ENHANCED SE-RESNET 101 FOR FOOD IMAGE SEGMENTATION

SUHAILA FARHAN AHMAD ABUOWAIDA

UNIVERSITI SAINS MALAYSIA

2022

ENHANCED SE-RESNET 101 FOR FOOD IMAGE SEGMENTATION

by

SUHAILA FARHAN AHMAD ABUOWAIDA

Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosophy

December 2022

ACKNOWLEDGEMENT

First of all, profuse thanks to Allah, who helped me and gave me the ability to achieve this work. I would like to thank my supervisor Dr. Huah Yong Chan from the School of Computer Sciences at Universiti Sains Malaysia, for his tireless help and support throughout my PhD and his advice, patient, and motivation, allows me to grow as a research scientist. As my PhD has come to an end, I would like to thank him for being a wonderful supervisor and a wonderful father. Also, I must express my very profound gratitude to my colleagues. Most of all, I would like to give a big thank to my mum, dad, husband, children, and aunt as well as my brother and sister for their help and love in my academic pursuits, In particular, the patience and understanding shown by the husband without him I would not have achieved any of them.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
AMI	Amazon Machine Image
ANN	Artificial Neural Network
AP	Averaged Precision over union section over IoU thresholds
AR	Aspect Ratio
AWS	Amazon Web Services
BMI	Body Mass Index
BN	Batch Normalization
BOW	Bag Of visual Words
CH	Co-occurrence Histogram
CLBP	Completed a Local Binary Pattern
CNN	Convolution Neural Network
COCO	Common Objects in Context
CRF	Conditional Random Field
DCD	Dominant Colour Descriptor
DCNN	Deep Convolution Neural Network
DNN	Deep Neural Network
FCIS	Fully Convolutional Instance-Aware Semantic Segmentation
FCN	Fully Convolutional Networks

FPN	Feature Pyramid Network
FV	Fisher Vector
GD	Gradient Descent
GLCM	Grey Level Co-occurrence Matrix
GPU	Graphics Processing Unit
HCRF	Hierarchical Conditional Random Field
HoG	Histogram of Oriented Gradient
ILSVRC	Image-Net Large-Scale Visual Recognition Challenge
IoU	Intersection Over Union
L	Large
LRN	Local Response Normalisation
M	Medium
MDL	Minimum Description Length
MKL	Machine kernel learning
MNC	Multi-task Network Cascades
MSGD	Mini-batch Stochastic Gradient Descent
NN	Neural Network
NNC	Nearest Neighbour Classifier
NYU	New York University
PCN	Point Completion Network
ReLU	Rectified Linear Units

RGB	Red, Green, and Blue
RGB-D	RGB-Depth
RoI	Regions of Interest
RPN	Region Proposal Network
SCD	Scalable Colour Descriptor
S	Small
SENet	Squeeze-and-Excitation Network
SGD	Stochastic Gradient Descent
SIFT	Scale Invariant Image Transform
STF	Semantic Texton Forest
SVM	Support Vector Machine
VGG	Visual Geometry Group
VIA	Visual Geometry Group
WHO	World Health Organisation
YOLACT	You Only Look At CoefficienTs
2D	2-Dimensional
3D	3-Dimensional

LIST OF APPENDICES

SE-RESNET 101 YANG DIPERTINGKATKAN UNTUK PENSEGMENAN IMEJ MAKANAN

ABSTRAK

Sejak kebelakangan ini, pembelajaran mendalam telah menunjukkan kegunaan dan keupayaannya dalam penglihatan komputer kerana ketepatan dan kebolehterimaannya yang tinggi. Kajian ini memfokuskan pada kaedah pensegmenan tika yang dipertingkatkan untuk pelbagai jenis makanan berdasarkan kaedah pembelajaran mendalam dan kaedah penganggaran volum makanan yang dipertingkatkan berdasarkan kaedah pensegmenan tika yang dipertingkatkan untuk penganggaran kalori makanan yang lebih baik. Pensegmenan tika untuk berbilang makanan mengalami prestasi yang lemah walaupun menggunakan pembelajaran mendalam disebabkan penggunaan ResNet-101 sebagai tulang belakang kerangka untuk pengekstrakan ciri. Masalah ResNet-101 masih wujud, seperti penentuan blok ResNet-101 yang sesuai, dan ciri-ciri kecil menjadi lebih kecil atau hilang semasa pensampelan turun. Kaedah pembelajaran mendalam sedia ada juga menggunakan lapisan pengumpulan RoI untuk mengekstrak ciri saiz tetap daripada peta ciri. Pengumpulan RoI memperkenalkan salah jajaran antara RoI dan mengekstrak ciri disebabkan penggunaan pengkuantuman yang mengurangkan ketepatan kaedah. Oleh itu, terdapat keperluan untuk mengatasi isu ini, terutamanya dalam pensegmenan tika makanan berbilang berdasarkan kaedah pembelajaran mendalam. Kaedah pensegmenan tika yang dipertingkatkan yang dicadangkan melibatkan dua langkah utama: (1) Tulang belakang ResNet-101 telah dipertingkatkan untuk pengekstrakan ciri yang lebih baik dengan mencari blok ResNet yang optimum dan menyambungkannya dengan modul Rangkaian Squeeze-and-Excitation (SENet) melalui Piramid Ciri yang disesuaikan Rangkaian (FPN). (2) Lapisan penggabungan RoI

telah dipertingkatkan untuk menyelesaikan persoalan salah jajaran melalui pembatalan pengkuantuman untuk pengekstrakan ciri-ciri kecil daripada setiap RoI. Kajian itu diikuti dengan menambah baik kaedah anggaran isipadu makanan melalui seni bina pengekod-penyahkod yang digunakan untuk mengendalikan anggaran kedalaman dan menggabungkan penapis penyingkiran terluar yang dicadangkan dan teknik hul cembung. Kajian menganggarkan kalori makanan dalam imej tanpa objek rujukan untuk memadankan item makanan dalam imej. Kaedah pensegmenan tika yang dicadangkan untuk pelbagai jenis makanan meningkatkan ketepatan sebanyak 96.23%. Cadangan anggaran isipadu yang dipertingkatkan untuk anggaran kalori yang lebih baik berdasarkan peningkatan kaedah pensegmenan tika mempamerkan ciri yang serupa dengan ketepatan 87.95%.

ENHANCED SE-RESNET 101 FOR FOOD IMAGE SEGMENTATION

ABSTRACT

In recent years, deep learning has demonstrated its usefulness and capability in computer vision due to its high accuracy and acceptability. This thesis focuses on the enhanced instance segmentation method for multiple types of food and the improved food volume estimation method for better food calorie estimation. The existing instance segmentation methods, for multiple foods, suffer from poor performance despite using deep learning. The poor performance is due to the adoption of ResNet-101 as a backbone for feature extraction. ResNet-101 problems still exist, such as determining suitable number of ResNet-101 blocks and small features becoming smaller or vanishing during downsampling. The existing instance segmentation methods for multiple foods also adopt a RoI pooling layer to extract a fixed-size feature from the feature map. The RoI pooling layer introduces misalignments between the RoI and extracts features because of applying quantisation, which reduces the method's accuracy. Therefore, this research aims to enhance methods for instance segmentation method. The proposed enhanced instance segmentation method involves two main steps: (1) Enhancing the ResNet-101 backbone for better feature extraction by finding the optimal ResNet blocks and connecting them with the Squeeze-and-Excitation Network (SENet) via adapted Feature Pyramid Network (FPN). (2) Enhancing the RoI pooling layer for solving the question of misalignment via annulling the quantisation for extraction of small features from each RoI. The study was followed by improving the food volume estimation method through an encoder-decoder architecture to handle the depth estimation. The proposed outlier removing the filter was combined with the convex hull technique. The proposed enhanced instance segmentation method for

multiple food types improved accuracy by 96.23%. The proposed improved volume estimation for better calorie estimation based on an enhanced instance segmentation method exhibits similar characteristics with 87.95% accuracy.

CHAPTER 1

INTRODUCTION

1.1 Motivation

The developments in computer vision and machine learning opened the way for the robustness of multiple food calorie estimation methods. Multiple food calorie estimation has been a popular research topic in health-related areas for years. The performance and effectiveness of the food calorie estimation method depend on two factors, include the following:

1.1.1 Food segmentation

Food segmentation of the image is the initial stage, where the segmentation for each food item in the image is recognised. Food segmentation is a significant challenge due to many challenges, such as different ingredients, sizes, shapes, colours, and a variety of food with similar shapes and appearances. The existing segmentation methods had a low segmentation rate because of the inability to deal well with food images containing more than one type of food, and the methods suffered from a loss of features at the instance level.

1.1.2 Food volume estimation

During this stage, the method calculates each item's volume in the food image once the food segmentation is part of stage one. Estimating volume also needs to calculate the volume in the food image by reconstructing a 3D shape from a 2- Dimensional (2D)

image to predict each food item's volume and calculate the calorie via deep learning methods. The existing food volume estimation methods had several challenges, because they depend on a reference object to be placed next to the food items to match the food items in the images, and in the stereo-based approach, participants must obtain various food images from different viewing angles. This approach may be tiresome. The reflecting light of an item might change depending on the viewing angle, making feature point matching and 3D reconstruction difficult (Lo *et al.*, 2018).

1.2 Problem statement and research questions

The multiple food calorie estimation is a significant challenge due to many challenges, such as different ingredients, shapes, and duplication. The variety of food with a similar shape and appearance leads to difficulty in calorie estimation of the various types of food via the image. Therefore, a segmentation method is needed to separate those types with irregular shapes, especially when there is an occlusion in the food image. The object segmentation is considered an open research problem in the computer vision field because it requires the correct detection of all objects in an image while precisely segmenting each instance (He *et al.*, 2017). The object segmentation methods exhibit systematic errors on overlapping instances and create spurious edges (He *et al.*, 2017). The existing segmentation methods based on traditional methods (Hoashi *et al.*, 2010; Chen and Ngo, 2016; Yang *et al.*, 2010; Siswantoro *et al.*, 2015) had a low segmentation rate because of the inability to deal well with food images that contained more than one type of food. In addition, the types of foods have a diversity of shapes, colours, and sizes. The instance segmentation for multiple foods suffers from poor performance despite the use of deep learning by Le (2020); Li *et al.* (2020); Pfisterer

et al. (2019); Ege and Yanai (2017) due to the adoption of ResNet-101 as a backbone for feature extraction (Le, 2020; Li *et al.*, 2020). The ResNet-101 problems still exist, such as determining suitable number of ResNet-101 blocks and small features becoming smaller or vanishing during downsampling (Lin *et al.*, 2017a). Additionally, the existing deep learning method adopts the RoI pooling layer (Pfisterer *et al.*, 2019; Ege and Yanai, 2017) to extract a fixed-size feature from the feature map. The RoI pooling introduces misalignments between the RoI and extracts features because of applying quantisation, which reduces the method's accuracy (He *et al.*, 2017). Therefore, there is a need to overcome these issues, especially in multiple food instance segmentation based on deep learning methods. Additionally, methods are needed to estimate food calories without a reference object after segmenting the type of food from the input image. The majority of studies used 3D for estimating the food volume estimation. They depend on a reference object to be placed next to the food items (Dehais *et al.*, 2016; Liang and Li, 2017; Pouladzadeh *et al.*, 2015; Puri *et al.*, 2009) to match the food items in the images. In the stereo-based approach (Dehais *et al.*, 2016; Liang and Li, 2017; Kong and Tan, 2012; Villalobos *et al.*, 2012; Puri *et al.*, 2009), participants must obtain various food images from different viewing angles. This approach may be tiresome. The reflecting light of an item might change depending on the viewing angle, making feature point matching and 3D reconstruction difficult (Lo *et al.*, 2018). Therefore, there is a need to develop food calorie estimation through 3D construction. This research aims to segment the food type based on enhancing deep learning methods to calculate food calories. This research is primarily aimed at estimating calories of food types: (1) without major exterior characteristics or features, (2) with occlusion, (3) without a fiduciary marker, (4) without depth-sensing cameras, and (5) with

an irregular shape.

Therefore, the research questions are:

1. How to enhance the quality of instance segmentation based on deep learning for better segmentation of multiple types of food?
2. How to improve volume method for better food calorie estimation?

1.3 Objective of the research

To recap, this study aims to provide a better calorie estimation through more recent and improved deep learning methods. The objectives of this research are:

1. To enhance the instance segmentation method by finding the suitable number of ResNet blocks and connecting them with the Squeeze-and-Excitation Network (SENet) architecture via an adapted Feature Pyramid Network (FPN).
2. To enhance the instance segmentation method by preserving the location of the feature map by stopping quantisation.
3. To improve the volume estimation method by combining the proposed outlier removing filter and the convex hull technique.

1.4 Contributions to this research

The contributions of this research include the following:

1. Enhanced instance segmentation method for multiple types of food based on

deep learning methods:

- ResNet-101 backbone has been enhanced for better feature extraction by finding the suitable number of ResNet blocks and connecting them with the SENet architecture via an adapted FPN.
- RoI pooling layer has been enhanced for solving the question of misalignment via annulling the quantisation for extraction of small features from each RoI.

2. Improved volume estimation for better calorie estimation based on enhanced instance segmentation method:

   - 3D reconstruction method has been improved by combining the proposed outlier removing filter and the convex hull technique.

1.5 Research approach

An overview of the research approach and flow is illustrated in Figure 1.1, which consists of a series of steps to meet the objectives established for this research. The first step covers aspects associated with enhancing instance segmentation methods for multiple types of food based on deep learning methods. The second step addresses improved volume estimation for better food calorie estimation.

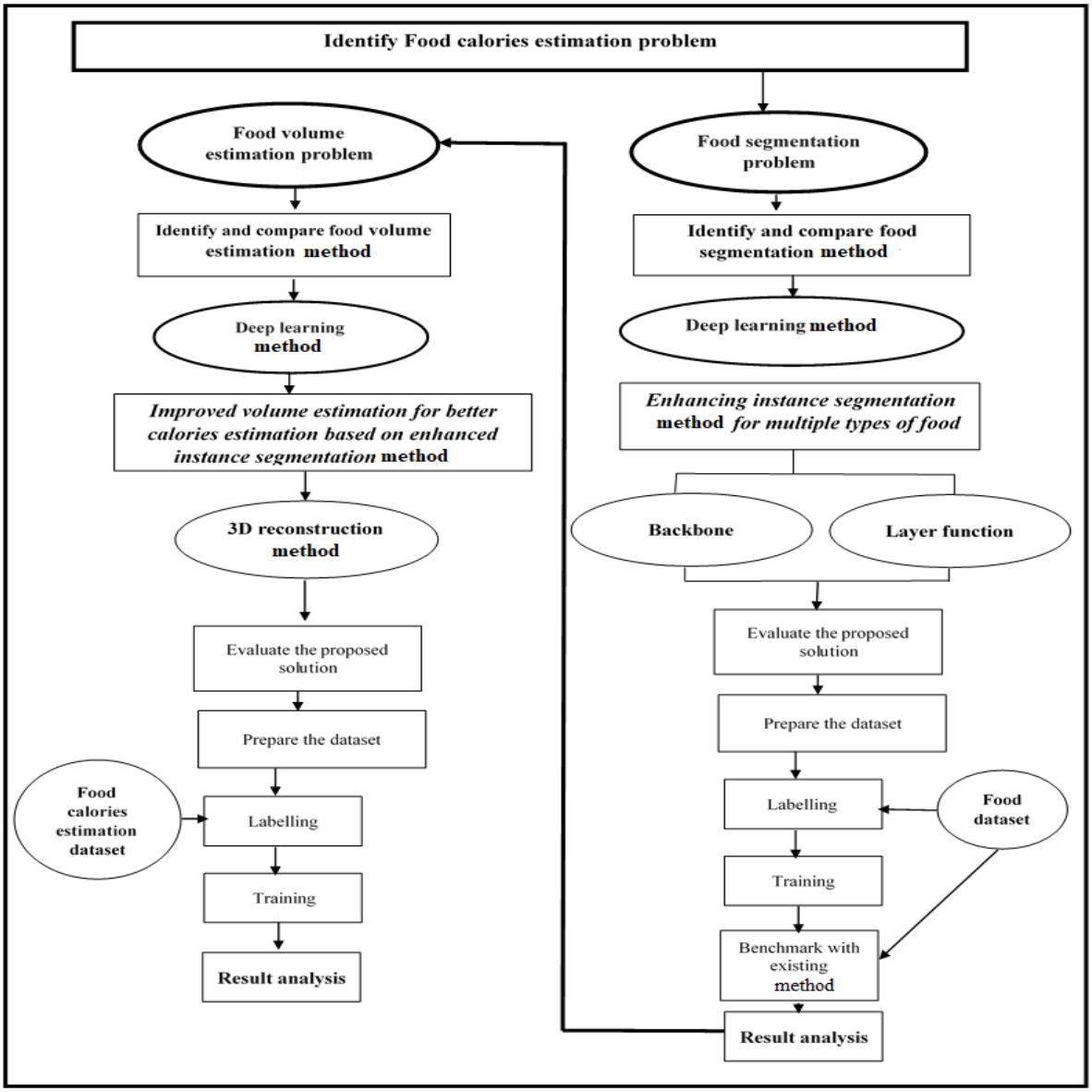

Figure 1.1: Overview of research approach

1.6 Thesis organisation

The organisation of this thesis is structured into five chapters:

- Chapter 1: The current chapter presents the motivation, objectives, research questions, problem statement, and contributions of this research.
- Chapter 2: In this chapter, a literature review is presented on various food seg-

mentation and volume estimation methods based on the traditional machine learning method and deep learning.

- Chapter 3: In this chapter explains the enhancement process involved in the food segmentation and volume estimation methods.
- Chapter 4: The results and discussion are presented in Chapter 4. This chapter includes the results of evaluating the proposed methods in this research.
- Chapter 5: This chapter presents this research's conclusion and future work.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

A review of existing methods found in previous literature is discussed in this chapter regarding their application in recognising, detecting, and segmentation objects. This is followed by further review and discussion on food segmentation and volume estimation based on deep learning and traditional machine learning methods. This is followed by further review and discussion of the point cloud completion. This research aims to enhance instance segmentation and improve volume estimation. This research refers to the food and volume estimation segmentation, discussed in detail in Section 2.3. The segmentation of food and volume estimation is one of this study's most important and prominent components. This section reviews the literature review based on traditional and deep learning methods. Section 2.4 describes a series of studies on 3D reconstruction, consisting of multiple methods that address depth estimation images based on traditional and deep learning methods and point cloud completion.

2.2 Object segmentation

Object segmentation is an essential task for computer vision. It requires splitting the visual data into segments to simplify the processing of the image. Segments reflect objects or portions of things and consist of pixel collections. Segmentation is a complex problem because it requires detecting all objects in an image and precisely segmenting. The segmentation aims to recognise individual objects and localise each

object.

2.2.1 Object segmentation based on traditional computer vision methods

Image segmentation is the process of splitting images into multiple image segments. The threshold method was a simplified method of image segmentation (Batenburg and Sijbers, 2009, 2008), which referred to the threshold value for converting a grey-scale picture into a binary image.

Histogram-based methods were one of the efficient and straightforward methods of segmentation. The histogram-based method was calculated from other pixels in the image. The histogram's tops and valleys were used to find the image clusters, which could be determined by colour or by strength (Delon *et al.*, 2006). The histogram quest's downside was that significant peaks and valleys were challenging to locate in the image.

A new segmentation method based on an edge detection method was proposed to enhance segmentation using region boundaries (Barghout, 2014). Lindeberg and Li (1997) established an integrated method that segmented edges into straight and curved edge segmented for part-based entity identification based on a Minimum Description Length (MDL). A split-and-merge process improved the method with nominee breakpoints collected from adjacent junction clues to gain more possible points to consider partitions into separate sets. Regional-growing methods were suggested to conduct segmentation based on the premise that neighbouring pixels within one area had identical values (Huang *et al.*, 2014), where one pixel was contrasted with its neighbours. If a similarity condition was reached, the pixel might be set to connect to the same cluster

as one or more of its neighbours. The selection of similarity criteria was crucial, and the effects were affected by noise in both situations.

It is evident from the methods addressed that there was a need for precise segmentation of objects. The previous analysis used traditional computer vision methods to segment object parts, but several problems could occur. This might influence the precision of objects' segmentation, such as overlapping objects, which cannot manage multiple objects.

2.2.1(a) Object recognition based on traditional machine learning methods

Object recognition is the process of identifying objects in images. Object detection takes an image as input and generates one or more bounding boxes, each with the class label attached. Both recognition and detection of the object are important in computer vision. Chang and Krumm (1999) used the Colour co-occurrence Histogram (CH) for recognition, which depended on the image. It allowed adding geometric details to the colour histogram to save pixel pairs' track. In the test stage, the method was matched in the sub-region to find all objects using a false alarm rate to choose the better parameters suitable for the method. Simultaneously, Ramesh and Mohan (2007) presented a method that applied many steps through a distributed system. The method was split into two levels, the upper and lower levels. The upper level was interested in the cognitive process, and the other bottom level was interested in the biological process in the human brain. Whereas Olaode and Naghdy (2019) detailed a shape context colour histogram and completed a local binary pattern (CLBP) method to recognise various classes of objects. Kim and Kweon (2007) used as a codebook method to minimise

the intra-classes. The method depended on a cookbook to reduce the surface marking effect. In this method, there were three stages. The first stage removed the surface marking part in the training stage, the second removed the codebook, and the last stage used Nearest Neighbour Classifier (NNC) and Support Vector Machine (SVM) to classify different matrices. The method applied to the Caltech-101 database has significant intraclass variations. Otoom *et al.* (2008) rated the performance of various feature groups for choosing the better feature group that was more suitable for recognizing objects. However, the experiment showed that using different recognitions and evaluation schemes, the method recognises the geometric primitives feature group's statistics better than the Scale Invariant Image Transform (SIFT) key points histogram. Wang *et al.* (2013) indicated that a method depended on comparative object similarity for learning object methods having less training in this stage. The method modified the detection and recognition methods to combine similarity constraints. Although, the method suffered drawbacks, given it was poorly adapted for large datasets. A separate study (Mokji and Bakar, 2007) presented a new method for the Grey Level Co-occurrence Matrix (GLCM) computation that depended on the Haar wavelet transform technique to minimise the computational problem through pixel entries and thereby increase the accuracy of the brodatz texture. Rockinger (1997) proposed a new technique based on a shift-invariant wavelet transform for the fusion of spatially registered images and image sequences. This technique showed better results in the image sequence problem. Moreover, the combined techniques showed an advantage in temporal stability and consistency. Dao and Vemuri (2002) proposed that the Neural Network (NN) method could apply controlled input data files for intrusion detection in the computer network. The method compared many different NN methods, such as the gradient descent, the

gradient descent with momentum, the learning rate, the conjugate gradient backpropagation, and the quasi-newton method, where the best method depended on the user's when logging into the computer network.

Cover and Hart (1967) suggested K-NN as a method to recognise the object. The method's notion was centred on the nearest feature space in the training process, considered the simplest method in traditional machine learning methods. The method consisted of two stages, the training stage and test training. In the training stage, the method kept featuring vectors for the label object, while in the test stage, the unlabelled object transferred to the nearest label, as shown in Figure 2.1.

The advantage of the K-NN method was used with various methods. However, the K-NN method's disadvantage was when the dataset was small and did not have many features, leading to an error in classifying the object.

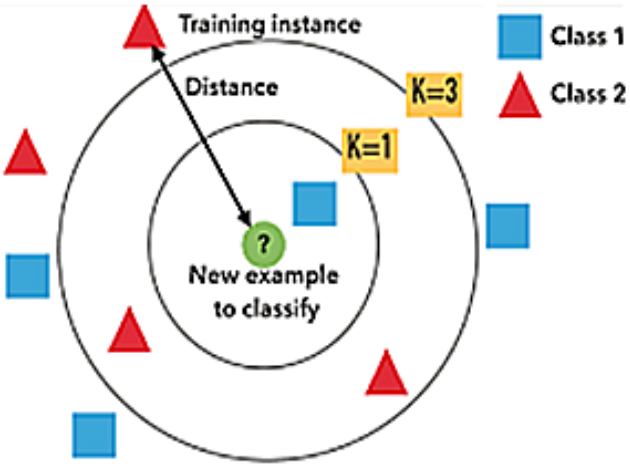

Figure 2.1: K-NN recognition (Kim[1] *et al.*, 2012)

Burges (1998) applied the SVM method using various levels of space to classify the object. The SVM method uses different perspective that maximises the edge when having different classes by dividing it. The SVM method consisted of two parts: the

training and test parts. In the training part, the method split the points to the nearest point of classes, while the test part predicted the point in space to which classes belong depending on the point's location, as shown in Figure 2.2.

The SVM method for recognition depended on the training data that predicted the class labels in the test stage. SVM recognition's advantage is that the SVM method provided a good result in different datasets, even with a small number of classes in the training stage. However, the SVM methods' disadvantages were the selected kernel parameters in the training and testing stage and the high computational time in the training and testing stage.

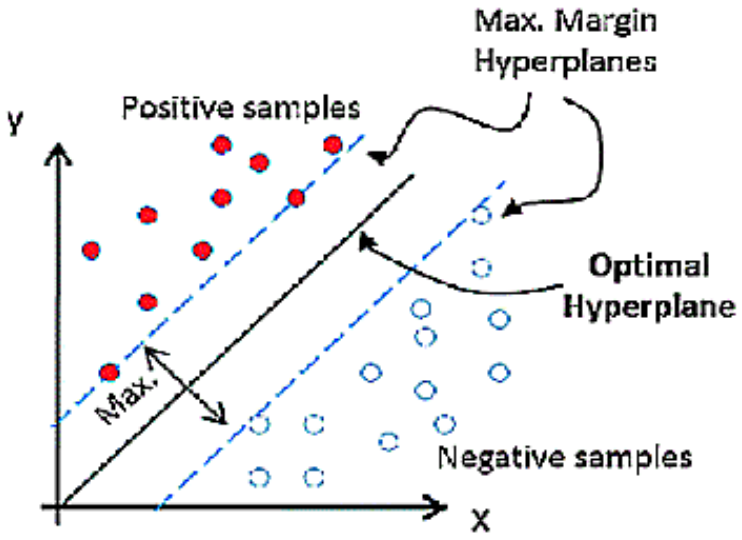

Figure 2.2: SVM recognition (Burges, 1998)

Nevertheless, Shah (2004) suggested an Artificial Neural Network (ANN) method to design and improve a hierarchical network depending on incorporating textural features. In this paper, the authors noted the importance of textural features to enhance image recognition using the ANN method. Opposingly, Haykin and Network (2004) proposed an ANN method to solve linearity and loss associated with mathematical problems. The ANN method used neurons to deal with available data following fea-

ture extraction from the image and a backpropagation method in the training stage to train, choose and update the better weight for neurons towards a suitable dataset, as illustrated in Figure 2.3. The ANN had many advantages since it could recognise or regress images but suffered from overfitting and vanishing gradient problems.

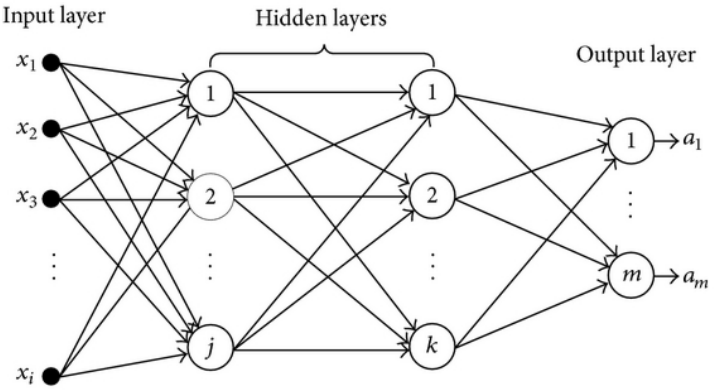

Figure 2.3: ANN Architecture (Shah, 2004)

A comparative study of various methods used for object recognition based on traditional machine learning methods such as SVM, KNN, and ANN were undertaken. Based on the analysis, each method had both advantages and disadvantages. Table 2.1 summarises the advantages and disadvantages of object recognition based on traditional machine learning methods.

Table 2.1: Summary of advantages and disadvantages of object recognition method based on traditional machine learning methods

Authors	**Method**	**Advantage**	**Disadvantage**
Chang and Krumm (1999).	Colour cooccurrence ogram.	Ability to work in unclear backgrounds.	Inability to deal well with images that contained more than one type of object.

Continued on next page

Table 2.1 – *Continued from previous page*

Authors	Method	Advantage	Disadvantage
Ramesh and Mohan (2007).	Cognitive process and biological process in the human brain.	Ability to work with different datasets.	Poorly adapted for large datasets.
Kim and Kweon (2007).	Codebook, NNC, and SVM.	Ability to work with the intra classes.	Poorly adapted for large datasets.
Wang *et al.* (2013).	Comparative object similarity to learning in training.	Could work with few training stages.	Inability to deal well with images that contained scale variation.
Mokji and Bakar (2007).	Grey Level Co-occurrence Matrix and Haar wavelet.	Minimised computational time.	Poorly adapted for large datasets.
Rockinger (1997).	Shift-invariant wavelet transform.	Ability to work with temporal stability and consistency.	Not flexible when handling occlusion image.
Dao and Vemuri (2002).	NN techniques.	Ability to work with different datasets.	Could not deal well with background clutter and could not avoid the vanishing gradient problem.
Burges (1998).	SVM.	Avoided overfitting problem and dealt efficiently with the complexity of decision rules.	Not linearly separable and could not deal well with the method's different structures.
Shah (2004).	ANN and textural Features.	Ability to work with different classes.	Could not deal well with the illumination problem and could not avoid the vanishing gradient problem.

Continued on next page

Table 2.1 – *Continued from previous page*

Authors	Method	Advantage	Disadvantage
Haykin and Network (2004).	ANN.	Solved linearity problem and dealt with loss or not clear mathematically.	Could not deal well with background clutter and could not avoid the vanishing gradient problem.

2.2.1(b) Food image segmentation based on traditional computer vision methods

Segmentation is a crucial phase in recognising various regions of an image and then extracting entity positions. A segmentation method for food identification is food should be placed in the image; other items like background or food containers are excluded (Hoashi *et al.*, 2010; Chen and Ngo, 2016). When properly applied, segmentation improved classification accuracy, mainly when many food items must be classified (Ciocca *et al.*, 2016; Kawano and Yanai, 2014). Segmenting food items is also difficult since some food images do not have clear features, for example, contours of form and food edges (Bosch *et al.*, 2011). The segmentation could be more difficult if food is sliced and occluded in food processing posts that cheat on each other and remove other pieces of food (Yang *et al.*, 2010; Siswantoro *et al.*, 2015). The research has been conducted to discuss issues relevant to the food segmentation method. Kawano and Yanai (2014) created a mobile application and proposed to develop a direct bounding box, as shown in Figure 2.4.

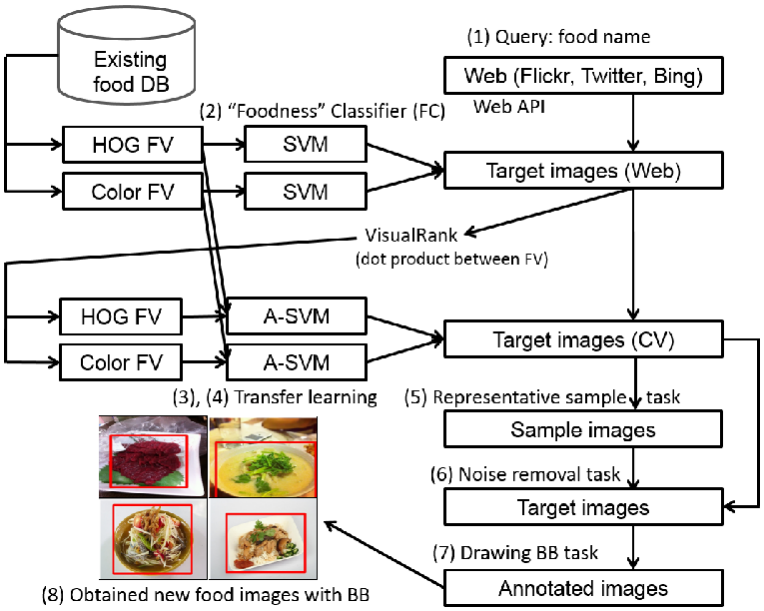

Figure 2.4: Process flow of mobile application (Kawano and Yanai, 2014)

These were segmented areas to remove selected regions using a GrabCut method. Their strategy increased the overall precision of classification. However, the user's right to choose output is also constrained by food products properly. Pouladzadeh *et al.* (2014) proposed a method using the GraphCut method to segment food. The two-set image graph representation (A, B) centred on the dissimilarity contained in the weight (w) of the boundary binding the pixels next to it (u, v) selected background food images, as shown in Figure 2.5.

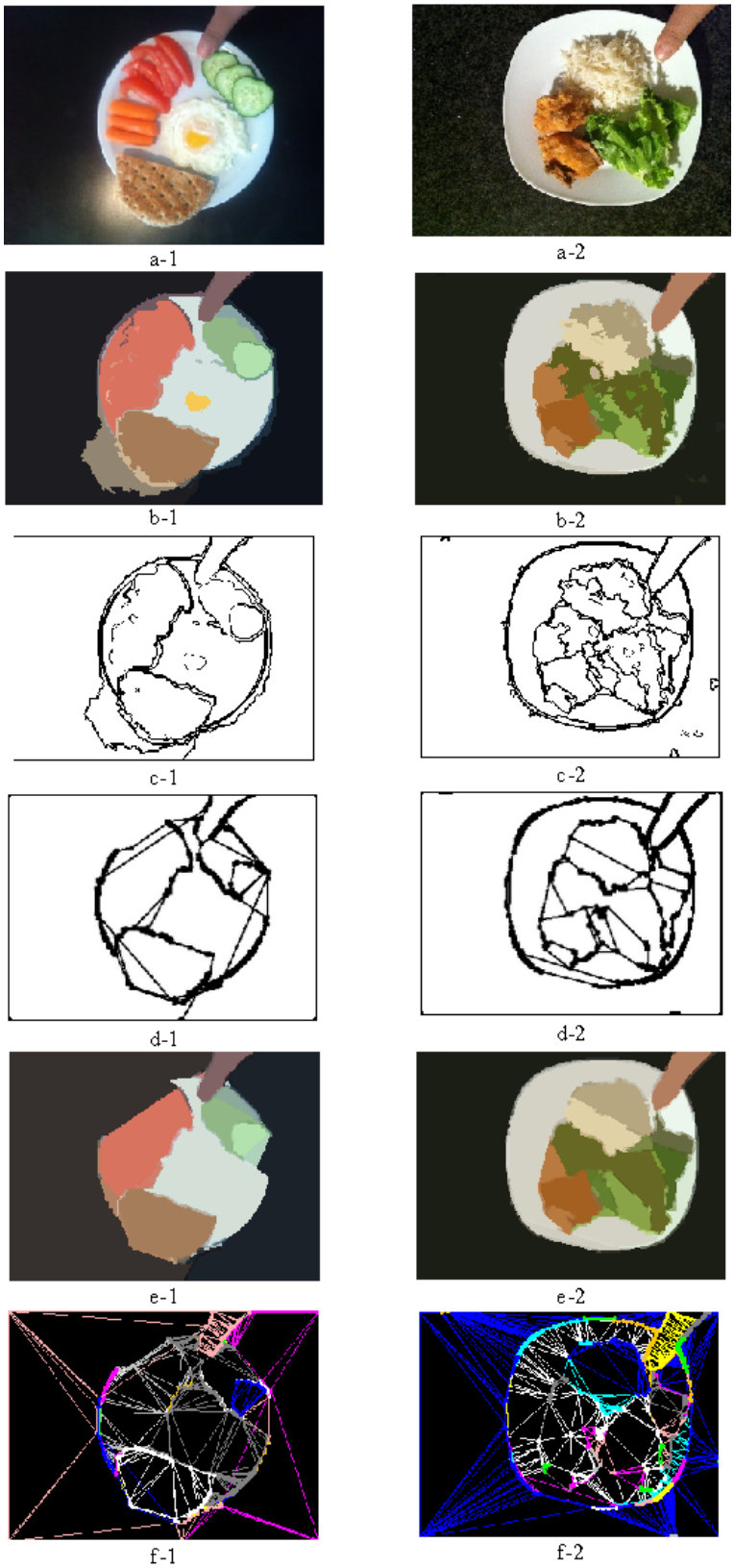

Figure 2.5: Process flow of GraphCut method (Pouladzadeh *et al.*, 2014)

Several segmentation methods merged image colour, saturation, JSEG segmentation, and noise removal to fix various food markers (Ciocca *et al.*, 2016). Multiple food groups contained in this research food tray provided in a canteen was found. Nevertheless, the tray photos were separated manually by polygonal boundaries. Furthermore, Yang *et al.* (2010) tried to separate the component based on the spatial connection between the items in the image using the Semantic Texton Forest (STF). Nevertheless, the composition of visually distinctive ingredients arranged in predictable space settings was the basis for this method. Zhu *et al.* (2014) also proposed several theories for segmentation with a class name by using the effects of the classification as input on the segmentation. The number of segments was calculated in the image, considering the trust values attributed to each segment.

He *et al.* (2013) adopted a local segmentation of shift method introduced along with a refinement of segmentation feedback to improve the categorised items' ranking. The overall ranking has been an enhanced method to normalise cuts (Shi and Malik, 2000; Kong and Tan, 2012) using a distance method of the viewpoint of three caught views and segmented food items by clustering functionality. Simultaneously, Fang *et al.* (2018) requested users to draw a wealth of details box and pick the correct food tag from the accessible collection segment immediately, utilising the GrabCut technique. It was noticed that the semi-automatic segmentation method was efficient when used on a broad image dataset.

A comparative study of various techniques used for food segmentation based on traditional machine learning methods such as STF and GrabCut methods were undertaken. Table 2.2 summarises the performance of food segmentation methods based on

a traditional machine learning method.

Table 2.2: Summarises the food image segmentation methods based on traditional machine learning methods

Authors	**Method**	**Performance**
Yang *et al.* (2010).	Spatial relationships and STF.	Inability to deal well with image parsing in most of the food types.
Fang *et al.* (2018).	Manually drawn bounding box, manually selecting food tag and GrabCut.	Could work efficiently when using a large dataset but is not flexible when dealing with types of the food because they have a diversity of shapes, colours, and sizes.
Kawano and Yanai (2014).	Bounding box and GrabCut segmentation.	Poorly adapted for large datasets.
He *et al.* (2013).	Local variation segmentation and GrabCut segmentation.	Inability to deal well with image parsing in most of the food types.
Pouladzadeh *et al.* (2014).	GrabCut segmentation.	Poorly adapted for large datasets.
Zhu *et al.* (2014).	Multiple segmentation hypothesis with assigned segment confidence scores.	Not flexible when dealing with types of the food because they have a diversity of shapes, colours, and sizes.
Ciocca *et al.* (2016).	A combination of colour, saturation, JSEG and noise removal.	A good result on particular types of foods but is not flexible when handling with the image contained more than one type of food.
Kong and Tan (2012).	Perspective distances method and cluster segmentation.	High accuracy in the simple method but poor performance when dealing with the image contained more than one type of food.

2.2.2 Object segmentation object based on deep learning methods

Image segmentation is a technique that divides the image into multiple subgroups called image segments. This reduces the complexity of the image and makes further processing or analysis easier. Semantic segmentation is the process of grouping elements of an image that belong to the same object class together. The proposed semantic segmentation method uses a combined spatial pyramid pooling module and encode-decoder. Long *et al.* (2015) suggested Fully Convolutional Networks (FCN) for the semantic segmentation method. The method's main idea is fully convolutional networks that accept input of the arbitrary size and generate correspondingly-sized output with efficient inference and learning. The FCN employed a skip method that creates precise segmentation by merging semantic information from a deep, coarse layer with appearance information from a shallow layer. One central disadvantage in FCN is the loss of detailed information due to downsampling operations.

Chen *et al.* (2018) proposed method encodes multi-scale contextual information by probing the incoming features with filters or pooling operations at multiple rates and practical fields of view. At the same time, the latter networks can capture sharper object boundaries by gradually recovering spatial information. The instance segmentation distinguished the class level of accuracy for each object. Pinheiro *et al.* (2015) proposed a deep mask method, for instance segmentation through learning to segment objects and then classifier using Fast RCNN. The deep mask is divided into two-part. The system's first part generates a class-independent segmentation mask, while the

second part generates the probability of the patch being centered on a whole object. The method is applied effectively to the whole image at test time, generating a series of segmentation masks, each awarded an item probability score. The deep mask still has issues such as slowness and less accuracy.

Dai *et al.* (2016) proposed a Multi-task Network Cascades (MNC) method, for instance segmentation. The MNC comprised three stages; each stage had a particular task to predict each object's instance level. The first stage proposed the bounding box for each object in the image, the second stage presented a mask for each bounding box and the third stage distinguished between each instance. However, the MNC had numerous gaps in predicting instance segmentation, was inflexible, and took much time when predicting instance segmentation. Also, the main problem in MNC was that the three stages did not work parallelly. It required many parameters for each stage, leading to a prolonged time to predict instance-aware semantic segmentation, as shown in Figure 2.6.

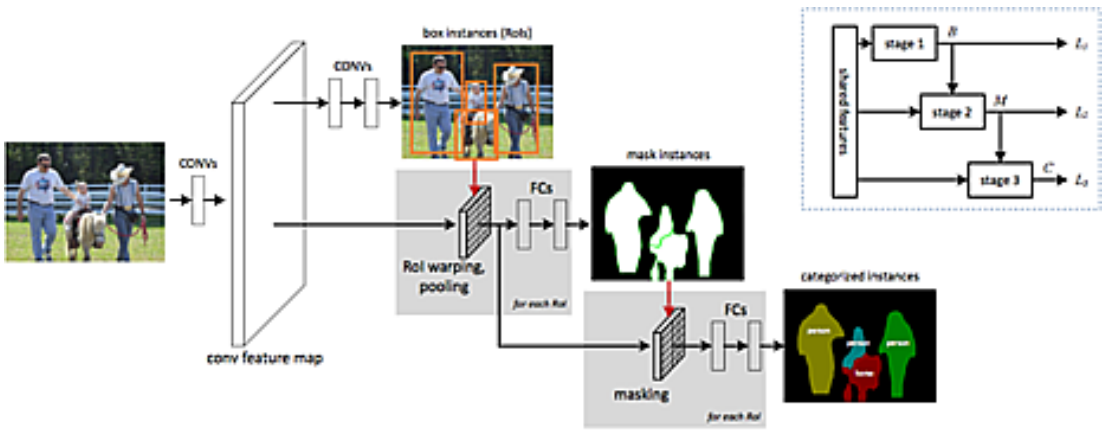

Figure 2.6: MNC architecture (Dai *et al.*, 2016)

Li *et al.* (2017) proposed a method to predict instance-level segmentation called Fully Convolutional Instance-Aware Semantic Segmentation (FCIS). The method's notion was to position sensitive inside or outside score maps for each object. The method predicted two maps, the RoI inside map and RoI outside the map, where the two maps inside and outside work together to expect and classify sub-tasks, as shown in Figure 2.7. The soft-max operation predicted each pixel's probability in the image where the high probability referred to the object category. The main advantage of this method was its flexibility in end-to-end testing and training. However, the method's central gap was overlapping and prediction errors, especially the edges.

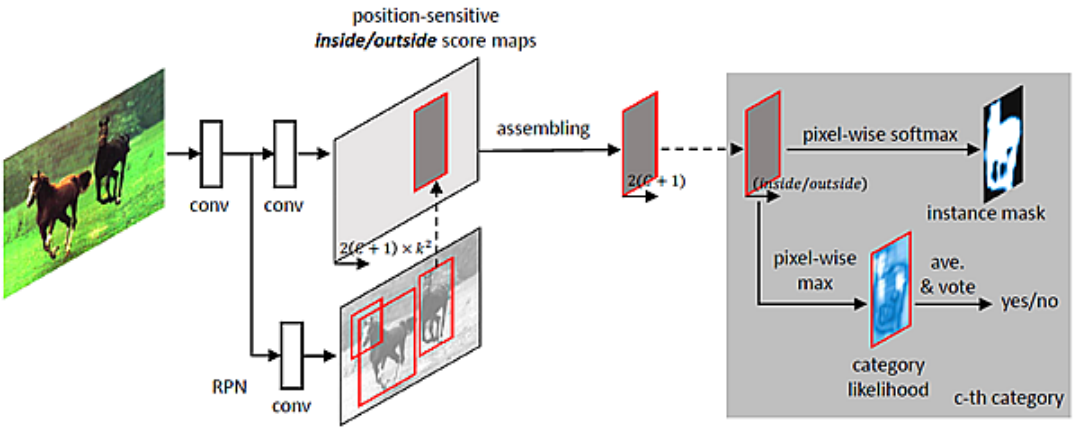

Figure 2.7: FCIS architecture (Li *et al.*, 2017)

A method was proposed, named a Mask Region-Convolution Neural Network (Mask R-CNN) (He *et al.*, 2017); it was used to predict instance-level segmentation. The method utilised the Faster R-CNN and Fully Convolutional Network (FCN) to predict class, bounding box, and mask for each object. The Mask R-CNN worked in parallel to decrease the training and testing time. However, Mask R-CNN's main limitation was losing some instance-level features.

However, Bolya *et al.* (2019) suggested You Only Look At CoefficienTs (YOLACT),

for instance segmentation in real-time processing, as illustrated in Figure 2.8. The YOLACT method worked parallel to split the YOLACT of instance segmentation into two main stages. The first stage created a multi-mask for the input image based on the FCN for semantic segmentation method (Long *et al.*, 2015). The second task added a new head to the branch, called coefficients, which was added to each mask in the first stage. The YOLACT produced the instance segmentation by combining the mask and coefficient. The main advantage of YOLACT was its ability to work in real-time, although the accuracy was exceptionally low compared with other methods. Regard-

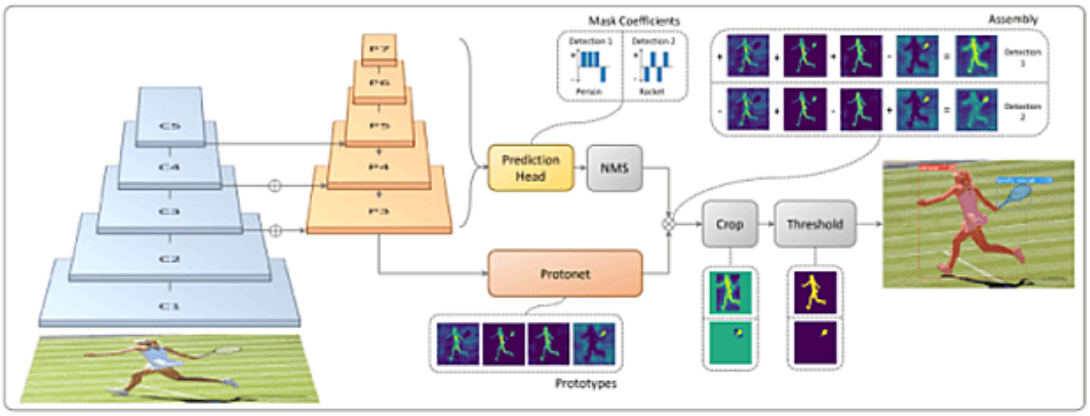

Figure 2.8: YOLACT architecture (Bolya *et al.*, 2019)

ing the architecture of YOLACT (Cai and Vasconcelos, 2019) suggested constructing a Cascade R-CNN that minimised the overfitting using a sequential threshold. However, the Cascade R-CNN increased the threshold in the training and testing time. The Cascade R-CNN architecture consisted of many stages. Each stage extended Faster R-CNN to localise each input image's object and extended Mask R-CNN to instance segmentation, as shown in Figure 2.9.

Cascade R-CNN's main advantage was decreasing overfitting through a sequence of detectors trained with increasing IoU thresholds and sequentially more selective