

Final Report

|  |
| --- |
| School of Computing  Faculty of Engineering AND PHYSICAL SCIENCES |

<Object Detection Website Based on Deep Learning>

<Yunjia Feng>

Submitted in accordance with the requirements for the degree of  
<BSc Computer Science>

**<2021/22>**

The candidate confirms that the following have been submitted*:*

*<As an example>*

|  |  |  |
| --- | --- | --- |
| **Items** | **Format** | **Recipient(s) and Date** |
| *Final Report* | *Hard copy and PDF file* | *Hard copy handed to SSO (DD/MM/YY); PDF uploaded to Minerva (DD/MM/YY)* |
| *Participant consent forms* | *PDF file / file archive* | *Uploaded to Minerva (DD/MM/YY)* |
| *Link to online code repository* | *URL* | *Sent to supervisor and assessor (DD/MM/YY)* |
| *User manuals* | *PDF and/or hard copy* | *Sent to client and supervisor (DD/MM/YY)* |

The candidate confirms that the work submitted is their own and the appropriate credit has been given where reference has been made to the work of others.

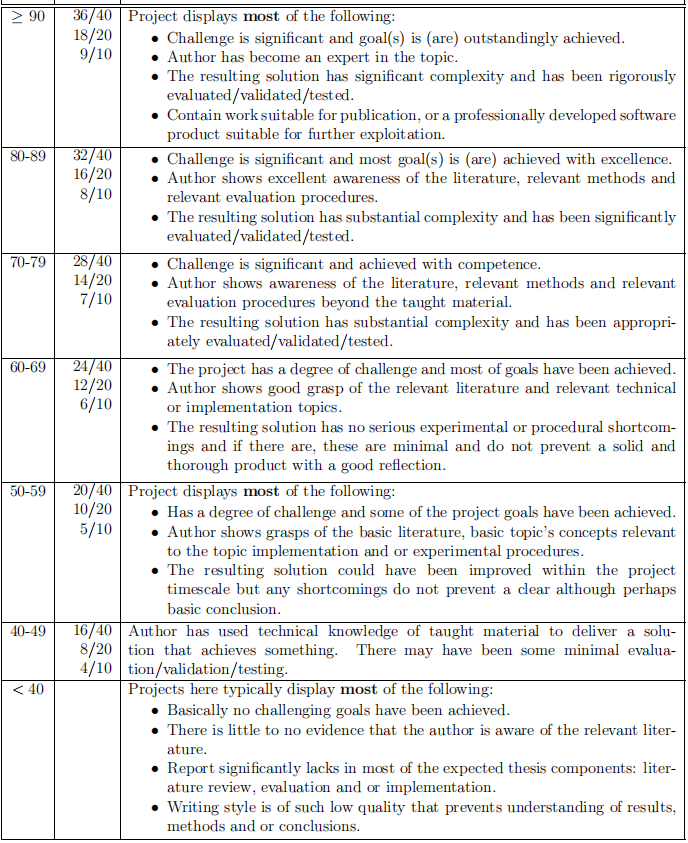
I understand that failure to attribute material which is obtained from another source may be considered as plagiarism.

(Signature of student) Yunjia Feng

© <2022> The University of Leeds and <Yunjia Feng>

# Summary

*<Concise statement of the problem you intended to solve and main achievements (no more than one A4 page)>*



# Acknowledgements

*<This page should contain any acknowledgements to those who have assisted with your work. Where you have worked as part of a team, you should, where appropriate, reference to any contribution made by others to the project.>*

*Note that it is not acceptable to solicit assistance on ‘proof reading’ which is defined as “the systematic checking and identification of errors in spelling, punctuation, grammar and sentence construction, formatting and layout in the text”; see*

https:://www.leeds.ac.uk/secretariat/documents/proof\_reading\_policy.pdf

# Table of Contents (example of how to format)

Summary iii

Acknowledgements iv

Table of Contents v

Chapter 1 Level 1 Heading with ‘heading 1’ Style 1

1.1 Level 2 Heading with ‘heading 2’ Style 1

1.1.1 Level 3 Heading with ‘heading 3’ Style 1

1.1.1.1 Level 4 Heading with ‘heading 4’ Style 1

Chapter 2 Tables and Figures 2

2.1 Tables using the ‘table caption’ and ‘table description’ Styles 2

2.2 Figures using the ‘figure caption’ and ‘figure description’ Styles 2

List of References 3

Appendix A Self-appraisal 4

A.1 Level 2 Heading with ‘heading 2’ Style 4

A.1.1 Level 3 Heading with ‘heading 3’ Style 4

A.1.1.1 Level 4 Heading with ‘heading 4’ Style 4

Appendix B External materials 4

B.1 Level 2 Heading with ‘heading 2’ Style 4

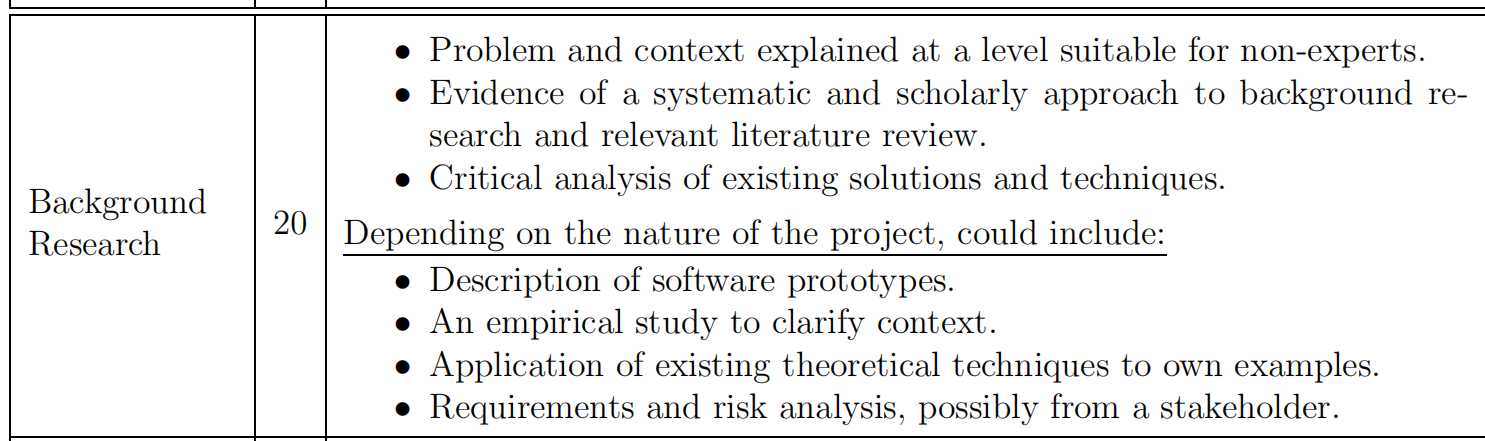
# Chapter 1 Introduction and Background Research

<Recommend using ‘Heading 1’ for chapter titles, ‘Heading 2’ style for section headings, ‘Heading 3’ for subsection headings, and ‘Heading 4’ for sub-subsection headings, but whatever you choose you must be consistent. Don’t forget that text (other than headings) should be 11 point.>

Problem and Context

Analysis of existing solutions and techiniques

Description of software prototypes,



## 1.1 Introduction

<A brief introduction suitable for a non-specialist, *i.e.* without using technical terms or jargon, as far as possible. This may be similar/the same as that in the 'Outline and Plan' document. The remainder of this chapter will normally cover everything to be assessed under the `Background Research` criterion in the mark scheme.>

Object detection has always been of great significance in the field of computer science, with the main objective to enable the computer to accurately classify the objects in a given picture and find the position of each object. In recent years, many computer vision researchers both at home and abroad had developed a large number of excellent object detection algorithms in a neural network, including Faster R-CNN, SSD, YOLO. The goal of this project is to select appropriate object detection algorithms and data sets to train a deep learning model, then develop a website for users which allows them to complete object detection tasks easily. Moreover, the website should provide additional functions that exclude basic object detection functionality to offer users a better experience, for instance, allowing them to change weights to suit different tasks, or mark recognition results after logging in.

From the perspective of research, the significance of object detection dues to it is one of the fundamental tasks in the field of computer vision, since it is the basis of many other high-level tasks, including image classification, face recognition, target tracking, pedestrian re-recognition. Meanwhile, there is a large number of well-known national and international research teams had been focused on the field of object detection: MIT Computer Science and Artificial Intelligence Laboratory, Stanford Computer Vision Lab, National Laboratory of Pattern Recognition of Chinese Academy of Sciences, LAMDA Institute of Nanjing University, who put enormous efforts into this task in order to make a better world.

While from the perspective of the application, object detection has shown a wide range of practical usages: face detection technology, pedestrian detection technology applied in video surveillance, entrance and exit statistics, traffic sign detection technology, vehicle detection technology applied in aided driving, automatic driving. All these technologies serve a common purpose to make people living in a more convenient life. At the same time, major technology companies, for example, Microsoft, Google, Ali, and Baidu, have also spent a lot of manpower and material resources to explore the object detection field, which indicates the significance and prospect of object detection.

## 1.2 Literature Review

<This section heading is purely a suggestion -- you should subdivide this chapter in whatever manner you think makes most sense for your project. It may also make sense to spread the `Background Research' over more than one chapter, in which case they should be named sensibly.>

更早之前的滑窗法？

Since 2006, in the lead of Hinton, Bengio, Lecun, and other researchers, an enormous number of deep neural network papers had been published, especially after Hinton's research group participated in the ImageNet image recognition competition in 2012 and won the championship using AlexNet[1], constructed by CNN (Convolutional Neural Network), then neural networks began to receive extensive attention from then on.

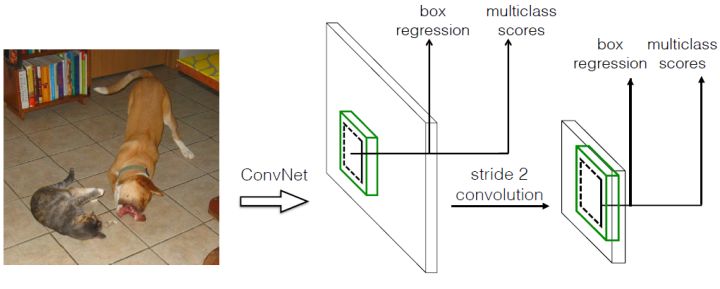
(Object detection 的Problems)Computer vision analysis of target motion can be roughly divided into three levels: motion segmentation and object detection, target tracking, action recognition and behavior description[2]. Object detection is not only one of the basic tasks to be solved in the field of computer vision but also the basis of video surveillance technology among the other tasks. However, object detection is still a challenging task with great potential and has great space for improvement, since the objects in the video have different poses and often appear to overlap, especially when their movements are irregular. Meanwhile, the resolution, weather, illumination, and other conditions of the surveillance video or images as well as the diversity of scenes should also be taken into consideration, which makes this task more challenging.

Overall, object detection is a task to find all the objects of interest in the image for two sub-tasks, including object positioning and object classification[3]. The traditional object detection method, for example, the sliding window algorithm is generally divided into three stages: firstly, select some candidate regions on a given image, then extract features from these regions, and finally classify them using trained classifiers.

At present, the mainstream object detection algorithms are mainly based on a deep learning model, which can be roughly divided into two categories:

* One-stage object detection algorithm. This kind of detection algorithm does not need the Region Proposal Stage and can directly generate the category probability and position coordinates of objects through only one stage. One-stage typical algorithms include YOLO, SSD, and CornerNet[4].
* Two-stage object detection algorithms divide detection problems into two stages, the first stage is the generation of Region Proposals, which includes the approximate location information of the object, and the second stage is the classification and location refinement of the candidate regions. The representatives of two-stages algorithms are R-CNN, Fast R-CNN, Faster R-CNN.

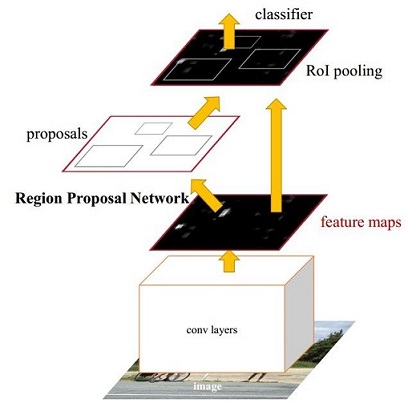
The main performance indicators of an object detection model are the accuracy and processing speed, while the accuracy mainly considers the positioning and classification of the object. In general cases, the two-stage algorithm has advantages in accuracy, whilst the one-stage algorithm has higher processing speed.

* SSD, stands for Single Shot MultiBox Detector, is a single-stage, multiple proposal object detection algorithm. SSD uses CNN with a multi-scale feature map to detect objects, the basic structure shows as the figure below:

**Figure 1.2.1** SSD basic structure

Additionally, SSD adopts VGG16 as the basic model, and then adds a convolution layer based on VGG16 to obtain more feature maps for detection[5]. There are also some improved algorithms based on SSD, for instance, DSSD[6] and FSSD[7], which have a different structure for their CNN module.

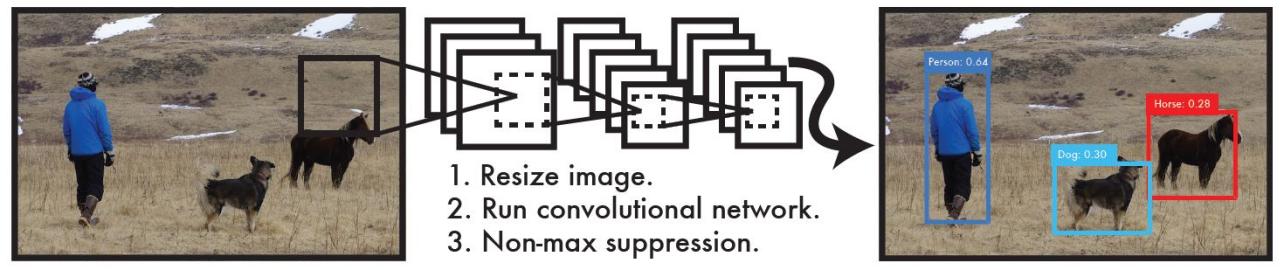
* Faster R-CNN, is a model after the evolvement of R-CNN and Fast R-CNN, when Ross B. Girshick proposed it in 2016[8]. In terms of structure, the Faster R-CNN integrated feature extraction, bounding box regression (rectangular refine), and classification into one network, which greatly improves the overall performance[9], especially in the detection speed. The structure of the network is shown as below:



**Figure 1.2.2** Faster R-CNN basic structure

Furthermore, Cascade R-CNN was proposed as a multiple-stage method on the basis of previous versions[10] but has better accuracy in detection.

* YOLO, stands for You Only Look Once, which adopts a separate CNN model to achieve end-to-end object detection. The whole system is shown in the figure below:



**Figure 1.2.3** YOLO basic structure

The input image is resized and sent to the CNN network, then the detected object results are obtained by processing network prediction. Compared with the R-CNN algorithm, YOLO is a unified framework with faster speed while the training process is also end-to-end(解释什么叫end-to-end)[11].

In general, object detection is an active research direction in the field of computer vision. Although one-stage detection algorithm and two-stage detection algorithm had achieved good results, there is still a great potential promotion for application in real scenes, and it will remain a meaningful and challenging topic for now and future.

Object detection had long been one of the most important and challenging task in computer vision, which detects instances of objects in a given image. It is also basis of other high-level significant computer vision tasks, instance segmentation[1-4], object tracking[8], etc.

Over twenty years, object detection had been developed rapidly especially due to the archivement of deep learning recently[9], tremendous imrpovements pushed many applications into usage, for instance, face detection technology, autonomous driving, pedestrian detection technology applied in video surveillance, entrance and exit statistics.

Looking back through the history of object detection, it could be divided into two periods: traditional method and deep learning method with the separate line of 2014.

Traditional methods including Viola Jones Detectors (2001)[10,11], HOG Detector(2005)[12] and DPM (2008). DPM, stands for Deformable Part-based Model, was the winner of the VOC-2007 competition, which also was the best technique in the end era of the traditional method, which focusing on detecting smaller parts of an object, for example, a human could be divided into arms, legs, head.

The efforts of traditional methods not only clear the path but also brought many inspirations to the researchers who study deep learning methods.

For the deep learning era, the methods were normally divided into two categories: one-stage methods and two-stage methods.

One-stage methods hold the philosophy of detect the objects in only one step, famous algorithms including YOLO, SSD, RetinaNet, while the two-stage methods embrace the idea of “from coarse to fine”[引用20years论文], representive methods for example, R-CNN, Fast R-CNN, Faster R-CNN, etc.

YOLO[YOLO 引用] stands for “You Only Look Once”, is one of the most famous and first one-stage object detection algorithm. YOLO-v0 originates from the idea of trasnforming the classificationing network directly into positioning network by dividing the image into serveral parts, then predicts with bounding boxes. Since YOLO both has rapid recognition speed and high accuracy, which is suitable for a lightweight website application, this project adopts YOLO-v5 as basis to develop a object detection website. More details will be concentrated on YOLO from v0 to v5 in later chapters.

SSD [21] stands for Single Shot MultiBox Detector, the uniqueness of SSD is its multi-scale detection techiniques which allow predicting different sizes of objects using different layers in the network, which significantly enhance the accuracy of single stage method.

RetinaNet was proposed in 2017[23], which deeply analyse the differences between one-stage methods and two-stage methods, and designed a new loss fucntion called “Focal Loss” in order to solve the problem of imblance between negative and positive samples.

R-CNN, Fast R-CNN, and Faster R-CNN are a series two-stage algorithms. R-CNN introduces selective search to generate 2000 proposals, then resizes them to fed in a CNN network for feature extraction, a SVM classifier will handle the features and give predictions.[R-CNN引用]

Since the 2000 proposals of R-CNN needs large computational resources, leading a really slow speed, Fast R-CNN extracts features from the whole image instead of proposal regions to update weights, and adopts selective search to the outputs of convolution rather than raw image which signiticantly reduces computation cost and enhance detecting speed.[Fast R-CNN引用]

The Faster R-CNN intergrates RPN (Region Proposal Network) to replace original selective search, reducing a large mount of computation redundancy and improve accuracy at the same time.[Faster R-CNN引用]

Excellent algorithms still need proper dataset to train ane evolve, competitions normally will be hold to test performance on datasets for famous algorithms, also serving as benchmark for later research. Among these datasets, Pascal VOC, MS-COCO and Open Images are frequently used for object detection tasks.

Pascal VOC stands for The PASCAL Visual Object Classes (VOC) Challenges[50,51], supports many computer vision tasks, for example, object detection, instance segmentation, [放图 link] object tracking. Since Pascal VOC is one of the most important dataset for object detection, so there are already benchmarks created by other famous algorithms. Moreover, with relatively smaller size than others, Pascal VOC is suitable for this lightweight project. This project finally selects VOC 2007 for training, because VOC 2007 focused on object detection tasks, but the latter version (VOC 2012) also used for other tasks (instance segmentation and object tracking). Despite its small size, it also has enough data to train a decent model, with 9963 images (train, validation and test), in total 24640 objects for 20 different classes.

(后期调整一下顺序？) MS-COCO is also a famous dataset that used in object detection task, which has more than 160 thousand images with about 900 thousand objects for 80 classes. Open Image dataset is normally used in two tasks: object detection and predict relationships between objects.

Within the appropriate algorithm and datasets for training, then a evaluation metrix is needed to test the performance of a method. In object detection task, mAP (mean Average Precision) is one of the most important indicator, which was first mentioned in VOC-2007. To define AP, PR (Precision-Recall) curve is requested with specified IoU (Intersection over Union) threshould (usually 0.5), the area under the PR curve means the AP value. After that, sum up all the AP for every class then divide class number can obtain the final mAP.

Among the history of object detection, there are some important concepts that brought huge influence to current object detection research need to be mententioned.

CNN, stands for convolutional neural networks, had been used since 1990[96 ?是否引用] in a variety of computer vision field. CNN had began the new era of deep learning by allowing parameter sharing and sparsity of connections, which serve as the basis of all the one-stage and two-stage methods that mentioned above.

NMS, stands for Non-Maximum Suppression, is a technique that used to eliminate redundant bounding boxes for the same object generated by the network. There are also many other NMS versions, for instance, Soft-NMS, DIoU NMS, Conv NMS, Learning NMS, etc. The invention of NMS had greatly enhanced accuracy of the algorithms, which make it remains necessary component of modern network.

Moreover, some new strategies that used to applied in other filed are now put into experiment in object detection and yield remarkable results.

Adversarial training, or GAN (Generative Adversarial Networks) [286] is quite popular these days especially in researches that allowing AI to generate their own production (composing songs, draw paintings, etc.). Typically, GAN includes two networks, one “teacher” network to criticise the AI production, while another one “student” network generate AI production, two networks will learn together and yields better results after training. For object detection task, GAN had been put into usage to enhance the performance when detecting small or overlapping objects.

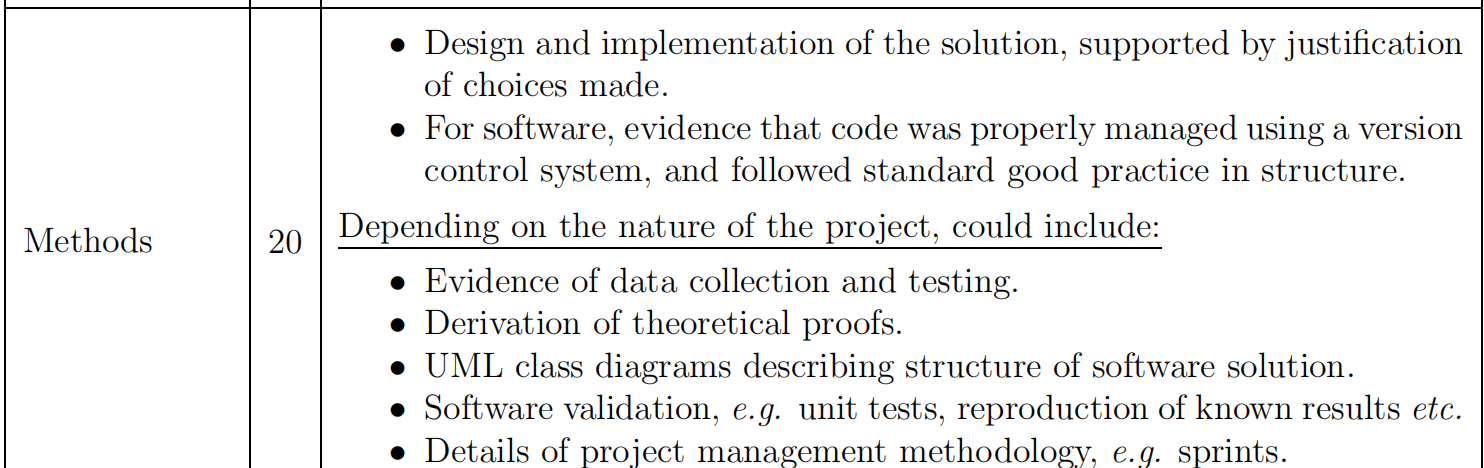
Inspite of the great progress that object detection ever had since start, there are still some serious challenges that had troubled many researchers till now, for instance, when detecting bad weather situation (strongly snowy, foggy, etc.), or many small targets overlapping (a really busy street with hundreds of pedestrians) performance of algorithms normally are [barely](javascript:;) [satisfactory](javascript:;). To solve these problems above, further researches are needed to enable object detection techniques to applied in more situations to provide people better lives. For the future, object detection researches may focued mroe on real-time techniques, which could provide video surveillance on road or autonomous cars. Therefore, video instead of images would become the mainstream media to apply object detection, which means more fast but accurate (YOLO) algorithms will be developed to serve the needs. Object detection application that in smartphones is also a promising direction, which may also trigger the development of more light-weight models that could used in mobile devices. In a netshell(是否合适说netshell?), object detection techniques are and will be more important in people’s lives than ever.

简单提一下技术？：CNN, NMS, ~~ROI Pooling?~~, ~~multi-scale?~~, ~~Using context~~, ~~cascade~~?,

New method: adverisal training, ~~weakly supervised object detection~~

# Chapter 2 Methods

<Everything that comes under the `Methods' criterion in the mark scheme should be described in one, or possibly more than one, chapter(s). Note that it is not normally relevant to include extensive code, but short snippets for key aspects may be suitable.>



Justification of choices made

Version control system evidence,

good practice in structure

Data collection(format transform, VOC2007), testing(Postman, Selenium, pytest...)

UML class diagrams describing structure of software solution

Software validation (unit tests, reproduction fo known results)

Details of project managemnet methodology (sprints) ? Agile???

## 2.1 Table example

|  |  |  |
| --- | --- | --- |
| **Heading One** | **Heading Two** | **Heading Three** |
| 1.1 | 1.2 | 1.3 |
| 1.21 | 1.22 | 12.3 |
| 12.31 | 12.32 | 12.33 |

Text before table. Text before table. Text before table. Text before table. Text before table. Text before table. Text before table. Text before table. Text before table. Text before table.

**Table 2.1** This is the table description in the ‘table description’ style.

## 2.2 Figure example

Figures can be added using the Illustrations section of the Insert tab.

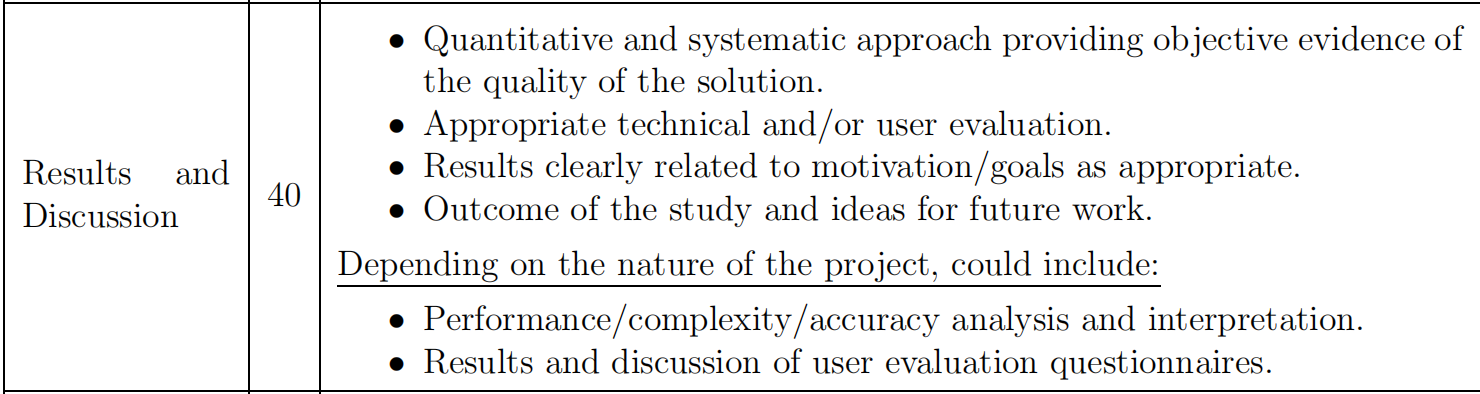


**Figure 2.1** This is the figure description in the ‘figure description’ style.

先写YOLO v0-v5进化

# Chapter 3 Results

<Results, evaluation (including user evaluation) *etc*. should be described in one or more chapters. See the `Results and Discussion' criterion in the mark scheme for the sorts of material that may be included here.>



Quantitative and systematic approach (mAP...)

Techinical /user evaluation ???

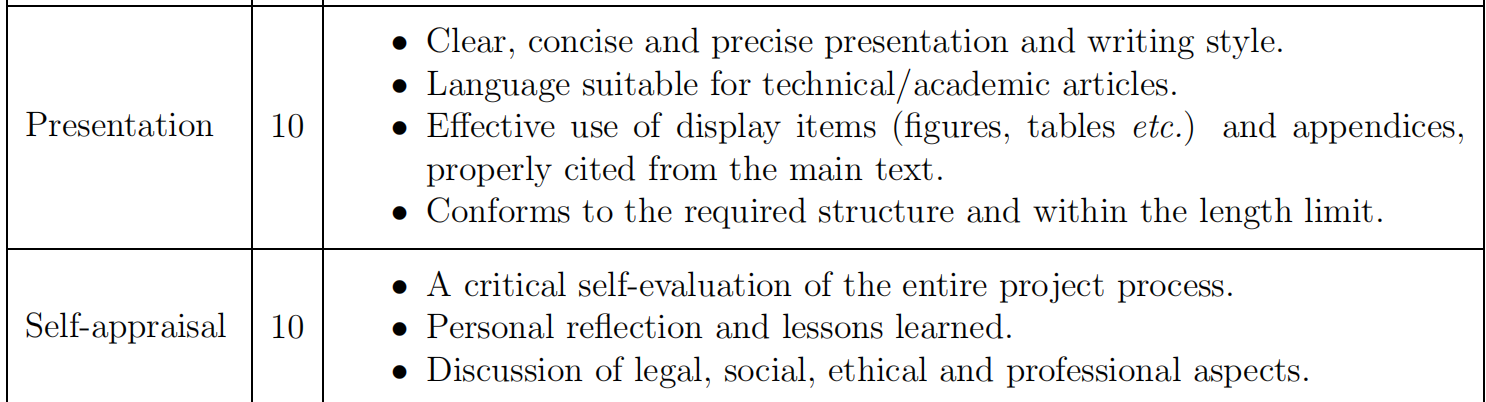
Outcome of the study and ideas for future work

Performance/complexity/ accurarcy analysis and inerpretation

Results and discussion of user evaluatin questionnaires ??? to be determined

# Chapter 4 Discussion

<Everything that comes under the `Results and Discussion' criterion in the mark scheme that has not been addressed in an earlier chapter should be included in this final chapter. The following section headings are suggestions only.>



## 4.1 Conclusions

<Text in 11-point size and 1.5 line spacing.>

## 4.2 Ideas for future work

<Text in 11-point size and 1.5 line spacing.>

# List of References

*<It is expected that the list would reflect the breadth and depth of scholarly research undertaken by the student during the course of the project.>*

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Advances in neural information processing systems, 2012, pp. 1097–1105.

[2] Vishwakarma S, Agrawal A. A survey on activity recognition and behavior understanding in video surveillance [J]. The Visual Computer, 2012: 1-27.

[3] Jiang, B., Luo, R., Mao, J., Xiao, T., & Jiang, Y. (2018). Acquisition of localization confidence for accurate object detection. In ECCV (pp. 784–799).

[4] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In Proc. Eur. Conf. Comp. Vis., pages 734–750, 2018.

[5] Liu, Wei, et al. "SSD: Single Shot MultiBox Detector." European Conference on Computer Vision Springer International Publishing, 2016:21-37.S

[6] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg. DSSD: Deconvolutional single shot detector. arXiv:1701.06659, 2016. 1, 2, 8

[7] Z. Li and F. Zhou. Fssd: Feature fusion single shot multibox detector. arXiv preprint arXiv:1712.00960, 2017.

[8] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection

and semantic segmentation. In CVPR, 2014.

[9] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In NIPS, 2015.

[10] Cai, Z., Vasconcelos, N.: Cascade r-cnn: Delving into high quality object detection. arXiv preprint arXiv:1712.00726 (2017)

[11] Redmon, Joseph, et al. "You Only Look Once: Unified, Real-Time Object Detection." (2015):779-788.

[12] Alexey Bochkovskiy, Chien-Yao Wang, and HongYuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934, 2020.

[13] Joseph Redmon and Ali Farhadi. YOLOv3: An incremental improvement. arXiv preprint arXiv:1804.02767, 2018.

[14] J. Redmon and A. Farhadi. YOLO9000: Better, faster, stronger. In CVPR, 2017.

[15] Zou, Zhengxia, et al. "Object detection in 20 years: A survey." arXiv preprint arXiv:1905.05055 (2019).

# Appendix A Self-appraisal

<This appendix must contain everything covered under the ’self-appraisal’ criterion in the mark scheme. Although there is no length limit for this section, 2-4 pages will normally be suﬃcient. The format of this section is not prescribed, but you may like to consider the following sections and subsections.>

## A.1 Critical self-evaluation

## A.2 Personal reﬂection and lessons learned

## A.3 Legal, social, ethical and professional issues

<Refer to each of these issues in turn. If one or more is not relevant to your project, you should still explain *why* you think it was not relevant.>

### A.3.1 Legal issues

<Discussion of legal issues>

### A.3.2 Social issues

### <Discussion of social issues>

### A.3.3 Ethical issues

### <Discussion of ethical issues>

### A.3.4 Professional issues

<Discussion of professional Issues>

# Appendix B External Materials

<This appendix should provide a brief record of materials used in the solution that are not the student's own work. Such materials might be pieces of codes made available from a research group/company or from the internet, datasets prepared by external users or any preliminary materials/drafts/notes provided by a supervisor. It should be clear what was used as ready-made components and what was developed as part of the project. This appendix should be included even if no external materials were used, in which case a statement to that effect is all that is required.>

Code: Yolov5 developed as part of the project

Dataset: VOC2007 ready-made components