INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM ENGENHARIA DE INTELIGÊNCIA ARTIFICIAL




# Benchmark aplicado à Deteção de Objetos de Mamoas Arqueológicas a partir de dados LiDAR

**MIGUEL RIBEIRO VILAR BRÁS DA SILVA**
Outubro de 2023

P.PORTO


# Benchmark applied to Object Detection of Archaeological Mounds from LiDAR data

**Miguel Ribeiro Vilar Brás da Silva**

**1210065**

**Dissertation to obtain the master's degree in Artificial Intelligence Engineering**

**Supervisor:**

**Doctor Maria Goreti Carvalho Marreiros, Coordinator Professor, Polytechnic of Porto – Scholl of Engineering**

**Co-supervisor:**

**Doctor Luís Manuel Silva Conceição, Researcher, Polytechnic of Porto – Scholl of Engineering**

**Doctor Luís Carlos Gonçalves dos Santos Seco, Assistant Professor, University of Maia**

**Jury**:

President:

Doctor António Constantino Lopes Martins, Assistant Professor, Polytechnic of Porto – Scholl of Engineering

Vocals:

Doctor Maria Goreti Carvalho Marreiros, Coordinator Professor, Polytechnic of Porto – Scholl of Engineering

Doctor André Miguel Pinheiro Dias, Assistant Professor, Polytechnic of Porto – Scholl of Engineering

Porto, September 2023

# Abstract

Human history and its archaeological evidence are priceless and should be preserved, esteemed and respected. However, the traditional work of an archaeologist is mainly manual labour, sluggish and requires specialized knowledge as well as considerable experience, which represents quite a limitation due to the available community of archaeologists. Besides this fact, concerns about global warming, the generalized rise of sea levels or destruction due to human activities, among others, contribute to a growing fear of losing some archaeological sites as the traditional method of identification and preservation of these sites can't keep up with the propagation speed of such problems.

Because of this, a growing willingness to implement Artificial Intelligence techniques has been evidenced, which allows some help to the archaeologist in several tasks, with particular focus to archaeological sitting identification, through remote detection.

Currently, there are no applications or tools that can execute such work, however, there has been a growing effort in studies and work on a scientific and academic level.

This thesis aims to implement a tool that, through LiDAR data readings, gathered from some geographical area, can perform object detection to specific archaeological findings (such as mounds), testing a variety of machine learning models to, assigning a classification, determine if it's in the presence of an archaeological mound.

The input of the work done for this thesis consists of a Digital Terrain Model (DTM), a Local Relief Model (LRM) and a Slope obtained from drone flights over Viana do Castelo, with the use of LiDAR sensors. The combination of these three images was processed to achieve a single image with higher identification of certain features for future model training. For comparison reasons, two datasets were built with different margin sizes around each archaeological mound.

The goal of the thesis is to perform tests on some object detection architectures, compare the efficiency of their evaluations and be able to determine which of the tested models performs a better prediction result on detecting the presence of an archaeological mound.

This study was able to perform the comparison of a total of nine Deep Learning (DL) architectures, testing four two-stage detectors and five one-stage detectors.

As expected, most of the two-stage detectors outperformed the one-stage detectors in terms of mean average precision for the detection of archaeological mounds, except for the one-stage detector Fully Convolutional One-Stage (FCOS), which achieved the highest mean average precision from all, showing results between 68.1% to 78.6% for both size dataset.

**Keywords**: Archaeology, mounds, LiDAR, object detection

# Resumo

A história da humanidade e as suas evidências arqueológicas são inestimáveis e devem ser preservadas, respeitadas e valorizadas. No entanto, o trabalho tradicional de um arqueólogo é principalmente uma tarefa manual, lenta e requer conhecimento especializado, bem como considerável experiência, o que representa uma limitação significativa devido à disponibilidade limitada de arqueólogos. Além disso, preocupações com o aquecimento global, o aumento generalizado do nível do mar ou a destruição devido a atividades humanas, entre outras, contribuem para um crescente receio de perder alguns sítios arqueológicos, já que o método tradicional de identificação e preservação desses sítios não consegue acompanhar a velocidade de propagação de tais problemas.

Decorrente destes factos, aliado a uma tendência generalizada e com sucesso no recurso a técnicas de Inteligência Artificial em outras especialidades, também na Arqueologia tem-se vindo a verificar uma adesão significativa. A adoção de técnicas de Inteligência Artificial tem permitido alguma ajuda aos arqueólogos em várias tarefas, com especial foco na identificação de sítios arqueológicos através do recurso a métodos de deteção remota.

Atualmente, não existem aplicações ou ferramentas que possam executar este trabalho, no entanto, tem-se verificado um esforço crescente de estudo e desenvolvimento de trabalho nesse sentido, quer ao nível académico quer científico.

Esta tese tem como objetivo implementar uma ferramenta que, através da leitura de dados LiDAR, coletados de uma determinada área geográfica, consiga efetuar uma deteção de objetos referentes a vestígios arqueológicos específicos (mamoas), recorrendo a uma variedade de modelos de *machine learning,* atribuindo uma classificação para determinar se identificou ou não com sucesso a presença de uma mamoa.

O ponto de partida do trabalho realizado nesta tese inicia-se com o acesso e trabalho realizado sobre três técnicas de visualização aplicada sobre dados LiDAR, nomeadamente consiste no acesso a ficheiros como *Digital Terrain Model* (DTM), *Local Relief Model* (LRM) e *Slope*. Estes dados LiDAR e consequente conversão nas técnicas de visualização anteriormente citadas ocorreram a partir de voos de *drones*, equipados com sensores LiDAR que, sobrevoando a zona de Viana do Castelo, proporcionou a obtenção de tais dados. Adicionalmente aos três ficheiros de técnicas de visualização, foi também disponibilizado um ficheiro *shape* que fornece informação georreferenciada da localização de mamoas na área sobrevoada pelos *drones*.

Com recurso ao software QGIS, foi possível identificar que, as localizações das mamoas encontravam-se relativamente concentradas numa parte específica das imagens. Desta forma, e considerando o tamanho dos ficheiros em questão, efetuou-se uma seleção nas imagens, cortando áreas que já apresentassem uma distância considerável da mamoa mais próxima, de forma a tornar mais ágil o processo de trabalho e treino dos modelos escolhidos.

Posteriormente, e com as imagens em tamanho mais reduzido, efetuou-se uma operação de combinação entre as três tipologias de imagens, obtendo uma única imagem onde, incorporando as características destas, permitiu realçar determinados aspetos com intuito de, posteriormente, auxiliar nas tarefas de treino e teste dos modelos de aprendizagem profunda a que foram aplicados.

Seguiu-se o processo de pré-processamento de dados tendo sido definido e trabalhado um programa que executasse a mesma tarefa, fornecendo como output um *dataset* em formato COCO, formato escolhido dada popularidade e sucesso verificado na aplicação a métodos de deteção de objetos. A construção deste *dataset* foi igualmente realizada de forma a criar estrutura de ficheiros que, respeitando na mesma o formato COCO, proporcionasse a aplicação da técnica de *leave-one-out cross-validation*, uma vez que, foi considerado a melhor opção dada existência de apenas 77 mamoas, de forma a evitar cenários de enviesamento de dados ou até *overfitting*. Para diversificar e enriquecer esta análise comparativa, foram criados dois *datasets* diferentes, cujas *bounding boxes* em volta das mamoas apresentavam tamanhos diferentes, nomeadamente 15x15 metros e 30x30 metros.

Como o objetivo da tese é a realização de testes em algumas arquiteturas de deteção de objeto, foi utilizada um projeto que, está precisamente preparado e desenvolvido para a realização de análises de *benchmark*, de várias metodologias de classificação de imagem, nas quais estão incluídas as de deteção de objeto. Esta biblioteca permitiu a realização do estudo comparativo não só entre as arquiteturas analisadas e identificadas como as mais promissoras e populares na análise de estado de arte, como ainda permitiu a comparação com outras arquiteturas dada a variedade de oferta de modelos que a mesma proporcionava.

Este estudo conseguiu realizar a comparação com um total de nove arquiteturas de aprendizagem profunda, testando quatro detetores *two-stage* e cinco detetores *one-stage*. Como era esperado, a maioria dos detetores *two-stage* superou os detetores *one-stage* em termos de precisão média de deteção de mamoas, com exceção do modelo *Fully Convolutional One-Stage* (FCOS), que alcançou a maior precisão média de todos os modelos testados, apresentando resultados entre 68,1% e 78,6% em ambos os *datasets*.

Igualmente esperado foi a confirmação do modelo *one-stage Single Shot Detector* (SSD) como sendo o modelo com mais rápido tempo de processamento de treino, apesar de, entre os restantes modelos, a diferença de tempo já ser menos significativa e não se notar uma supremacia dos modelos *one-stage* como seria inicialmente esperado.

**Palavras-chave**: Arqueologia, mamoas, LiDAR, deteção de objeto

# Acknowledgements

I would like to express my heartfelt gratitude to the individuals who played pivotal roles in the successful completion of my master's dissertation. Their unwavering support, guidance, and encouragement were instrumental throughout this challenging journey.

First and foremost, I extend my deepest appreciation to my supervising professors, Doctor Goreti Marreiros and Doctor Luis Conceição. Their consistent understanding and invaluable feedback were a beacon of guidance during unforeseen circumstances and multiple shifts in the direction of my thesis. I thank both of you for your keen insights and suggestions, which enriched the depth of my research.

I am profoundly indebted to everyone in Feralbyte, Unipessoal Lda., specially to Doctor Luís Seco and Hugo Freire for providing the support and even some labour time to fully dedicate myself to this thesis and bring it to fruition.

A huge thank you also to Doctor Jorge Garcia and Doctor Manuel Carranza for their interest shown in the execution of my thesis. Their invaluable inputs, whether they were advice, guidelines, corrections, or every other feedback that was provided were always much appreciated and crucial to the work done on this thesis.

I extend my thanks to my fellow master's colleagues, Bruno Ribeiro, Bruno Veiga, and Carlos Coelho. Together, as part of the [AI]nergy group, we formed an amazing support system. Even while each of us was immersed in our own theses, their willingness to offer help and words of encouragement were immensely appreciated.

I want to acknowledge the unwavering support of my family. Their patience and understanding during the most stressful periods of my master's degree were a source of strength.

Finally, but certainly not least, I want to express my deep appreciation to my companion, Sandra. Throughout this journey, she displayed incredible understanding, support, and care. Her constant presence, encouragement, and ability to provide moments of joy and relaxation were invaluable and difficult to put into words.

Thank you all for being part of this significant chapter in my academic journey. Your contributions, whether big or small, have left an indelible mark on my work and my life.

# Table of Contents

# List of Figures

# List of Tables

# Acronyms

| | |
|---|---|
| **AI** | Artificial Intelligence |
| **CNN** | Convolutional Neural Network |
| **COCO** | Common Objects in Context |
| **DL** | Deep Learning |
| **DTM** | Digital Terrain Model |
| **FCOS** | Fully Convolutional One-Stage |
| **FPN** | Feature Pyramid Network |
| **GPS** | Global Positioning System |
| **IoU** | Intersection over Union |
| **JSON** | JavaScript Object Notation |
| **LiDAR** | Light Detection and Ranging |
| **LOOCV** | Leave-One-Out Cross-Validation |
| **LRM** | Local Relief Model |
| **mAP** | mean Average Precision |
| **ML** | Machine Learning |
| **NMS** | Non-Maximum Suppression |
| **RoI** | Region of Interest |
| **RPN** | Region Proposal Network |
| **SSD** | Single Shot Detector |
| **TOOD** | Task-aligned One-stage Object Detection |
| **VT** | Visualization Technique |
| **YOLO** | You Only Look Once |

# 1 Introduction

This initial chapter aims to provide a general contextualization of the work undertaken in accordance with the identified issues, as well as outline the main objectives to be addressed and the chapter structure that will be employed in this dissertation.

## 1.1 Contextualization

Just like in various fields and specialties, we have also witnessed a significant increase in the use and application of Artificial Intelligence (AI) techniques in Archaeology, aiding in the execution of various tasks associated with this discipline.

In the technical approach to archaeological work across vast areas of terrain under study, in addition to maps and other relevant documentation, photographic surveys of archaeological sites are also conducted.

Classic photographs obtained by different cameras and photographers during field campaigns often do not reach meaningful conclusions. Even when digitized, older photos archived may prove challenging for interpretation and analysis, apart from the time-consuming process involved in their analysis.

AI now allows for the analysis of cluster of pixels from one or multiple photographs of a site, enabling the recognition of patterns and the extraction of features from specific objects that may hold meaningful significance. This characterization and differentiation can contribute to understanding and distinguishing the presence of archaeological mounds (Fiorucci et al., 2022).

Currently there are a multitude of available models that are suited to realize the task of object detection through the analysis of remotely obtained data, acquired from drone or airplane flights, allowing an automated detection of the presence of archaeological mounds from sensors such as Light Detection and Ranging (LiDAR) (Guyot, Lennon, Lorho, et al., 2021).

The work developed in this thesis, is inspired and originates from the project Odyssey – Platform for Automated Sensing in Archaeology[1], funded by Portugal 2020, counted with a partnership from ERA – Arqueologia S.A., Maiêutica Cooperativa de Ensino Superior C.R.L. and Universidade de Aveiro and started in May 2021. The project Odyssey originated due to the lack of reliability of patrimonial information as well as a reliance on intensive workforce and traditional processes. It was estimated that the mistake of wrongly identifying an archaeological site could reach up to 90%. On the other hand, it has been identified that, despite the technical evolution applied to Archaeology in recent years, it's still a very traditional area with high dependency and utilization of workforce.

There are mostly three main objectives to this project that starts by developing an integrated platform of geographical information, destined to be used from archaeologists and patrimonial technicians allowing to consolidate and access several sources of patrimonial information, using remote sensing technologies like LiDAR or multispectral images. Another objective is to establish an automatization applied to such images to optimize the process of identification of archaeological findings. A more articulated process of ground truthing for confirmation on field, if the predictions were accurate or not, is also an objective to improve the annotation process and consequently allow the improvement of the predicting algorithms.

The execution of the Odyssey project, among other outputs, produced the publication of the paper (Canedo et al., 2023), which has a lot that has been learned and used in this thesis.

Following these objectives from project Odyssey, the author aims to continue with the strategy of, making use of image processing techniques and AI tools, realize a benchmark analysis that will establish a comparison between some of the most popular architectures for object detection, applied to the use case of detecting archaeological mounds. This study will use a remote sensing technique, to gather LiDAR data, collected previously by drone, from flights done over the Alto Minho region, mainly in Viana do Castelo, in Portugal.

## 1.2 Problem Description

The rhythm of discoveries through traditional methods is relatively slow, often taking months or even years before new findings emerge. This slow pace is a cause for concern, especially considering the numerous factors that can threaten or even destroy these findings and their valuable archaeological elements. These factors include issues related to global warming, the widespread rise in sea levels, human activities, or simply natural disasters like earthquakes.

The shortage of qualified personnel in archaeology is also a concern. As it is well known, the research and detection work related to archaeological sites is traditionally manual and time-consuming, requiring a high level of knowledge and expertise in the field. This limitation

[1] https://odyssey.pt/

further hampers the recruitment of qualified human resources for this role, consequently further slowing down the entire process of archaeological mound detection.

There are also concerns about invasive and destructive methods for detecting and identifying archaeological elements (D. Davis et al., 2018), with excavations being highlighted as the most invasive technique. On the other hand, the technical activities carried out during archaeological campaigns in the field often involve the use of heavy equipment to facilitate access to locations of interest. This can include earth-moving operations to create access points, levelling natural slopes, deforestation, or land clearing for new temporary installations and equipment storage, thus contributing to the alteration or damage of archaeological findings.

Additionally, there are locations that are challenging to access, either due to the natural characteristics of the sites or because they are considerably far away from the researcher, making it difficult to carry out on-site analysis and work.

To address this latter aspect, the option of manually conducting visual image analysis, collected by satellites, aircraft, or drones, has been employed. However, given the quantity, diversity, and size of available images, it remained a highly time-consuming task.

It is important to pursue elaborated learning methods to detect and characterize traces of archaeological structures to accurately identify and characterize patterns and obtain a robust and effective tool for detecting various types of concealed and sometimes imperceptible archaeological structures at the sites under analysis. In this way, by harnessing the potential of AI automation, we aim to achieve security and reliability in the results, enabling the inventory and mapping of archaeological structures on a large scale. For this purpose, the analysis of images through AI represents a determining factor that provides archaeologists with an additional tool to help overcome some limitations and constraints, such as slow stereoscopic reading and photointerpretation. This contributes to defining and obtaining clues or evidence, both in the discovery and validation of new archaeological sites and in the task of finding new objects of potential interest.

## 1.3 Objectives

The use of AI and its algorithms can relieve the archaeologists from tiring and time-consuming tasks by performing predictions, being able to detect and localize archaeological findings of potential interest, such as detecting archaeological mounds.

The dissertation will contribute to minimizing the problems mentioned in the previous section, using remote sensing techniques based on LiDAR data, which will enable a faster and automatic analysis of vast areas in a less invasive manner, much faster than traditional methods, without the need for physically visiting the location.

Some object detection techniques will be tested, performing a benchmark analysis, to assess which of them will be able to achieve a better average precision, attempting to determine the presence of archaeological mounds, since this analysis has yet to be done on the field of archaeology.

## 1.4 Document Organization

The document of this thesis is organized into five chapters. The first chapter, as an introductory section, presents and describes the context in which this thesis is situated, the problem it aims to help solve, and establishes the definition of the objectives to be achieved in this work.

The second chapter, the literature review, begins by describing the methodology used to identify the solutions and techniques that have been employed, presenting the most promising results as inspiration for the thesis.

In the third chapter, in addition to presenting and explaining the data that will be used, the methods and tools to be used in the thesis are also discussed, based on the analysis conducted in the previous chapter, with the expectation of providing the best results.

A more detailed and objective explanation of how the solution presented in this thesis was implemented is carried out in the fourth chapter. It's also when a description of the process used for experimentation and the evaluation of the results obtained from the developed solution are presented through an analysis and discussion of these results.

Finally, in the fifth and last chapter, a comprehensive summary of the thesis is provided, with a particular focus on whether the initially proposed objectives have been met based on the results obtained. Considering this comparative analysis and the work accomplished, the next steps for future work are also outlined.

# 2 Literature Review

In this chapter, a description will be provided about the adopted literature research strategy, as well as the presentation of the data sources that were used for this purpose. Afterwards, there will be a presentation of papers, as well as their respective techniques, that were considered the most interesting and relevant for the purpose of the theme of this thesis.

## 2.1 Methodology

The literature review conducted for the development of this dissertation was inspired by some of the techniques used in the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) methodology. However, due to time constraints that would hinder a complete systematic review, only a few of the steps and procedures of this methodology were implemented.

To better guide the research methodology and its corresponding results, it was important to begin with the definition of the main research questions. It was established that it would be important to address two questions, as can be seen in Table 1.

Table 1 – Research questions

| **Identifier** | **Research question** |
|---|---|
| RQ1 | What are the main applications of LiDAR data in archaeology? |
| RQ2 | What object detection techniques are used for the detection of archaeological mounds? |

The first research question (RQ1) is important for a better understanding of the current use of remote detection methods in the field of Archaeology, mainly with the use of LiDAR sensors

and understand the possibilities of use and processing of such data. The second research question (RQ2) will allow an evaluation of the most popular and used object detection techniques in the field of archaeology.

Afterwards, some research keywords were chosen to aid the bibliographic research as presented in the Table 2.

Table 2 – Research keywords

| **Identifier** | **Keywords** | **Relation with the Research Question** |
|---|---|---|
| KW1 | ("*archaeology*" OR "*archaeological*" OR "*archeologist*") AND | RQ1, RQ2 |
| KW2 | "LiDAR" AND | RQ1 |
| KW3 | ("*artificial intelligence*" OR "*machine learning*" OR "*neural network*") AND | RQ1, RQ2 |
| KW4 | "*object detection*" | RQ2 |

After the definition of these research keywords, the decision was made to apply them in searches across online bibliographic data sources, with the choices as listed in Table 3.

Table 3 – Research data sources

| **Identifier** | **Data Source** | **URL** |
|---|---|---|
| DS1 | IEEE Xplore | https://ieeexplore.ieee.org/ |
| DS2 | MDPI | https://www.mdpi.com/ |
| DS3 | Science Direct | https://www.sciencedirect.com/ |
| DS4 | Web of Science | https://www.webofscience.com/ |

The data extraction process of articles containing the keywords earlier defined, was performed through queries used on the data sources. This process provided a result of a total of 101 articles, 1 from IEEE (DS1), 46 from MDPI (DS2), 36 from Science Direct (DS3) and 18 from Web of Science (DS4). With the help of Mendeley Reference Manager[2], 6 articles were identified as duplicates, which reduced the total articles to 95 articles. Afterwards these articles went through a filtering process with some inclusion and exclusion criteria mentioned on Table 4 and Table 5, respectively, by performing a screening of the abstract and, if any

[2] https://www.mendeley.com/

doubt occurred, a diagonal reading of the paper would be done to further understand if the article would be suitable for the theme of this thesis.

Table 4 – Inclusion criteria

| **Identifier** | **Inclusion criteria** |
|---|---|
| IC1 | The article shows contribution to the fields of study |
| IC2 | The article describes a tool or an application scenario |
| IC3 | The article describes an archaeology object detection model from LiDAR data |

Table 5 – Exclusion criteria

| **Identifier** | **Exclusion criteria** |
|---|---|
| EC1 | The article is over 6 years old |
| EC2 | The article is a book chapter, dissertation, systematic review, or thesis |
| EC3 | The article is not written in English |
| EC4 | The article focus describes related work with theme but over focus on aspects not required for the objective of the thesis |

After undergoing this process, 40 articles were excluded due to falling in the exclusion criteria of EC1 and EC2. Besides these articles, another 43 were excluded due to the EC4, specifically 10 being unrelated with Archaeology, 6 with no use of airborne LiDAR data, 11 note making use of Deep Learning techniques and 16 not using object detection techniques, with the majority of these choosing to use segmentation techniques. Overall, this left a total of 13 remaining articles which were studied and incorporated in the following state of the art section.

Despite accessing the data sources from Table 3, on some occasions, to have complementary research, or for specifics subjects, the online libraries such as “B-On” and “Google Scholar” were also used for that very purpose.

## 2.2 State of the Art

This section will address the scientific articles that have been found and considered relevant for gaining an understanding of currently existing technologies in use. The section is divided into two modules: the remote sensing module, which will place special emphasis on LiDAR

data, and the object detection module which will focus on the architectures for object detection that will be tested.

### 2.2.1 Remote Sensing

In recent years, there has been a growing trend of appreciation and popularization of remote sensing techniques in the field of Archaeology, strongly influenced by the increasing awareness and preference for non-invasive or non-destructive methods (Argyrou & Agapiou, 2022).

Among the various remote sensing methods, it is worth highlighting LiDAR technology. Despite having been first applied in Archaeology in the 1990s, it has since evolved into an increasingly valuable tool for discovering and visualizing archaeological resources and sites. Notable instances include the discovery of new monuments at Stonehenge and the identification of Mayan cave entrances in Belize using this technology. What's particularly intriguing and innovative is the application of a transfer learning approach, utilizing a deep Convolutional Neural Network (CNN). Initially trained on LiDAR data acquired from lunar observations, this approach, when adapted to the archaeological context, provided equally positive outcomes in the detection of archaeological objects (Gallwey et al., 2019).

Traditionally, from LiDAR data, a Digital Terrain Model (DTM) is created, followed by the application of Visualization Techniques (VT) that best suit the identification or discovery objectives. In the article (Guyot, Lennon, Lorho, et al., 2021), the authors compared 13 different VTs to determine the most suitable or effective one for subsequent use in a deep CNN. The Multiscale Topographic Analysis (MSTP) and enhanced MSTP ($e^2$MSTP) stood out as the ones that produced the best results.

### 2.2.2 Object Detection

Object Detection is a powerful computer vision technique that goes beyond traditional image classification. It’s capable of, not only identifying what objects are present in an image, but precisely locate their positions by detecting and delineating a bounding box around one or more objects. The Deep Learning (DL) object detection methods are usually divided into two categories: one-stage and two-stage detectors.

#### 2.2.2.1 Two-Stage Detectors

Two-stage frameworks divide the object detection procedure into two distinct shapes: the initial region proposal phase and the subsequent classification stage. In this approach, the models initially suggest multiple potential objects, referred to as regions of interest (RoI), by employing reference boxes (anchors). In the subsequent step, these proposals undergo classification, and their positional accuracy is improved through some refinement operations (Carranza-García et al., 2020).

**Faster Region-based Convolutional Neural Network (R-CNN)**

There are a vast number of examples where the application of the architecture of Faster R-CNN has been successfully applied to airborne LiDAR data for object detection. A good example of this is reflected on the paper (Ø. D. Trier et al., 2021), where they aim to find and map the presence of cultural heritage, in Norway, such as grave mounds, charcoal kilns and pitfall traps for deer hunting systems. It was shown that, with an accuracy between 84% to 96%, the model successfully identified the objects as intended.

In a central part of the Netherlands, another study (Verschoof-van der Vaart & Lambers, 2019) used this architecture to identify the presence of Celtic fields, barrows and charcoal kilns. Despite some changes performed on the hyper-parameters such as number of epochs or the sizes on the anchor boxes, the performance was more inconsistent than the previous study, showing precision values between 0.26 to 0.9, however, not being able to detect charcoal kilns.

A slightly different issue addressed, also from the Netherlands, on the western part of the province of Gelderlan, the authors in the paper (Fiorucci et al., 2022), introduce two automated evaluation measures, namely centroid-based and pixel based to encode aspects of the archaeologists thinking process. They consider that the standard metric for measuring object detection in deep learning methods, Intersection over Union, isn't quite the most suitable metric due to geographical positioning, shapes or even areas of the bounding boxes when comparing with the actual area containing archaeological. In the region of Veluwe, testing and area around 2.200 $Km^2$, the test to identify pre-historic barrows and Celtic fields is done resorting to the deep learning architecture Faster R-CNN. The study shows and focuses on the fact that the suggested metrics to achieve a more accurate evaluation of the actual position of the archaeological object and more similar to an archaeologist way of thinking. They achieve a discrepancy lower than 1% and 3% for both suggested metrics when compared to the semi-automatic GIS-based measurement.

A comparative analysis between automatic detection and on-screen detection from an expert is the objective of the paper (Oliveira et al., 2021). The region of Meuse, in the northeast of France, was the choice to test these methods on identifying historical charcoal kilns, an area around 230 $Km^2$. The architecture chosen for the process of the automatic detection was Faster R-CNN, with the use of a Resnet 50 backbone, since it's "one of the smallest of the available models that have proven to be efficient at detecting small objects", according to the authors. It was interesting to realize that, because the expert was evaluating with a high level of confidence, incurred in missing a significant number of kilns, as proven from the 63,9% of recall, on the other hand, allowed to have an amazing precision of 98,4% (only 1 false positive). The automatic detection, was not able to match this precision, attaining 84,4% (15% false positives), but this was compensated with a significant increase of the recall, succeeding in identifying 90% of the kilns.

The development of a Workflow for Object Detection of Archaeology in the Netherlands (WODAN) is proposed in (Lambers et al., 2019). Considering that there doesn't exist a variety of datasets with archaeological objects identified and annotated to perform the activities, the authors suggest the adoption of citizen science. This basically means that volunteers, non-professional scientists, but potentially researchers, would help with scientific inquiries to help identify archaeological objects. The project Heritage Quest helps exactly with this initiative, since it offers a web platform to allow its participants to help identify potentials barrows, Celtic fields and charcoal kiln within LiDAR images. With these newly generated datasets, WODAN then applies an adaptation of Faster R-CNN to perform the tasks of automated object detection verifying its predictions against the information gathered from the citizen science on the Heritage Quest.

**CNN**

The use of a CNN pre-trained with a ResNet18 model was chosen, in the paper (Ø. Trier et al., 2018), to automatic detection, from LiDAR data, to identify the presence of prehistoric roundhouses, shieling huts of medieval or post-medieval date and small clearance cairns. This study was performed on the west of Scotland, precisely the island of Arran, involving an area around 432 Km2. The first class to be identified did have an interesting performance since it was able to identify 73% of the total existing roundhouses, although it still presented 87% of false positives. The other two classes did not perform as expected, although the authors do understand that the variety in form might contribute to be easier to mistake it for other features. Specifically, only 20% and 26% of existing small cairns and shieling hut, respectively, were successfully identified. The detection of false positives was also high, particularly the shieling hut achieving a value of 189%, while the false positives for the small cairns reached 90%.

#### 2.2.2.2 One-Stage Detectors

The one-stage frameworks, despite its usual lower accuracy when compared with the two-stage detectors, present a lower computational cost. This occurs mostly because one-stage detectors perform both proposal and classification in a single operation (Zou et al., 2023).

**You Only Look Once (YOLO)**

The architecture YOLO is probably the most used of the one-stage detectors and due to its popularity has an extensive diversity of versions of the model.

The previous project of Odyssey, in fact, used the variant of YOLOv5 to attempt to identify the presence of new burial mounds. The paper (Canedo et al., 2023), worked on the same region as this thesis, namely region of Alto Minho, in Portugal, using the full LiDAR data covering an area around approximately 2220 $Km^2$ having access to DTM and consequently LRM of the regions of Viana do Castelo, Paredes de Coura, Arcos de Valdevez and Parque Nacional da Peneda-Gerês Since the dataset available was still small, a copy-paste data augmentation methodology was chosen. The application of the model YOLOv5 to the augmented dataset

provided promising results since it achieved a 72,53% of positive rate of identification of burial mounds.

Another interesting and slightly different application is the use of LiDAR on a YOLOv3 architecture, but to identify archaeological shipwrecks underwater. The paper (Character et al., 2021), also combined Sonar to the LiDAR data, however, the results obtained were very promising, achieving a precision of 90% of success, being a significant improvement of previous studies, which only achieved results between 29% to 80%.

The approach on the paper (Arnoldussen et al., 2023) is slightly different as the authors focus resides more on the preparation of data to identify the land usage as well as palaeodemography. This study is done in the regions of Zeijen, Putten, Riethoven and Posterholt, in Netherlands. For the land usage it was usual to resort to information on settlements, funerary sites, or a combination of these, mostly resorting to human expert interpretation. However, it's suggested that the presence and coverage of Celtic fields (prehistorical field systems) would provide a better analyses and information as an AI-assisted mapping of prehistoric field system from LiDAR data, would be a better solution to be used for the deep learning architecture chosen, YOLOv4. This solution allowed to detect a higher area of Celtic fields, which, is an interesting result, although if it weren't for factors like human recent constructions, believe it would be even bigger. As for population sizes, the model gave an estimation with a higher deviation, possibly influenced by the present-day land usage. Despite this, the authors still consider this method to provide a more robust and reliable result.

In Spain, specifically in the region of Galicia, was also conducted a study (Berganzo-Besga et al., 2021), to detect burial mounds through the combined use of LiDAR data and multispectral satellite data. The region chosen comprised an area around 30.000 $Km^2$ with the presence of 10.527 burial mounds. Despite claiming that region-based CNN (R-CNN) have been more used recently, the authors chose to implement the YOLOv3 architecture since it’s faster than R-CNN such as Faster R-CNN, for example. This study identified that they were having some problems with false positive identifications so decided to combine a Random Forest algorithm for soil classification. The use of pre-treatment of the LiDAR dataset with Multi-Scale Relief Model (MSRM), improved visibility of features in the DTM, which allowed a reduction of false positives. Having defined an Intersection over Union threshold of 0.5, they managed to achieve a detection rate of 89.5%.

**RetinaNet**

This popular one-stage detector was used in the paper (D. S. Davis & Lundin, 2021), to identify charcoal production sites in Sweden. The choice to use the RetinaNet DL model rested on the fact that they intended to use a less computationally expensive model then the most popular architectures (such as R-CNNs) as thus ensuring a higher usability for a more diverse scenario. The accuracy achieved on the training samples were between 55% and 63%.

**Single Shot Detector (SSD)**

SSD is another one-stage detector with a good popularity and respective usage in various disciplines and archaeology isn't an exception.

In the Sechura desert, in Peru, can be found various Nasca geoglyphs which are considered as a UNESCO[3] World Heritage Site. The paper (Sakai et al., 2023), inclusively deals with the same limitation that the work done in this thesis also faces, namely the limitation of existing only thirty-two verified geoglyphs images to train, validate and test. Because of this fact, the twenty-one geoglyphs with the clearest traces and easier identified as such, were used as train, while the next four more clearer were chosen for validation and the remaining seven as test. The model still performed positively identifying five out of the seven geoglyphs. The most significant is that when the model tried to identify new and undiscovered geoglyphs it successfully identified four new geoglyphs that were later confirmed as such by an on-site survey.

## 2.3 Discussion

The attempt to answer the research questions presented in Table 1 were slightly harder than initially expected.

Although it became clear that an increase in last years has been made in the use of remote sensing technologies and deep learning in the field of Archaeology, the number of papers found that combined the three big subjects of this thesis (Archaeology, LiDAR data and deep learning object detection), ended up being a relatively low number than what was expected.

It was possible to establish that LiDAR has really been adopted by the academic and scientific community, since, if not the main, it's used on a high majority of papers for the use of remote sensing technologies. Despite cases where they were also combined with some other techniques, it was easy to ascertain to be extremely popular. The RQ1, was easily answered as it has a variety of use for different kinds of objects to be detected, to detect landscapes like Celtic fields, or even to be used on underwater detection.

It was the attempt to answer the RQ2, that remained the whish for a higher diversity of deep learning techniques tested with LiDAR data for archaeological object detection. However, it became clear that a model of each of the one-stage or two-stage detector stands out as a lot more popular than the alternatives. For the two-stage detector, Faster R-CNN shown a significant popularity and implementation, on the other hand, for the one-stage detector, the architecture of YOLO stood out from the remaining deep learning of its category.

These two deep learning models will obviously be tested on this thesis to establish a comparison of performance, but also to understand if the object to be detected, namely

[3] https://www.unesco.org/

archaeological mounds, do affect to good performance that these models usually achieve through all the papers observed.

To have a more diverse comparison between deep learning models, the architectures of RetinaNet, SSD and CNN's will also be subject to analyses.

Despite these choices, due to the variety of object detection deep learning models available on the project that will be explained further ahead in this thesis, other models will also be chosen to enrich the benchmark analyses.

# 3 Methods and Materials

In this chapter, the focus will be on the selected models and methods for the development of this thesis, including the dataset that will be used to train the model. Additionally, an examination of data protection issues, security concerns, and ethical considerations will be conducted.

## 3.1 Dataset

The dataset that will be used in the present thesis will consist mostly of data obtained from the LiDAR technology which, from its point cloud source, allowed us to get some images from a few VT's such as DTM, Local Relief Model (LRM) or Slope.

### 3.1.1 Understanding LiDAR

LiDAR technology consists essentially of a laser, a scanner, and a specialized sensor/receiver with Global Positioning System (GPS).

LiDAR can also be used for three-dimensional topographic surveys, using an infrared laser that, when scanning, maps the terrain. Another type of survey performed by LiDAR is bathymetric, using a green light beam that, when penetrating the water, takes elevation measurements of both the riverbed and the seabed.

By using laser pulses, as can be seen in Figure 1, LiDAR sends beams of light outside the visible spectrum and records how long each pulse takes to be reflected off an object (such as an archaeological site) and return to the sensor. In this way, the direction and distance obtained for each recorded return allow the LiDAR system to progressively build up a collection of data in the form of a point cloud.

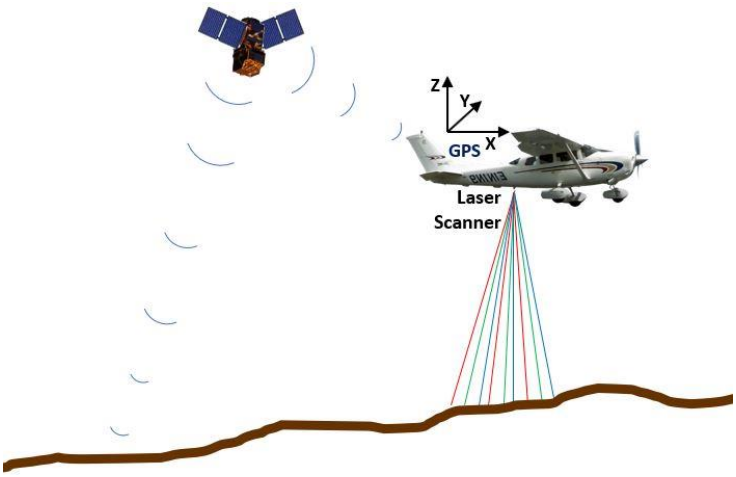


Figure 1 – Example of a LiDAR system

While LiDAR equipment can follow different scanning methodologies due to hardware manufacturers requirements, they generally conduct site surveys in a circular pattern, like a radar dish, while simultaneously moving the laser beam up and down. After the scanning operations are completed, the readings obtained during the survey are processed and organized, forming the 3D point cloud of LiDAR data (Marcos et al., 2010).

This is how these point clouds form extensive collections of 3D elevation points, including spatial coordinates x, y, and z, along with additional attributes, such as the recorded GPS time, if available.

Depending on the characteristics of each site, the processing of the LiDAR point cloud allows for the subsequent classification of detected surface types, such as buildings or ruins, tree canopies, shrubs, paths, roads, viaducts, bridges, rocky outcrops, slopes, or even some discrepancies, natural or artificial changes in the terrain that the laser beam detected during the survey scan (Fiorucci et al., 2022).

### 3.1.2 Identification and Data Collection

The LiDAR aerial data for this project were provided by the Comunidade Intermunicipal do Alto Minho – CIM Alto Minho, as presented in the Table 6 (Fonte et al., 2021).

Table 6 – Summary of LiDAR data characteristics

| | |
|---|---|
| Date of the flights | Between January 28th and 31st, 2018 |
| Region of the flights | Viana do Castelo (Alto Minho region) |
| Number of scan strips | 96 strips |
| Covered area | 2260 $km^2$ |
| Average measure density | 2 pulses/$m^2$ |

| | |
|---|---|
| Sensor type | Leica ALS80-HP |
| Laser wavelength | 1064 nm |
| Average flight altitude | 2628 masl (meters above sea level) |
| Beam divergence | 0,26 mrad |
| Variable pulse frequency | Between 200 khz and 225 khz |
| Variable scan frequency | Between 42 Hz and 45 Hz |
| Variable field of view | Between $20^{o}$ and $45^{o}$ |

The LiDAR data in this project were acquired through specific flights conducted between January 28th and 31st, 2018, utilizing a Leica ALS80-HP sensor. These flights covered the entire territory corresponding to the district of Viana do Castelo (Alto Minho region). The area covered by the flights was approximately 2260 $km^2$, which translated to 96 scan strips, resulting in an average measurement density of around 2 pulses/$m^2$ when considering only the last recorded returns.

During this survey, the Leica ALS80-HP sensor operated at a laser wavelength of 1064 nm, with an average flight altitude of 2628 meters above sea level (masl). The beam divergence was 0.26 mrad, with a pulse frequency ranging between 200 kHz and 225 kHz, a scan frequency varying between 42 Hz and 45 Hz, and a field of view ranging from $20^{o}$ to $45^{o}$. Up to five return pulses were recorded.

It's worth noting that LiDAR aerial surveys captured by drone flights achieve a high level of precision (approximately 20-35 mm) while scanning areas with dense forest and shrub cover. The output is obtained in the form of point clouds, with an approximate density of 250 points/$m^2$.

Through the application of some VTs, the output of a DTM, a LRM and a Slope image of the area were generated, which consisted of the base for the work undergone on this thesis, as show on Figure 2.

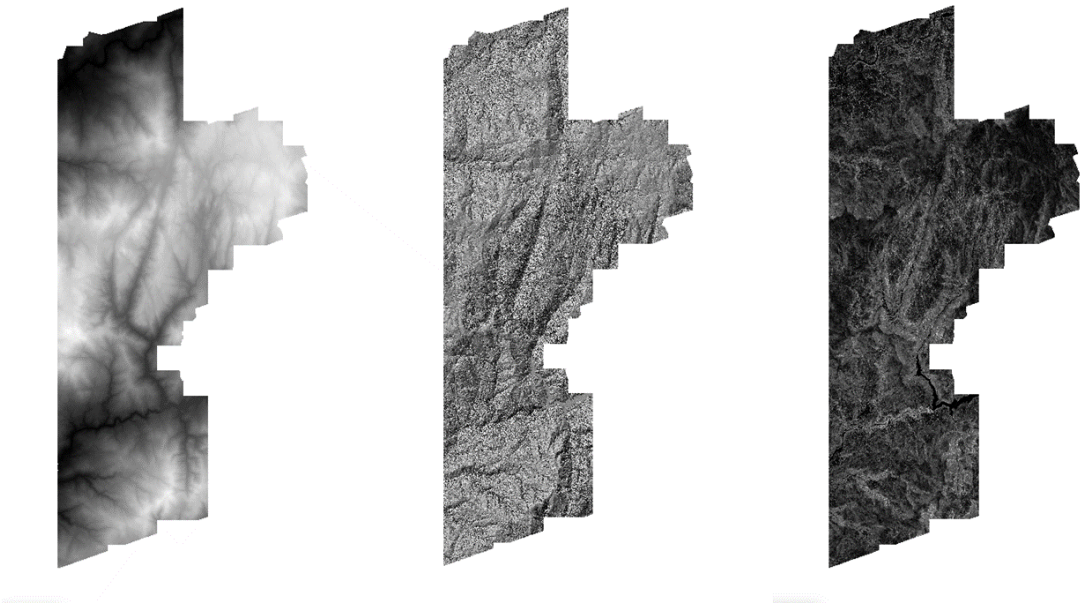

Figure 2 – DTM, LRM and Slope of the area flown by drones, over Viana do Castelo

### 3.1.3 Pre-processing

The starting geospatial data, previously shown in Figure 2, needed to be processed in a manner that it would be easier to work with. The conversion of the geospatial data to an annotation format, namely to a JavaScript Object Notation (JSON) Common Objects in Context (COCO) format (Lin et al., 2015), was chosen given the popularity and success cases such as proven from the papers (Guyot, Lennon, & Hubert-Moy, 2021), (Carranza-García et al., 2020) and (Gallwey et al., 2019), just to mention a few.

For this step, the software QGIS[4], version 3.22.4, was used since it's an amazing and popular tool to work with geospatial data.

It is not unusual or surprising that all these VTs, consist of heavy files, with the MDT file of 1.4GB, the LRM of 3.6GB and the Slope file with 3,7GB. Given the size of each of these images, and since the shape file containing the information of the location of the 77 archaeological mounds were relatively concentrated on a part of the images (as shown on Figure 3 as orange dots) it was thought of a strategy to crop the images to include simply the area with the presence of archaeological mounds.

[4] https://qgis.org

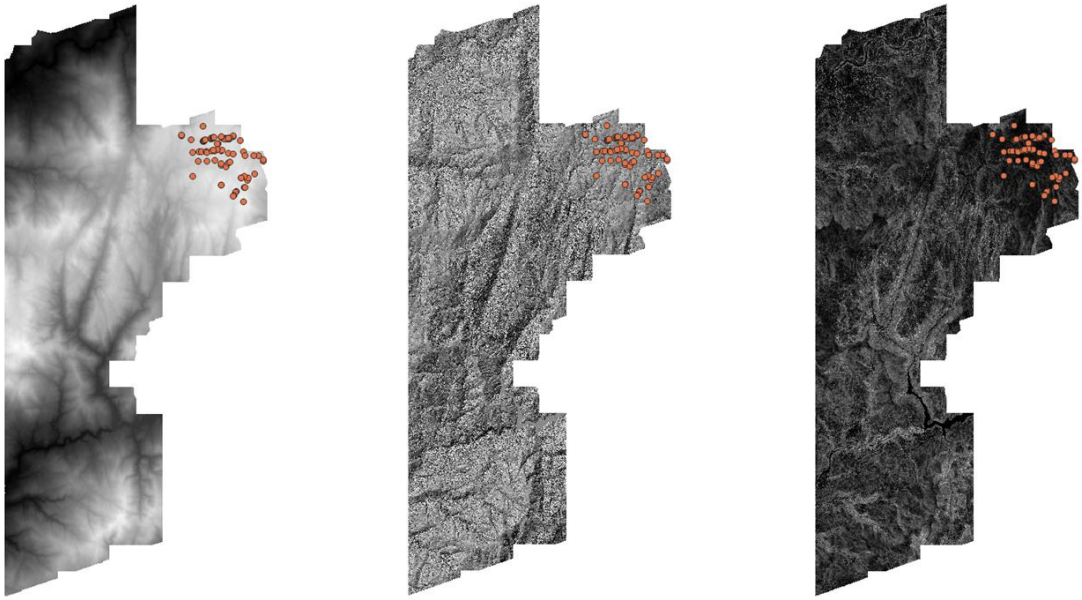

Figure 3 - DTM, LRM and Slope with shapefile showing archaeological mounds

The decision to give a margin of 1000x1000 meters around each mound was made, resulting in cropping the above images, with significant margin around each mound, but still accomplishing the intended reduction of size while including all that area around each mound.

For design purposes of the pre-processing project, an image with three bands was required, usually Red-Green-Blue (RGB). However, due to having three different VTs, it was chosen to combine all of them into a single image. The DTM and the Slope images were already in grayscale (single band), however, the LRM file had three bands for RGB, so this image had first to be converted to a grayscale image to match the DTM and Slope of single band images. This conversion was performed with the application of one of the most common RGB-to-grayscale methods, namely with the use of the National Television Standards Committee (NTSC) formula, as can be seen in the formula (1).

$$Y = 0{,}299 * R + 0{,}587 * G + 0{,}114 * B \quad (1)$$

With all the images in grayscale, was then possible to use QGIS combination tool to achieve a new output image which, each of the bands represented one of the VTs images, achieving the following result, still containing all the points of the mounds, as show on Figure 4, where it can be seen the combination image with the shape file of the mounds.

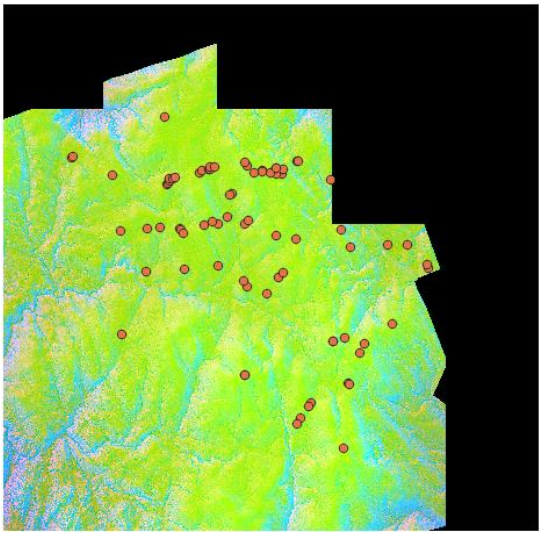

Figure 4 – Combination of DTM, LRM and Slope with identified points of mounds

Despite the shape file consisting of points, for the object detection purpose, it was required for it to be identified with the shape of a rectangle, so two rectangle shape files were generated, one with 15x15 meters and another with 30x30 meters.

This way, achieving two different datasets of different sizes, would also allow to perform some tests of the models on different dataset sizes, observing and comparing how the size difference would or would not be an influential factor for the prediction of the presence of archaeological mounds. For the remainder of this thesis each dataset with be named and referred to as mounds-30 or mounds-15, respectively for the 30x30 meters and the 15x15 meters.

It was then possible to run the pre-processing project to build a COCO format dataset for both sizes to, later, perform tests and experiments with the two and establish comparison and conclusions in terms of the model's performance.

Since there only existed 77 archaeological mounds on the dataset, which recognizably is a low base number for ML training, this COCO dataset was built to ensure the option of using a leave-one-out cross-validation (LOOCV) method, to enrich the options of training, while also avoiding or mitigating the risk of overfitting the models.

An alternative to the LOOCV was the k-fold cross-validation, however, from an already small dataset, it was thought that dividing to even smaller parts could further hinder the training process. Besides this, there would also be some risk of some of these smaller parts being comprised with the easiest classification images or the hardest images to classify which could lead to performance affected by this.

On the other hand, with the LOOCV, the training process would incur on all images except for one, ensuring the training on the maximum available images, with the obvious exception of a single image which would be later used for validation. The LOOCV would then iterate through

all the images replacing the image to be validated for a new one, as well as removing that new image from the training set but restoring the previous image left out for validation purposes.

For this purpose, the COCO format of the dataset consisted of a data folder with two subfolders, specifically an annotations and an images folder.

The “images” folder was basically the folder to store all the tiff image files of the dataset.

The “annotations” folder was used to store the JSON files for the dataset.

A global JSON file was kept which contained lists of properties for the categories, images and annotations. The categories list consisted of a single entry, since this thesis aimed to identify just archaeological mounds, however, if it would be a requirement to identify other archaeological findings, a newer category entry would have to be created for each of them. The images list consisted of objects with properties to identify an image by id, as well as save information of the file name of the respective image. Finally, the annotations list was created to establish a relation between the image, the category, the file name and the coordinates of the bounding box for the archaeological mound.

Besides this, a JSON file, with the same COCO format as the one explained on the global file, was created for each of the images of the dataset as validation set, each of them, also having a corresponding training set on another JSON file which contained information on all the dataset images except for the one left out for validation purposes.

This structure under the COCO format is what will enable the training of the models explained on the following section while being applied to the LOOCV technique to properly evaluate each of the model’s performance.

## 3.2 Benchmark and Models

This section will introduce and describe the DL models that have been selected for this thesis as well as the project chosen to help with its train and test.

Having realized that there already existed a project that helps with such benchmark analysis, even being used in the paper (Carranza-García et al., 2020), came as proof that it could also be used to perform the intended benchmark analysis of object detection for archaeological mounds. Besides this fact, the Python public repository named MMDetection (Chen et al., 2019) also had the upside of using the same COCO format that had been previously prepared, as input dataset for the DL models. The execution of this project was done resorting to PyTorch 2.0.1, with CUDA 11.8.

Another decisive fact was that this project already included the implementation of the most popular architectures for object detection that have been address earlier, while still displaying several other options.

This allowed the selection and testing of the DL models that were approached on the State of the Art chapter and, due to the diverse offer of models available on MMDetection, also allowed to perform tests on a few other models to enrich even further the benchmark analyses.

To establish a fair parametrization for all models, some parameters were determined to be fixed across all models, namely the number of classes to be classified as one (archaeological mound) and the use of a learning rate schedule equal across comparison. However, for experimentation purposes, two tests were performed, one on a 1x learning schedule and another one on a 2x learning schedule, which basically implied, a training process for 12 and 24 epochs, respectively. The remaining hyperparameters of each of the models were kept unchanged, preserving the default values of the respective papers which developed them. In terms of feature extractors for the models, mostly were preserved the use of the default backbone suggested in the respective papers, believing them to be the ones with the best performance, having used the ResNet50 for all of them, with exception to the use of DarkNet53 for YOLOv3 and VGG16 for SSD.

To also ensure a fairness in the comparison for all methods, every experimentation was performed on the same computer with an AMD Ryzen 7 5800X 8-Core processor and a NVIDIA GeForce RTX 3090 24GB GPU.

### 3.2.1 Two-Stage Detectors

Starting with the Two-Stage Object Detectors would have to undoubtedly talk about the Faster R-CNN architecture, since it's probably the most popular of them.

#### 3.2.1.1 Faster R-CNN

The architecture of Faster R-CNN (Ren et al., 2016), is basically the combination of a Region Proposal Network (RPN) with the model Fast R-CNN (Girshick, 2015) into a single network. The RPN divides an image into small regions which are potential areas where the objects might be present. This smart filtering analyses features extracted from the image using a CNN, which for this study will be a ResNet50. It will then undergo a technique of RoI pooling by resizing all the regions to a fixed size (200x200 pixels) which then will go through the process of classification and definition of the most suitable bounding box area. This process can be summarized as shown on Figure 5.

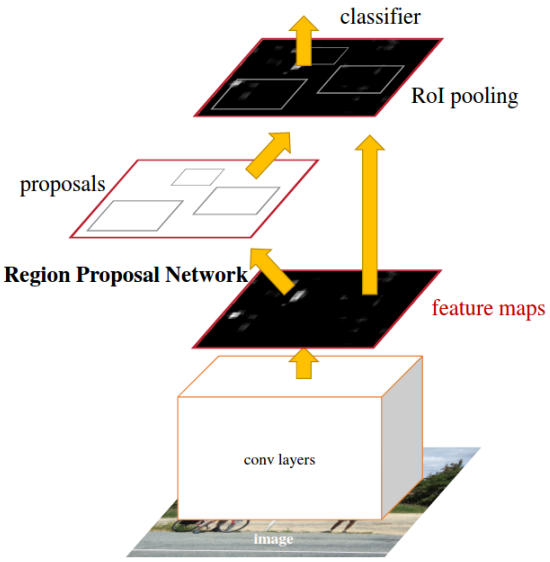


Figure 5 – Summarized example of Faster R-CNN (Ren et al., 2016)

#### 3.2.1.2 RPN

Although the Faster R-CNN already includes in its process the execution of the task of the RPN, since the benchmark repository included the option to run just himself as a model felt interesting to evaluate its performance and verify how much differs from the Faster R-CNN. It's expected to perform with an inferior accuracy, but still felt it could be an interesting test and comparison opportunity.

#### 3.2.1.3 Cascade R-CNN

This architecture is developed aiming to solve or minimize the problem of overfitting and quality mismatch between the detector and test hypotheses. This is suggested in the paper (Cai & Vasconcelos, 2019), by training sequentially detectors with increasing Intersection over Union (IoU) thresholds where each detectors output is provided as a training set to the next detector thus being able to increase the performance of object detection while also minimizing the risk of overfitting.

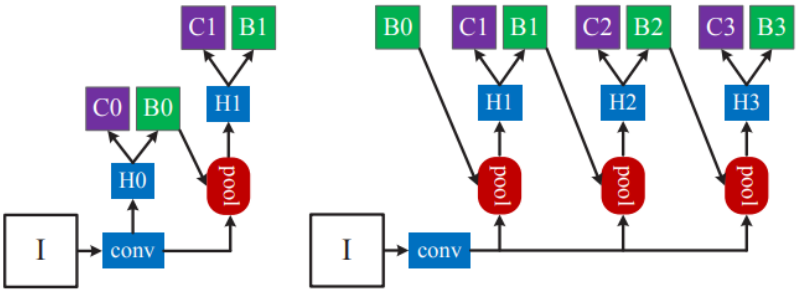


Figure 6 – Comparing architectures of Faster R-CNN (on the left) and Cascade R-CNN (on the right) (Cai & Vasconcelos, 2019)

#### 3.2.1.4 Dynamic R-CNN

The paper (Zhang et al., 2020), identifies a problem that affects the precision performance of two-stage object detection, such as an inconsistency problem between the fixed network settings and the dynamic training procedure. The Dynamic R-CNN, based on the statistics obtained on the training phase, suggests, automatically, as can be seen on Figure 7, an adjustment to the label assignment criteria (IoU threshold) and the shape of regression loss function to achieve an improvement of the average precision.

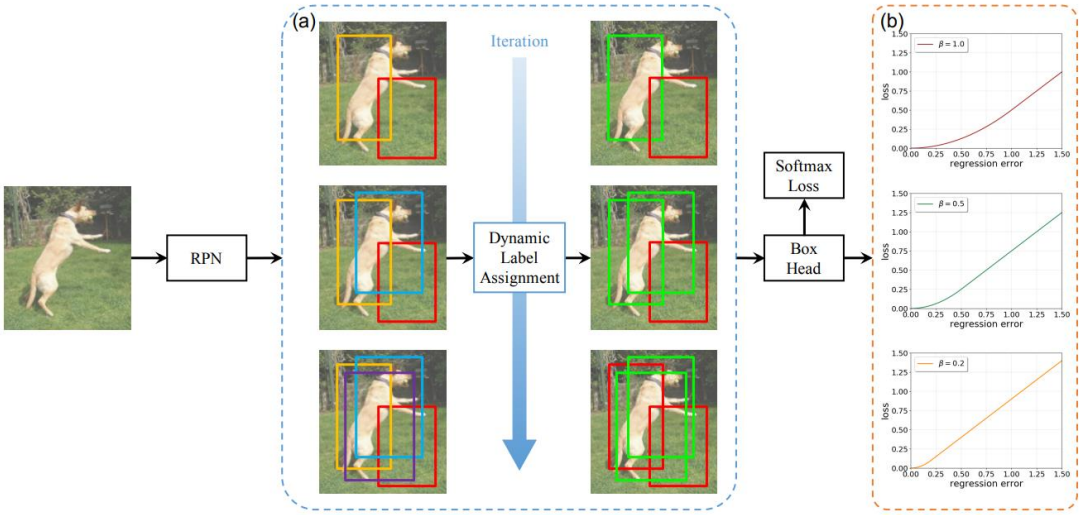


Figure 7 – Example of the pipeline of Dynamic R-CNN (Zhang et al., 2020)

### 3.2.2 One-Stage Detectors

For a good comparison, the One-Stage Object Detectors were subject to analyses so, similarly to the previous one-stage detectors, the choice to begin with probably the most popular architecture YOLO was made.

#### 3.2.2.1 YOLOv3

The YOLO architecture uses a grid-like approach by dividing an image into various cells where each of these cells are responsible for predicting a set of bounding boxes and probabilities to belong a specific class. The YOLOv3, according to the paper (Redmon & Farhadi, 2018), promises to be an improved version of its predecessors since it applies multi-scale detection, changes in its loss function and a more robust backbone since Darknet-53 is a better feature extractor when comparing to DarkNet-19 that was used on previous version of YOLO, as show on the summarized flow of operations, on Figure 8.

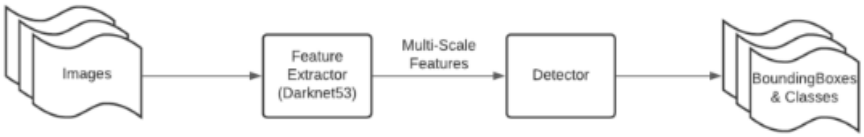


Figure 8 – Flow of work of YOLOv3 (Character et al., 2021)

#### 3.2.2.2 RetinaNet

The RetinaNet architecture is another very popular model of the one-stage detectors with the ability to detect objects and determine a bounding box around it on a single passage to the image. As shown in Figure 9, it combines a convolutional backbone for feature extraction (in the current case, RestNet50) followed by a Feature Pyramid Network (FPN), which provides a higher quality feature map. This model, according to (Lin et al., 2018), by introducing a new Focal Loss allowed the model to achieve the top speed of the one-stage predecessors detectors while still performing prediction accuracy to match the level of the state-of-the-art of the two-stage detectors.

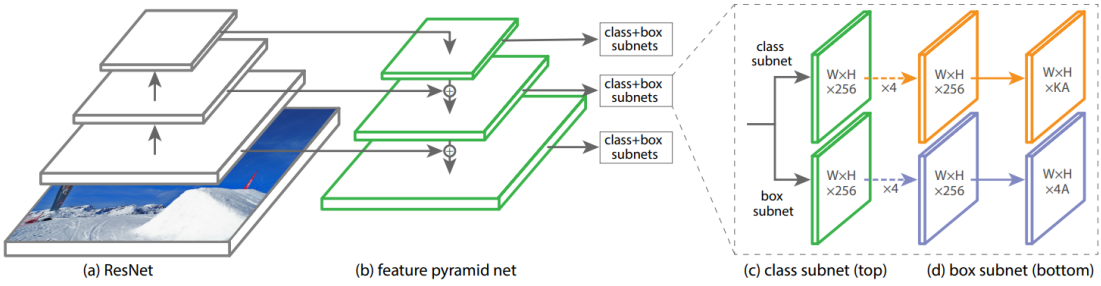


Figure 9 – Example of the architecture of RetinaNet (Lin et al., 2018)

#### 3.2.2.3 SSD

Using a single deep neural network to detect object in images is how easily can be characterized the model SSD. As can be seen in Figure 10, instead of using a grid-based approach like YOLO, anchor boxes are employed at different aspect ratios and scales to identify the presence of objects while generating scores for each box and thus performing adjustments to improve the identification of the objects full shape. According to the paper (Liu et al., 2016), SSD models rival in terms of accuracy to models that use the additional step of object proposal while achieving a much faster performance for training and inference.

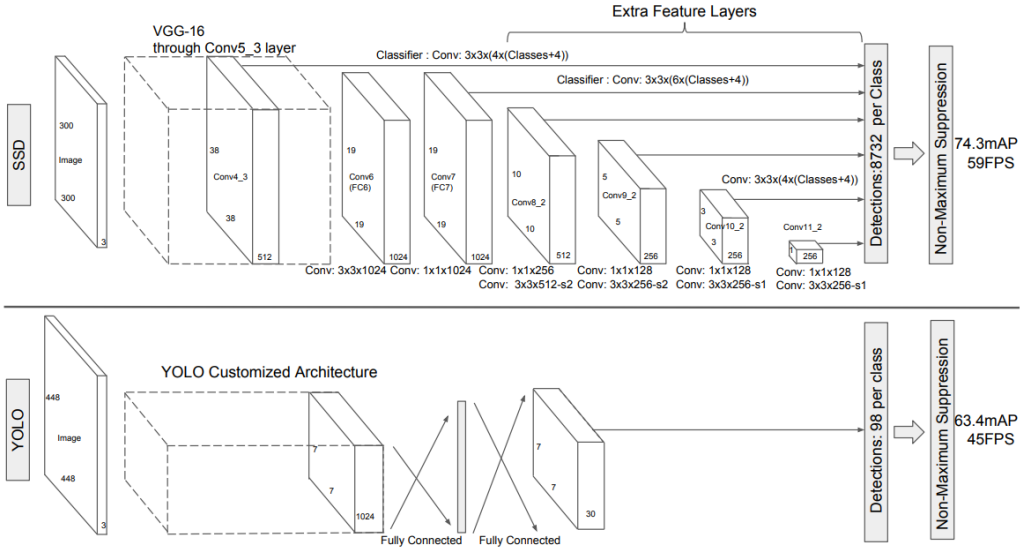


Figure 10 – Comparing the SSD and YOLO architecture (Liu et al., 2016)

### 3.2.2.4 Fully Convolutional One-stage Object Detection (FCOS)

The FCOS architecture, proposed in the paper (Tian et al., 2019) and shown in Figure 11, similarly to a semantic segmentation method, attempts to perform the object detection task in a per-pixel predicting analysis. What is the major difference from this method to the other mentioned one-stage detectors, and even the two-stage detector Faster R-CNN, is that FCOS does not rely on proposals or even anchor boxes. This avoids the traditional sensitiveness of the hyper-parameters around the anchor box definition, instead, only performing a post-processing non-maximum suppression (NMS), but still achieving an improvement to object detection accuracy.

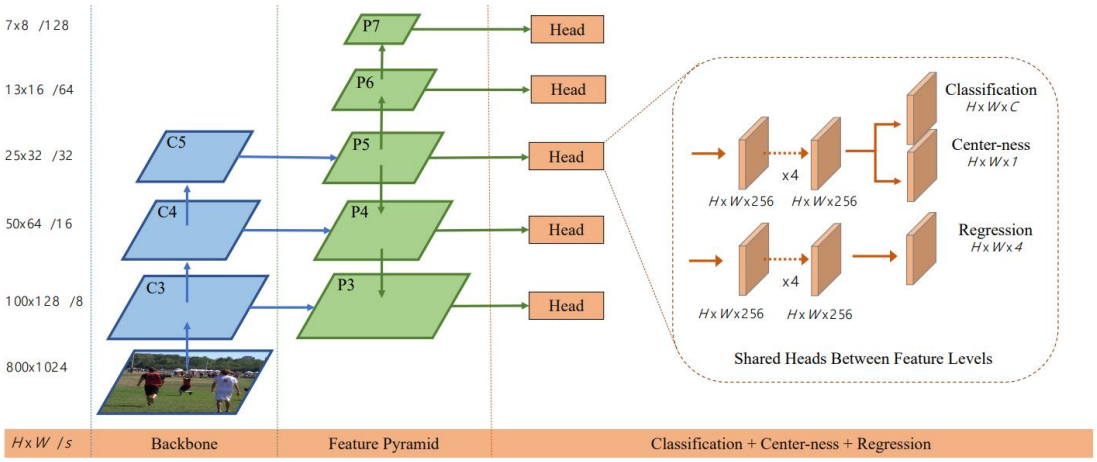


Figure 11 – Example of an FCOS architecture (Tian et al., 2019)

#### 3.2.2.5 Task-aligned One-stage Object Detection (TOOD)

Most of the one-stage object detection methods, focuses on optimizing two sub-tasks such as object classification and localization, by using two distinct and parallel branches, which can lead to a divergence of predictions between the two tasks. TOOD, in the paper (Feng et al., 2021), proposes an alignment to these tasks to improve its learning capability, as can be seen in Figure 12. This is possible by the development of a new task-aligned head better balancing the task-specific features and the learning task-interactive. The proposed Task Alignment Learning also provides an approximation or even unification of the optimal anchors of both tasks providing the intended better alignment which allows to surpass other one-stage object detection in terms of average precision.

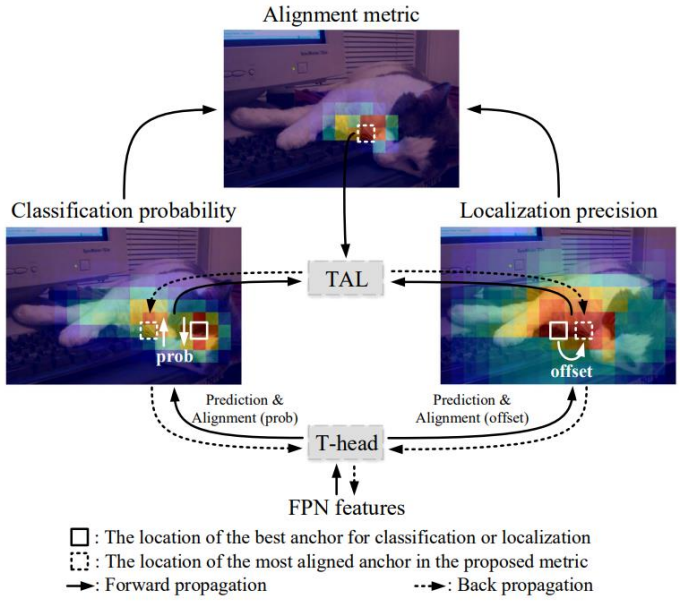


Figure 12 – Exemplification of TOOD learning mechanism (Feng et al., 2021)

## 3.3 Data Privacy, Security Analysis and Ethical Issues

The increasingly rapid pace of importance and utilization of the internet and digital and online tools has also been accompanied by a growth in the risks of cybercrime (Hunton, 2009). In fact, it was noticeable that, as a result of the consequences of the Covid-19 pandemic, with a greater reliance on computer systems and tools, there has also been an increase in the number of attempts at various types of cybercrimes, such as phishing, malware, denial-of-service (DoS) attacks, malicious social media messaging (MSMM), among others (Alawida et al., 2022).

Despite AI bringing many benefits and tools that aid in combating cybercrime, the truth is that it also has some vulnerabilities that need to be properly safeguarded because at various stages of its life cycles, it can be susceptible to different types of attacks (Hu et al., 2021).

However, in accordance with the growing concerns about security associated with the application of AI techniques or tools, this project will also take precautions. The project will be developed in a Bitbucket[5] repository, which will be set to private to limit and prevent unauthorized access. This way, it will be possible to maintain control over any changes made to the project, as well as over the data through the visualization and tracking of changes made to the project's versions in the repository.

As was evident in the previous chapter, the data to be processed in this thesis will consist of readings from LiDAR, which do not contain any sensitive, private, or personal information about any specific individuals or entities. These data are exclusively comprised of LiDAR readings from areas that are not even private or residential properties, so there are no concerns regarding data protection or ethical issues.

[5] https://bitbucket.org/

# 4 Results and Discussion

This chapter is going to address the metrics used to evaluate and establish a comparison between the different chosen architectures. Their performance will also be discussed and analysed when comparing themselves on the two available datasets to perform the objective of this thesis: object detection benchmark analysis.

## 4.1 Evaluation Metrics

There are several metrics that can be used to evaluate the performance of a model. Since this thesis address object detection from image, the choice fell for two of the most popular evaluation techniques for this scenario: Confusion Matrix and mean Average Precision (mAP).

### 4.1.1 Confusion Matrix

The confusion matrix is one of the several forms to perform an evaluation to a Machine Learning (ML) model classification. A comparison is realized between the prediction class of the ML model and actual value of the object to check if the classification process was accurate.

Depending on how this classification is accurate or not, a determination if it was a True Positive (TP), True Negative (TN), False Positive (FP) or False Negative (FN) is made.

TP means that the model classified the object with the right class of object.

TN on the other hand, is attributed when the model correctly states that the object does not belong to the class of object.

FP occurs when the model predicts the object as a part of the evaluating class, however that assessment is wrong.

FN is when the model predicts that the object doesn't belong to the class when in fact it does.

An example of a type of Confusion Matrix can be like shown in Figure 13.

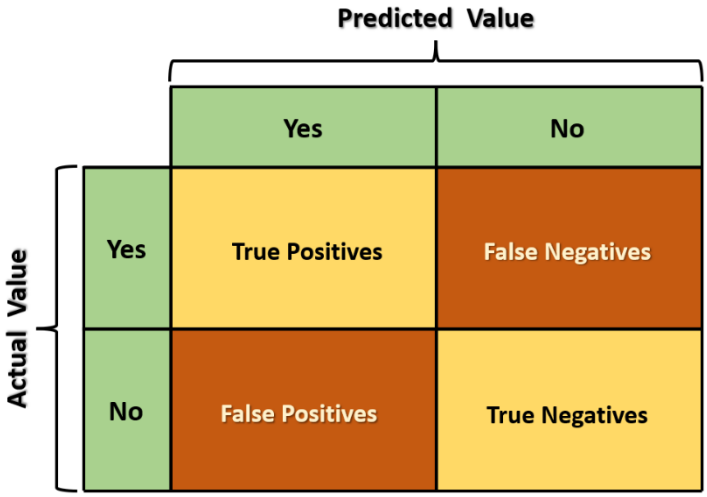


Figure 13 – Possible example of a Confusion Matrix

It's easily understandable that the desired result is to have the highest possible scores of TP and TN while minimizing the occurrence of FP and FN.

### 4.1.2 Mean Average Precision (mAP)

To further help evaluate the accuracy of the model's prediction, of archaeological mound detection, the metric of mAP was chosen since it's one of the most used and common metrics to establish evaluation to object detection problems.

To help with the quantification of the precision, an IoU metric is used to quantify how much a prediction area overlaps the actual value area to assess its precision of the prediction, which can be calculated as shown in (2).

$$IoU\ (A, B) = \frac{\|A\ \cap B\|}{\|A\ \cup\ B\|} \tag{2}$$

The prediction is considered correctly depending on the percentage of overlap that occurs from the prediction bounding box, to the actual bounding box of the object. The Figure 14, provides an example for this, where if this percentage is higher than the threshold defined for the IoU, it's considered to be a good prediction.

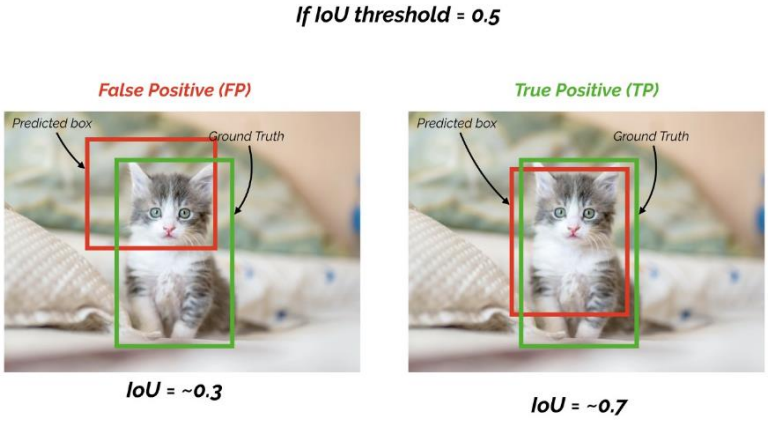


Figure 14 – Example of IoU application to evaluate prediction accuracy

Usually, and for this thesis purposes, the mAP will be calculated considering IoU threshold of 0.5, of 0.75 and the standard between 0.5 and 0.95.

### 4.1.3 Recall

Given that the confusion matrix provides with the number of True Positives, False Negatives and False Positives, this enables the calculation of the Recall metric.

This metric is also one of the most used in object detection since it provides information of how many of the intended objects were identified out of the total of those objects present in the dataset.

As can be seen in (3), this metric evaluates the True Positives comparatively to all of the True Positives and the False Negatives.

$$Recall = \frac{True\ Positives}{(True\ Positives + False\ Negatives)} \quad (3)$$

## 4.2 Benchmark Analysis

Overall, the results obtained were mostly as expected for both datasets, either with the archaeological mounds of size 30x30 meters (mounds-30) or of 15x15 meters (mounds-15). Most of the two-stage object detectors out-performed the one-stage detectors in terms of average precision, while most of the one-stage object detectors implied lesser time of computational processing. For each of the datasets, a comparative analysis will be conducted firstly on the 1x learning rate schedule of 12 epochs and afterwards, verify the difference of results from the application of the 2x learning rate schedule of 24 epochs.

### 4.2.1 Dataset of mounds-30 with 1x learning rate schedule of 12 epochs

A comparative analysis of the performance of the DL models that were tested on this thesis can be observed in Figure 15, since it shows the combination of the Confusion Matrix of each of the models, evaluating the prediction capacity of them.

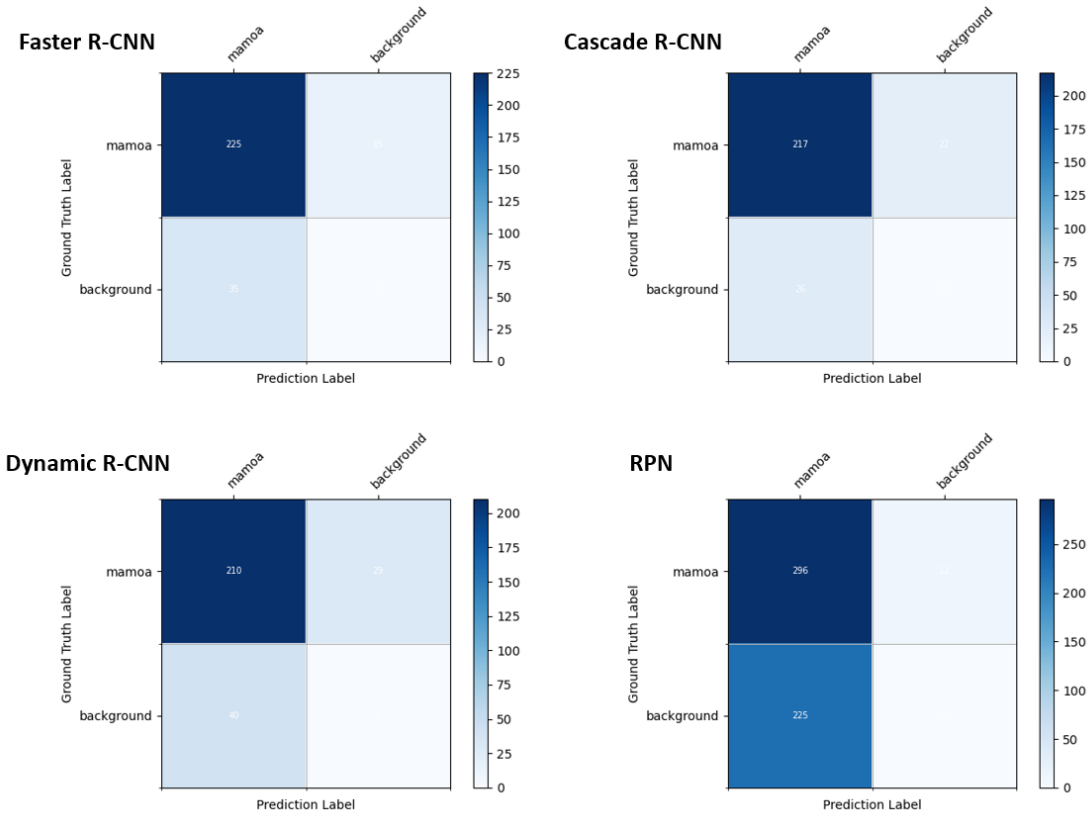


Figure 15 - Confusion Matrix for the two-stage detectors, executed on dataset mounds-30, for 12 epochs

The Faster R-CNN architecture provides a combination of results that attest to the expected good performance of this model. It accomplishes a total of 225 true positive identifications against 15 false negatives and 35 false positives. An example of this model's successful identification can be observed on the image Figure 16.

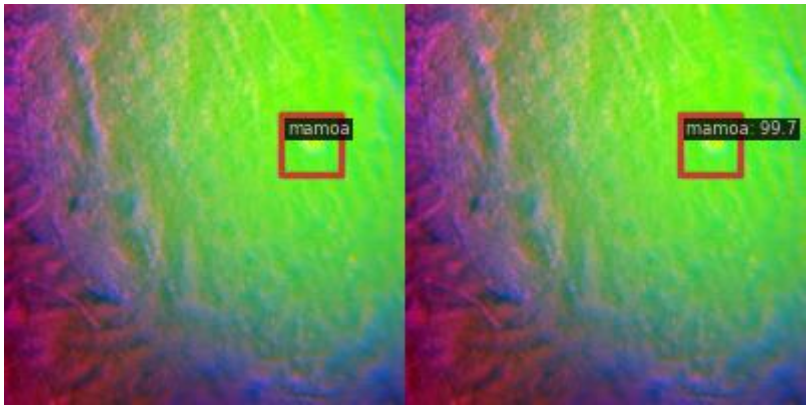


Figure 16 - Example of a good prediction with Faster R-CNN, on dataset mounds-30, for 12 epochs

The architecture of Cascade R-CNN is another model with a fairly good performance and an occurrence of errors not very focused on a type. It generates a total of 217 true positives, while incurring in 22 false negatives and 26 false positives. One example of its good predictions can be viewed in Figure 17.

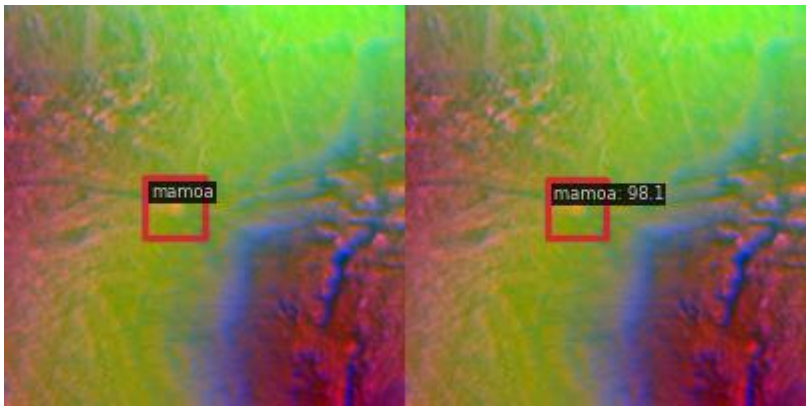


Figure 17 - Example of a good prediction with Cascade R-CNN, on dataset mounds-30, for 12 epochs

The Dynamic R-CNN, although it still accomplishes a significant number of true positives identifications, to a total of 210, incurs in a more significant error of 40 false positives against 29 false negatives. The Figure 18, shows an example where the model wrongly identifies an archaeological mound on the lower-left side of the image, however, our ground-of-truth image shows no existing mound on that area.

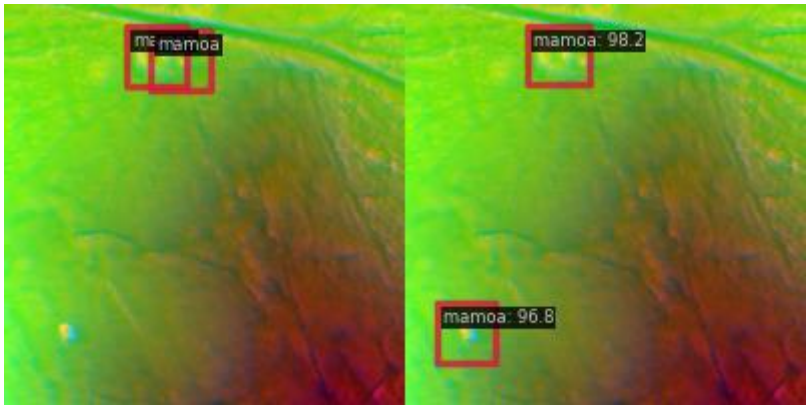


Figure 18 - Example of a bad prediction with Dynamic R-CNN, on dataset mounds-30, for 12 epochs

The last of the two-stage object detectors, RPN, although achieves a high true positive identification, shows to have a serious problem when identifies 225 false positives, against just 12 false negatives. The fact that achieves a value of 296 true positives, also shows that has a problem typical with the usage of this model alone by overlapping bounding boxes. This still happens even with the application of an NMS to attempt to avoid or minimize this. The Figure 19, shows an example of a false positive by identifying a mound where none exists on the ground-of-truth as well as showing the overlapping issue mentioned before.

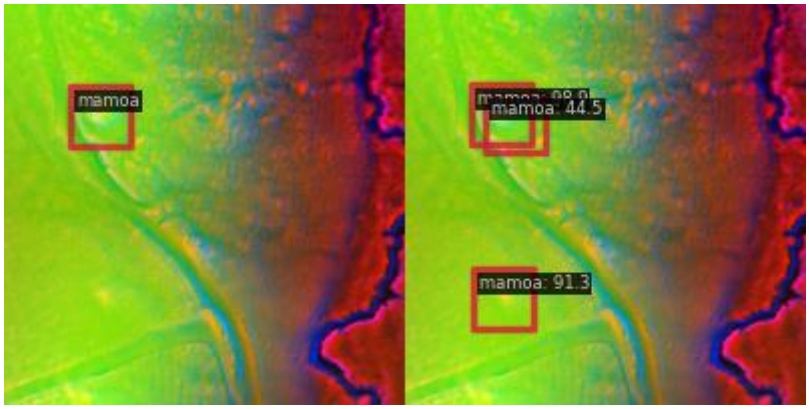


Figure 19 – Example of a bad prediction with RPN, on dataset mounds-30, for 12 epochs

A similar test on the dataset mounds-30, for 12 epochs, was performed on the one-stage object detectors, which originated the confusion matrix presented in Figure 20.

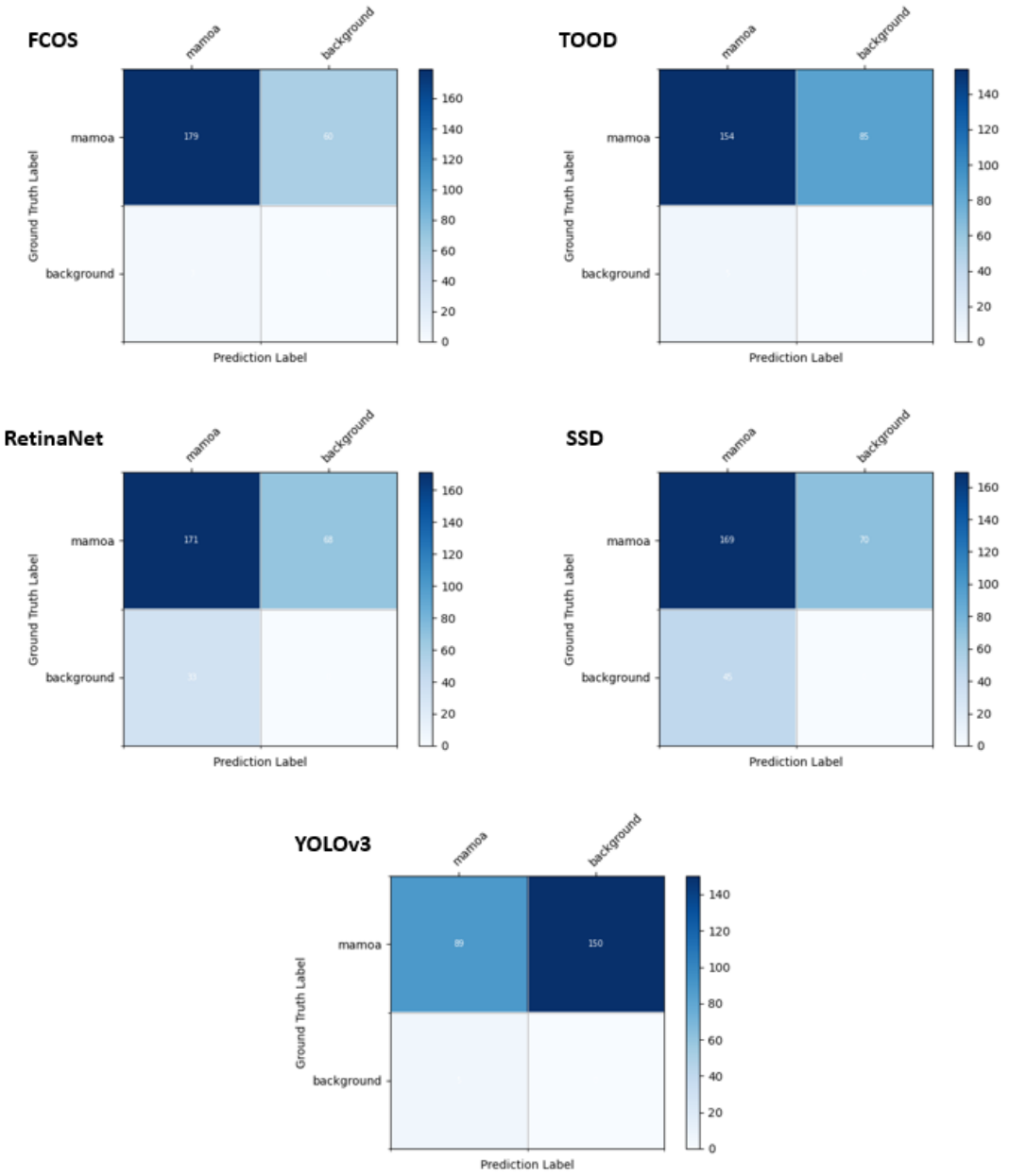


Figure 20 – Confusion Matrix for the one-stage detectors, executed on dataset mounds-30, for 12 epochs

Despite encountering correctly 179 true positive identifications of archaeological mounds, the FCOS model shows a tendency to incur in false negative mistakes since it happened 60 times, only having 3 false positives. The Figure 21, shows an example of an archaeological mound that was wrongly overlooked by the model when the ground-of-truth confirms the existing of the archaeological object.

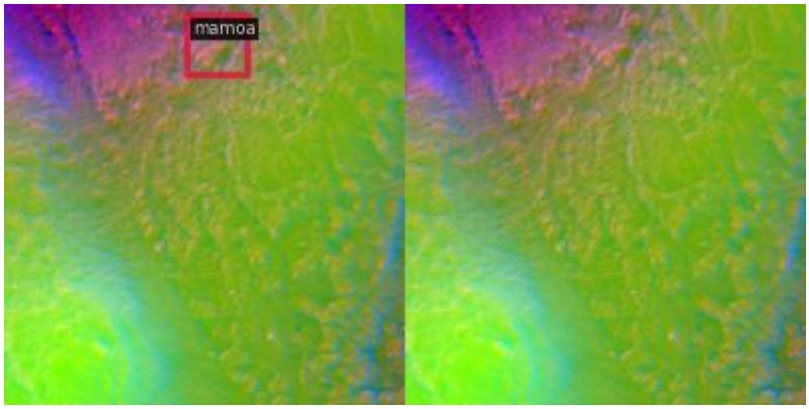


Figure 21 – Example of a bad prediction with FCOS, on dataset mounds-30, for 12 epochs

Similarly to FCOS, the TOOD architecture also presents the same problems of incurring most frequently in false negatives on 85 times, against 5 false positives. Despite these numbers, the model still achieves a total of 154 true positive identifications. Likewise, the Figure 22, also shows the presence of a mound on the ground-of-truth that failed to be identified by the model.

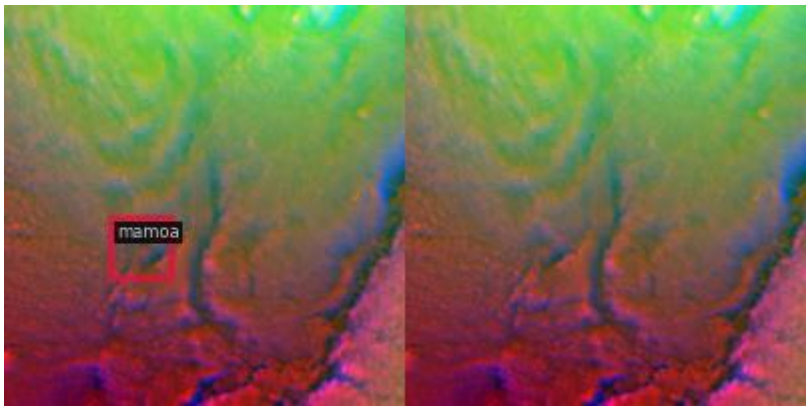


Figure 22 – Example of a bad prediction with TOOD, on dataset mounds-30, for 12 epochs

The model RetinaNet, comparatively to the other models, shows an increase in the occurrences of false positives, incurring in a total of 33. It still shows 68 false negatives, although achieving a total of 171 true positives. In the Figure 23, can be seen that, although correctly identifies an archaeological mound on the top-right side of the image, the centre of it, identifies wrongly the presence of archaeological mounds, since not even one exists on the ground-of-truth part of the image.

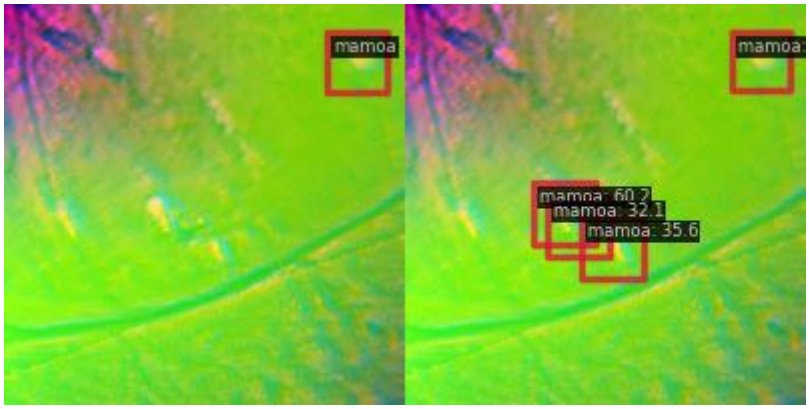


Figure 23 – Example of a bad prediction with RetinaNet, on dataset mounds-30, for 12 epochs

The SSD model continues with the tendency of increasing in the false positive identifications to totalize a number of 45 false positives. It still incurs in 70 false negatives while achieving 169 true positive identifications of archaeological mounds. The Figure 24, shows a successful identification of an archaeological mound, but on bottom-right a wrongly identified archaeological mound is also identified by the model.

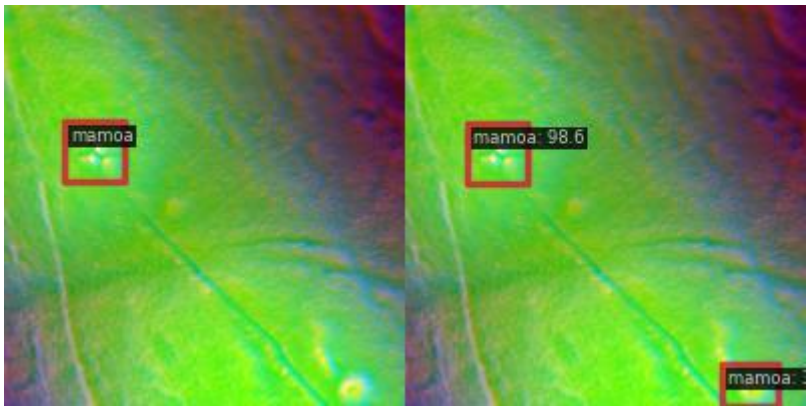


Figure 24 – Example of a bad prediction with SSD, on dataset mounds-30, for 12 epochs

Finally, the YOLOv3 model has proven to have difficulty in avoiding false negative predictions since it incurred in 150. It's true that it only identified 5 false positives, but it also achieved a considerably low value of 89 true positives. The Figure 25, is a good example of this model's tendency for false negatives since it shows on the ground-of-truth two archaeological mounds, but the model failed to identify a single one of them.

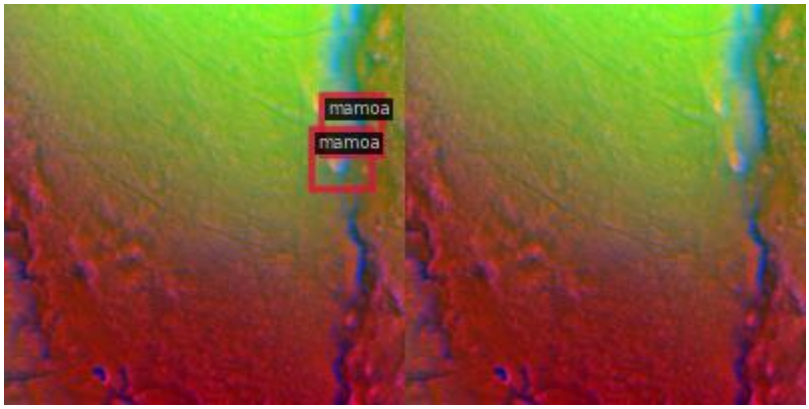


Figure 25 - Example of a bad prediction with YOLOv3, on dataset mounds-30, for 12 epochs

A different option for comparison of these models is through the interpretation of mAP with different threshold of IoU, the processing time to run the models, or the recall metric. These calculated values can be observed in Table 7.

Table 7 – Results obtained of all models with dataset mounds-30 and trained for 12 epochs (best results in bold)

| **mounds-30 (12 epochs)** | | | | | | |
|---|---|---|---|---|---|---|
| **Architecture** | **Category** | **Full processing time (hh:mm)** | **mAP** | **$mAP_{0.5}$** | **$mAP_{0.75}$** | **Recall** |
| Faster R-CNN | Two-stage detector | 02:57 | 66.6 | 90.8 | **80.4** | 93.8 |
| Cascade R-CNN | Two-stage detector | 04:21 | 67.7 | 90.9 | 79.9 | 90.8 |
| Dynamic R-CNN | Two-stage detector | 03:00 | 60.9 | 86.3 | 72.7 | 87.9 |
| RPN | Two-stage detector | 02:19 | 46.6 | 83.2 | 47.0 | **96.1** |
| FCOS | One-stage detector | 03:11 | **69.2** | **91.0** | **80.4** | 74.9 |
| TOOD | One-stage detector | 04:36 | 51.7 | 80.9 | 61.3 | 64.4 |
| RetinaNet | One-stage detector | 02:59 | 48.2 | 76.0 | 57.8 | 71.5 |
| SSD | One-stage detector | **02:11** | 44.5 | 70.2 | 56.0 | 70.7 |
| YOLOv3 | One-stage detector | 03:06 | 30.9 | 64.7 | 23.2 | 37.2 |

Surprisingly, it wasn't a two-stage detector to achieve the better mAP. Instead, the model FCOS, proven himself to be the model with the highest mean average precision of the different IoU threshold tested. Only for the IoU of 0.75, was the model Faster R-CNN able to equal its value to 80.4. Apart from this exception, the remaining one-stage models were not able to match the superiority of the performance of the two-stage detectors when analysing the metric of mAP.

As for the Recall metric, the prediction of the two-stage models being of a higher performance to identify objects is proven, since all of them achieve a higher Recall value compared to the one-stage detectors. For this specific study, the architecture of RPN proved to be the one to successfully identify most of the existing archaeology mounds.

In terms of processing time, the one-stage model SSD proved to be the fastest one of all the tested models and despite this, still managed to present higher mAP than the model YOLOv3, for example.

### 4.2.2 Dataset of mounds-30 with 2x learning rate schedule of 24 epochs

This section, the analysis is conducted on the same dataset, however changing the learning rate schedule, increasing from the 12 epochs to 24 epochs, to verify if originates improvement on the deep learning models.

A possible analysis to the performance of the DL models can be done through the interpretation of the Confusion Matrix of each of the models, as can be seen in Figure 26, for the two-stage object detectors trained on the dataset mounds-30.

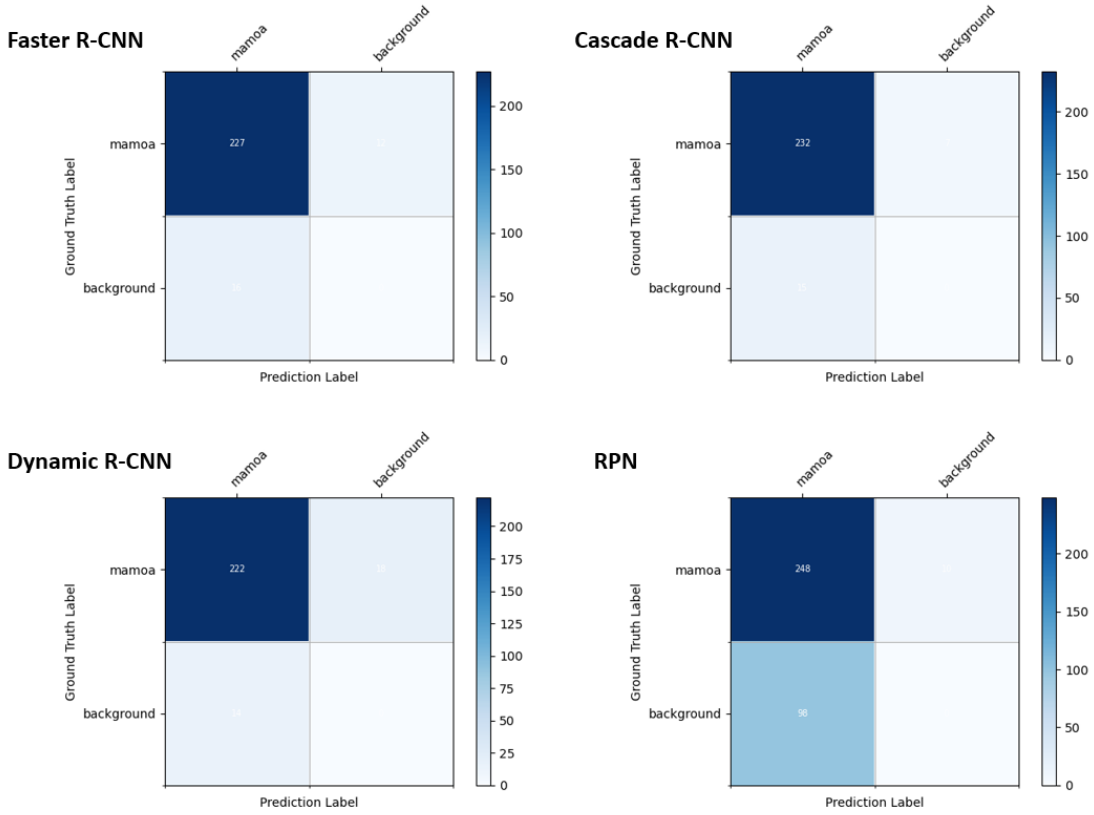


Figure 26 – Confusion matrix for the two-stage detectors, executed on dataset mounds-30, for 24 epochs

The model Faster R-CNN doesn't show a significant tendency of errors since it presents 227 true positive predictions against just 12 false negative and 16 false positive, so the mistakes, not only are residual but are also similarly dispersed between false negative and false positives. A good example of prediction from this modal can be seen on the Figure 27, chosen with a 100% precision since on this and future examples, on the left side is the ground of truth, while on the right side is the models prediction.

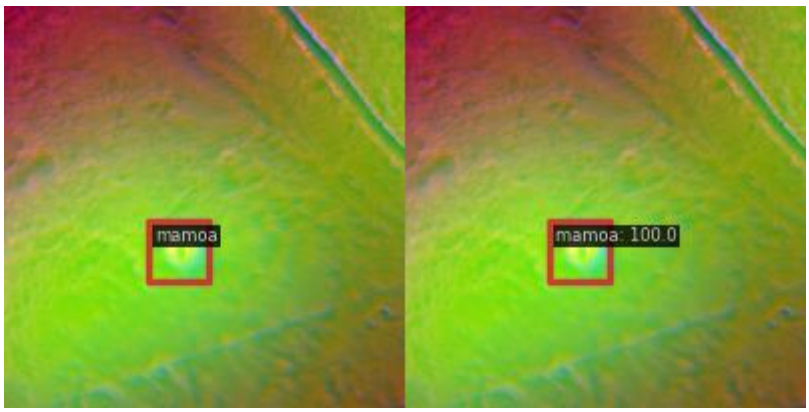


Figure 27 – Example of a good prediction with Faster R-CNN, on dataset mounds-30, for 24 epochs

The Cascade R-CNN model could accurately predict 232 occurrences as true positive against 7 false negatives and 15 false positives, so there is a slightly bigger tendency to incur in false positive predictions like shown in the Figure 28, since it predicted with 62.7% the presence of an archaeological mound (right side of the figure) when it isn't confirmed with our ground of truth (left side of the figure). This image is also a good example of a false negative since there actually existed a verified archaeological mound (left side of the figure), but the model failed to predict it, considering wrongly to be negative to the presence of a mound.

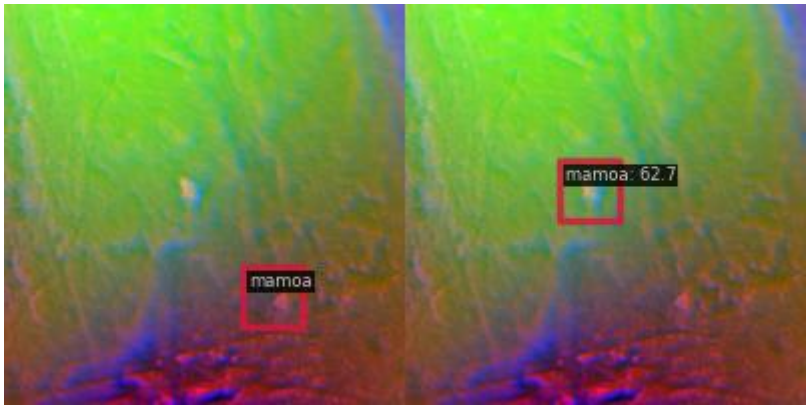


Figure 28 – Example of a bad prediction with Cascade R-CNN, on dataset mounds-30, for 24 epochs

The Dynamic R-CNN model achieves a total of 222 true positive predictions while producing 18 false negatives and 14 false positive predictions, so in this case, a slightly higher tendency to incur in false negatives, just like observed in Figure 29.

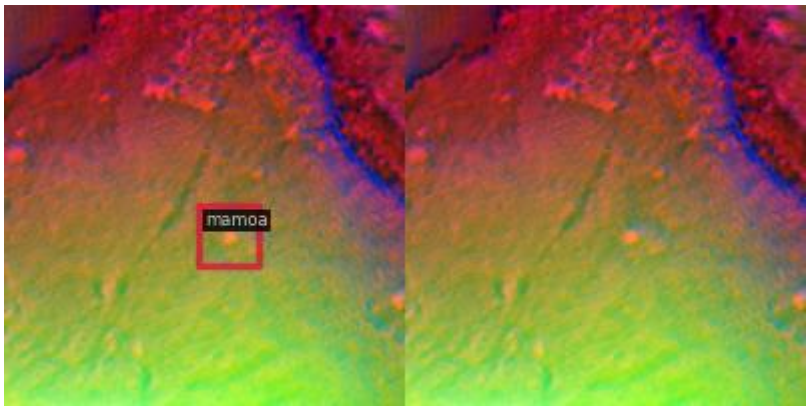


Figure 29 – Example of a bad prediction with Dynamic R-CNN, on dataset mounds-30, for 24 epochs

The last model of the two-stage object detectors, namely RPN, despite encountering 248 true positive results, also found 10 false negatives and high 98 of false positives which clearly stands-out as the highest problem of this model. As can be seen in Figure 30, the left side, which is our ground of truth, reports the presence of two archaeological mounds while the

RPN model, on the right side, predicts the presence of five mounds, meaning it incurs in 3 false positive evaluations.

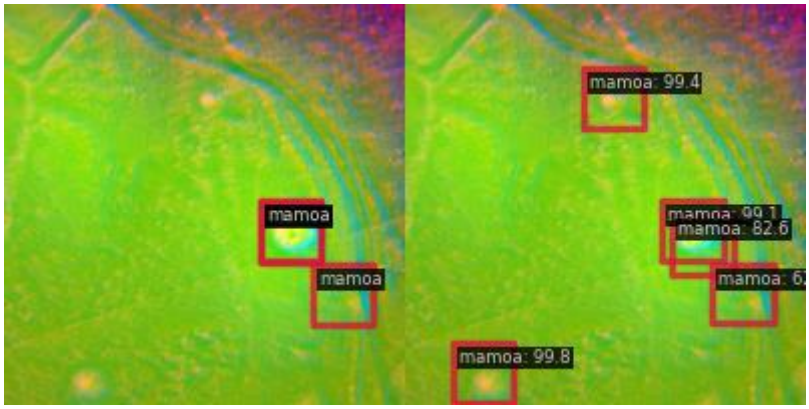


Figure 30 – Example of a bad prediction with RPN, on dataset mounds-30, for 24 epochs

Similarly, for the one-stage object detectors tested in this thesis, the confusion matrix for the dataset of mounds-30 were generated as observed in the Figure 31.

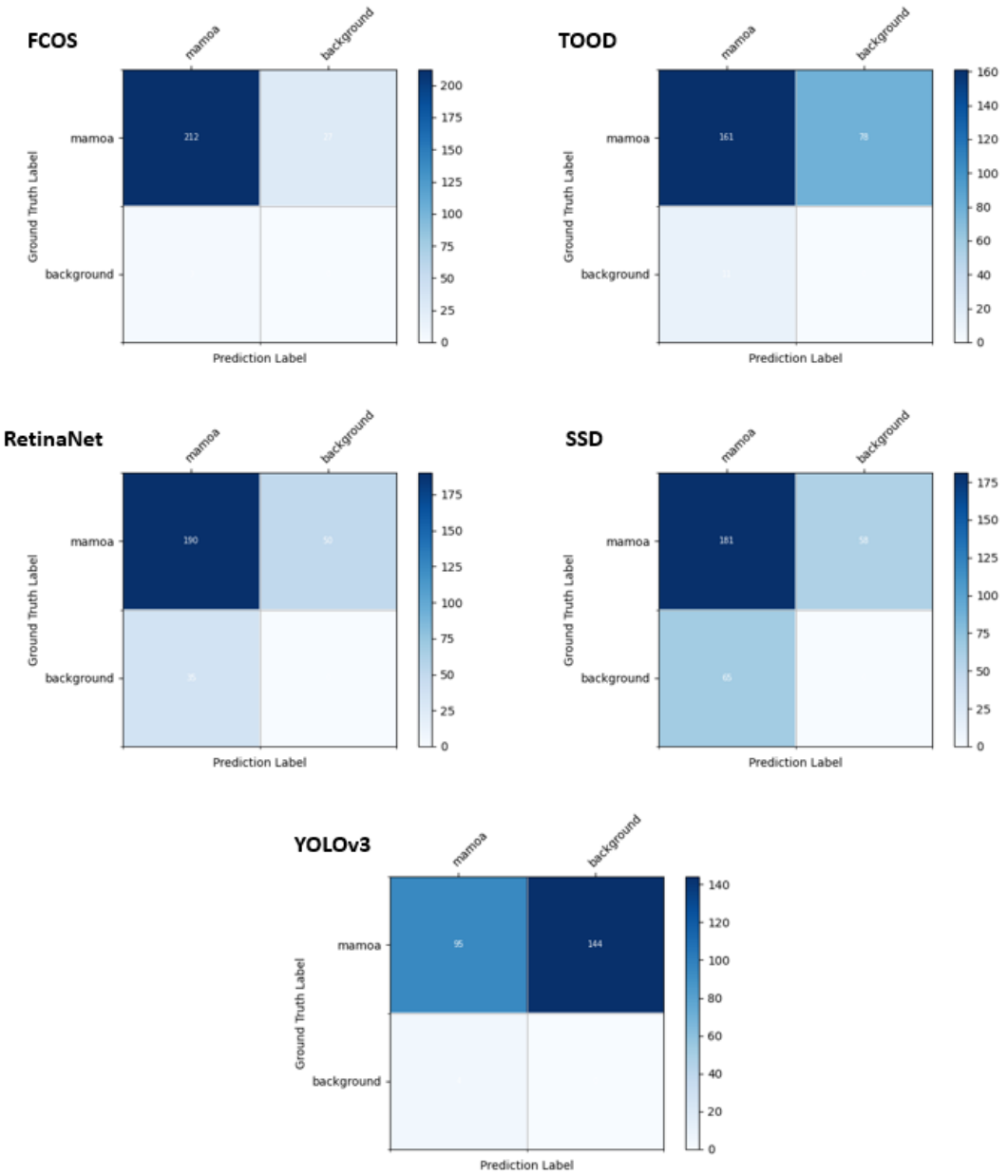


Figure 31 - Confusion matrix for the one-stage detectors, executed on dataset mounds-30, for 24 epochs

The FCOS model, despite encountering correctly 212 true positive occurrences of archaeological mounds, has a distinct tendency to incur in mistakes of false negatives since it happened 27 times when compared with just 3 false positives. An example of one of these false negative mistakes can be observed in the Figure 32.

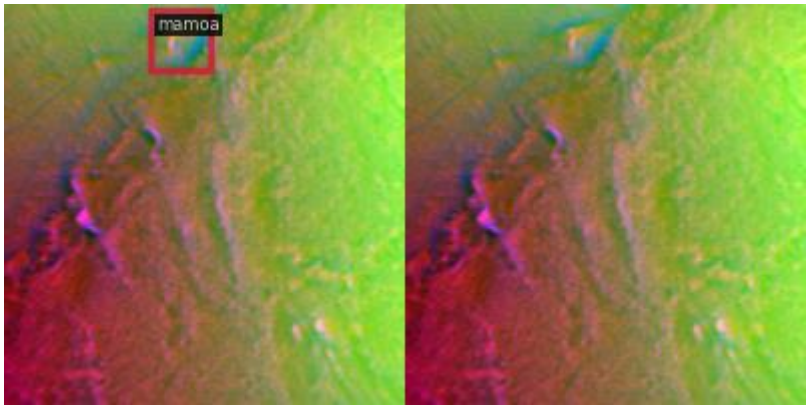


Figure 32 - Example of a bad prediction with FCOS, on dataset mounds-30, for 24 epochs

TOOD's architecture, presents a similar difficulty than the one observed in FCOS but with a higher impact since it presents 78 false negatives (like the example shown in Figure 33), just 11 false positives while being able to achieve 160 true positives identified.

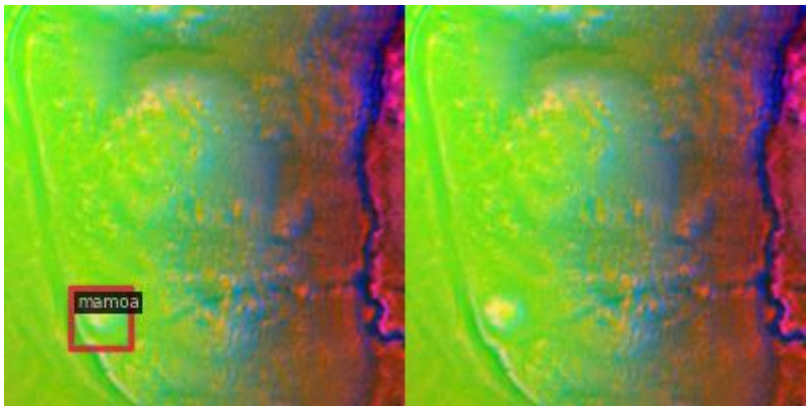


Figure 33 - Example of a bad prediction with TOOD, on dataset mounds-30, for 24 epochs

The RetinaNet model has a more balanced distribution of errors between false negatives (50 occurrences) and false positives (35 occurrences) while being able to present 190 true positives, a higher number when compared to TOOD. The Figure 34, is a good example since it shows both mistakes. The right side of the figure predicts the presence of a mound on the bottom left of the screen but incurs in a false positive since no mound exists in our ground of truth (left side of the figure). Besides this, the ground of truth identifies two mounds, but the lower one is not identified on the model prediction, incurring in a false negative.

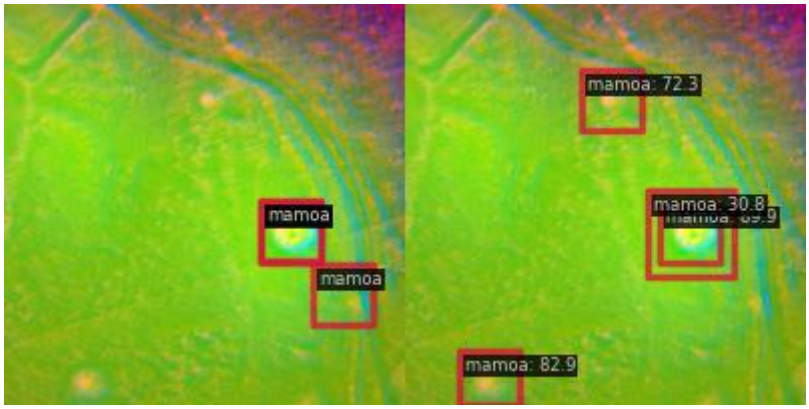


Figure 34 - Example of a bad prediction with RetinaNet, on dataset mounds-30, for 24 epochs

A similar behaviour is observed on the confusion matrix for the model SSD, although with less true positives (181 predictions) and higher false negatives (58 predictions) and false positives (65 predictions) where an example of a false positive can be verified in Figure 35.

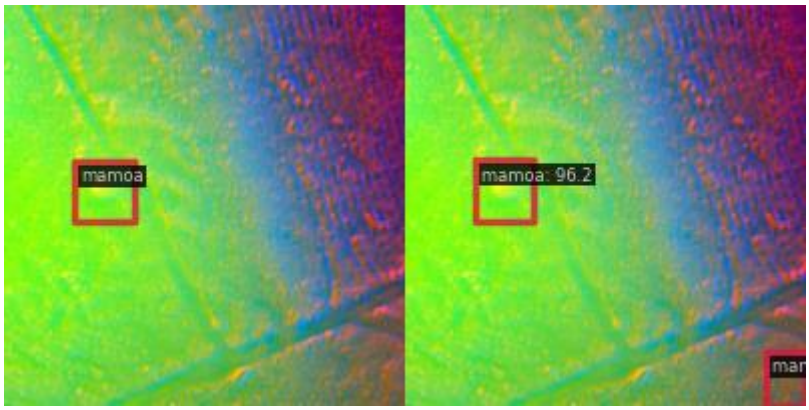


Figure 35 - Example of a bad prediction with SSD, on dataset mounds-30, for 24 epochs

The last of the one-stage detectors, namely YOLOv3, presents the lowest ability to predict true positives (just 95 accurate predictions) and while almost with no false positives (just occurring 4 times), displays 144 predictions of false negatives. An example of a false negative from this model can be viewed as presented in Figure 36.

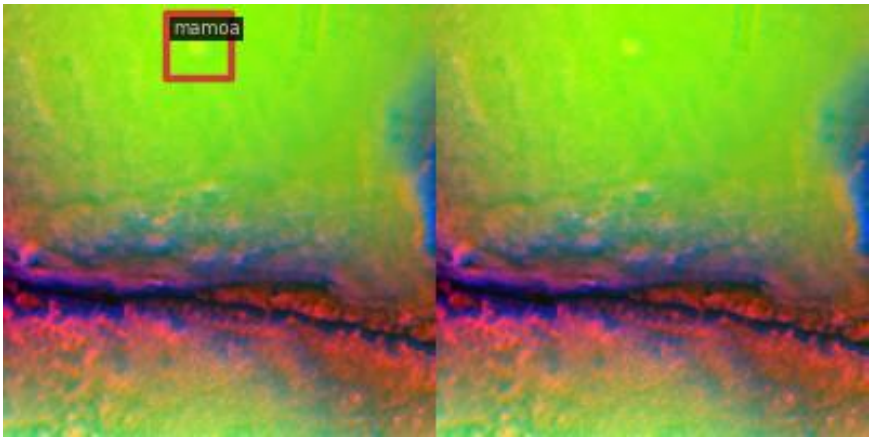


Figure 36 - Example of a bad prediction with YOLOv3, on dataset mounds-30, for 24 epochs

On a different analysis approach, these models can also be compared and analysed through the interpretation of mAP with different threshold of IoU or even of the processing time to run the models. This comparison information can be view in Table 8.

Table 8 – Results obtained of all models with dataset mounds-30 and trained for 24 epochs (best results in bold)

| **mounds-30 (24 epochs)** | | | | | | |
|---|---|---|---|---|---|---|
| **Architecture** | **Category** | **Full processing time (hh:mm)** | **mAP** | **$mAP_{0.5}$** | **$mAP_{0.75}$** | **Recall** |
| Faster R-CNN | Two-stage detector | 09:51 | 77.2 | 93.5 | **90.5** | 95.0% |
| Cascade R-CNN | Two-stage detector | 08:07 | 76.0 | 93.1 | 87.1 | **97.1%** |
| Dynamic R-CNN | Two-stage detector | 05:44 | 70.0 | 90.8 | 83.4 | 92.5% |
| RPN | Two-stage detector | 04:08 | 54.7 | 89.1 | 57.1 | 96.1% |
| FCOS | One-stage detector | 07:19 | **78.6** | **93.9** | 89.7 | 88.7% |
| TOOD | One-stage detector | 08:42 | 55.0 | 79.5 | 68.1 | 67.4% |
| RetinaNet | One-stage detector | 07:52 | 54.4 | 81.0 | 64.3 | 79.2% |
| SSD | One-stage detector | **03:47** | 46.8 | 71.7 | 58.3 | 75.7% |
| YOLOv3 | One-stage detector | 08:40 | 36.5 | 69.8 | 34.5 | 39.7% |

There were some surprises, mostly since the model that achieved the highest average precision turned out to be a one-stage object detector, namely the FCOS. It wasn't however a value so much higher than, for example, the two-stage object detector Faster R-CNN, which was the highest performance of their category. The difference between them were merely between 0,4 to 1,6 percentual points.

The increase in the number of training epochs to 24, did not affect the hegemony of the two-stage model detectors as they still prove to achieve a higher performance to identify objects, since all of them achieve a higher Recall value compared to the one-stage detectors. It did, however, switch the leading architecture from the RPN to the Cascade R-CNN as the best performance on this metric.

The model which achieved the lowest requirement of time was the one-stage object detector SSD, requiring a total of 03h47m. Even though it was the model with the lowest requirement for processing time it still provided better average precision when compared to another one-stage detector such as YOLOv3.

As expected, the model with the highest need of time for the dataset mounds-30 was a two-stage object detector, specifically the Faster R-CNN, with a total of 09h51m spent.

### 4.2.3 Dataset of mounds-15 with 1x learning rate schedule of 12 epochs

Starting with the analysis of the confusion matrixes generated from the tested DL models, specifically for the two-stage detectors, when executed on the dataset mounds-15 with a 1x learning rate schedule of 12 epochs, the result obtained was as can be observed in Figure 37.

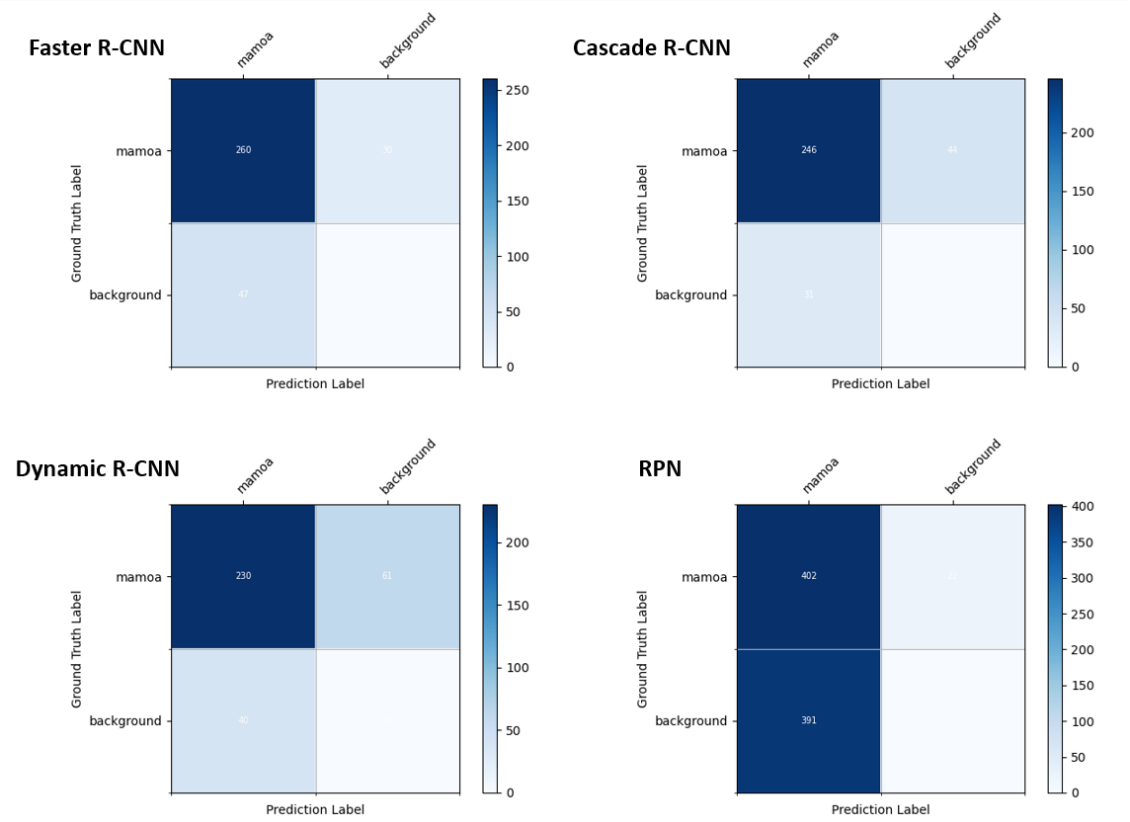


Figure 37 – Confusion matrix for the two-stage detectors, executed on dataset mounds-15, for 12 epochs

Starting with the model Faster R-CNN, it presents some balance on the incurrence of errors, not being too evident on a certain one. If incurs in 30 false negatives and 47 false positives. Despite these values, it accomplishes 260 true positive identifications of archaeological mounds. A good example of these successful identifications can be proven by observing Figure 38, as it shows a mound identified by the model that actually is present on the ground-of-truth side of the image.

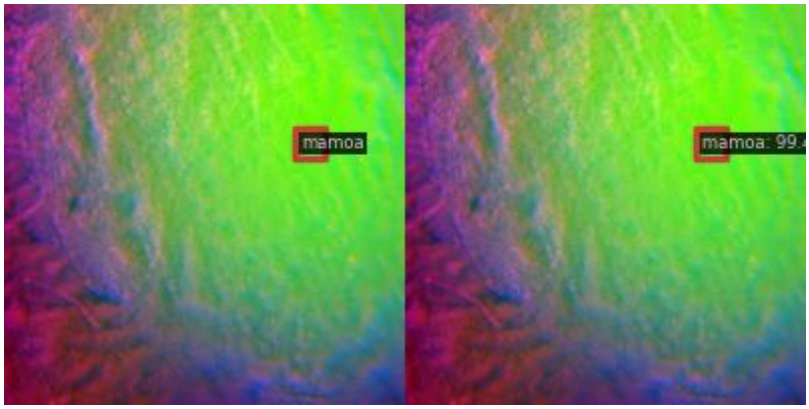


Figure 38 – Example of a good prediction with Faster R-CNN, on dataset mounds-15, for 12 epochs

The Cascade R-CNN model has a similar behaviour than the Faster R-CNN. It achieves a total of 246 true positive identifications, against 44 false negative and 31 false positives. Figure 39, presents an example of one of the good predictions that this model performed on its training.

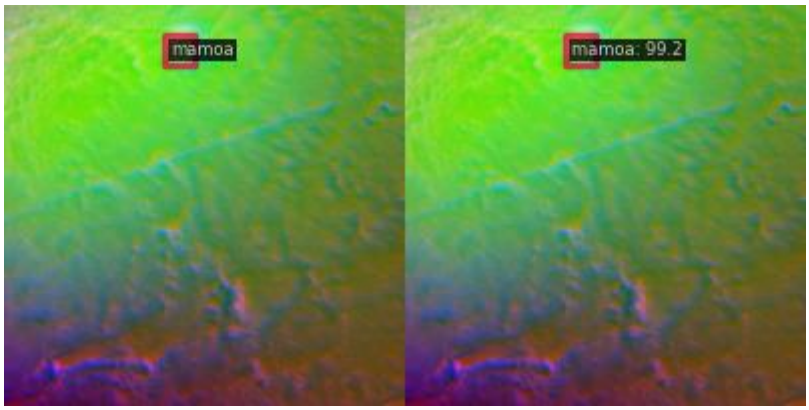


Figure 39 – Example of a good prediction with Cascade R-CNN, on dataset mounds-15, for 12 epochs

The Dynamic R-CNN model achieves a lower true positive identification of 230, while increasing the incurrence of false negatives to 61 and verifies 40 false positives. The example on Figure 40, demonstrates an example of an archaeological mound present in the ground-of-truth that failed to be identified by the model.

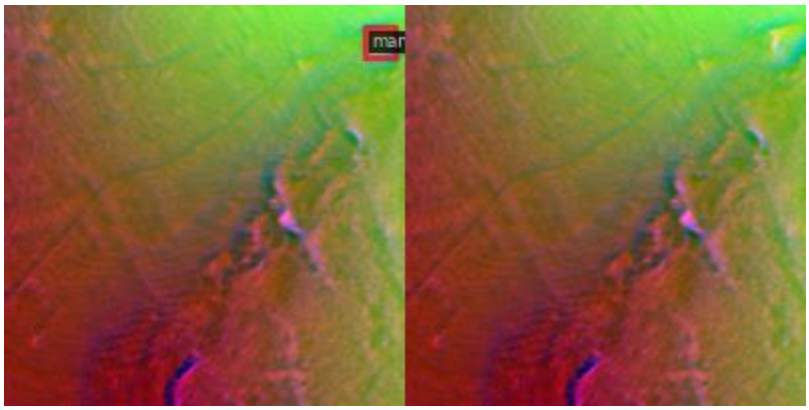

Figure 40 – Example of a bad prediction with Dynamic R-CNN, on dataset mounds-15, for 12 epochs

The last of the two-stage detectors, as confirmed before, maintains the problems of incurring in a high number of false positives (391) while having exaggerated identifications of true positives (402) due to overlapping of the bounding boxes. Besides this it presents 22 identifications of false negatives. The Figure 41, is a good example of both problems, since it identifies wrongly, an archaeological mound, on the right middle-top image where none exists on the ground-of-truth. Besides this, the actual mound that is correctly predicted by the model is presented with overlapping of bounding boxes.

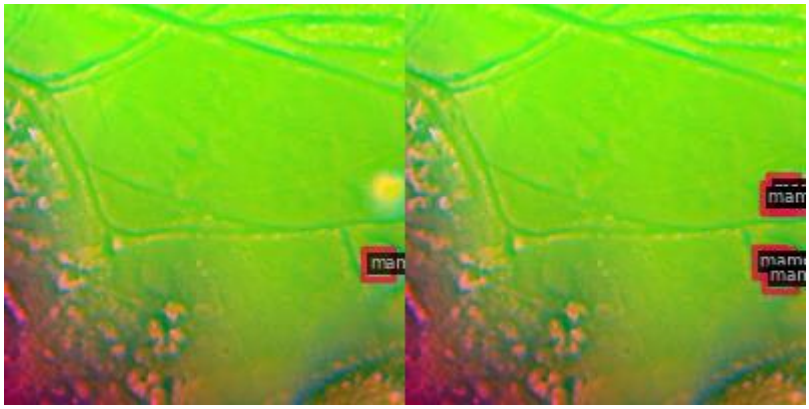

Figure 41 – Example of a bad prediction with RPN, on dataset mounds-15, for 12 epochs

Likewise, the execution of the one-stage detectors, on the dataset mounds-15, for 12 epochs were also performed which generated the following confusion matrixes, as seen in Figure 42.

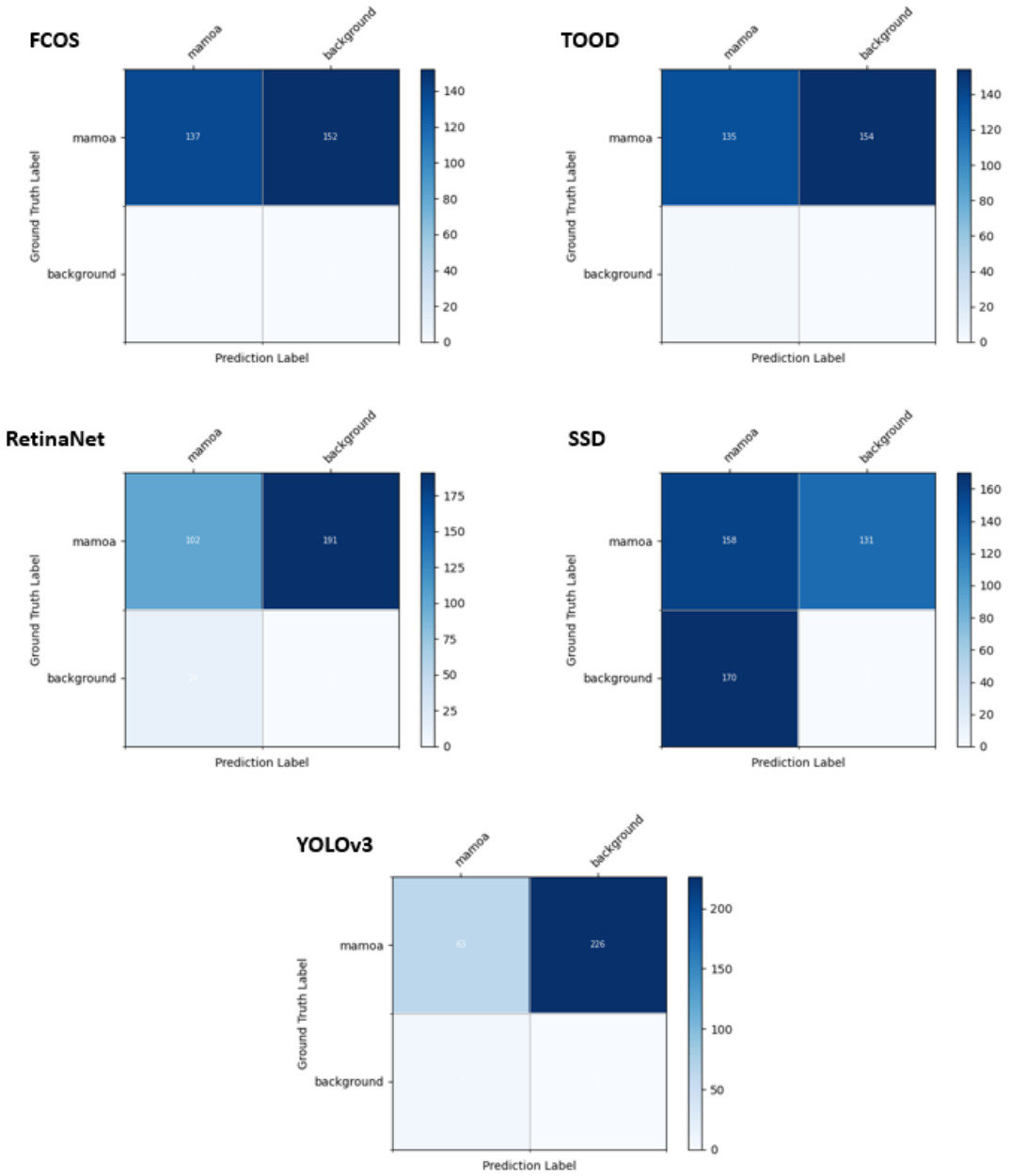


Figure 42 – Confusion matrix for the one-stage detectors executed on dataset mounds-15, for 12 epochs

It's evident that, comparatively to the two-stage detectors that have just been addressed, the one-stage models present a significant lower identification of true positives.

The model FCOS, although it has no incurrence of false positives, could only identify 137 true positive identifications of archaeological mounds. And it still incurred in a total of 152 false negative identifications. A good example of these false negatives can be verified in Figure 43, as the archaeological mound present on the ground-of-truth was not identified as such by the model.

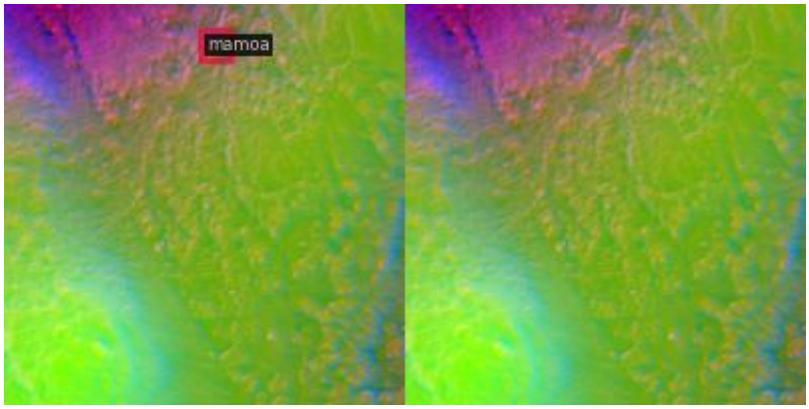


Figure 43 – Example of a bad prediction with FCOS, on dataset mounds-15, for 12 epochs

The architecture for the TOOD model has an almost equal behaviour than the one verified on the FCOS, since it has 135 true positives, just 3 false positives and a total of 154 false negatives. Figure 44, is a good example of one of these mistakes, since the model failed to identify the archaeological mound present on the ground-of-truth side of the image.

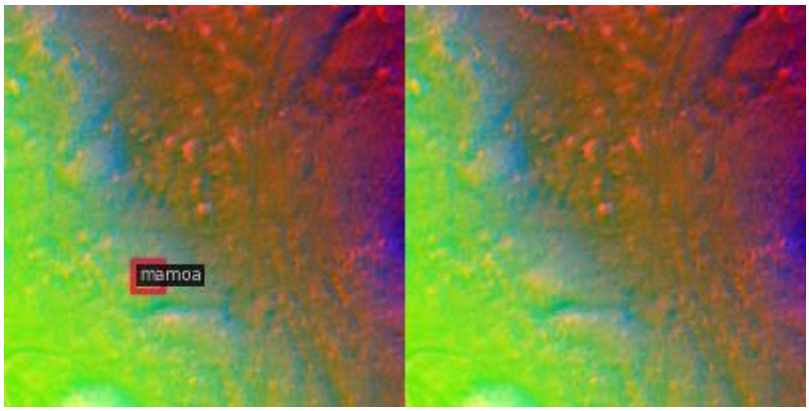


Figure 44 – Example of a bad prediction with TOOD, on dataset mounds-15, for 12 epochs

The RetinaNet model increases the verified problems identified on the two other one-stage models already mentioned, since it identifies 191 false negatives and 14 false positives. The detection of true positives also shows a negative evolution since it lowers to 102 identifications. The inability to some identification of archaeological mounds is also verified in Figure 45, as the ground-of-truth shows evidence of a mound that the model fails to identify.

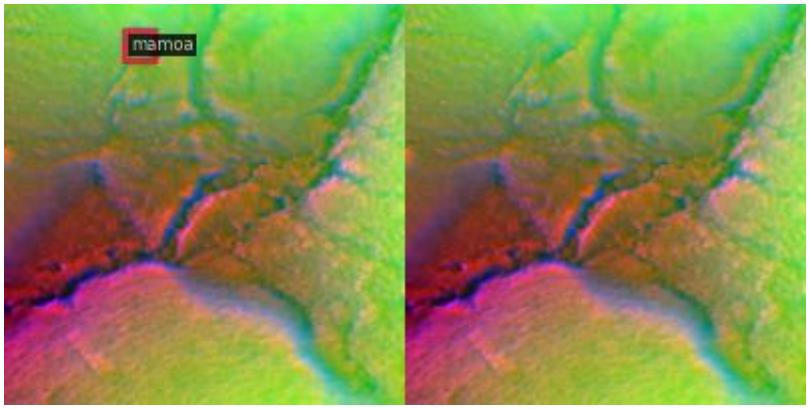


Figure 45 – Example of a bad prediction with RetinaNet, on dataset mounds-15, for 12 epochs

The SSD model shows a behaviour quite different from the ones observed thus far. It's able to achieve a higher number of true positives (158), although it has a significant increase in false positives (170). It still shows a significant number of 131 identifications of false negatives. The Figure 46 is a good example since it shows two prediction mistakes, since on top fails to identify an archaeological mound present on the ground-of truth. On the other hand, on bottom-right, identifies wrongly the presence of an archaeological mound that does not exist on the ground-of-truth side of the image.

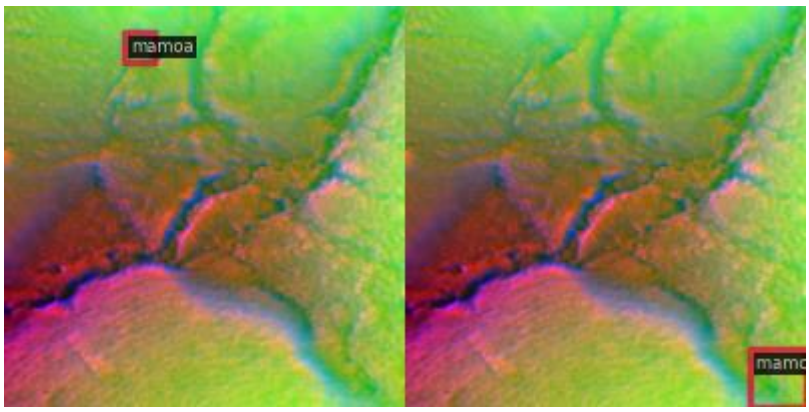


Figure 46 – Example of a bad prediction with SSD, on dataset mounds-15, for 12 epochs

The last of the one-stage models, YOLOv3, presents a significant problem as it incurs in a total of 226 false negative identifications. It has a low record of 7 false positives, but it also is just able to predict 63 true positives. Figure 47, serves as a good example of the inability to the model to successfully identify an archaeological mound present in the ground-of-truth side of the image.

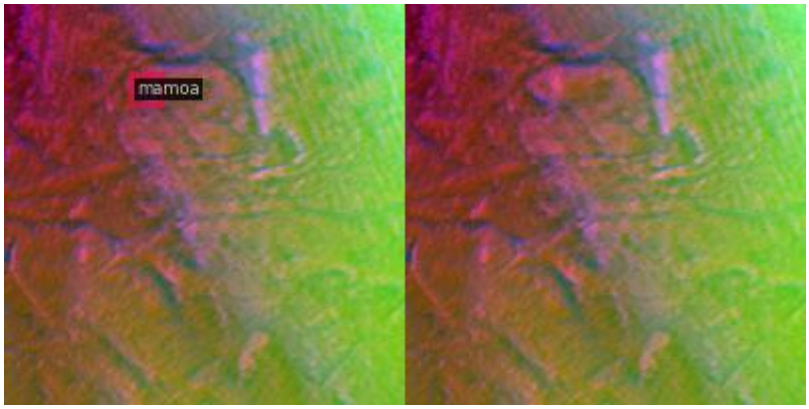


Figure 47 - Example of a bad prediction with YOLOv3, on dataset mounds-15, for 12 epochs

The observation of the metrics of mAP, processing time to run models or recall, also allows to provide important information when comparing the deep learning models. To have an easier reading of these values, the Table 9, allows for a better analysis of them.

Table 9 – Results obtained of all models with dataset mounds-15 and trained for 12 epochs (best results in bold)

| **mounds-15 (12 epochs)** | | | | | | |
|---|---|---|---|---|---|---|
| **Architecture** | **Category** | **Full processing time (hh:mm)** | **mAP** | **$mAP_{0.5}$** | **$mAP_{0.75}$** | **Recall** |
| Faster R-CNN | Two-stage detector | 04:15 | 53.8 | 88.4 | 57.1 | 89.7 |
| Cascade R-CNN | Two-stage detector | 06:20 | 52.0 | 84.3 | 55.5 | 84.8 |
| Dynamic R-CNN | Two-stage detector | 04:20 | 46.4 | 79.3 | 50.2 | 79.0 |
| RPN | Two-stage detector | 03:18 | 36.3 | 78.5 | 27.9 | **94.8** |
| FCOS | One-stage detector | 04:37 | **58.1** | **91.6** | **66.6** | 47.4 |
| TOOD | One-stage detector | 06:44 | 35.8 | 74.4 | 30.6 | 46.7 |
| RetinaNet | One-stage detector | 04:21 | 23.5 | 59.0 | 10.4 | 34.8 |
| SSD | One-stage detector | **03:06** | 14.4 | 37.1 | 7.4 | 54.7 |
| YOLOv3 | One-stage detector | 04:29 | 13.4 | 48.7 | 1.6 | 21.8 |

The model FCOS, although being a one-stage object detector which, in theory, should not outperform some of the two-stage object detector, still proves to be the model with the highest mAP regardless of the IoU threshold that is observed.

For this dataset, even though the two-stage detectors continue to outperform the one-stage detectors regarding the Recall metric, it even widens the difference between them. Just like for the dataset mounds-30 when trained for 12 epochs, the same architecture of RPN proves to be the one which can successfully identify most of the existing archaeologic mounds.

In terms of processing time, with no surprise, the one-stage detector SSD, continues to be the model with the lower time required (03h06m). Interestingly, the model with the higher requirement for processing time isn't one of the two-stage detectors, but it's the architecture TOOD, spending a total of 06h44m.

### 4.2.4 Dataset of mounds-15 with 2x learning rate schedule of 24 epochs

To perform an analysis on the performance of the DL models that were tested, the confusion matrix generated for each of the models was a good starting point. The two-stage object detectors trained on the dataset mounds-15, for 24 epochs, generated their respective confusion matrixes, as can be viewed in Figure 48.

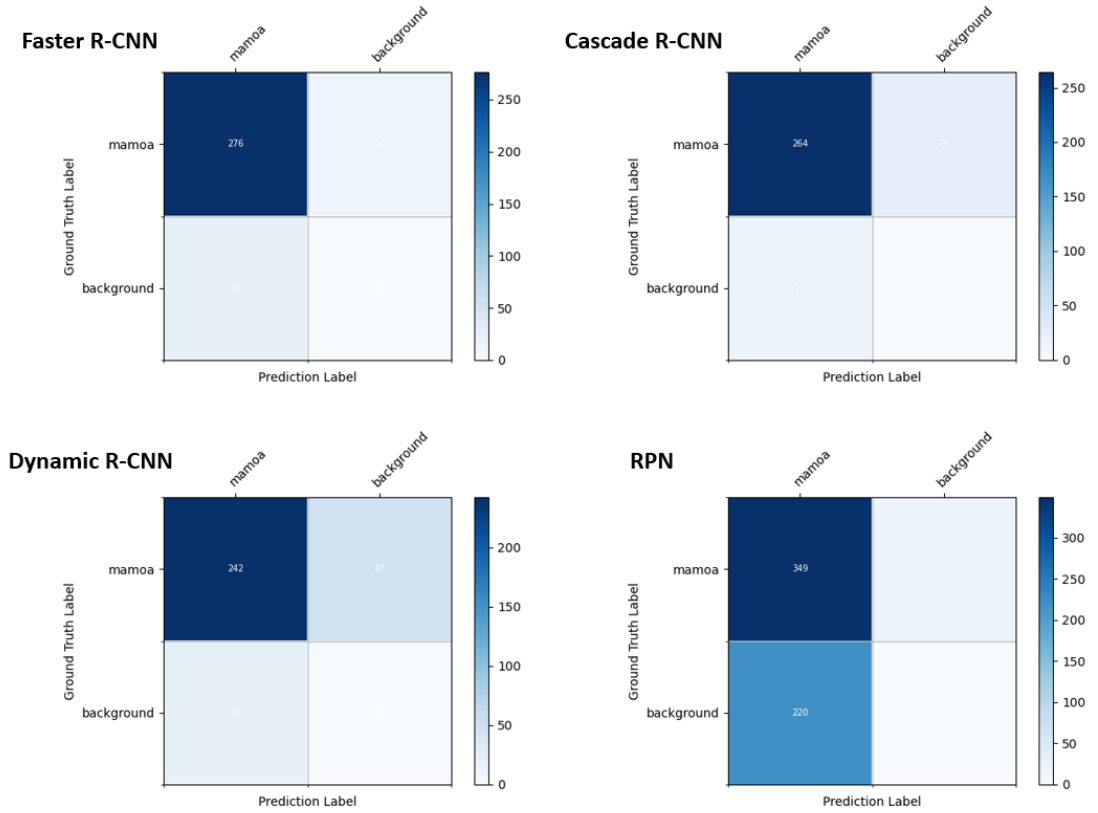


Figure 48 - Confusion matrix for the two-stage detectors, on dataset mounds-15, for 24 epochs

For this dataset, the model Faster R-CNN shows a slight tendency for a more significant error occurrence of type false positive, since it occurred 23 times, when comparing to 14 times of false negatives. Despite these errors still shows a significant number of 276 predictions with a true positive evaluation. The example in Figure 49, represents a good example of a wrong evaluation by the model predicting a mound (right side of figure) where none existing in the ground of truth (left side of figure) incurring in a false positive while also incurring in a false negative since the ground of truth shows the presence of a mound that was not identified by the model.

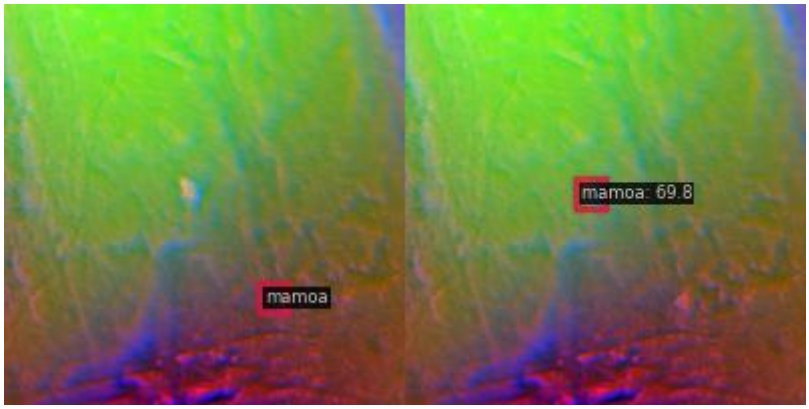


Figure 49 - Example of a bad prediction with Faster R-CNN, on dataset mounds-15, for 24 epochs

Inversely, the Cascade R-CNN presents the inverse tendency for a slightly higher occurrence of false negatives (26 occurrences) when compared to the false positives (16 occurrences). It still achieves a total of 264 true positive predictions of archaeological mounds. An example of these false negative prediction can be viewed on Figure 50.

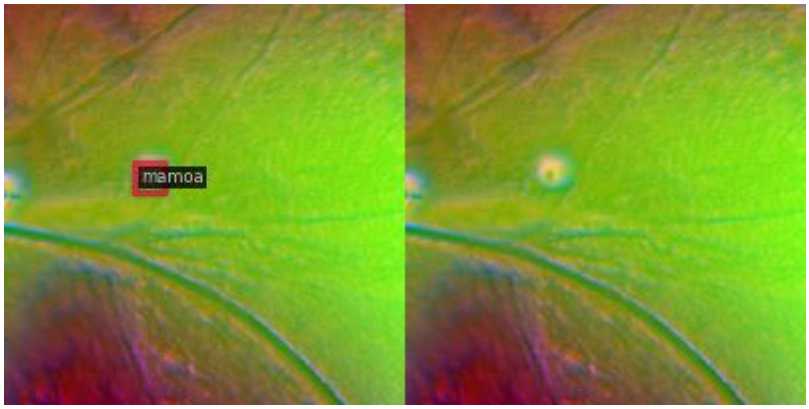


Figure 50 - Example of a bad prediction with Cascade R-CNN, on dataset mounds-15, for 24 epochs

The Dynamic R-CNN model shows the same tendency as the Cascade R-CNN, however with a higher weight of prediction error of false negative (47 occurrences) when compared with false positive (20 occurrences). Despite this increase, it's still able to achieve 242 true positive

predictions. This is proven as shown on the Figure 51, as the ground-of-truth shows an image identifying an archaeological mound that was not predicted by the model.

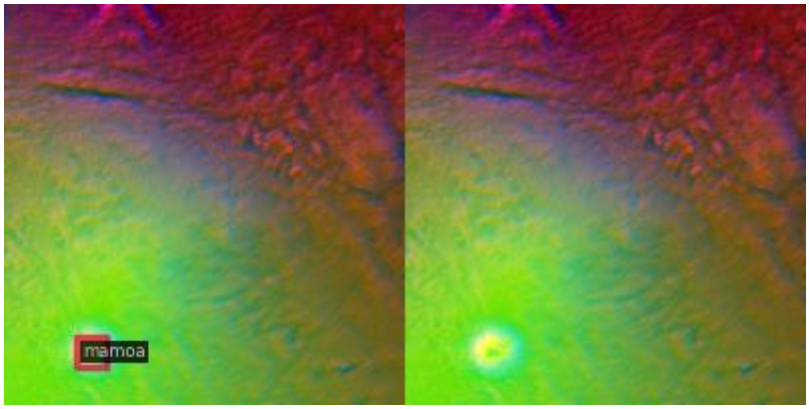


Figure 51 - Example of a bad prediction with Dynamic R-CNN, on dataset mounds-15, for 24 epochs

Lastly the RPN model, on this dataset size has a problem of some overlapping bounding boxes which generates a higher number of values when comparing to other models. This still happens even with the application of an NMS to attempt to avoid or minimize this. This can be seen as the example present in Figure 52.

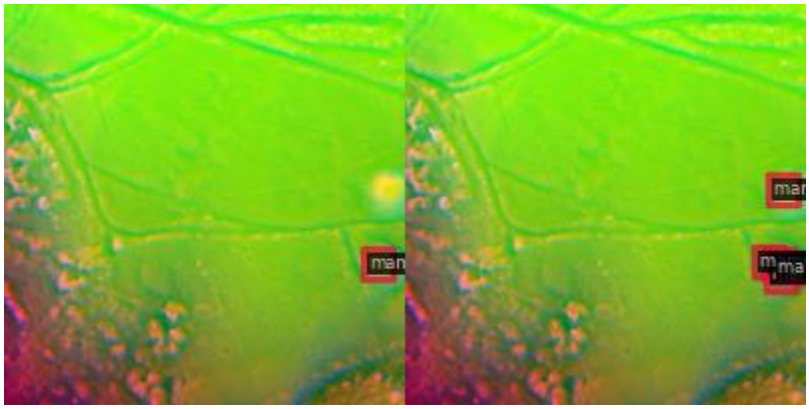

Figure 52 – Example of a bad prediction, with overlapping bounding box with RPN, on dataset mounds-15, for 24 epochs

As was performed for the two-stage object detectors, the same testing was performed in this thesis to the one-stage object detectors, with the dataset of mounds-15, which generated the respective models confusion matrix as observed in Figure 53.

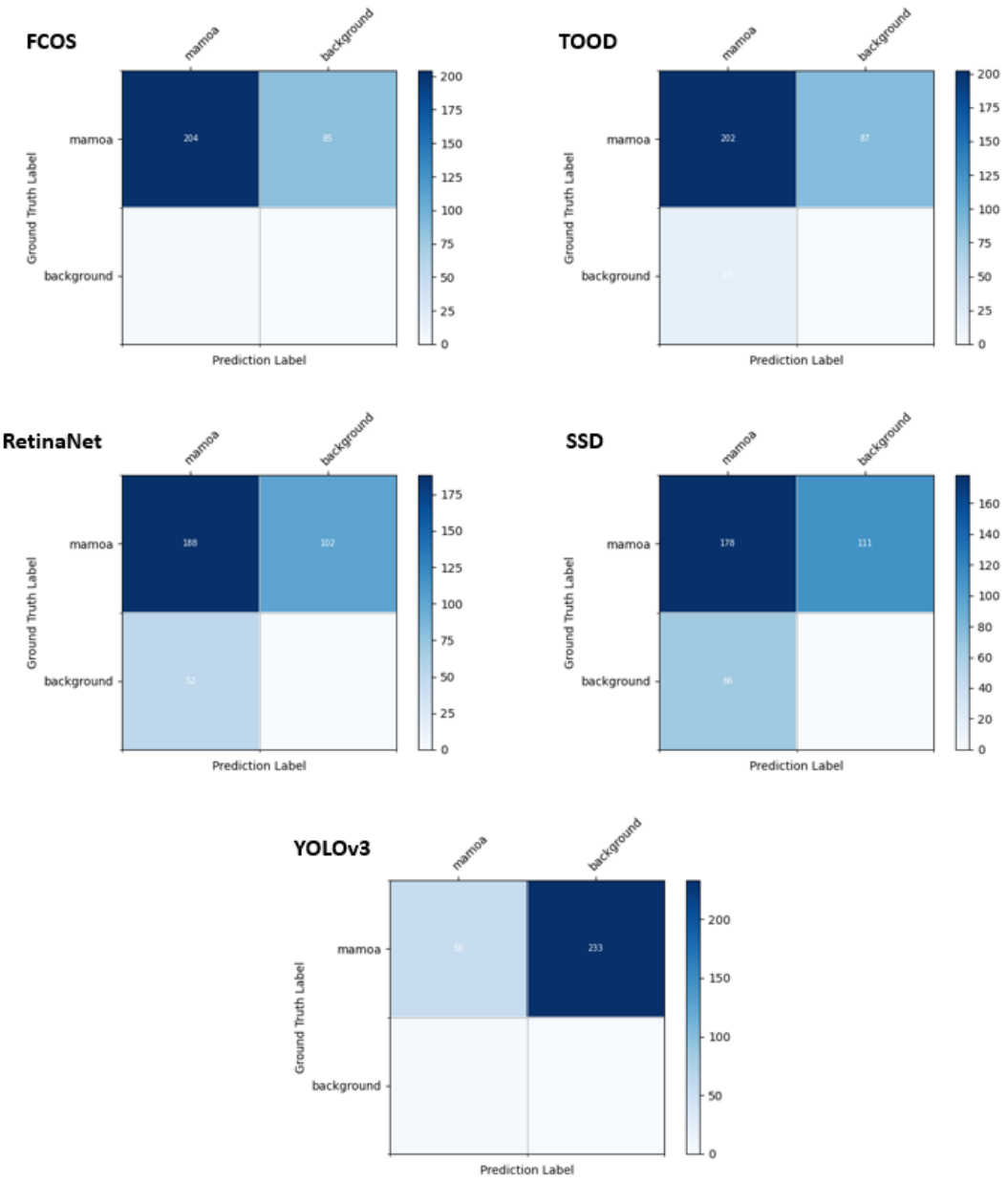


Figure 53 - Confusion matrix for the one-stage detectors, executed on dataset mounds-15, for 24 epochs

The one-stage detectors for this dataset of mounds-15 all present the same tendency of error, specifically for incurring in false negative predictions.

Looking at the architecture for the FCOS, despite accurately predicting 178 true positives (shown in Figure 54), it incurs in 111 false negatives. It's a huge contrast when realizing that it has only a single false positive prediction.

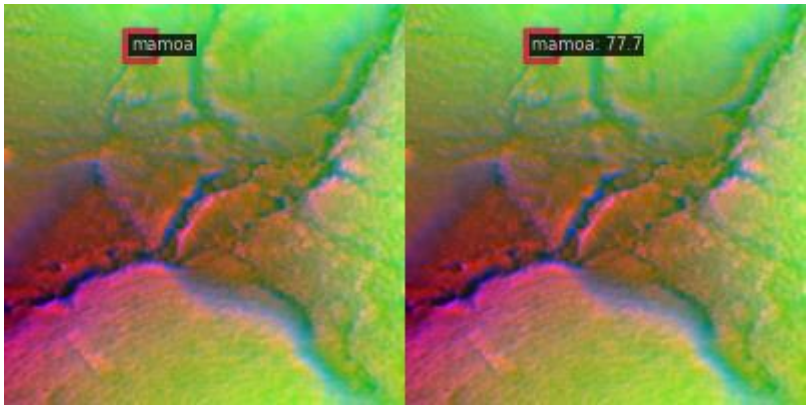


Figure 54 - Example of a good prediction with FCOS, on dataset mounds-15, for 24 epochs

The TOOD model presents results very similar to FCOS, since it has the same 178 true positive as well as the same 111 false negative (example shown in Figure 55). The main difference, although slight is that it incurs in 7 false positives.

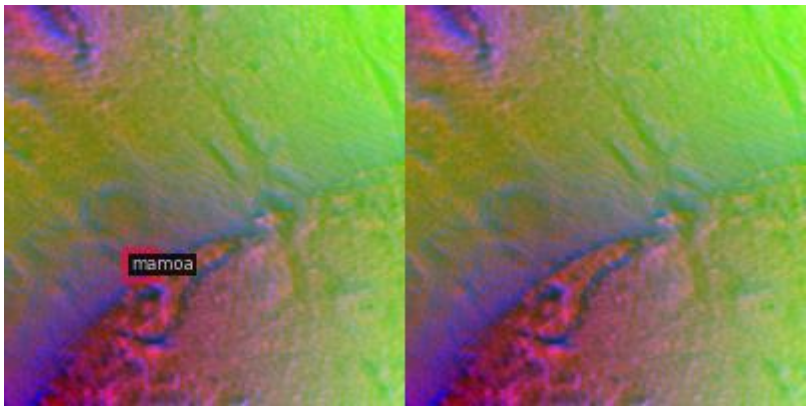


Figure 55 - Example of a bad prediction with TOOD, on dataset mounds-15, for 24 epochs

The RetinaNet model maintains the tendency of a higher false negative (as exemplified in Figure 56) with a total of 102 while showing a significant increase in the number of false positives to a total of 52. It still achieves 188 true positive occurrences.

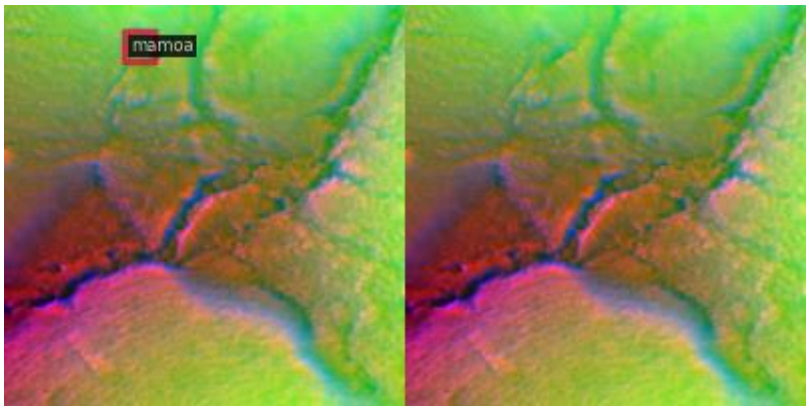


Figure 56 - Example of a bad prediction with RetinaNet, on dataset mounds-15, for 24 epochs

Similar behaviour is also present on the model SSD. Reducing the true positives to a value of 178, while increasing the false negatives to 111 and the false positives to 66. An example of a false negative, as a representation of the biggest problem of the model, is shown in Figure 57.

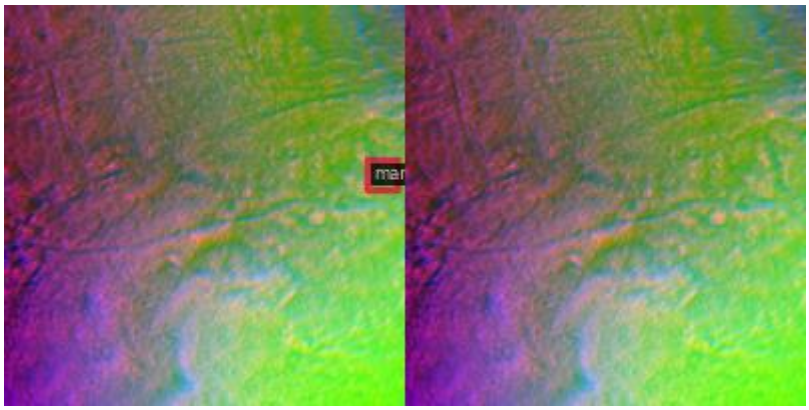

Figure 57 - Example of a bad prediction with SSD, on dataset mounds-15, for 24 epochs

Finally, and the one-stage detector with the worst performance, the YOLOv3 model can only accurately predict 56 true positives, incurring in a massive mistake for a total of 233 false negatives and 3 false positives. In Figure 58 can be seen an example of a false negative since the model failed to identify the mound present in the ground of truth.

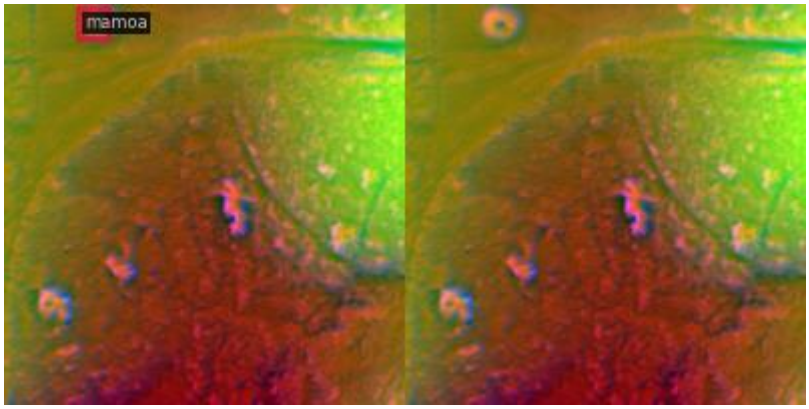


Figure 58 - Example of a bad prediction with YOLOv3, on dataset mounds-15, for 24 epochs

Another metric that can be used for comparison and analysis consist of the interpretation of mAP with different threshold of IoU or even the processing time to run the models. This comparison information can be viewed in Table 10.

Table 10 – Results obtained of all models with dataset mounds-15 and trained for 24 epochs (best results in bold)

| **mounds-15 (24 epochs)** |
|---|

| Architecture | Category | Full processing time (hh:mm) | mAP | $mAP_{0.5}$ | $mAP_{0.75}$ | Recall |
|---|---|---|---|---|---|---|
| Faster R-CNN | Two-stage detector | 12:30 | 66.8 | 93.2 | 74.2 | **95.2%** |
| Cascade R-CNN | Two-stage detector | 12:00 | 64.4 | 89.5 | 69.7 | 91.0% |
| Dynamic R-CNN | Two-stage detector | 08:13 | 56.9 | 84.3 | 63.7 | 83.7% |
| RPN | Two-stage detector | 06:07 | 42.9 | 84.3 | 39.0 | 94.6% |
| FCOS | One-stage detector | 10:19 | **68.1** | **94.0** | **75.8** | 70.6% |
| TOOD | One-stage detector | 12:58 | 41.8 | 78.4 | 38.2 | 69.9% |
| RetinaNet | One-stage detector | 10:37 | 29.8 | 63.4 | 21.6 | 64.8% |
| SSD | One-stage detector | **05:30** | 26.2 | 63.5 | 14.2 | 61.6% |
| YOLOv3 | One-stage detector | 11:29 | 16.1 | 50.7 | 2.6 | 19.4% |

It's also clear that the smaller the size of the dataset tiles, the more negatively it affects the performance of most models. Only the case of model FCOS, when evaluating the $mAP_{0.5}$, achieves an improvement, although an almost insignificant of 0.1 percentual points. Another note of exception for the model Faster R-CNN which, again on the evaluation of the $mAP_{0.5}$, shows a close evaluation between datasets since the mounds-15 is only 0.3 percentual points inferior when comparing to the dataset mounds-30.

For the remaining models, the difference obtained in mAP calculated for the different dataset sizes, provided a tendency of the mounds-15 showing an inferior predicting accuracy as its values where always lower than the ones obtained for mounds-30. The minimum variance verified was of 1.1 percentual points, but its maximum variance was of 44.1 percentual points.

On an average analysis, the predicting capacity of object detection models showed that when attempting to detect archaeological mounds on the dataset mounds-30 achieved an average precision superior by 16.5 percentual points when compared to the dataset mounds-15.

The increase for the 24 epochs does allow for a generalized increase in the recall metric on all the models, except for the RPN, which lowered from 94.8 to 94.6, although it's a minor difference. It's interesting to realize that the Recall metric performs better when trained on the dataset mounds-30, since the difference of recall between training epochs is lower.

Training on 12 epochs widens the difference and generates a difference in recall metric considerably higher between the two datasets.

In terms of processing time, it's easily understandable that when dividing the same image in smaller tiles to form the dataset mounds-15, would generate a higher number of tiles to analyse and as such originate a higher need of processing time.

The model which achieved the lowest requirement of time was the one-stage object detector SSD, requiring a total of 05h30m. Even though it was the model with the lowest requirement for processing time it still provided better average precision when compared to another one-stage detector such as YOLOv3.

A bit surprisingly, it was a one-stage object detector to require the highest processing time, namely the TOOD, having spent a total of 12h58m.

# 5 Conclusions

This chapter will be focusing on the main conclusions obtained through the course of this thesis as well as underline what next steps can follow the current work to complement or further this benchmark analysis.

## 5.1 Main Conclusions

In Archaeology, the limited number of professional experts of the discipline, the physical distance that they might be from archaeological findings or even the difficulty of accessibility to archaeological sites are just a few mentionable facts that hinder traditional method to evaluate and assess new archaeology findings.

Through this thesis, it was shown that, recent advances in technology for remote sensing, such as LiDAR, can provide invaluable assistance to remotely perform a mapping of the new archaeological findings and allow a first analysis without having to be present in the actual site.

There exists a variety of VT's that can process the LiDAR data to improve and enhance the image as was experienced with DTM, LRM and Slope. The combination of these three VT's allowed to further enhance the image to became easier to be applied later to the DL methods. The datasets were then generated on a COCO format, on two different sizes (15x15 meters and 30x30 meters), and to ensure compatibility for an application of a LOOCV strategy.

Experiments were conducted on both types of object detection categories. For the two-stage object detectors, was chosen to test the Faster R-CNN, Cascade R-CNN, Dynamic R-CNN and RPN. As for the one-stage object detectors, the chosen were YOLOv3, RetinaNet, SSD, FCOS and TOOD.

A detailed comparison was then possible to establish between all these models on the two generated datasets, where it became clear that the precision on the smaller dataset mounds-15 where generally worse than the precision obtained on the dataset mounds-30.

Overall, and as expected, the two-stage detectors performed better in terms of mean average precision while the one-stage detectors performed faster with some exceptions on both.

Note however to the fact that the model with the highest mean average precision was the one-stage object detector FCOS, but very close to the values of two-stage detector Faster R-CNN.

As expected, the faster model to be executed was a one-stage detector, namely SSD.

The work done on this thesis allowed to perform a diverse analysis of some of the most popular DL methods developed for object detection, to have a benchmark analysis to determine some evaluation points to assess the better models, whether to focus simply on maximizing the identification precision or to be able to perform a fast classification, or even a balance between the two possible requirements.

However, it should be mentioned that the choice of the best model cannot be bluntly stated since it would depend on the use case to be applied. If time limitations aren't a factor, in general terms, the two-stage object detectors outperform the one-stage object detectors. On the other hand, if there is a time limitation, one-stage object detector can be a better choice since, in general terms, they prove themselves to be faster, however, globally performing a worse average precision to accurately predict the presence of archaeological mounds.

So, the combination and this trade-off between predictive precision and time limitation needs to be well though and considered when attempting to make a choice on the application of a certain architecture.

## 5.2 Future Work

Despite having achieved a good comparison base for object detection in the search of archaeological mounds with LiDAR data, having tested most of the architectures with the highest popularity and precision records, there is still a wide variety of models that could have been experimented if no time limitation were at stake.

Unfortunately, the time limitation, combined with the high demand of computational processing time to run the models, hindered further exploration of the remaining models that the MMDetection repository had to offer.

Special note to the fact that the model YOLO, has more recent versions, which promises to be even more accurate and faster to make their predictions, although, as far as it could be ascertained, no application of the most recent version, namely YOLOv8, has yet to be done on the field of object detection for archaeological purposes. So, testing this model would be a priority and a personal wish to further enrich this benchmark analysis.

It would also be interesting to deepen the benchmark analysis to consider, for each model, the experimentation of different feature extractors, similarly to what was explored in (Carranza-García et al., 2020). For example, the model Faster R-CNN could have been combined with a variety of ResNet. For this thesis, the feature extractor chosen for most of the models was the ResNet-50, but the Faster R-CNN could have also been tried with the ResNet-101 or the Res2Net-101, just to mention two different options.

# References


Alawida, M., Omolara, A. E., Abiodun, O. I., & Al-Rajab, M. (2022). A deeper look into cybersecurity issues in the wake of Covid-19: A survey. *Journal of King Saud University - Computer and Information Sciences*, *34*(10, Part A), 8176–8206. https://doi.org/https://doi.org/10.1016/j.jksuci.2022.08.003

Argyrou, A., & Agapiou, A. (2022). A Review of Artificial Intelligence and Remote Sensing for Archaeological Research. *Remote Sensing*, *14*(23). https://doi.org/10.3390/rs14236000

Arnoldussen, S., der Vaart, W. B., Kaptijn, E., & Bourgeois, Q. P. J. (2023). Field systems and later prehistoric land use: New insights into land use detectability and palaeodemography in the Netherlands through LiDAR, automatic detection and traditional field data. *Archaeological Prospection*, *30*(3), 283–300. https://doi.org/https://doi.org/10.1002/arp.1891

Berganzo-Besga, I., Orengo, H. A., Lumbreras, F., Carrero-Pazos, M., Fonte, J., & Vilas-Estévez, B. (2021). Hybrid MSRM-Based Deep Learning and Multitemporal Sentinel 2-Based Machine Learning Algorithm Detects Near 10k Archaeological Tumuli in North-Western Iberia. *Remote Sensing*, *13*(20). https://doi.org/10.3390/rs13204181

Cai, Z., & Vasconcelos, N. (2019). *Cascade R-CNN: High Quality Object Detection and Instance Segmentation*.

Canedo, D., Fonte, J., Seco, L. G., Vázquez, M., Dias, R., Pereiro, T. Do, Hipólito, J., Menéndez-Marsh, F., Georgieva, P., & Neves, A. J. R. (2023). Uncovering Archaeological Sites in Airborne LiDAR Data With Data-Centric Artificial Intelligence. *IEEE Access*, *11*, 65608–65619. https://doi.org/10.1109/ACCESS.2023.3290305

Carranza-García, M., Torres-Mateo, J., Lara-Benítez, P., & García-Gutiérrez, J. (2020). On the Performance of One-Stage and Two-Stage Object Detectors in Autonomous Vehicles Using Camera Data. *Remote Sensing*, *13*(1), 89. https://doi.org/10.3390/rs13010089

Character, L., Ortiz JR, A., Beach, T., & Luzzadder-Beach, S. (2021). Archaeologic Machine Learning for Shipwreck Detection Using Lidar and Sonar. *Remote Sensing*, *13*(9). https://doi.org/10.3390/rs13091759

Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., Zhang, Z., Cheng, D., Zhu, C., Cheng, T., Zhao, Q., Li, B., Lu, X., Zhu, R., Wu, Y., … Lin, D. (2019). MMDetection: Open MMLab Detection Toolbox and Benchmark. *ArXiv Preprint ArXiv:1906.07155*.

Davis, D. S., & Lundin, J. (2021). Locating Charcoal Production Sites in Sweden Using LiDAR, Hydrological Algorithms, and Deep Learning. *Remote Sensing*, *13*(18). https://doi.org/10.3390/rs13183680

Davis, D., Sanger, M., & Lipo, C. (2018). Automated mound detection using lidar and object-based image analysis in Beaufort County, South Carolina. *Southeastern Archaeology*, *38*, 23–37. https://doi.org/10.1080/0734578X.2018.1482186

Feng, C., Zhong, Y., Gao, Y., Scott, M. R., & Huang, W. (2021). *TOOD: Task-aligned One-stage Object Detection*.

Fiorucci, M., der Vaart, W. B., Soleni, P., Le Saux, B., & Traviglia, A. (2022). Deep Learning for Archaeological Object Detection on LiDAR: New Evaluation Measures and Insights. *Remote Sensing*, *14*(7). https://doi.org/10.3390/rs14071694

Fonte, J., Meunier, E., Gonçalves, J. A., Dias, F., Lima, A., Gonçalves-Seco, L., & Figueiredo, E. (2021). An Integrated Remote-Sensing and GIS Approach for Mapping Past Tin Mining Landscapes in Northwest Iberia. *Remote Sensing*, *13*(17). https://doi.org/10.3390/rs13173434

Gallwey, J., Eyre, M., Tonkins, M., & Coggan, J. (2019). Bringing Lunar LiDAR Back Down to Earth: Mapping Our Industrial Heritage through Deep Transfer Learning. *Remote Sensing*, *11*(17). https://doi.org/10.3390/rs11171994

Girshick, R. (2015). *Fast R-CNN*.

Guyot, A., Lennon, M., & Hubert-Moy, L. (2021). Objective comparison of relief visualization techniques with deep CNN for archaeology. *Journal of Archaeological Science: Reports*, *38*, 103027. https://doi.org/10.1016/J.JASREP.2021.103027

Guyot, A., Lennon, M., Lorho, T., & Hubert-Moy, L. (2021). Combined Detection and Segmentation of Archeological Structures from LiDAR Data Using a Deep Learning Approach. *Journal of Computer Applications in Archaeology*, *4*, 1. https://doi.org/10.5334/jcaa.64

Hu, Y., Kuang, W., Qin, Z., Li, K., Zhang, J., Gao, Y., Li, W., & Li, K. (2021). Artificial Intelligence Security: Threats and Countermeasures. *ACM Comput. Surv.*, *55*(1). https://doi.org/10.1145/3487890

Hunton, P. (2009). The growing phenomenon of crime and the internet: A cybercrime execution and analysis model. *Computer Law & Security Review*, *25*(6), 528–535. https://doi.org/https://doi.org/10.1016/j.clsr.2009.09.005

Lambers, K., der Vaart, W. B., & Bourgeois, Q. P. J. (2019). Integrating Remote Sensing, Machine Learning, and Citizen Science in Dutch Archaeological Prospection. *Remote Sensing*, *11*(7). https://doi.org/10.3390/rs11070794

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2018). *Focal Loss for Dense Object Detection*.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., & Dollár, P. (2015). *Microsoft COCO: Common Objects in Context*.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., & Berg, A. C. (2016). SSD: Single Shot MultiBox Detector. In *Computer Vision ECCV 2016* (pp. 21–37). Springer International Publishing. https://doi.org/10.1007/978-3-319-46448-0_2

Marcos, G., Henrique Soares, K., do Amaral, M., Flavio Felipe, K., & Marco, M. (2010). LiDAR: princípios e aplicações florestais. *Pesquisa Florestal Brasileira*, *30*(63). https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=edsdoj&AN=edsdoj.6af51612162446497d918e73921a331&lang=pt-pt&site=eds-live&scope=site

Oliveira, C., Aravecchia, S., Pradalier, C., Robin, V., & Devin, S. (2021). The use of remote sensing tools for accurate charcoal kilns' inventory and distribution analysis: Comparative assessment and prospective. *International Journal of Applied Earth Observation and Geoinformation*, *105*, 102641. https://doi.org/https://doi.org/10.1016/j.jag.2021.102641

Redmon, J., & Farhadi, A. (2018). *YOLOv3: An Incremental Improvement*.

Ren, S., He, K., Girshick, R., & Sun, J. (2016). *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks*.

Sakai, M., Lai, Y., Olano Canales, J., Hayashi, M., & Nomura, K. (2023). Accelerating the discovery of new Nasca geoglyphs using deep learning. *Journal of Archaeological Science*, *155*, 105777. https://doi.org/https://doi.org/10.1016/j.jas.2023.105777

Tian, Z., Shen, C., Chen, H., & He, T. (2019). *FCOS: Fully Convolutional One-Stage Object Detection*.

Trier, Ø., Cowley, D., & U. Waldeland, A. (2018). Using deep neural networks on airborne laser scanning data: Results from a case study of semi-automatic mapping of archaeological topography on Arran, Scotland. *Archaeological Prospection*, *26*. https://doi.org/10.1002/arp.1731

Trier, Ø. D., Reksten, J. H., & Løseth, K. (2021). Automated mapping of cultural heritage in Norway from airborne lidar data using faster R-CNN. *International Journal of Applied Earth Observation and Geoinformation*, *95*, 102241. https://doi.org/https://doi.org/10.1016/j.jag.2020.102241

Verschoof-van der Vaart, W. B., & Lambers, K. (2019). Learning to Look at LiDAR: The Use of R-CNN in the Automated Detection of Archaeological Objects in LiDAR Data from the Netherlands. *Journal of Computer Applications in Archaeology*. https://doi.org/10.5334/jcaa.32

Zhang, H., Chang, H., Ma, B., Wang, N., & Chen, X. (2020). *Dynamic R-CNN: Towards High Quality Object Detection via Dynamic Training*.

Zou, Z., Chen, K., Shi, Z., Guo, Y., & Ye, J. (2023). *Object Detection in 20 Years: A Survey*.