analyses, InSPyReNet with the Res2Net50 backbone performed considerably better than all others. On the other hand, Deepgaze and Itti were found to perform the worst. Deepgaze achieved an AP of 0.08% and 0.17% for whole-only artefacts, likely due to its lower resolution and detail, which hinder its ability to detect fine-grained features. Next, Itti performed slightly better with an AP of 0.14% and 0.29% for whole-only, yet still fell short of the more advanced methods. The best min and max areas for Itti were 5,000 and 20,000, respectively, indicating it works better with medium-sized artefacts but struggles with larger or smaller ones.

In contrast, InSPyReNet with the Res2Net50 backbone achieved a considerable improvement with an AP of 5.91% and 9.29% for whole-only artefacts, with the best min and max areas being 5,500 and 45,000, respectively. This suggests that it effectively captures a wide range of artefact sizes, making it more versatile. The SWIN variant of InSPyReNet performed less effectively than the Res2Net50 variant, with an AP of 1.53% and 1.60% for whole-only, likely due to different feature extraction characteristics. Depth

Table 5.5: Best Min, Best Max, Best AP_{50}, and Best AP_{50} Whole Only for each method.

Method	Best Min	Best Max	AP_{50}	AP_{50} Whole Only
Deepgaze	10,000	20,000	0.08	0.17
Itti	5000	20000	0.14	0.29
InSPyReNet - Res2Net50	5,500	45,000	5.91	**9.29**
InSPyReNet - SWIN	10,000	70,000	1.53	1.60
Depth	8000	30000	2.87	3.57

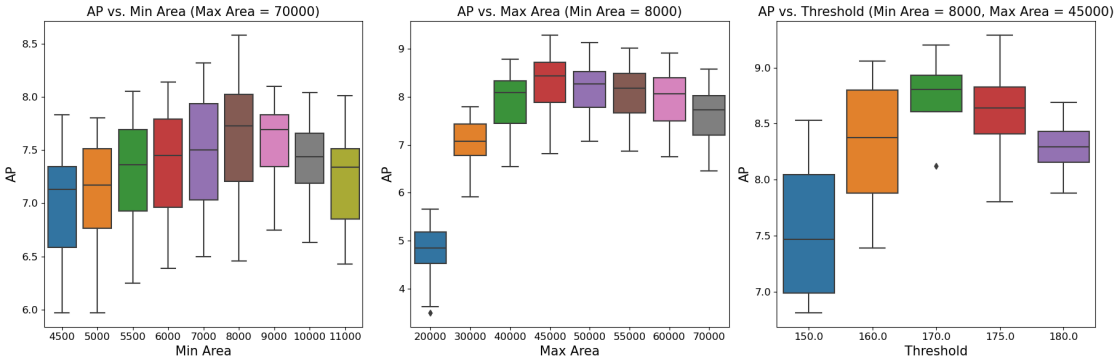


Figure 5.4: Average Precision obtained by Salient Object Detection grid search over a range of minimum bounding box areas, maximum bounding box areas and intensity thresholding values.

Table 5.6: Ablation study showing the effect of each parameter on AP_{50} and Whole AP_{50} on InSPyReNet (Res2Net50) saliency maps.

Method	AP_{50}	Whole AP_{50}	Best Value (Whole)
No Filtering	1.82	0.94	None
Thresholding	4.56	5.64	175.0
Threshold + Min	5.29	8.08	8000
Threshold + Min + Max	5.63	8.66	45000
Threshold + Min + Max + Erosion	5.91	8.65	5.0
Threshold + Min + Max + Erosion + Dilation	5.91	9.29	10.0

maps also showed reasonable performance with an AP of 2.87% and 3.57%, indicating that structural information from depth contributes positively to artefact detection.

Given the superior performance of InSPyReNet with the Res2Net50 backbone, further investigation on the effect of proper parameter tuning was conducted, through several ablation studies. The box plots presented in Figure 5.4 provide a visual comparison of the effects of varying parameters on AP for whole objects scores. The first plot shows that increasing the minimum area generally improves AP, but there is a point of diminishing returns beyond 7000-8000. This indicates that while filtering out smaller, irrelevant regions is beneficial, excluding too many regions can also reduce performance. The second plot shows a similar trend with the maximum area parameter, where performance peaks around 40000-45000. This suggests that focusing on appropriately sized artefacts, rather than very large or very small regions, yields better detection performance. The third plot highlights the threshold parameter, where an optimal threshold of around 175 yields the best performance, indicating that a moderate level of thresholding effectively isolates salient regions without missing subtle details.

Another ablation study is presented in Table 5.6, which highlights the incremental benefits of each preprocessing step. The initial no-filtering approach yields very low AP scores, suggesting that raw saliency maps on their own are not sufficient to directly utilise with our SOD technique. Introducing thresholding substantially improves performance, indicating the importance of isolating salient regions and omitting noise. Further enhancements with minimum and maximum area constraints reflect the need to focus on appropriately sized regions, and carefully consider the targets being detected. Finally, Erosion and dilation steps provide marginal improvements, helping refine the contours of detected artefacts.

These findings illustrate the importance of carefully tuning preprocessing parameters to maximise the effectiveness of salient object detection techniques. The success of InSPyReNet with the Res2Net50 backbone, combined with the insights from the parameter analysis, underscores the potential for saliency maps to contribute meaningfully to artefact detection tasks.

5.3.2 Map-Channel Fusion technique Comparison

In this experiment, we aim to compare the efficacy of different channel augmentations using varied fusion strategies, specifically Multiplication and Weighted Mean, in conjunction with several saliency estimation and depth techniques. Our objective is to understand how these combinations influence the performance of object detectors, specifically YOLOv5m and YOLOv7-tiny. The fusion techniques are assessed based on their impact on different channels (R, G, B), merged by different estimation techniques, including Itti, Deepgaze, ResNet (InSPyReNet), SWIN (InSPyReNet), and Depth. The weight of the weighted mean is set to a constant of 0.5, aiming to utilise a consistent value for reliable comparisons, later experiments expand on the impact of this parameter.

Table 5.7 presents the initial results of this comparison on the holdout dataset. For YOLOv7-tiny, the Mean fusion technique proved to be particularly effective. The highest accuracy was observed with the Itti model on the R channel (88.0%). Additionally, other models such as ResNet and SWIN also performed well with accuracies around 87.5% when using red channel means. This suggests that the averaging technique is more effective for YOLOv7-tiny, particularly when applied to the red channel. These results also exceed the top result from the original RGB evaluation, which had a maximum mAP_{50} of 87.25%, showing promise in improving performance.

When considering YOLOv5-m, the highest two results were achieved by ResNet multiplication by the red channel, and tied with blue and deepgaze mean, with a mAP_{50} of 86.8%. Although most YOLOv5 results were similar, and all within the margin of error, the

Table 5.7: Fusion Technique Comparison based on mAP_{50} over holdout test dataset.

	Multiply			Mean		
Model-Size	R	G	B	R	G	B
YOLOv5-m						
- Itti	86.3	84.1	86.6	86.0	86.0	85.9
- Deepgaze	84.8	85.6	85.2	86.6	85.9	<u>86.8</u>
- ResNet	<u>86.8</u>	86.4	86.3	85.5	86.7	85.4
- SWIN	86.2	85.4	86.2	85.8	86.3	85.6
- Depth	85.8	85.1	86.3	85.3	84.7	85.3
YOLOv7-tiny						
- Itti	85.6	86.2	84.8	**88.0**	86.4	86.1
- Deepgaze	86.8	87.1	86.7	86.4	86.8	86.1
- ResNet	85.9	86.0	86.6	87.5	86.5	86.4
- SWIN	86.2	86.0	86.9	87.5	85.8	86.6
- Depth	86.4	87.0	85.4	85.7	85.3	85.1

peak 85.38% mAP_{50} achieved on the RGB dataset was considerably exceeded by several enhanced configurations, still pointing towards the improvement of top models through fusion.

From the consistent gains in YOLOv7-tiny performance, and mostly similar results for YOLOv5m, we conclude that the red channel combined with mean fusion provides the highest potential for added performance. These particular improvements in the R channel accuracy are attributed to the lack of original usage of the channel. Specifically, since deep water sites lack a lot of red light, it leaves unused bandwidth, which may be leveraged by fusion techniques to inject additional information, without much loss to the original channel. Although the results were still extremely close, new heights in accuracy were achieved, calling for further investigation.

5.3.3 Detailed Estimation Technique Comparison

The unique effect of each technique on the red-channel performance will be further validated through 5-fold cross-validation. This additional testing phase will assist in verifying the consistency and reliability of the observed trends, particularly in the comparison between the several estimation and fusion techniques.

The results of this validation are presented in Table 5.8. For YOLOv7-tiny, the Mean SWIN achieved the highest mAP_{50} (86.34%), while the Mean Depth model recorded a notable performance in $mAP_{50:95}$, particularly through the noticeable improvement in broken

Table 5.8: Comparison of resultant detection performance between mean and multiplication channel fusion based on 5-Fold CV.

Model	mAP_{50}	$mAP_{50:95}$			
		Avg	Whole	Broken	Buried
YOLOv7-tiny (R)					
- Mean SWIN	**86.34**	51.02	66.36	42.96	43.72
- Mean Depth	86.18	<u>51.64</u>	67.08	43.76	44.04
- Mean ResNet	86.08	50.76	66.24	42.88	43.20
- Mean Deepgaze	86.06	50.96	66.30	43.18	43.34
- Mean Itti	86.00	50.72	66.28	43.20	42.68
- Multiply Itti	86.12	50.90	66.16	42.96	43.56
- Multiply Deepgaze	86.08	50.80	66.12	43.04	43.32
- Multiply SWIN	85.80	51.06	66.44	43.30	43.38
- Multiply ResNet	85.80	51.12	66.38	43.34	43.64
- Multiply Depth	85.66	51.32	66.18	43.14	44.02
YOLOv5-m (R)					
- Mean Depth	85.66	52.75	68.42	44.76	45.06
- Mean Itti	85.16	52.62	68.16	44.91	44.77
- Mean SWIN	85.14	52.41	68.00	44.76	44.46
- Mean ResNet	85.12	52.55	68.21	44.38	45.05
- Multiply Itti	<u>85.69</u>	**52.80**	67.97	45.57	44.86
- Multiply ResNet	85.36	52.57	68.14	44.85	44.73
- Multiply Depth	85.34	52.68	68.36	44.58	45.10
- Multiply SWIN	85.12	52.36	68.08	44.65	44.35

and buried object accuracy. On a more general note, comparing mean and multiplication results, the mean tends to be marginally higher than the multiplication.

For YOLOv5m, the Itti multiplication technique achieved the highest $mAP_{50:95}$, closely followed by the mean depth technique. The results among the fusion techniques were very similar with YOLOv5, indicating that the initial observation of minimal impact from fusion methods is consistent.

Comparing these results to the RGB Cross-validation results, YOLOv7-tiny originally achieved mAP_{50} (86.14%) which through enhancement had a very slight increase to mAP_{50} (86.34%). On the other hand, YOLOv5-m had an original top 85.38% mAP_{50} which became 85.69% and retained the 52.80% $mAP_{50:95}$.

5.3.4 Weighted Mean Parameter Comparison

The effects of varying the weighting of maps for the weighted mean were analysed with particular attention to YOLOv7-tiny due to its particular preference towards mean fusion. For comparison, the YOLOv5m depth channel was also included, being the top mean fusion result. The results are presented in Table 5.9, which displays weighted means at intervals of 0.25, 0.5 and 0.75 for the three top estimation techniques, namely Depth, InSPyReNet-Swin and InSPyReNet-ResNet.

Table 5.9: 5-Fold Results by variation of map weight for weighted mean fusion.

Model	mAP_{50}	$mAP_{50:95}$			
		Avg	Whole	Broken	Buried
YOLOv7-tiny (R)					
- 0.75 Depth	86.34	51.04	66.42	43.00	43.66
- 0.75 ResNet	86.00	50.90	66.32	42.84	43.56
- 0.75 SWIN	85.92	50.78	66.38	43.18	42.80
- 0.50 Depth	86.18	<u>51.64</u>	67.08	43.76	44.04
- 0.50 ResNet	86.08	50.76	66.24	42.88	43.20
- 0.50 SWIN	85.88	51.08	66.46	43.28	43.50
- 0.25 Depth	**86.38**	51.14	66.40	43.28	43.76
- 0.25 ResNet	85.84	50.80	66.02	42.80	43.50
- 0.25 SWIN	86.14	50.74	66.00	42.54	43.64
YOLOv5-m (R)					
- 0.75 Depth	85.64	**52.92**	68.35	45.10	45.31
- 0.50 Depth	<u>85.66</u>	52.75	68.42	44.76	45.06
- 0.25 Depth	85.40	52.84	68.44	44.95	45.14

For the YOLOv7-tiny (R) model, the 0.25 depth weight configuration yields the highest mAP_{50} with a value of 86.38%, being the highest value achieved for fold-based techniques. However, the 0.50 weight configuration achieves the highest $mAP_{50:95}$ value at 51.64%. In the case of YOLOv5m, the 0.75 depth weight configuration stands out with the highest $mAP_{50:95}$ average of 52.92%, being the highest $mAP_{50:95}$ score achieved.

In a similar pattern to the previous sections, although the difference between techniques is still within the margin of error, this fine-tuning still allows the model to reach new highest scores, emphasising the value in correctly finding the sweet spot for fusing the information.

Table 5.10: Comparison of mAP_{50} and $mAP_{50:95}$ for 4-Channel techniques methods using modified YOLOv7 model.

Model	mAP_{50}	$mAP_{50:95}$			
		Avg	Whole	Broken	Buried
YOLOv7					
- BGRD	85.45	50.51	65.26	40.99	45.28
- BGRS SWIN	84.35	49.02	63.29	39.56	44.19
- BGRS ResNet	84.95	50.36	63.96	42.20	44.93
- BGRS Itti	85.20	50.52	65.28	40.67	45.60
- BGRS Deepgaze	**85.65**	**51.04**	64.62	41.46	47.04

5.3.5 RGB-D Model Comparison

Following the successful improvement in accuracy, a four-channel version of YOLOv7 was implemented, integrating the three original RGB channels with a fourth depth channel. From the results shown in Table 5.10, this configuration generally did not achieve better results, whereas the individualised fusion techniques were typically better across the board. There are two primary reasons which are linked to this decrease in performance. Firstly, several image augmentation procedures had to be turned off to accommodate the four-channel inputs due to their assumption of standard image shapes. Augmentations are crucial for improving model robustness and generalisation, and their absence likely will have a contribution to the lower performance. Additionally, the different architecture of the four-channel model could not leverage pretrained weights, which are generally available for three-channel configurations, which as was seen in our second experiment (Table 5.3) generally improved the performance of the models.

In summary, although the inclusion of four channels meant none of the original image channels were sacrificed, the practical limitations of implementing this different architecture resulted in decreased performance. These results increase the importance of the dedicated fusion techniques, which inherently guide the model into better utilising the auxiliary features available, whilst retaining the original structures.

5.4 Stage 3 (04) - Detection Localisation

This final stage targets the evaluation and analysis of the 2D-3D projection procedures, which considerably increase the interpretability and overall value of the detection techniques. The evaluation is primarily targeted at calculating the retained performance after the base models have gone through the projection. This also compares the effects of

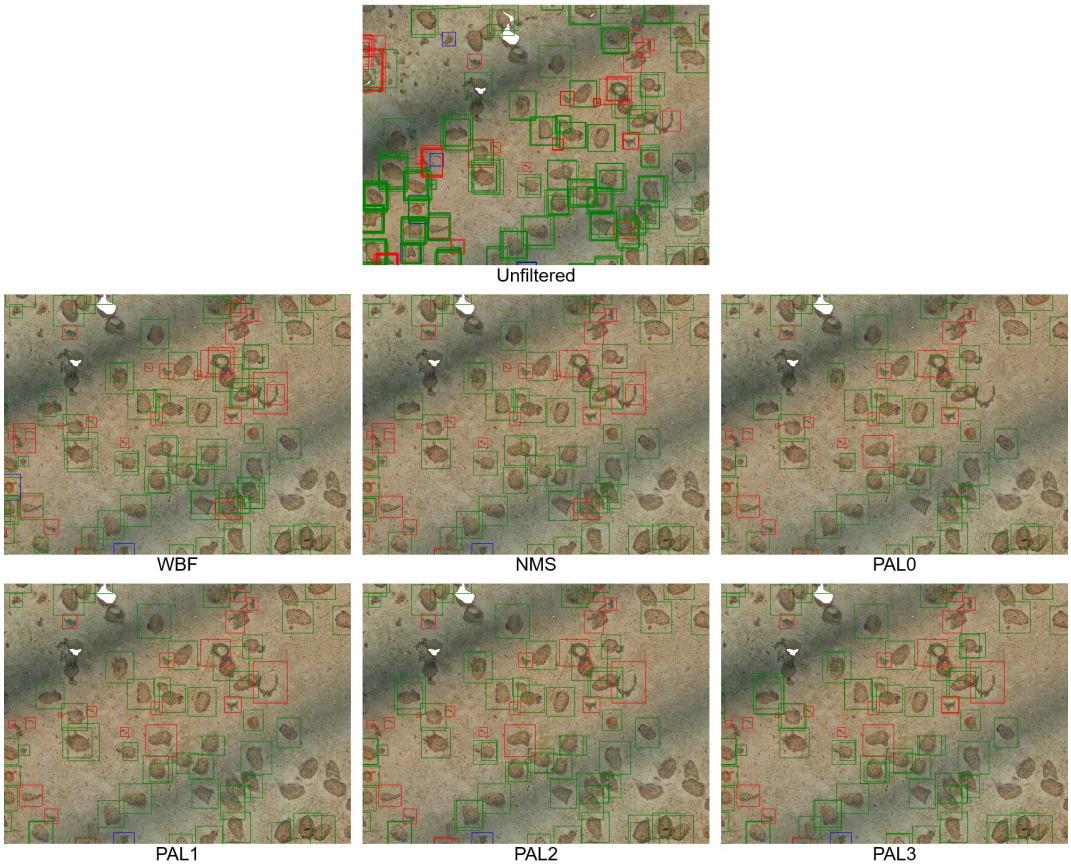


Figure 5.5: Comparison of Box suppression techniques to remove duplicate detections.

the several suppression techniques mentioned, requiring the balance between retained samples and removing duplicates.

5.4.1 Suppression Technique comparison

In this section, we evaluate various suppression techniques applied to the YOLOv7 model to manage overlapping detections and reduce duplicates in projected annotations. These techniques aim to retain the most relevant detections while minimising redundancy, which is crucial given the absence of absolute ground truth data. The key metrics considered in this evaluation include mAP at different IoU thresholds, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the number of retained ground truth (GT) annotations and predictions. A visual representation of these results may also be seen in Figure 5.5, whilst the metrics are presented in Table 5.11.

The unfiltered method, shows a good mAP_{50} of 72.1%, yet a sharp drop in $mAP_{50:95}$

Table 5.11: Comparison of different suppression techniques based on best YOLOv7.

Method	mAP_{50}	$mAP_{50:95}$	Error (mm)		Retained	
			MAE	RMSE	GT	PRED
Unfiltered	72.1	10.9	5.15 ± 4.04	6.54	2328	2240
NMS	53.0	15.1	5.15 ± 4.04	6.54	767	714
WBF	40.9	11.8	5.23 ± 4.30	6.77	1067	894
PAL0	56.1	24.0	4.33 ± 3.04	5.29	571	551
PAL1	56.6	22.2	4.37 ± 3.13	5.37	650	628
PAL2	58.9	20.8	4.33 ± 3.16	5.36	753	728
PAL3	61.3	20.0	4.32 ± 3.14	5.34	833	824

at 10.9%. Without any suppression, all detections are retained, including duplicates and overlapping predictions, which results in a higher detection accuracy at a 50% IoU threshold due to the leniency in matching multiple detections to the same ground truth instance. However, the accuracy considerably drops at stricter IoU thresholds because the overlap and duplicates lead to false positives, thus reducing the overall precision. The error metrics (MAE of 5.15 ± 4.04 and RMSE of 6.54) serve as a baseline reference, reflecting the inherent noise and redundancies in the unfiltered dataset.

NMS results in a substantially lower mAP_{50} (53.0) but a higher $mAP_{50:95}$ (15.1). This reduction in mAP_{50} is caused by NMS removing many overlapping detections, and hence the inflated numbers due to duplicated correct detections are removed. However, the increase in $mAP_{50:95}$ suggests improved performance at stricter IoU thresholds. Even confirmed visually, NMS effectively eliminates many redundant detections, leading to cleaner and more accurate predictions under strict conditions. However, there are still some notable duplicates, particularly in cases where one box is completely encapsulated in another. The error metrics remain similar to the unfiltered method, indicating that NMS does not affect localisation precision. The substantial reduction in retained counts (32.9% GT, 31.9% predictions) demonstrates NMS's aggressive approach to eliminating duplicates, which is beneficial in reducing redundancy but may risk losing some true positives.

WBF shows a further reduction in mAP_{50} (40.9) and in this case slightly lower $mAP_{50:95}$ (11.8) than NMS. The lower mAP_{50} indicates that WBF might change the boxes too much, resulting in fewer total detections and potentially missing some true positives. The slightly higher error metrics (MAE of 5.23 ± 4.30 and RMSE of 6.77) suggest that WBF introduces some localisation inaccuracies during the fusion process. WBF retains more annotations than NMS (38.0% GT, 33.2% predictions), indicating a more conservative approach, which is also seen in the many duplicates present in the visualisation. This method was not found

to be as performant as NMS.

The Polygonal Area Locking at different thresholds, (PAL0, PAL1, PAL2, PAL3) show progressive retention rates according to the leniency. The varying retention rates among PAL methods reflect their incremental approach in handling overlaps, PAl0 being the strictest, keeping the least amount of detections from any method. On the other hand, PAL3 is the most lenient, retaining many detections, whilst potentially keeping redundant ones. However, these are all done whilst retaining the highest mAP values amongst other suppression techniques, indicating higher consistency. The PAL methods generally improve distance metrics as well, with PAL0 showing the lowest MAE (4.33 ± 3.04) and RMSE (5.29), indicating better localisation accuracy than all other techniques.

These results have been further considered using a visual analysis on Figure 5.11. We may see that with PAL0, the overlap is practically all removed, yet there is also visible loss in detections compared to the unfiltered and other methods, suggesting it may be too strict. On the other hand, PAL2 and PAL3 retain a noticeable amount of duplicates, which suggests that these may be too lenient, and allow unneeded overlap. For this reason, PAL1 was found to provide the best balance between retention, accuracy and removal of duplicates for our use case. That being said, other methods or PAL thresholds might still be useful in scenarios requiring more aggressive duplicate removal, despite the potential loss in overall detection accuracy and retention rates, where the use case at hand needs to be considered. It is important to note that these metrics are not absolute values, as the ground truth annotations undergo the same suppression process. Thus, these results should be interpreted as relative comparisons derived from 2D detection performance. Further refinement and validation against more robust ground truth data would be necessary to draw definitive conclusions on absolute accuracy, yet with the results obtained, we can confidently say these results are satisfactory to our intended use case.

5.4.2 IoU Distance Relation Analysis

With the aim of better interpreting the relationship between IoU and the Haversine distance, an analysis was performed to study what a specific IoU value translates to in metric distance deviation. This analysis helps us understand how the spatial accuracy of detected objects correlates with their overlap ratio with ground truth annotations.

Figure 5.6 presents a scatter plot illustrating the relationship between the Haversine distance and the IoU of boxes on the orthomodel. Each point in the scatter plot represents a matched pair of a predicted bounding box and its corresponding ground truth box. These matches were carried out using an IOU threshold of 50%, meaning this is the minimum IoU considered. This allows us to know how mAP on the orthomosaic translates to distance

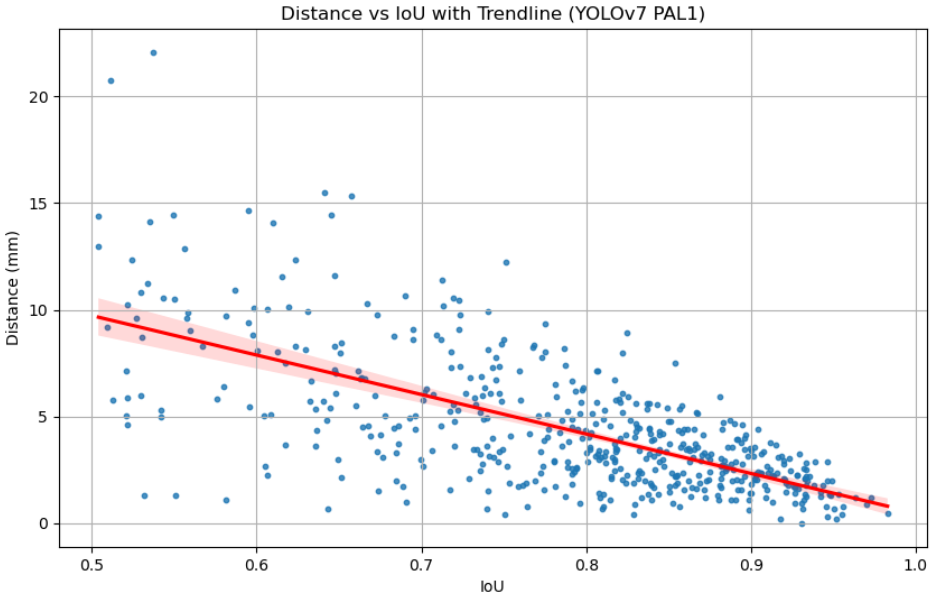


Figure 5.6: Scatter plot showing the relationship between Haversine distance and IoU.

deviation between centres of projected bounding boxes.

From the scatter plot, several observations can be made. There is an expected inverse correlation between IoU and Haversine distance. Higher IoU values are generally associated with smaller Haversine distances, indicating that well-overlapping boxes are also closer in spatial terms. Clusters of points at high IoU and low distance suggest that many predictions are both accurate and well-localised. Outliers, on the other hand, represent cases where predictions, despite having a decent IoU, have spatially distant centre points.

The observed inverse correlation between IoU and Haversine distance is expected due to several reasons. Higher IoU values indicate that the predicted and ground truth boxes share a significant overlap area, which naturally reduces the spatial distance between their centres. Additionally, a well-localised prediction will have its centre close to the ground truth centre, resulting in both a high IoU and a low Haversine distance. In conclusion, still being on a millimetric scale, these values are found promising, where even bounding boxes which partially overlap will still be in close proximity, and provide useful localisations.

5.5 Discussion

The section focuses on the presentation and analysis of additional results, aimed at better understanding the relative strengths and limitations experienced by these techniques. This is first achieved through a more thorough analysis of the best model results, presenting a confusion matrix linking back to per-class analysis. These findings are then compared to similar works to analyse how our findings align with the rest of the research. Subsequently, a gradient visualisation is performed meant to study which specific parts the model focuses on, and how this information propagates through the different layers. Finally, a set of sample prediction images is presented, aimed at visually studying how the object detector handles the many scenarios present in the dataset.

5.5.1 Per-Class Detection Analysis

Referring back to the definition of classes in Section 4.2.2, the metrics obtained confirm the expected differences between classes. Whole amphorae were found to be the easiest to detect, achieving mAP_{50} close to 90% and $mAP_{50:95}$ around 67%. Being the most valuable artefact to detect, these results are a very good sign of the real-world applicability of these techniques. Broken and buried pieces typically saw similar $mAP_{50:95}$ results of around 43%, which albeit being considerably less than whole amphorae, their difficult and fuzzy nature still makes this a satisfactory result.

Figure 5.7 presents a confusion matrix for predicted classes based on the holdout test set, and using the YOLOv7-tiny architecture using Red-Depth Weighted mean with $0.25w$. This confusion matrix reveals several key points about the model's classification performance, where notably it should be interpreted column-wise, where the weight of each row depends on the distribution of the column and adds up to a total of 1.

On a general note, the model shows high accuracy in predicting the correct classes, with values being all over 84%. As may be noted, the whole class possesses the highest accuracy. It was wrongly predicted as buried 6% of the time, as broken 3% of the time, and completely missed a prediction 3% of the time. This shows that the class is relatively distinct. As expected slightly higher confusion with buried is seen, where the fuzzy definitions between completely buried and completely unburied may contribute to this. When it comes to the buried class, this saw 84% of the correct class, whilst seeing noticeable confusion with whole predictions, and the background. These results also make sense considering the definitions, where very protruding buried amphorae may be interpreted as whole whilst very buried ones and barely visible may be missed by the model. Moreover, analysing the broken class, this saw the highest confusion with the background. Due

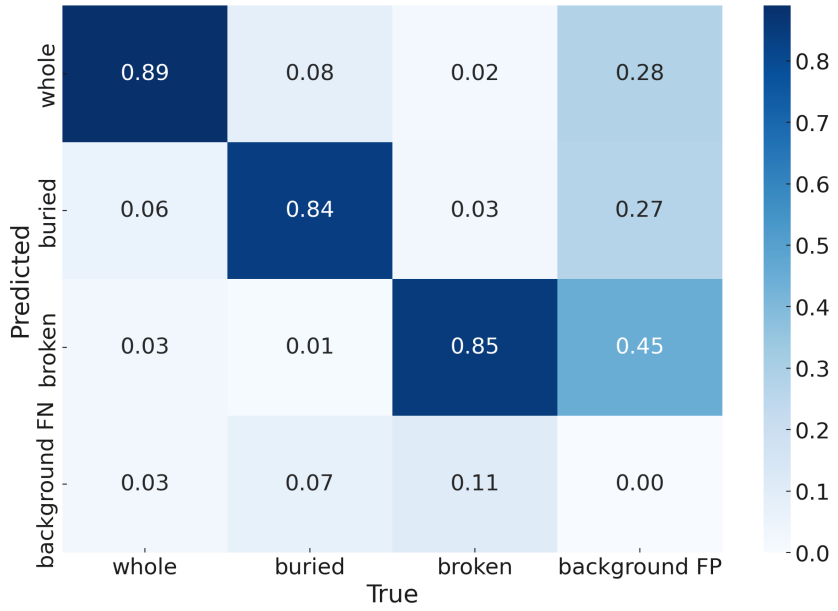


Figure 5.7: Confusion Matrix For Best YOLOv7 Configuration in mAP_{50}.

to the smaller shard-like nature of some artefacts, this is also an understandable result, in that they are considerably harder to find a pinpoint than say whole amphorae. Finally, analysing the background false positives, the highest value of this was on the broken class, where 45% of false positives (not representative of the amount) the background was wrongly predicted as a broken piece. This is likely linked to noise and debris being wrongly detected as small shards. All in all, the results are good, and confirm our initial expectations about the difficulty of each class and the fuzzy definition between classes.

5.5.2 Literature Comparison

In this section, we compare our results with those from existing literature to contextualise our findings and highlight similarities and differences.

Our best-performing model was YOLOv7-tiny, which achieved an 86.38% mAP_{50} and 51.14% $mAP_{50:95}$ on our original dataset. This performance is consistent with findings in other studies, especially considering the relatively limited size of our dataset.

Our work was first compared to the study by Paraskevas *et al.* (2023) investigated various YOLOv8 model sizes for detecting underwater pottery and found that the *small* variant performed best with a 75.5% mAP. Similarly, we observed that smaller models, such as YOLOv7-tiny, performed better in our context, likely due to the smaller dataset size, which aligns with the author's findings. Our accuracy was found to be consider-

ably higher, yet this is specific to the dataset being used. In another study, Yang *et al.* (2023) developed a custom feature extraction network, MDLA-DCN, achieving a 92.8% mAP on an original underwater cultural artefact dataset. Despite the difference in specific application and dataset, our results of 86.38% mAP_{50} are within a comparable range. Furthermore, Al-anni and Drap (2024) used YOLOv4 for detecting underwater archaeological artefacts, reporting around 85-87% performance in a single-class detection task, though without specifying the metric used. Our results with YOLOv7-tiny fall within this range, yet ours was performed on multi-class objects with difficult shapes, being a considerably harder problem to tackle, making our results better.

Separately, Ophoff *et al.* (2019) explored the integration of depth information with state-of-the-art object detectors, finding that mid-to-late layer fusion considerably improved detection results over traditional models like YOLOv2. In our study, incorporating depth information yielded minor improvements, particularly in the case of YOLOv7-tiny with a 0.25 depth weighting. This supports the findings, in that depth information can enhance detection performance, yet it still calls for better methods of fusion, perhaps more neurally integrated within the model itself. On the other hand, saliency-based object detection, such as the approach by Katyal *et al.* (2018), found saliency to enhance detection performance in foggy conditions. In our experiments, integrating saliency methods such as SWIN and ResNet with YOLOv7-tiny resulted in comparative performance, though the depth-enhanced models generally performed better. This suggests that while saliency can be beneficial, this discrepancy is marginal, and whether the small accuracy gain is worth the performance trade-off should be considered.

5.5.3 Gradient Visualisation

Figure 5.8 shows specific activations at the particular layers within the model as presented by XGrad-CAM. At layer 5, the information is still at a relatively high resolution, and we may see smaller and more accurate highlights towards the edges of objects and visually salient objects such as the brightly lit rope and noise. At the tenth layer, the resolution is considerably decreased. Heavier focus is put towards edges, and noise is less highlighted than before. We may also see a preference towards the edges of broken pieces, likely linked to the differentiation of classes. At the final layer, the focus is shifted towards global features, where larger chunks of activation are seen. Interestingly, being in the head, this seems to be a negative layer, where the main artefacts are not highlighted, but surrounding areas are. This may be attributed to the model finding where the bounds of the entire objects are, and hence focusing on the gaps between them.

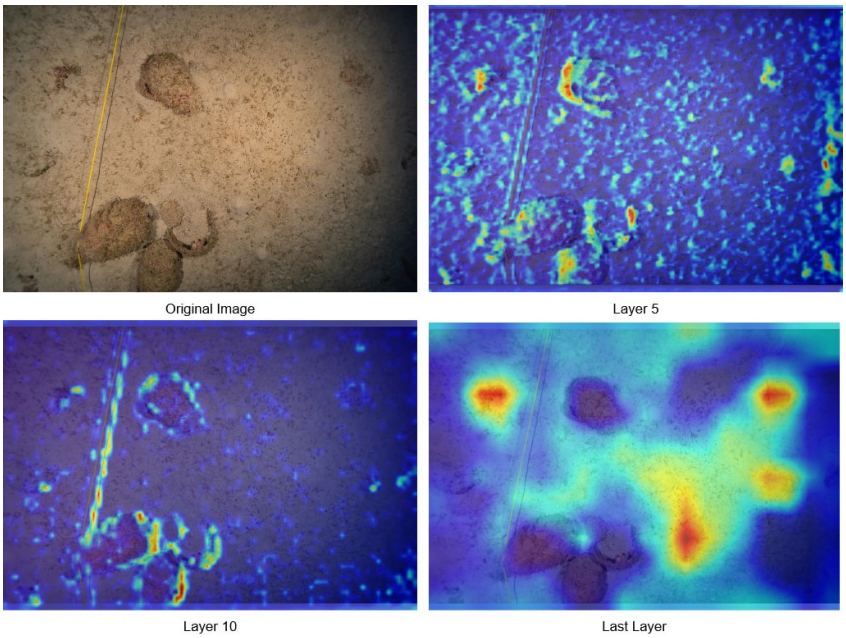


Figure 5.8: Gradient Activation Visualisation at Specific Layers.

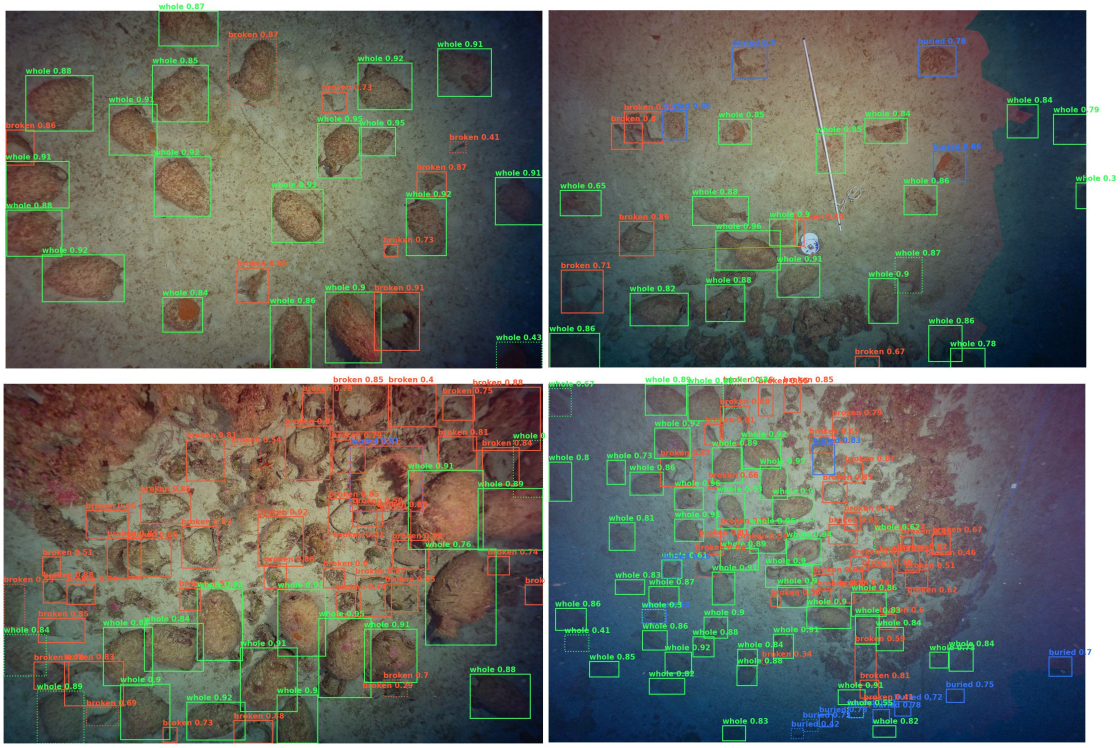


Figure 5.9: YOLOv7-tiny on Depth-Mean Fusion Predictions across several scenarios. Correct detections are indicated with solid lines, while incorrect predictions are shown in dashed boxes.

5.5.4 Visual Result Analysis

Figure 5.9 presents a sample of prediction results carried out by the YOLOv7-tiny model utilizing the 0.25w mean fusion between the red channel and depth. Four images are depicted, each highlighting notable scenarios.

The top-left image depicts several clearly distinguishable amphorae. The model localises and identifies most artefacts, achieving a precision of 88.89% and recall of 96.00%. This indicates the model's effectiveness in clear and unobstructed scenarios, where objects are well-defined and free from occlusion. In the top-right image, the introduction of debris and minor rock morphology increases detection complexity. Despite this, the model maintains high performance, with both precision and recall at 96.55%. This suggests the model's capability to handle moderately complex scenes effectively.

The bottom-left image presents a more challenging scenario with rocky outcrops and numerous broken pieces. Here, the model's performance drops slightly, with a precision of 83.61% and recall of 89.47%. The difficulty in distinguishing between rocks and broken pieces highlights the challenge in such environments. Nonetheless, the model correctly identifies a good number of objects in this complex scene. The bottom-right image is the most challenging case, taken at a higher altitude and different angle, resulting in varied lighting conditions. This image features a high number of objects, with 87 ground truth instances, making it highly convoluted. Despite these challenges, the model achieves a precision of 82.11% and recall of 89.66%. The good number of correct detections in such a complex scene indicates the model's ability to generalise to diverse examples.

Overall, the results demonstrate that the YOLOv7-tiny model performs well across a range of scenarios, from clear and unobstructed scenes to more complex environments. While the performance slightly decreases with increasing scene complexity, this is expected and still shows robust detection capabilities.

Finally, a full-scale orthomosaic visualisation presenting the predictions from YOLOv7-tiny on the best enhancement configurations, against the annotations and passed through PAL1 suppression is shown in Appendix A. This is based on what was deemed to be the best architecture, size, enhancement technique and suppression method, forming the best-observed set of annotations on the orthomosaic to maximise interpretability. Although some inconsistencies do exist, given the complexity of all the processes at work behind this representation, this is a great result, achieved through a completely automated process.

5.6 Summary

This chapter presented a detailed evaluation of the experiments, structured in a multi-stage approach to identify the most suitable model configurations, enhance detection performance using auxiliary information, and analyse the projection and suppression techniques for better interpretability of detection results.

In **Stage 1 (O2)**, various YOLO models (YOLOv5, YOLOv7, YOLOv8, YOLOv9), Faster R-CNN, and transformer-based models (RT-DETR, Deformable DETR) were evaluated on colour-corrected datasets with pretrained weights. YOLOv7-tiny achieved the highest performance with an mAP_{50} of 86.14%, while YOLOv5m provided the highest $mAP_{50:95}$ of 52.80%. Transformer-based models and Faster R-CNN underperformed compared to YOLO models. Colour correction and pretrained weights consistently improved performance, establishing YOLOv7-tiny and YOLOv5m as the top-performing architectures.

In **Stage 2 (O3)**, we integrated visual saliency and depth maps to enhance detection. Salient object detection provided an interesting baseline, where InSPyReNet with Res2Net50 backbone particularly outperformed other saliency techniques. Fusion techniques, specifically mean fusion on the red channel, improved performance, with YOLOv7-tiny achieving an mAP_{50} of 86.34%. Weight parameter analysis showed the 0.25 depth weight configuration for YOLOv7-tiny reached the highest mAP_{50} of 86.38%, and the 0.75 weight for YOLOv5m provided the best $mAP_{50:95}$ of 52.92%. These provided marginal yet consistent improvements over base models, showing their applicability to squeeze out more performance out of readily available information.

In **Stage 3 (O4)**, we focused on the retention of detection performance and suppression techniques in 3D projection. NMS reduced duplicates but retained fewer true positives, whereas WBF lost localisation information. Our hand-crafted PAL technique, particularly PAL1, was deemed to be the most appropriate technique for our use case providing an ideal balance, improving $mAP_{50:95}$ while reducing error metrics.

In conclusion, this comprehensive evaluation demonstrated the effectiveness of single-shot models for underwater artefact detection, the benefits of integrating auxiliary depth and saliency information, and the importance of effective suppression techniques for accurate 3D localisation. These findings provide a robust foundation for further refinements and real-world applications in underwater archaeology.

6 Conclusion

This chapter presents final reflections on the research conducted, summarising the key findings, addressing the challenges and limitations, and proposing directions for future work to enhance the field of underwater archaeological object detection.

6.1 Summary

This research aimed to address the problem of detecting underwater archaeological objects. This was further expanded to explore methods to enhance detection accuracy and contextual understanding by leveraging advanced object detection models and integrating auxiliary information such as localisation, depth, and saliency maps.

To achieve this, a comprehensive dataset was first compiled by collaborating with the Department of Classics and Archaeology to obtain images and a photogrammetric model from the Tower Wreck in Xlendi, Gozo. Annotation sessions with field experts resulted in a multi-class dataset covering 625m² using 864 images. Artefacts were classified according to their level of preservation, namely whole, buried, or broken. This dataset provided a solid foundation for training and evaluating detection models.

Various object detection models were evaluated, including YOLO architectures (YOLOv5, YOLOv7, YOLOv8, YOLOv9), transformer based models (DETR, RT-DETR, Deformable DETR), and two-stage techniques (Faster R-CNN). The purpose of these comparisons was to identify the most effective model types for our specific use case of underwater archaeological object detection. The findings revealed that YOLOv7-tiny achieved the highest overall performance with an mAP_{50} of 86.14%, while YOLOv5m provided the highest $mAP_{50:95}$ of 52.80%. Transformer based models and Faster R-CNN generally underperformed compared to single-shot YOLO models, exhibiting longer training times and unreliable convergence. These comparisons helped determine the most suitable models for the specific challenges of underwater environments.

Next, the impact of integrating visual saliency and depth maps on detection performance was explored. Saliency estimation techniques included Itti, Deepgaze, and In-

SPyReNet, whereas the depth was estimated using the photogrammetric model. A salient object detection baseline was performed, which, although displayed elementary localisation ability, highlighted the requirement for specialised detection tools. Following this, several fusion techniques to include these into imagery for object detectors were investigated, including channel multiplication and weighted channel means. The purpose was to understand how auxiliary information could enhance detection performance in challenging underwater conditions. Marginal yet consistent gains in performance over base models were observed, particularly when averaging the depth and the red channels. This was theorised to stem from the underutilisation of the red channel in deep water scenes, which was used as bandwidth to integrate the structural depth information.

Finally, efforts were focused on increasing the interpretability of detections using automated localisation techniques. Photogrammetric projection was utilised to extract geographic coordinates and contextual information from detected objects. Furthermore, an integrated visualisation tool to show all the detections on an orthomosaic of the area was developed, presenting its own set of difficulties and duplicate detections. Various suppression techniques, namely NMS, WBF, and the proposed PAL method, were compared on their ability to reduce duplicates. PAL was found to provide the best balance, retaining more detections while improving $mAP_{50:95}$ compared to others. These comparisons aimed to enhance the practical applicability of the methods towards end users by providing more digestible results.

6.2 Critique and Limitations

While the research presented notable advancements, several limitations were encountered. The absence of detailed ground truth data on orthomosaic annotations limited the depth of our projection analysis. Having precise ground truth annotations would have allowed for more rigorous evaluation of the projection accuracy, and localisation retention. Additionally, other fusion methods, perhaps more learnable or tunable, could have been explored. Custom auxiliary branches within models dedicated to additional channels might have provided better integration and performance such as found in similar literature. Additionally, the computational resources required for training and evaluating deep learning models were substantial, constraining the extent and speed of experimentation. Finally, albeit showing gains in accuracy, the overhead involved in generating these auxiliary maps limits the feasibility of real-time application, which in certain scenarios is necessary for dynamic underwater exploration missions or automated operations.

6.3 Objectives Achieved

O1: Compile a Maritime Archaeology dataset. A multi-class object detection dataset based on the Tower Wreck archaeological site in Xlendi, Gozo was formulated. Field experts annotated these images, covering various classes of artefacts according to their level of preservation, namely whole, buried, or broken. The dataset further covered varying scenarios, including dense, sparse and rocky areas, all with their own implications. This comprehensive dataset provided a solid foundation for training and evaluating detection models, ensuring the diversity and quality required for robust model development.

O2: Evaluate and compare object detection techniques. Various object detection models, including YOLO architectures, transformer models, and two-stage techniques, were evaluated. The comparison concluded that YOLOv7-tiny was the most effective model for underwater archaeological object detection, outperforming transformer models and two-stage techniques in terms of resultant predictive performance. Transformer models and two-stage techniques generally require larger datasets to fully leverage their capabilities, which made them not as compatible with our smaller data size. Additionally, the use of colour correction and pretrained weights were found to typically be beneficial for training performance, enhancing the overall detection accuracy. This evaluation helped determine the suitability of different model architectures for the specific challenges posed by underwater environments.

O3: Explore the impact of auxiliary data. The integration of visual saliency and depth maps into object detection models was explored. Marginal yet consistent improvements in performance were observed, particularly with the YOLOv7-tiny model, when auxiliary data was included. This indicated that auxiliary information such as saliency and depth maps can enhance detection accuracy in challenging underwater conditions, albeit the gains were modest.

O4: Increase interpretability using localisation techniques. Automated localisation techniques were developed using photogrammetric projection to extract geographic coordinates and contextual information from detected objects. An integrated visualisation tool was developed to display all detections on an orthomosaic of the area, facilitating the practical use of detection results in archaeological analysis.

6.4 Future Work

Whilst bridging the gap between archaeology and automated interpretation tools, several paths for future work emerge, which have the potential to increase both the efficacy and the practicality of these techniques.

One promising direction of research is the utilisation of the 3D model generated from our photogrammetric data to create synthetic views or virtual data, similar to techniques used in multi-view fusion. This approach could enrich the dataset and improve model robustness. Additionally, collecting diversified data with multiple views will help generalise the models and enhance their performance across different archaeological contexts. Exploring the applicability of neural 3D reconstruction, as opposed to traditional photogrammetric techniques, could also provide an interesting research path. Recent advancements propose the ability to cut down processing time while increasing the visual fidelity of the final models.

Integrating saliency and depth estimation directly into the detection models, rather than as separate techniques merged during preprocessing, is another promising direction. This could enhance performance by promoting more learnable enhancements, potentially improving both speed and accuracy. Additionally, comparing the estimated depth maps to those collected by specialised equipment could provide more accurate data than ones estimated by photogrammetry.

Comparing these results with SAHI or another method of performing detection directly on the orthomosaic could provide valuable insights into the relative strengths and weaknesses of our approach, which relies on projection. Such comparative analysis could help potentially circumvent the many processes and corrections associated with accurately projecting data from 2D images to 3D scenery.

Finally, the next logical step for increasing the archaeological value of our work is to attempt to identify detections by their amphora typology. This would heighten the analytical value and reduce manual labour considerably. In fact, work has already started on building upon the findings of this study to target more detailed typology identification of whole artefacts, providing a clear path towards increased real-world impact and utilisation.

References

AgiSoft LLC. Agisoft metashape, version 2.0.2, 2023. URL https://www.agisoft.com/.

Akkaynak, D. and Treibitz, T. Sea-thru: A method for removing water from underwater images. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1682–1691, 6 2019. ISSN 10636919. doi: 10.1109/CVPR.2019.00178.

Akyon, F. C., Altinuc, S. O., and Temizel, A. Slicing aided hyper inference and fine-tuning for small object detection. *Proceedings - International Conference on Image Processing, ICIP*, pages 966–970, 2022. ISSN 15224880. doi: 10.1109/ICIP46576.2022.9897990.

Al-anni, M. K. and Drap, P. Efficient 3d instance segmentation for archaeological sites using 2d object detection and tracking. *International Journal of Computing and Digital Systems*, 15:1333–1342, 3 2024. ISSN 2210-142X. doi: 10.12785/IJCDS/150194.

Anastasi, M., Capelli, C., Gambin, T., and Sourisseau, J. C. The xlendi bay shipwreck (gozo, malta): A petrographic and typological study of an archaic ceramic cargo. *Libyan Studies*, 52:166–172, 11 2021. ISSN 20526148. doi: 10.1017/LIS.2021.16.

Anichini, F. and Gattiglia, G. Big archaeological data. the archaide project approach. *Associazione consortium GARR, Pisa*, page 3, 2018.

Athira, P., Haridas, T. P. M., and Supriya, M. H. Underwater object detection model based on yolov3 architecture using deep neural networks. *2021 7th International Conference on Advanced Computing and Communication Systems, ICACCS 2021*, pages 40–45, 3 2021. doi: 10.1109/icaccs51430.2021.9441905.

Bainbridge, S. and Gardner, S. Comparison of human and camera visual acuity–setting the benchmark for shallow water autonomous imaging platforms. *Journal of Marine Science and Engineering 2016, Vol. 4, Page 17*, 4:17, 2 2016. ISSN 2077-1312. doi: 10.3390/JMSE4010017.

Barber, C. B., Dobkin, D. P., and Huhdanpaa, H. The quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software (TOMS)*, 22:469–483, 12 1996. ISSN 00983500. doi: 10.1145/235815.235821.

Beijbom, O., Edmunds, P. J., Kline, D. I., Mitchell, B. G., and Kriegman, D. Automated annotation of coral reef survey images. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1170–1177, 2012. ISSN 10636919. doi: 10.1109/cvpr.2012.6247798.

Bennett, J. Less than 1 percent of the world's shipwrecks have been explored. *Popular mechanics*, 2016.

Bianco, G., Gallo, A., Bruno, F., and Muzzupappa, M. A comparative analysis between active and passive techniques for underwater 3d reconstruction of close-range objects. *Sensors 2013, Vol. 13, Pages 11007-11031*, 13:11007–11031, 8 2013. ISSN 1424-8220. doi: 10.3390/S130811007.

Božić-Štulić, D., Kružić, S., Gotovac, S., and Papić, V. Complete model for automatic object detection and localisation on aerial images using convolutional neural networks. *Journal of Communications Software and Systems*, 14, 2018. ISSN 18466079. doi: 10.24138/JCOMSS.V14I1.441.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

Buslaev, A., Iglovikov, V. I., Khvedchenya, E., Parinov, A., Druzhinin, M., and Kalinin, A. A. Albumentations: Fast and flexible image augmentations. *Information*, 11, 2020. ISSN 2078-2489. doi: 10.3390/info11020125.

Cane, T. and Ferryman, J. Saliency-based detection for maritime object tracking. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 6 2016.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. End-to-end object detection with transformers. *European Conference on Computer Vision, Proceedings, Part I 16*, pages 213–229, 2020.

Chen, L., Jiang, Z., Tong, L., Liu, Z., Zhao, A., Zhang, Q., Dong, J., and Zhou, H. Perceptual underwater image enhancement with deep learning and physical priors. *IEEE Transactions on Circuits and Systems for Video Technology*, Pp:1, 1 2020a. doi: 10.1109/tcsvt.2020.3035108.

Chen, L., Zhou, F., Wang, S., Dong, J., Li, N., Ma, H., Wang, X., and Zhou, H. Swipenet: Object detection in noisy underwater scenes. *Pattern Recognition*, 132:108926, 12 2022. ISSN 0031-3203. doi: 10.1016/j.patcog.2022.108926.

Chen, Z., Gao, H., Zhang, Z., Zhou, H., Wang, X., and Tian, Y. Underwater salient object detection by combining 2d and 3d visual features. *Neurocomputing*, 391:249–259, 5 2020b. ISSN 0925-2312. doi: 10.1016/J.NEUCOM.2018.10.089.

Cheng, J., Dong, L., and Lapata, M. Long short-term memory-networks for machine reading. *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, pages 551–561, 1 2016. doi: 10.18653/v1/d16-1053.

Chu, J., Zhang, Y., Li, S., Leng, L., and Miao, J. Syncretic-nms: A merging non-maximum suppression algorithm for instance segmentation. *IEEE Access*, 8:114705–114714, 2020. ISSN 21693536. doi: 10.1109/ACCESS.2020.3003917.

Costa, E., Balletti, C., Beltrame, C., Guerra, F., and Vernier, P. Digital survey techniques for the documentation of wooden shipwrecks. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLI-B5: 237–242, 6 2016. ISSN 1682-1750. doi: 10.5194/ISPRS-ARCHIVES-XLI-B5-237-2016.

Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1:4171–4186, 10 2018.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*, 2021.

Dragomir, Dumitru, E., Christian, S., Scott, R., Cheng-Yang, F., Wei, B. A. C. L., and Anguelov. Ssd: Single shot multibox detector. *European Conference on Computer Vision*, pages 21–37, 2016.

Drap, P., Merad, D., Mahiddine, A., Seinturier, J., Peloso, D., Boï, J.-M., Chemisky, B., and Long, L. Underwater photogrammetry for archaeology. what will be the next step? *International Journal of Heritage in the Digital Era*, 2:375–394, 9 2013. ISSN 2047-4970. doi: 10.1260/2047-4970.2.3.375.

Drap, P., Merad, D., Hijazi, B., Gaoua, L., Nawaf, M. M., Saccone, M., Chemisky, B., Seinturier, J., Sourisseau, J. C., Gambin, T., and Castro, F. Underwater photogrammetry and object modeling: A case study of xlendi wreck in malta. *Sensors 2015, Vol. 15, Pages 30351-30384*, 15:30351–30384, 12 2015. ISSN 1424-8220. doi: 10.3390/s151229802.

Edward, T. R. and Drummond. Machine learning for high-speed corner detection. *European Conference on Computer Vision*, pages 430–443, 2006.

Erdem, E. and Erdem, A. Visual saliency estimation by nonlinearly integrating features using region covariances. *Journal of Vision*, 13:11, 4 2013. ISSN 1534-7362. doi: 10.1167/13.4.11.

Ertan, Z., Korkut, B., Gördük, G., Kulavuz, B., Bakırman, T., and Bayram, B. Enhancement of underwater images with artificial intelligence. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVIII-4-W9-2024:149–156, 3 2024. ISSN 1682-1750. doi: 10.5194/ISPRS-ARCHIVES-XLVIII-4-W9-2024-149-2024.

Espinosa, A. R., McIntosh, D., and Albu, A. B. An efficient approach for underwater image improvement: Deblurring, dehazing, and color correction. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pages 206–215, 1 2023.

Fang, Y., Liao, B., Wang, X., Fang, J., Qi, J., Wu, R., Niu, J., and Liu, W. You only look at one sequence: Rethinking transformer in vision through object detection. *Advances in Neural Information Processing Systems*, 34:26183–26197, 2021.

Fayaz, S., Parah, S. A., and Qureshi, G. J. Underwater object detection: architectures and algorithms – a comprehensive review. *Multimedia Tools and Applications*, 81:20871–20916, 6 2022. ISSN 15737721. doi: 10.1007/s11042-022-12502-1.

Feng, P. and Tang, Z. A survey of visual neural networks: current trends, challenges and opportunities. *Multimedia Systems*, 29:693–724, 4 2023. ISSN 14321882. doi: 10.1007/S00530-022-01003-8/TABLES/6.

Foresti, G. L. and Gentili, S. A vision based system for object detection in underwater images. *International Journal of Pattern Recognition and Artificial Intelligence*, 14:167–188, 2000. doi: 10.1142/S021800140000012X.

Fu, C.-Y., Liu, W., Ranga, A., Tyagi, A., and Berg, A. C. Dssd : Deconvolutional single shot detector. 1 2017.

Fu, C., Liu, R., Fan, X., Chen, P., Fu, H., Yuan, W., Zhu, M., and Luo, Z. Rethinking general underwater object detection: Datasets, challenges, and solutions. *Neurocomputing*, 517:243–256, 1 2023. ISSN 0925-2312. doi: 10.1016/j.neucom.2022.10.039.

Fu, R., Hu, Q., Dong, X., Guo, Y., Gao, Y., and Li, B. Axiom-based grad-cam: Towards accurate visualization and explanation of cnns. *31st British Machine Vision Conference*, 8 2020.

Gambin, T., Drap, P., Cheminsky, B., Hyttinen, K., and Kozak, G. Exploring the phoenician shipwreck off xlendi bay, gozo. a report on methodologies used for the study of a deep-water site. *Underwater Technology*, 35:71–86, 2018. ISSN 17560551. doi: 10.3723/ut.35.071.

Gambin, T. *A drop in the ocean - Malta's trade in olive oil during the Roman period*, pages 86–97. 1 2012.

Gambin, T., Hyttinen, K., Sausmekat, M., and Wood, J. Making the invisible visible: Underwater malta–a virtual museum for submerged cultural heritage. *Remote Sensing 2021, Vol. 13, Page 1558*, 13:1558, 4 2021. ISSN 2072-4292. doi: 10.3390/RS13081558.

Gambin, T., Sausmekat, M., Wood, J., and Hyttinen, K. When time is of the essence–recording an underwater excavation at 110 m. *Journal of Marine Science and Engineering 2023, Vol. 11, Page 1835*, 11:1835, 9 2023. ISSN 2077-1312. doi: 10.3390/JMSE11091835.

Gené-Mola, J., Sanz-Cortiella, R., Rosell-Polo, J. R., Morros, J. R., Ruiz-Hidalgo, J., Vilaplana, V., and Gregorio, E. Fruit detection and 3d location using instance segmentation neural networks and structure-from-motion photogrammetry. *Computers and Electronics in Agriculture*, 169:105165, 2 2020. ISSN 0168-1699. doi: 10.1016/J.COMPAG.2019.105165.

Girshick, R. Fast r-cnn. *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 580–587, 11 2013. ISSN 10636919. doi: 10.1109/CVPR.2014.81.

Grzadziel, A. Application of remote sensing techniques to identification of underwater airplane wreck in shallow water environment: Case study of the baltic sea, poland. *Remote Sensing 2022, Vol. 14, Page 5195*, 14:5195, 10 2022. ISSN 2072-4292. doi: 10.3390/rs14205195.

Guan, X., Luo, X., and Wang, D. Yolov5-uod: Underwater object detection method based on improved yolov5. *2023 13th International Conference on Information Science and Technology (ICIST)*, pages 157–166, 2023. doi: 10.1109/ICIST59754.

2023.10367136.

Guo, Y., Li, H., and Zhuang, P. Underwater image enhancement using a multiscale dense generative adversarial network. *IEEE Journal of Oceanic Engineering*, 45:862–870, 7 2020. ISSN 15581691. doi: 10.1109/joe.2019.2911447.

Hartley, R. and Zisserman, A. Multiple view geometry in computer vision. *Cambridge University Press*, 2:672, 2004. ISSN 05215405.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 770–778, 12 2015. ISSN 10636919. doi: 10.1109/CVPR.2016.90.

Heritage Malta. Tower wreck project, 6 2023. URL https://underwatermalta.org/discover/tower-wreck/.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 4 2017.

Hummel, R. Image enhancement by histogram transformation. *Computer Graphics and Image Processing*, 6:184–195, 4 1977. ISSN 0146-664X. doi: 10.1016/S0146-664X(77)80011-7.

Iglhaut, J., Cabo, C., Puliti, S., Piermattei, L., O'Connor, J., and Rosette, J. Structure from motion photogrammetry in forestry: a review. *Current Forestry Reports*, 5:155–168, 9 2019. ISSN 21986436. doi: 10.1007/S40725-019-00094-3.

Islam, M. J., Wang, R., and Sattar, J. Svam: Saliency-guided visual attention modeling by autonomous underwater robots. *Proceedings of Robotics: Science and Systems*, 11 2020a. doi: 10.48550/arxiv.2011.06252.

Islam, M. J., Xia, Y., and Sattar, J. Fast underwater image enhancement for improved visual perception. *IEEE Robotics and Automation Letters*, 5:3227–3234, 4 2020b. ISSN 23773766. doi: 10.1109/lra.2020.2974710.

Itti, L., Koch, C., and Niebur, E. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20:1254–1259, 1998. ISSN 01628828. doi: 10.1109/34.730558.

Jocher, G. Ultralytics yolov5, 2020. URL https://github.com/ultralytics/yolov5.

Jocher, G., Chaurasia, A., and Qiu, J. Ultralytics yolov8, 2023. URL https://github.com/ultralytics/ultralytics.

Jones, C. A. and Church, E. Photogrammetry is for everyone: Structure-from-motion software user experiences in archaeology. *Journal of Archaeological Science: Reports*, 30:102261, 4 2020. ISSN 2352-409X. doi: 10.1016/J.JASREP.2020.102261.

Katyal, S., Kumar, S., Sakhuja, R., and Gupta, S. Object detection in foggy conditions by fusion of saliency map and yolo. *Proceedings of the International Conference on Sensing Technology, ICST*, pages 154–159, 7 2018. ISSN 21568073. doi: 10.1109/ICSENST.2018.8603632.

Kerbl, B., Kopanas, G., Leimkuehler, T., and Drettakis, G. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42, 7 2023. ISSN 0730-0301. doi: 10.1145/3592433.

Kim, T., Kim, K., Lee, J., Cha, D., Lee, J., and Kim, D. Revisiting image pyramid structure for high resolution salient object detection. *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pages 108–124, 12 2022.

Kızıldağ, N. Mapping and recording of ancient shipwrecks by using marine remote sensing techniques: Case studies from turkish coasts. *Marine Science and Technology Bulletin*, 11:331–342, 2022. ISSN 2147-9666. doi: 10.33714/masteb.1144180.

Koch, C. and Ullman, S. Shifts in selective visual attention: towards the underlying neural circuitry. *Human neurobiology*, 4:219–227, 1985. ISSN 0721-9075. doi: 10.1007/978-94-009-3833-5_5.

Koot, R., Hennerbichler, M., and Lu, H. Evaluating transformers for lightweight action recognition. *arXiv preprint arXiv:2111.09641*, 2021.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 2012.

Kuhn, H. W. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2:83–97, 1955.

Lei, F., Tang, F., and Li, S. Underwater target detection algorithm based on improved yolov5. *Journal of Marine Science and Engineering 2022*, *Vol. 10*, *Page 310*, 10:310, 2 2022. ISSN 2077-1312. doi: 10.3390/jmse10030310.

Li, X., Shang, M., Qin, H., and Chen, L. Fast accurate fish detection and recognition of underwater images with fast r-cnn. *OCEANS 2015 - MTS/IEEE Washington*, 2 2016. doi: 10.23919/OCEANS.2015.7404464.

Liang, T., Chu, X., Liu, Y., Wang, Y., Tang, Z., Chu, W., Chen, J., and Ling, H. Cbnet: A composite backbone network architecture for object detection. *IEEE transactions on image processing*, 31:6893–6906, 2022. ISSN 1057-7149.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. Feature pyramid networks for object detection. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

Lingua, A., Marenchino, D., and Nex, F. Performance analysis of the sift operator for automatic feature extraction and matching in photogrammetric applications. *Sensors 2009*, *Vol. 9*, *Pages 3745-3766*, 9:3745–3766, 5 2009. ISSN 1424-8220. doi: 10.3390/S90503745.

Liu, C., Li, H., Wang, S., Zhu, M., Wang, D., Fan, X., and Wang, Z. A dataset and benchmark of underwater object detection for robot picking. *2021 IEEE International Conference on Multimedia and Expo Workshops*, *ICMEW 2021*, 2021a. doi: 10.1109/ICMEW53276.2021.9455997.

Liu, J., Liu, S., Xu, S., and Zhou, C. Two-stage underwater object detection network using swin transformer. *IEEE Access*, pages 1–1, 11 2022a. ISSN 21693536. doi: 10.1109/access.2022.3219592.

Liu, K., Sun, Q., Sun, D., Peng, L., Yang, M., and Wang, N. Underwater target detection based on improved yolov7. *Journal of Marine Science and Engineering*, 11, 2023. ISSN 2077-1312. doi: 10.3390/jmse11030677.

Liu, S., Zhou, H., Li, C., and Wang, S. Analysis of anchor-based and anchor-free object detection methods based on deep learning. *2020 IEEE International Conference on Mechatronics and Automation, ICMA 2020*, pages 1058–1065, 10 2020. doi: 10.1109/ICMA49215.2020.9233610.

Liu, Y., Sangineto, E., Bi, W., Sebe, N., Lepri, B., and Nadai, M. D. Efficient training of visual transformers with small datasets. *Neural Information Processing Systems*, 2021b.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021c.

Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., Wei, F., and Guo, B. Swin transformer v2: Scaling up capacity and resolution. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11999–12009, 2022b. doi: 10.1109/CVPR52688.2022.01170.

Lowe, D. G. Object recognition from local scale-invariant features. *Proceedings of the IEEE International Conference on Computer Vision*, 2:1150–1157, 1999. doi: 10.1109/ICCV.1999.790410.

Lu, H., Li, Y., Xu, X., Li, J., Liu, Z., Li, X., Yang, J., and Serikawa, S. Underwater image enhancement method using weighted guided trigonometric filtering and artificial light correction. *Journal of Visual Communication and Image Representation*, 38:504–516, 7 2016. ISSN 1047-3203. doi: 10.1016/J.JVCIR.2016.03.029.

Ma, D., Fang, H., Wang, N., Pang, G., Li, B., Dong, J., and Jiang, X. A low-cost 3d reconstruction and measurement system based on structure-from-motion (sfm) and multi-view stereo (mvs) for sewer pipelines. *Tunnelling and Underground Space Technology*, 141:105345, 11 2023. ISSN 0886-7798. doi: 10.1016/J.TUST.2023.105345.

Mallet, D. and Pelletier, D. Underwater video techniques for observing coastal marine biodiversity: A review of sixty years of publications (1952–2012). *Fisheries Research*, 154:44–62, 6 2014. ISSN 0165-7836. doi: 10.1016/j.fishres.2014.01.019.

Malumbres, M. P., Garrido, P. P., Calafate, C. T., and Gil, J. O. Underwater wireless networking techniques. *Encyclopedia of Information Science and Technology, Second Edition*, pages 3958–3864, 1 2009.

Marc, Gérard, S., Sébastien, V., Thomas, C., Sébastien, M. D. V., and Chaumont. Coral reef fish detection and recognition in underwater videos by supervised machine learning: Comparison between deep learning and hog+svm methods. *Advanced Concepts for Intelligent Vision Systems*, pages 160–171, 2016.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., and Ng, R. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 2020.

More, B. and Bhosale, S. A comprehensive survey on object detection using deep learning. *Revue d'Intelligence Artificielle*, 37:407–414, 4 2023. ISSN 0992499X. doi: https://doi.org/10.18280/ria.370217.

Nassar, A. S., D'Aronco, S., Lefèvre, S., and Wegner, J. D. Geograph: Graph-based multi-view object detection with geometric cues end-to-end. *European Conference on Computer Vision, Proceedings, Part VII*, pages 488–504, 2020. doi: 10.1007/978-3-030-58571-6_29.

OpenAI *et al.* Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2024.

Ophoff, T., Beeck, K. V., and Goedemé, T. Exploring rgb+depth fusion for real-time object detection. *Sensors 2019, Vol. 19, Page 866*, 19:866, 2 2019. ISSN 1424-8220. doi: 10.3390/S19040866.

Pan, X., Ge, C., Lu, R., Song, S., Chen, G., Huang, Z., and Huang, G. On the integration of self-attention and convolution. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 805–815, 11 2021. ISSN 10636919. doi: 10.1109/cvpr52688.2022.00089.

Pan, X., Ge, C., Lu, R., Song, S., Chen, G., Huang, Z., and Huang, G. On the integration of self-attention and convolution. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–825, 6 2022.

Paraskevas, K., Mariolis, I., Giouvanis, G., Peleka, G., Zampokas, G., and Tzovaras, D. Underwater detection of ancient pottery sherds using deep learning. *International Journal on Cybernetics & Informatics (IJCI) Vol.12, No.6*, 12:1, 10 2023. ISSN 2277548X. doi: 10.5121/IJCI.2023.120601.

Patacchiola, M. and Cangelosi, A. Head pose estimation in the wild using convolutional neural networks and adaptive gradient methods. *Pattern Recognition*, 71:132–143, 11 2017. ISSN 0031-3203. doi: 10.1016/J.PATCOG.2017.06.009.

Pedersen, M., Haurum, J. B., Gade, R., and Moeslund, T. B. Detection of marine animals in a new underwater dataset with varying visibility. *CVPR Workshops*, 2019.

Pizer, S. M., Amburn, E. P., Austin, J. D., Cromartie, R., Geselowitz, A., Greer, T., ter Haar Romeny, B., Zimmerman, J. B., and Zuiderveld, K. Adaptive histogram equalization and its variations. *Computer Vision, Graphics, and Image Processing*, 39: 355–368, 9 1987. ISSN 0734-189X. doi: 10.1016/S0734-189X(87)80186-X.

Qin, P., Li, C., Chen, J., and Chai, R. Research on improved algorithm of object detection based on feature pyramid. *Multimedia Tools and Applications*, 78:913–927, 1 2019. ISSN 15737721. doi: 10.1007/S11042-018-5870-3.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. You only look once: Unified, real-time object detection. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 779–788, 6 2015. ISSN 10636919. doi: 10.1109/CVPR.2016.91.

Reggiannini, M. and Moroni, D. The use of saliency in underwater computer vision: A review. *Remote Sensing 2021, Vol. 13, Page 22*, 13:22, 12 2020. ISSN 2072-4292. doi: 10.3390/rs13010022.

Ren, S., He, K., Girshick, R., and Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1137–1149, 6 2015. ISSN 01628828. doi: 10.1109/TPAMI.2016.2577031.

Schwarz, M., Schulz, H., and Behnke, S. Rgb-d object recognition and pose estimation based on pre-trained convolutional neural network features. *Proceedings - IEEE International Conference on Robotics and Automation*, pages 1329–1335, 6

2015. ISSN 10504729. doi: 10.1109/ICRA.2015.7139363.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 10 2017.

Seychell, D. and Debono, C. J. Ranking regions of visual saliency in rgb-d content. *2018 International Conference on 3D Immersion (IC3D)*, pages 1–8, 2018.

Sharma, V. K. and Mir, R. N. A comprehensive and systematic look up into deep learning based object detection techniques: A review. *Computer Science Review*, 38:100301, 11 2020. ISSN 1574-0137. doi: 10.1016/J.COSREV.2020.100301.

Solovyev, R., Wang, W., and Gabruseva, T. Weighted boxes fusion: Ensembling boxes from different object detection models. *Image and Vision Computing*, 107:104117, 3 2021. ISSN 0262-8856. doi: 10.1016/J.IMAVIS.2021.104117.

Spampinato, C., Chen-Burger, Y.-H. J., Nadarajan, G. D., and Fisher, R. B. Detecting, tracking and counting fish in low quality unconstrained underwater videos. *International Conference on Computer Vision Theory and Applications*, 2008.

Sun, Y., Wang, X., Zheng, Y., Yao, L., Qi, S., Tang, L., Yi, H., and Dong, K. Underwater object detection with swin transformer. *2022 4th International Conference on Data Intelligence and Security (ICDIS)*, pages 422–427, 8 2022. doi: 10.1109/icdis55630.2022.00070.

Tan, M. and Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. *Proceedings of the 36th International Conference on Machine Learning*, 97:6105–6114, 1 2019.

Tan, M. and Le, Q. Efficientnetv2: Smaller models and faster training. *Proceedings of the 38th International Conference on Machine Learning*, 139:10096–10106, 1 2021.

Tan, M., Pang, R., and Le, Q. V. Efficientdet: Scalable and efficient object detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10781–10790, 2020.

Tian, M., Wan, S., Ji, Y., and Yue, L. Salient objects detection in time sequenced images. *Proceedings of the International Joint Conference on Neural Networks*, pages 321–326, 2008. doi: 10.1109/IJCNN.2008.4633811.

Urick, R. J. *Principles of underwater sound*. 1975.

Vaswani, A., Brain, G., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

Wang, C.-Y., Bochkovskiy, A., and Liao, H.-Y. M. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7464–7475, 2023.

Wang, C.-Y., Yeh, I.-H., and Liao, H.-Y. M. Yolov9: Learning what you want to learn using programmable gradient information. *arXiv preprint arXiv:2402.13616*, 2 2024.

Wang, W., Lai, Q., Fu, H., Shen, J., Ling, H., and Yang, R. Salient object detection in the deep learning era: An in-depth survey. *IEEE transactions on pattern analysis and machine intelligence*, 44:3239–3259, 2022. ISSN 0162-8828.

Wegstein, H. J., Shimrat, M., Floyd, R. W., and Stockmal, F. Algorithm 112: Position of point relative to polygon. *Communications of the ACM*, 5:434, 8 1962. ISSN 15577317. doi: 10.1145/368637.368653.

Wen, G., Li, S., Liu, F., Luo, X., Er, M. J., Mahmud, M., and Wu, T. Yolov5s-ca: A modified yolov5s network with coordinate attention for underwater target detection. *Sensors 2023, Vol. 23, Page 3367*, 23:3367, 3 2023. ISSN 1424-8220. doi: 10.3390/S23073367.

Wilson, D., Zhang, X., Sultani, W., and Wshah, S. Image and object geo-localization. *International Journal of Computer Vision*, 132:1350–1392, 4 2023. ISSN 15731405. doi: 10.1007/S11263-023-01942-3.

Xu, S., Zhang, M., Song, W., Mei, H., He, Q., and Liotta, A. A systematic review and analysis of deep learning-based

underwater object detection. *Neurocomputing*, 527:204–232, 3 2023. ISSN 0925-2312. doi: 10.1016/j.neucom.2023.01.056.

Xu, X., Li, Y., Wu, G., and Luo, J. Multi-modal deep feature learning for rgb-d object detection. *Pattern Recognition*, 72: 300–313, 12 2017. ISSN 0031-3203. doi: 10.1016/J.PATCOG.2017.07.026.

Yamafune, K., Torres, R., and Castro, F. Multi-image photogrammetry to record and reconstruct underwater shipwreck sites. *Journal of Archaeological Method and Theory*, 24:703–725, 9 2017. ISSN 15737764. doi: 10.1007/S10816-016-9283-1.

Yang, Y. ., Liang, W. ., Zhou, D. ., Zhang, Y. ., Xu, G., Yang, Y., Liang, W., Zhou, D., Zhang, Y., and Xu, G. Object detection for underwater cultural artifacts based on deep aggregation network with deformation convolution. *Journal of Marine Science and Engineering 2023*, *Vol. 11*, *Page 2228*, 11:2228, 11 2023. ISSN 2077-1312. doi: 10.3390/JMSE11122228.

Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L. M., and Shum, H.-Y. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *The Eleventh International Conference on Learning Representations*, 2022.

Zhao, S., Kang, F., and Li, J. Concrete dam damage detection and localisation based on yolov5s-hsc and photogrammetric 3d reconstruction. *Automation in Construction*, 143:104555, 11 2022. ISSN 0926-5805. doi: 10.1016/J.AUTCON.2022.104555.

Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., and Chen, J. Detrs beat yolos on real-time object detection. *arXiv preprint arXiv:2304.08069*, 2023.

Zhu, X., Su, W., Lu, L., Li, B., Wang, X., and Dai, J. Deformable detr: Deformable transformers for end-to-end object detection. *International Conference on Learning Representations*, 2021.

Zou, Z., Chen, K., Shi, Z., Guo, Y., and Ye, J. Object detection in 20 years: A survey. *Proceedings of the IEEE*, 111:257–276, 3 2023. ISSN 15582256. doi: 10.1109/JPROC.2023.3238524.

Zuiderveld, K. *VIII.5. - Contrast Limited Adaptive Histogram Equalization*, pages 474–485. Academic Press, 1994. ISBN 978-0-12-336156-1. doi: https://doi.org/10.1016/B978-0-12-336156-1.50061-6.

A Full-Scale Orthomosaic

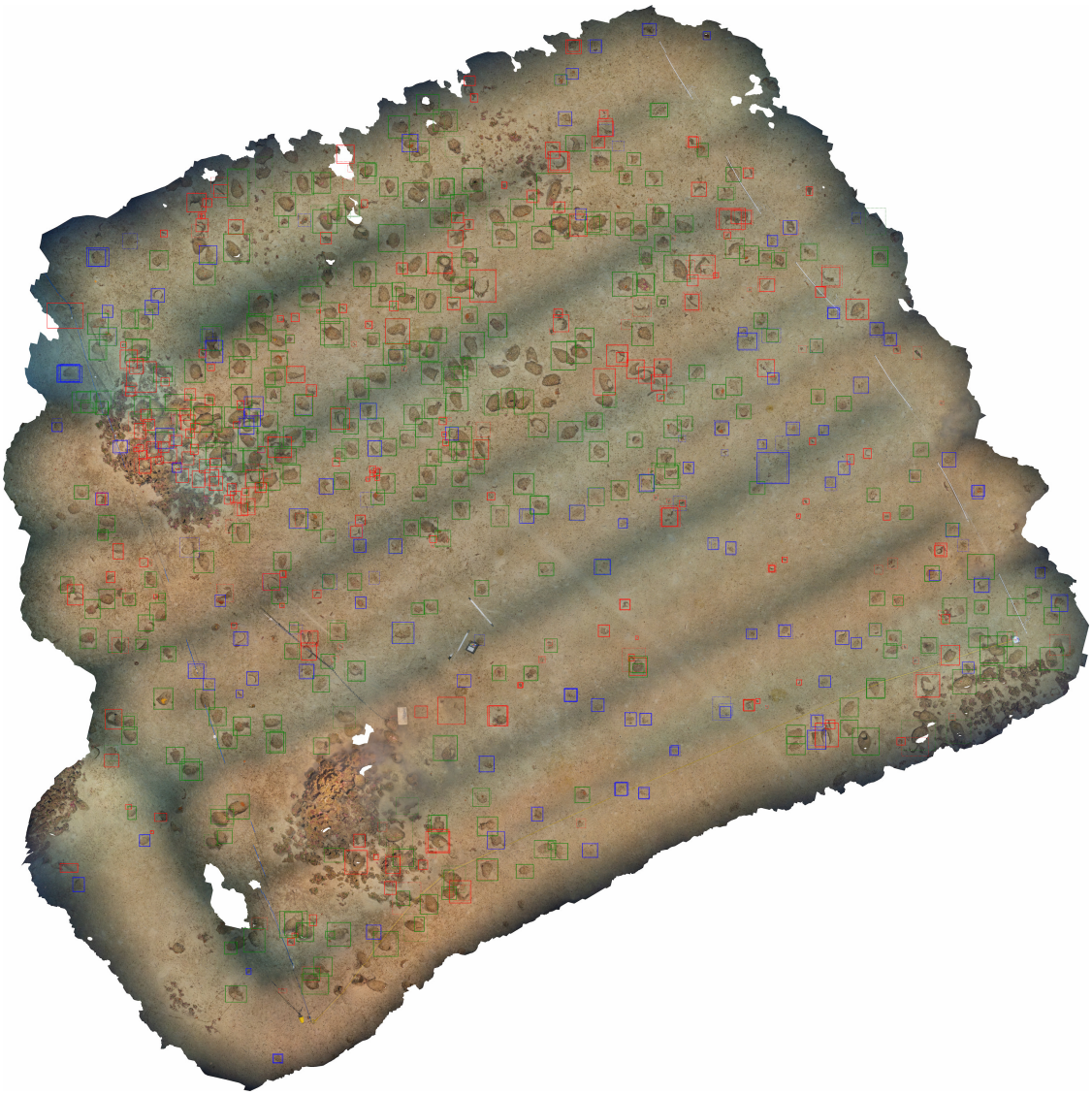


Figure A.1: Orthomosaic Projected and Normalised annotations (dashed) and predictions (solid) Suppressed using PAL1.