Watermark Identification from Currency Applying Deep Learning Techniques

Duo Tong

A project report submitted to Auckland University of Technology
in partial fulfilment of the requirements for the degree of
Master of Computer and Information Sciences (MCIS)

2021

School of Engineering, Computer & Mathematical Sciences

Abstract

Banknote identification plays an increasingly important role in financial fields due to the pervasion of e-commerce and the diffusion of automatic bank systems in terms of vending machines. Deep learning, in recent years, has become the mainstream direction in the object detection domain because of its advantages in accuracy and performance compared to machine learning techniques.

These days, the fifth generation of You Only Look Once (YOLOv5) has become a state-of-the-art detector because of its relatively outstanding accuracy at high speed. Ruminating that attention mechanisms have never been fused with YOLOv5 and its ability for performance overhaul, we add the squeeze-excitation (SE) attention module at the terminal of the backbone to further improve the watermark recognition skills.

In this report, both YOLOv5 and YOLOv5-SE are implemented to detect paper money including 10NZD, 50NZD, 100NZD. These methods not only slash human labour but also promote the precision of recognition. The contributions of this report are: (1) We generate a relatively comprehensive dataset regarding different situations including different distances, angles, and various augmentation skills. (2) We acquire extremely satisfactory outcomes whose precision reaches 99.99%. (3) By comparing different versions of YOLOv5 in terms of YOLOv5s, YOLOv5m, YOLOv5l and their variants with the auxiliary SE block, we find that more network layers propel higher precision and a better GIoU loss with the sacrifice of training speed.

Keywords: banknote identification, deep learning, YOLOv5, attention, SE, watermark recognition

Table of Contents

List of Figures

List of Tables

Attestation of Authorship

I hereby declare that this submission is my work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person (except where explicitly defined in the acknowledgments), nor material which to a substantial extent has been submitted for the laurel of any other degree or diploma of a university or other institution of higher learning.

Signature: Date: 01 June 2021

Acknowledgment

This research was accomplished as an indispensable segment during my study of Master of Computer and Information Science (MCIS) at Auckland University of Technology (AUT), New Zealand.

First and foremost, I especially appreciate the supervisor named Wei Qi Yan from the bottom of my heart since he can effectively answer the questions that I confronted within the experimental experience. Because of his patience, inspiration and instruction, I continuously pursue a higher target and devote myself to the completion of this report. Besides, I am inclined to thank the faculty of the university due to their assistance on the issues I met such as the selection of curriculum as well as the provision of abundant resources.

Duo Tong

Auckland, New Zealand

June 2021

Chapter 1
Introduction

This chapter can be divided into five segments: the background and motivations, research questions, contributions, objectives, and the structure of this paper.

1.1 Background and Motivation

Despite the escalating commence of electronic currency, these days, banknotes remain galore due to their indispensability in circulation, which means currency issuers have still confronted the menace of forging. With the prevalence of automated systems such as vending machines, currency recognition has become increasingly significant in a number of financial fields in terms of currency exchange centres, shopping malls, banking systems and ticket counters (Mittal & Mitta, 2018). Meanwhile, fraud techniques are becoming progressively more advanced, contributing to the plight of recognizing fake currency (Zhang & Yan, 2018). Besides, massive nations suffer from the forged currency on a large scale due to its ease of printing (Trinh, Vo, Pham, Nath & Hoang, 2020). Hence the identification of counterfeit currency has become one of the most tropical topics.

As a genre of image classification in the computer vision field, currency recognition can be defined as the process of identifying the denomination and even the authenticity of currency (Singh, Tiwari, Shukla & Pateriya, 2010). In order to effectively determine its credibility, it is necessary for banknotes to be inspected for several specialized security features involving serial number, puzzle number, the colour-changing bird, raised ink and transparent window, which are depicted in Figure1.1. There are a variety of methods to detect currency that majorly consists of image processing skills, machine learning and deep learning algorithms.

In recent years, deep learning technology has boomed in image classification and detection areas. As a genre of machine learning, they use a neural network framework consisting of multiple layers that are mainly constructed to perform classification tasks directly from sound, images and texture. Deep learning techniques exceed traditional machine learning algorithms in performance, accuracy and tuning capability, though they require much more data and training time. Another contributing factor of deep learning for being a popular technology in computation is that the complexity is increasingly declined with the enhancement of data and the layers of a neural network. There are various deep learning architectures in terms of VGG, YOLOv5, Faster R-

CNN, AlexNet and GoogleNet, which are utilized to produce learning patterns and relevance among data.

However, we would like to select a state-of-the-art algorithm to perform this task. YOLOv5 is an appropriate model because of its excellent performance on object detection, relatively agreeable precision, the first implementation on currency recognition and never synthesization with an attention mechanism.

Therefore, the focus of this research is watermark recognition of New Zealand currencies through implementing deep learning algorithms involving YOLOv5 and its variant (YOLO-SE), which synthesize the SE attention block. Noticeably, the experiment must ponder the size of the dataset to improve the precision and generalization ability. Hence we complete data augmentation containing cropping, flipping, rotation, colour modification, and noise addition.

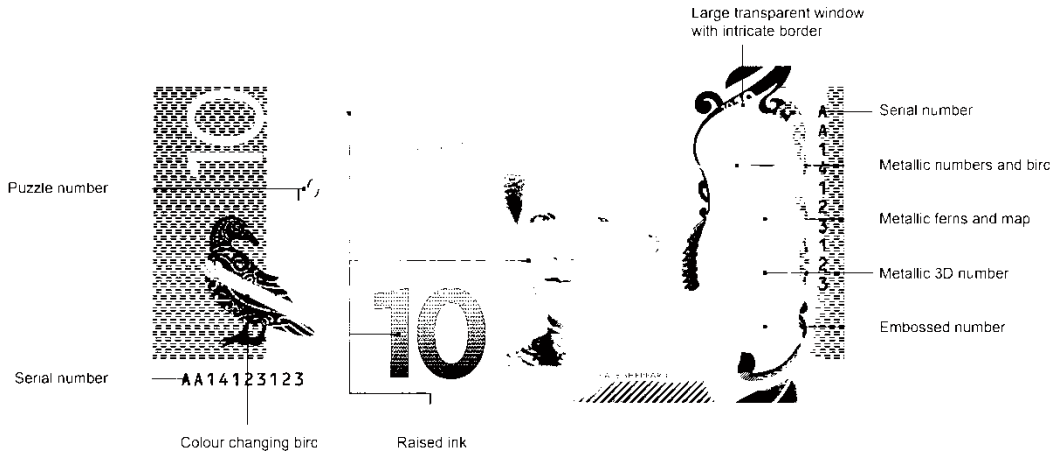

Figure 1.1 Different security features of currency.
The highlighted window is our detecting target.

1.2 Research Questions

This research work contributes to recognize transparent windows of NZD based on deep learning. The questions of this paper are,

(1) Compared to the baseline YOLOv5, how does the proposed algorithm (YOLOv5-SE) perform on the precision, recall and loss value and which one is optimal?

(2) What are the impacts of additional noises on the detection task?

(3) Can the depth of networks affect the performance by the comparison among the different versions of YOLOv5 with or without the SE module?

More specifically, the fundamental idea of this paper is watermark detection by integrating attention mechanism into YOLOv5. This project involves a number of techniques such as data labelling and augmentation. Additionally, the data set that we create is required to train to get the most gratifying outcomes. Moreover, it is so indispensable to evaluate these two models according to the performance that we can discover the optimal algorithm for this task. Finally, we would like to prove the influence of different numbers of network layers and extra noises.

1.3 Contributions

This paper aims to achieve currency identification based on the transparent window whose phases can be separated into data collection, data augmentation, denomination recognition and the analysis of the outcomes. The contributions in this research can be majorly concluded into three aspects the construction of comprehensive samples, the proposal of YOLOv5-SE and the resultant analysis.

First of all, regarding the requirement of data set in deep learning, the samples in this experiment involve the front and back sides, the changes of location, gyration, size and other augmentation skills. As a repercussion, we create a relatively full-scale data set, which is beneficial for the execution of the experiment.

Also, YOLOv5-SE, the amalgamation of YOLOv5 and Squeeze-and-Excitation (SE) attention mechanism, is proposed to identify currency watermarks. Moreover, we compare the YOLOv5 and YOLOv5-SE and analyse the likely reasons according to the experimental outcomes and the analysis from existing researches, which can effectively evaluate their advantages and disadvantages. Last but not least, we conduct complementary experiments to attest the relationship between the depth of networks and

noises and performance in this case.

1.4 Objectives of This Report

The main objective of this paper is to detect the transparent window of NZD through employing two deep learning classifiers including YOLOv5 and YOLO-SE and ascertain the optimal algorithm in this implementing scenario. The specific goals are illustrated as follows:

- To conduct a study in-depth, which assists to identify the suitable algorithm and formulate the process of denomination recognition.
- To implement two kinds of deep learning algorithms and achieve high performance on currency identification.
- To execute a comparative analysis and discern the more appropriate model based on the evaluation metrics.
- To recognize the influence of appended noises on detection.
- To verify the effect of different numbers of layers on currency identification.

Inaugurally, in order to achieve higher performance in currency identification, it is indispensable to enrich the size of the sample containing the front and back sides and the fruition of data augmentation. Secondly, we need to concentrate on the feature extraction on the basis of identifiable points to espouse the process of detection.

As for this experiment, the comparison among the selected algorithms will also be accomplished, which is beneficial to conclude the most appropriate method for the currency identification task. The complimentary experiments on the number of layers and additional noises are also crucial as they can further validate the appropriateness of the elected algorithm and its generalization capability.

1.5 Structure of This Report

The structure of this report is divided into the following five chapters:

- In Chapter2, the literature related to currency identification are reviewed. Furthermore, we discussed YOLO algorithms and attention mechanism.
- In Chapter 3, the research methodology will be introduced, which contains data

aggregation, data marking, data augmentation and data training. We will comprehensively describe the design of the experiment in this part, as well.

- In Chapter 4, the proposed models will be implemented. In addition, we will demonstrate and visualize the experimental outcomes into tables and line charts. Also, the limitations of the project will be depicted in detail.
- In Chapter 5, based on the experimental outcomes, the analysis and comparison will be accomplished.
- In Chapter 6, along with the conclusion, the future work is envisioned.

Chapter 2
Literature Review

In this chapter, we mainly present related work on currency identification. We extract the more comprehensive and novel idea through the review of the history of currency recognition in depth. We achieve the tasks by implementing diverse methods, which also provide further comprehension of currency detection and the knowledge of deep learning.

2.1 Introduction

With the continued development of technology, deep learning has become one of the most promising keys for diverse tasks relevant to computer vision containing object detection, image retrieval, and recognition (Singh et al., 2010).

In order to accomplish this topic, the main priority is that we need to elucidate the overall progress of currency identification including data construction, augmentation, training and resultant analysis. As stated in the Chapter1, deep learning techniques require considerable data and different data extensions, which must be considered to evade overfitting during the experiment. The selection of the appropriate algorithms is the second priority. The upcoming content is separated into four sections containing object detection, currency recognition, YOLO and attention mechanism.

2.2 Object Detection

Playing an essential role in the computer vision field, object detection handles detecting instances of objects from given categories in digital images and outputs the spatial location through bounding boxes (Zou, Shi, Guo & Ye, 2019). As a fundamental issue in image comprehension, it constructs the basis of diverse computer vision tasks in terms of image captioning, objects tracking and event detection (Liu et al., 2020). Object detection has become a prevalent technique supporting a number of applications such as autonomous driving, robot vision and intelligent video surveillance (Liu et al., 2020).

Generally, object detection can be split into three steps in terms of classification, detection and segmentation (Yang et al., 2020). Concerned with the comprehensive, classification is used to describe the image with a predefined class by structuring the input into a particular type of information. In contrast, detection offers the comprehension of image foreground and background. The detection model produces a list in which every object uses a data contingent to indicate the category and location of the detected target.

Segmentation is the description of images at a pixel level, which provides meaning to every pixel category. It is appropriate for scenes that require high-level comprehension.

Object detection is classified into two groups: the detection of particular instances and broad categories (Liu et al., 2020). The first genre focuses on detecting cases of a specific object in terms of the Eiffel Tower, while the other rivets the detection of some predefined object classes such as humans and cars. Previously, a variety of researchers devote themselves to detect a single category such as face, pedestrian and currencies. There are currently two representative types of detectors involving one-stage and two-stage networks, represented by YOLO and Faster-RCNN, respectively.

2.3 Currency Recognition

Throughout the pertinent historical works, there are various successful methods for currency identification containing image processing skills such as MATLAB technique, machine learning models containing SVM, K-means and HMM, and deep learning algorithms in terms of SSD and CNN.

In 1996, a BANK architecture was designed (Banknote Acceptor with Neural Kernel) relying on neural networks for currency identification as two steps: banknote recognition and verification (Frosini, Gori & Priami, 1996). For banknote perception, they adopted low-cost sensors that transmit the information from the light reflected by currency. The process of verification can be divided into dimensional check, thresholding criterion and auto authentication. The experimental results indicate that it is an effective method to detect forged banknotes. However, one of the premises is that users are not granted to perform tuning phases through the collection of banknotes, which might conspicuously bias the behaviour of machines.

In 2014, the MATLAB platform was applied as one of the most important methods for currency recognition (Alekhya, Prabha & Rao, 2014). Each picture can be firstly segmented into red, green and blue components (R1, G1 and B1), then name the

corresponding notes to authentic currency (R2, G2 and B2) and finally generating a new image to be tested by the combination of these two sets of parameters. To improve the efficiency of identifying fake banknotes, the system prefers to construct a new image based on R1, G2, B1 due to the sensibility of green components for humans. When the equivalence is more than 40% via calculating the standard deviation, the currency is regarded as the original banknote. Otherwise, it can be considered as a fake or damaged currency. The deployment of this technique in mobile systems enables citizens and laymen to effectively identify fake currency with a scanner and camera assistance, which is beneficial to decrease corruption to some extent. The reason for designing the brink value as 0.4, nonetheless, has not been explained clearly. As a sequel, the recognizing outcomes maybe not correct. For instance, the value is 0.45, so that the tested image is considered authentic, whereas it could be a fake banknote.

The recognition of Indian currency also utilized MATLAB techniques based on PCA (principal component analysis) and LBP (local binary patterns) for the objectives of training and matching separately (Gautam, 2020). They converted the images taken by the digital camera under ultraviolet light into grayscale ones. Based on image processing techniques achieved by MATLAB, the division of multiple parts by cropping and the intensity of each feature was calculated to confirm the incredibility of currency. The system acquires a high accuracy of 100% when testing the images from the dataset. Nevertheless, it fails to identify the hidden features involving latent images and watermarks. The given six features are not precisely extracted because of the variance of the size for each currency note, either.

Moreover, research focused on the conclusion of comprehensive features in Indian currency, composing twelve unique characteristics such as security ink, watermark and serial number (Upadhyaya, Shokeen & Srivastava, 2018). It is also necessary to extend the techniques taking data mining methods such as map-reduce to address this financial and economic challenge resulted from fake currency. Both machine learning and image processing are considered effective solutions for currency detection (Upadhyaya et al., 2018). Digital image processing refers to extracting features of images and acquiring

promoted quality ones by performing image operations, which can mainly dispose edge identification and feature extraction. The steps of the achievement of this technique contain the procurement of images, its enhancement in quality, segmentation, feature extraction and further analysis.

In 2015, an intelligent system aiming to recognize paper money was explored (Sarfraz, 2015). This method is entirely automatic, which means it can effectively save human labour. More specifically, based on features and correlation among instances, the paper currency recognition (PCR) system utilized Radial Basis Function Network to detect Saudi Arabian currency. In the research dealing with the mixture of noisy and normal images, the average recognition rate reached 91.5%. Nevertheless, in the proposed algorithm, the weights of the connections among the neurons fail to converge into the optimal rate. In other words, its accuracy has space to be promoted.

In the same year, another method was presented to detect currency on the basis of the frequency domain feature extraction, which applies the spatial characteristics in banknote images to accomplish the task (Shah, Vora & Mehta, 2015). The classifying process involves four phases in terms of the pre-processing for the optimal, the implementation of a two-dimensional discrete wavelet transforms, the extraction of coefficient statistical moments from the approximate efficient matrix and the utilization of serial number extraction through the deployment of OCR to detect fake currency.

Machine learning classifiers such as SVM, K-means, HMM, KNN and Bayes, are feasible methods to recognize currency. In 2015, a novel key was employed to identify paper money by applying a modular approach (Kamal, Chawla, Goel & Raman, 2015). First and foremost, the distinct features in Indian currency, including identification mark and central number, were extracted by utilizing the SURF descriptor, followed by K-means algorithm used to cluster similar characteristics. Eventually, an SVM classifier was adopted to train the dataset that achieved 95.11% accuracy.

In 2019, HMM was employed as a robust currency detection algorithm (Kamble,

Bhansali, Satalgaonkar & Alagundgi, 2019). The proposed method can be utilized to differentiate paper currency from various nations via modelling their texture features as a random process. To assess the performance of the algorithm, beyond 100 denominations from distinct countries were involved in the experiment, whose outcomes indicated 98% precision for currency detection.

Deep learning-based models have been also commonly applied in currency identification. The features of currency are extracted by utilizing Convolution neural network (CNN) algorithm under the framework of Single Shot Multi-Box Detector (SSD), which achieved 96.6% accuracy (Zhang & Yan, 2018). They successfully identified the denomination of NZD and discovered that recognition can reach the optimal performance when currencies stay clear and entirely presented before the camera with a parallel angle.

Indian currency recognition employed transfer learning where an extensive CNN pre-trained on enormous natural images is implemented to classify pictures from new classes (Mittal & Mittal, 2018). To avert the paucity of data, they prepared the dataset by preprocessing and augmentation of authentic banknotes obtained in different conditions of viewpoints, light, pose and quality. A new softmax layer on the top of the convolutional foundation of a pre-trained MobileNet was trained for multiple epochs to enhance the experimental accuracy.

Deep CNN was implemented as a feature extractor in currency identification without image processing technology and manually confirmation of the existence of security notes (Bharati & Pramanik, 2020). A pre-trained model was proposed whose architecture comprised five convolution layers followed by flattening and four dense layers. Besides, they presented the training, validation and testing accuracy that was up to 98.57%, 96.55% and 85.6%, separately. Additionally, the outcomes can be promoted through the implementation of edge detection and the input of cropped images.

CNN was applied to identify folded currency which involves angle folded, damaged and standard images (Jiao, He & Zhang, 2018). They highlighted that the precision of nine layers model is higher than the seven layers one reaching 96.46%. Nonetheless, this

experiment has some restrictions involving the lack of samples, exclusion of other national denominations and insufficient layers in CNN.

2.4 YOLO

Resulting from the necessity of being trained separately in every single component, the detection process of most approaches such as R-CNN is complicated, slow and difficult to optimize (Redmon, Divvala, Girshick & Farhadi, 2016). These restrictions motivate the emergence of YOLO algorithms, which is derived from the GoogleNet model famous for excellent performance in preceding works. Different from other models, such as two-stages algorithms that divide a picture into several parts and review it by segments, YOLO, as the name described, scans an image once to predict objects (Onyango, 2018). It regards object detection as a regression issue ranging from image pixels to the coordinates of bounding boxes as well as the probabilities of classes (Redmon et al., 2016). YOLO objectives to detect objects by precisely predicting the bounding box, including the instance and localizing it according to the bounding box coordinates.

YOLO has become one of the most popular models in object detection due to its superiority to other conventional algorithms. First and foremost, since only neural networks are required to run on a new image to predict detection at test time in lieu of a complicated pipeline, it performs extremely fast whose mean average precision (mAP) is over two times than other real-time models (Redmon & Farhadi, 2017). Second, when predicting objects, YOLO infers globally about images, which means it can view the whole image in the period of training and testing. The final advantage of YOLO is the high generalization, which is beneficial to apply in a new field or unexpected inputs.

Currently, the YOLO family has included versions from one to five. The specific development of them is introduced in Table 2.1. In 2016, Redmon et al. released a novel object detection system (YOLOv1) adopting Darknet architecture and P-ReLU activation function whose detecting speed reaching 45 frames each second. Although YOLOv1 can quickly recognize targets, its precision about localizing small objects is scant (Redmon et

al., 2016).

To improve the poor performance on recall and localizing accuracy of YOLO, in 2017, it was upgraded to YOLOv2, which utilizes Darknet-19 as the network for feature extraction and incorporates batch normalization that improves mAP and regularizes the model (Redmon & Farhadi, 2017). Differentiate from the first generation using fully connected layers to predict bounding boxes, YOLOv2 introduces an anchor mechanism and computes a superior anchor template in training sets (Redmon & Farhadi, 2017). The improvement of predicting bounding boxes benefits from the incorporation of anchor boxes in convolutional layers. Besides, YOLOv2 promotes the ability to predicting small-size objects by combining them with the fined-grain features.

However, when forwarding to deeper layers, the input is downsampled, which leads to the loss of fined-grain features in YOLOv2 (Wang & Yan, 2021). Hence it usually suffers from the issue of small object detection. This obstacle inspires the idea of using Residual networks (ResNet) that skips connections so that the propagation through deeper layers is activated under no gradient vanishing. YOLOv3 regards the fusion of YOLOv2, Darknet-53 and ResNet as the feature extractor (Redmon & Farhadi, 2018). The ascendancy of applying ResNet is that the network performance will not be deteriorated in overlaying layers and massive fine-grained features will not be lost in deeper layers since they can directly acquire information from the shallower layers (Thuan, 2021). As the other novelty of YOLOv3, the multiscale detector, which is formed by the last three residual blocks, is an effective approach for detecting small objects. It is divided into three different types including 13 × 13, 26 × 26 and 52 × 52. The small-scale detecting layer (13× 13) is responsible for large objects, while the large one (52 × 52) detects the small targets as a larger-size feature map has more details. In summary, the preservation of fine-grained features from previous layers assists large-scale detection layers in detecting small targets.

An advanced version for the YOLO family known as YOLOv4 was released, whose AR and FPS were increased by 10% and 12% on COCO datasets, respectively

(Bochkovskiy, Wang & Liao, 2020). Firstly, Distinct to YOLOv3, the fourth version applies the synthesis of Darknet-53 and dense blocks not residual ones as its backbone (CSPDarknet53). It is beneficial for maintaining features, the reuse of features, the decrease of the number of parameters, and the more efficient conservation of fine-grained features (Bochkovskiy et al., 2020). However, in order to ensure the high detection speed, YOLOv4 only updates the last convolutional block to be a dense block. Secondly, YOLOv3 utilizes Feature Pyramid Network (FPN) as its neck, where the fine-grained features take a long route to travel from low-level to high-level layers because of its top-down path in the structure (Thuan, 2021). Instead, as for the neck, YOLOv4 implements an updated Path Aggregation Network (PAN) architecture that concatenates a bottom-up augmentation path next to the original one based on FPN. This modification allows the ease of information transition and the elusion of information omission on FPN or the newly introduced (bottom-up augmentation path) features.

In 2020, Ultralytics published YOLOv5 with an architecture similar to YOLOv4, which engendered altercation in the computer vision field. The notable difference is the dissection of structural code. The model contains the backbone using Focus structure and CSP network, the neck utilizing SPP block and PANet, and the head applying YOLOv3 one with Generalized Intersection over Union (GIoU) loss. In spite of the similarity of architecture, its performance is excellent, especially for YOLOv5s.

Table 2.1 The development of YOLO algorithms

YOLO Family	Reference	Backbone	Neck	Head	Loss	Improvements
YOLOv1	(Redmon et al., 2016)	GoogleNet	/	Fc → 7x7(5+5+20)	MSE	• Direct fitness to the location of bounding boxes • Batch normalizaton • High-resolution classifier • Convolutional with anchor box • Dimension clusters • Direct location prediction • Fine-Grained Features • Multi-scale Training • Hierarchical classification
YOLOv2	(Redmon & Farhadi, 2017)	Darknet19	/	Conv → 13x13x5(5+20)	MSE	
YOLOv3	(Redmon & Farhadi, 2018)	Darknet53	FPN	Conv → 13x13x5(5+80) → 26x26x5(5+80) → 52x52x5(5+80)	MSE	• Multi-scale detector • Better network with ResNet • Binary cross-entropy loss
YOLOv4	(Bochkovskiy et al., 2020)	CSPDarknet53	SPP+PAN	Conv → 13x13x5(5+80) → 26x26x5(5+80) → 52x52x5(5+80)	CIoU	• Mosaic augmentation • Using multi-anchors for single ground truth • Eliminating grid sensitivity(sigmoid)
YOLOv5	(Ultralytics, 2020)	Focus CSP Darknet53	SPP+PAN	Conv → 13x13x5(5+80) → 26x26x5(5+80) → 52x52x5(5+80)	GIoU	• Adaptive anchor strategy • Adopting Focus and CSP structure

2.5 Attention mechanism

In human cognizance, attention acts as a significant part. For instance, individuals usually selectively concentrate on salient sections, which notably assist them to capture a better visual structure. Attention mechanism, which mimics human cognition, can not only highlight the position that we need to focus on but also present interests well (Woo, Park, Lee & Kweon, 2018). In accordance with attention targets, it can be generally classified into vanilla attention and self-attention. Since all useful information from the input sequence must be compressed into a fixed-length vector, standard encoder-decoder suffers from long sentence processing. Correspondingly, this shortcoming propels the generation of vanilla attention that syndicates the standard encoder-decoder and the capability of learning joint alignment as well as translation (Bahdanau, Cho & Bengio, 2014).

Self-attention refers to the particular attention mechanism related to different positions in a single sequence to compute the representation, which comprises source2token and token2token two types (Vaswani et al., 2017). Source2token self-attention was applied to suss the significance of every token to the gamut of sequences in the representation of sentences (Lin et al., 2017). The language translation models could fulfil the-state-of-art performance through implementing token2token self-attention (Vaswani et al., 2017).

Inspired by the achievement of self-attention in the NLP area, it has been one of the most predominant techniques in the computer vision domain. The category of attention-based algorithms application can be mainly split into the altered transformers, the integration of convolutional networks and a pure attention network.

Firstly, there are numerous researches associated with the transformer applied in image classification tasks. An image transformer, which incorporates self-attention into an autoregressive model, was proposed for image generation (Parmar et al., 2018). By diminishing a number of hand-designed elements, a detection transformer (DETR) was

introduced for end-to-end object detection (Carion et al., 2020). The vision transformer (ViT), which regards each image as a sequence of patches, implements the basic encoder with a supplementary learnable classification vector for image recognition (Dosovitskiy et al., 2020). It, nevertheless, has an impediment to pixel-level dense detection, which contributes to the production of the pyramid vision transformer (PVT) (Wang et al., 2021). PVT combines the gradual declining pyramid and spatial-reduction attention to acquire feature maps in multiple scales. The dense prediction transformer (DPT) was proposed to compensate for the drawback of omitting feature resolution as well as granularity in deeper layers for convolutional networks (Ranftl et al., 2017). It considered the vision transformer and convolutional networks as an encoder and decoder separately.

Secondly, it is prevalent for attention modules that coalesce into convolutional networks. Squeeze-and-Excitation (SE) demonstrates vast potential in advancing performance by reducing dimensionality (Hu, Shen & Sun, 2018). SE, however, considerably hoists the computational complication, which arises from the capture of dependencies across all channels. Efficient Channel Attention (ECA) is the complementary approach to solve the above issue (Wang et al., 2020). ECA can effectively shun the decline of channel dimensionality while obtaining cross-channel interaction via an extraordinary lightweight way. Both SE and ECA are channel-wise attention block unmatched with the demand of computer vision tasks as images are considered as the inputs with spatial architecture. Along with Bottleneck Attention Module (Park et al., 2018), Convolutional Block Attention Module (Woo et al., 2018) refines convolutional features through adopting channel and spatial attention.

Although the aggregate of self-attention and convolutional networks has been widely applied, it is not inextricably bonded with convolutional networks in the success of computer vision. pairwise attention (PSA) network, which is a variant of self-attention, can be an independent block for image identification models. Emanated from dot-product attention, PSA is regarded as a set of operators instead of a sequence. Its footprint is escalated or generated irregularly without effects on the number of parameters due to not

adhering stationary weights to particular positions, constant permutation, and cardinality. PSA is flexible to accommodate various functions and auxiliary inputs, and engender attention weights varying along spatial and channel dimensions. The final advantage of PSA is that computational expenditure is considerably sank because of the reduction of dimensionality by mappings.

Compared to recurrent neural networks (RNN) and convolutional neural networks (CNN), self-attention is much flexible since it can be modelled either long-range or local dependencies (Shen Zhou, Long, Jiang & Zhang, 2018). Another obvious superiority of self-attention is the ease of being facilitated due to its highly parallelizable computation (Vaswani et al., 2017). Self-attention also has several limitations because of the complexity of memory and quadratic computation. In computer vision, the input with considerable spatial dimensions further engenders the tremendous cost of global self-attention implementation.

Chapter 3
Methodology

We, in this chapter, mainly explain the specific method implemented in currency identification. From data construction to model training, each phase will be described in detail, which takes great advantage of the comprehension of the fundamental research.

3.1 Introduction

The research design is the most prominent since we will have a direction to start our work. Corresponding to the research questions and objectives, we design three experiments that are recorded in Table 3.1.

Experiments	Objectives	Models	Parameters	Dataset
Exp1	To prove the effectiveness of the proposed algorithm.	YOLOv5l & YOLOv5l-SE	Listed in Table 4.1	The generated data
Exp2	To substantiate the affect of attached noises.	YOLOv5-SE	Listed in Table 4.1	The generated data with extra nosies
Exp3	To confirm the impact of different numbers of layers.	YOLOv5s, YOLOv5s-SE, YOLOv5m, YOLOv5m-SE, YOLOv5l & YOLOv5-SE	Listed in Table 4.1	The generated data

Table 3.1 The list of experiments

To complete the task of currency detection, the secondary work we need to cogitate is the particular stages of research from data generation, training data and resultant analysis. The experimental implementation can considerably benefit from the basic idea about the specific process of currency identification.

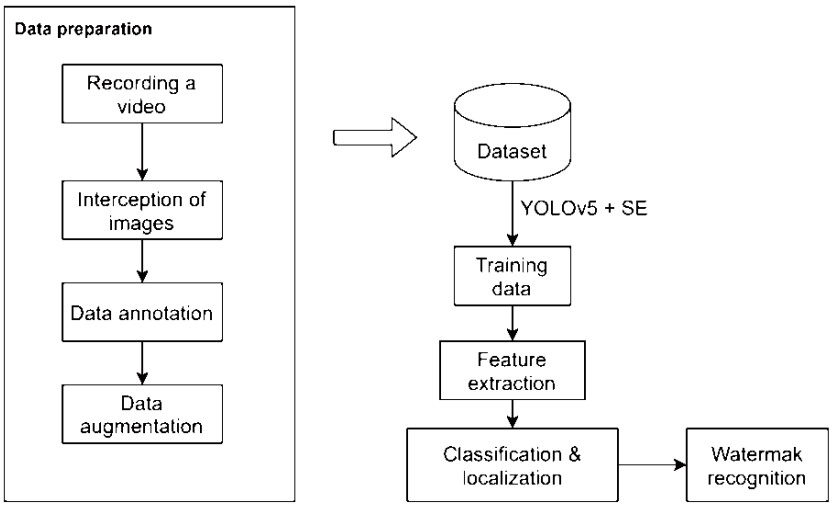

Figure 3.1 The phases of the experiment

In Figure 3.1, the first stage is the preparation of data, which can be separated into four parts, including shooting a video, the split of images by frames, label marking, and augmentation. The first two actions are used to acquire original images for the experiment. The next process is for the computers to recognize the inputs, while the final one is to address the scantiness of data and buttress the generalization capability. After that, the outcomes of data augmentation are utilized as the input of the designed model to complete

the transparent window recognition of currency, which includes training, feature extraction, dense detection, and the acquisition of the results. Specifically, the features of input images are compressed down through the backbone to complete feature extraction and next forward to detection neck and head to accomplish feature aggregation and detection, respectively. In particular, we merge localization and classification into one step since YOLOv5 is a one-stage detector in which these two operations for every bounding box are implemented simultaneously.

3.2 Data Preparation

3.2.1 Data Collection

In the beginning, we intend to find an existing dataset related to our topic in some official websites such as Kaggle, UCI Machine Learning repository and Roboflow. Nonetheless, beyond our expectations, there is no dataset about currency recognition. Then we exclude downloading images from the Internet and taking pictures through a camera since they are not convenient to acquire massive instances.

Therefore, we finally choose to manually produce a video through using the camera of the Iphone7s with the resolution of 1080 pixels at 30fps and then split it into images by each frame regarding the requirement of data volume in deep learning. The instances involve 10NZD, 50NZD and 100NZD. Each monetary denomination has front and back sides, hence we have six classes (including 10F, 10B, 50F, 50B, 100F and 100B) in this dataset. During the dataset generation, we are demanded to be careful in the following aspects to enhance the experimental precision. If images are inconsistent with these criteria, it is indispensable to doff them from the dataset.

- The images should be in high resolution.
- The currencies must be flat and the transparent window must be completely displayed in each image.
- The object should be displayed in the centre of the image and have sufficient space to ensure a complete appearance when cropping and resizing.

- The currencies should be rotated in different angles and distances between the camera and the object when shooting a video.

3.2.2 Data Labeling

Data labelling refers to the action of marking the object to be trained in the picture with a bounding box whose four coordinates are reserved in the corresponding “xml” or “txt” file (Lee, Im & Shim, 2019). The marked images demonstrate the recognizable patterns and tell machines which object they are required to detect. It plays an imperative role in deep learning as computers have no target to recognize without image annotation.

The labelling bounding box is generally classified into a rotated rectangle, straight rectangle and curved one. Regarding the similar shape between transparent windows and rectangle and the necessity of rotating currencies when shooting the video, the first type is more appropriate to the others in this case. Via rotated bounding boxes, the whole window is ably covered in complicated backgrounds and efficiently overcome the dilemma of object extraction (Liu, Wang, Weng & Yang, 2016). Furthermore, as Zhang and Yan (2018) pointed out, this way takes great advantage of marking the position as well as the state of currencies. However, the complexity will be highly promoted in practice since YOLOv5 has no parameter about rotating angle, which means the original code is required to be modified considerably.

Consequently, we utilize the annotation tool called LabelImg rather than roLabelImg, which can label the object as a straight or rotated rectangle. We can first draw a rectangle bounding box and then make adjustments to the window by dragging it. The process of adaption is beneficial to promote labelling accuracy and slump the effect of surrounding objects.

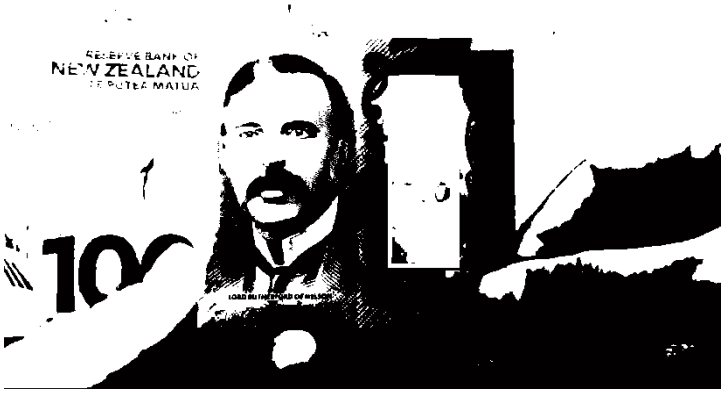

Figure 3.3 The example of data labelling

3.2.3 Data Augmentation

It is noted that the construction of efficient and robust deep learning networks covets massive high-quality data, particularly in the situation of sharing features amongst the involved classes (Rey-Area, Guirado, Tabik & Ruiz-Hidalgo, 2020). Due to the heavy reliance on big data, deep learning algorithms are able to effectively eschew overfitting, which is defined as the phenomenon when the model perfectly performs the training data while it fails to fit supplementary data (Shorten & Khoshgoftaar, 2019). Big data, unfortunately, is not accessible in some fields in terms of analysing medical images.

However, data augmentation is a feasible approach to remedy the scarcity of data and class imbalance. Data augmentation aims to bolster the variability of the original data so that the proposed model acquires higher robustness to the input images collected from various environments (Bochkovskiy et al., 2020). Data augmentation can enlarge the volume and reinforce the quality of datasets by encompassing a set of traditional transformation techniques containing geometric and photometric augmentation (Iwana & Uchida, 2021). The former refers to rotation, flipping, cropping, scaling and zooming, while the latter means colour modification such as colour jittering and manipulation, edge improvement and PCA. These two types are effective ways to prevent overfitting in deep learning. Nonetheless, Taylor and Nitschke (2017) designed a comparative experiment whose outcomes demonstrate that the geometric augmentation outstrips the photometric one in enhancing classification strength. They further illustrated that the cropping

operation plays the most critical role in this augmentation task resulting in an enhancement of 13.82% on classifying accuracy.

Furthermore, both affine image modification and colour alteration are the most common current tactics for data augmentation, which can escalate the number of instances, counterweight the size of classes and evoke experimental efficiency (Mikolajczyk & Grochowski, 2018). The performance is remarkably promoted by implementing expedient data augmentation even on complicated object detection tasks (Saeed, Li, Ozcelebi & Lukkien, 2020).

The size and colour probably affect the feature extraction when the note is either torn or dirty (Agasti, Burand, Wade & Chitra, 2017). The generation of images with diverse brightness can further improve classification accuracy (Zhang, Kinoshita & Kiya, 2020). By utilizing black-box optimization to discover the most suitable transformations including flip, mask and noise. They achieved the outperforming F-score about 1.5 to 10 points compared with the baseline. Generative Adversarial Network (GAN) was applied to learn the appropriate augmentation operations. (Zhu, Liu, Qin & Li, 2017). An effective strategy is proposed for data augmentation, divided into three phases involving traditional transformation, GAN and learning the augmentation through a neural net (Perez & Wang, 2017). They indicated that the coordination of conventional techniques and neural augmentation could further promote the precision of classification. Hence, neural augmentation is a crucial gauge to elevate the robustness of the simple, as well.

In this experiment, we espouse the traditional augmentation skills to complete the task of data preparation. The augmentation scheme we applied including flipping, rotating, cropping, colour tweak and noise injection. The details are described as follows:

- Flipping. It is achieved by using axis flipping on horizontal, vertical and diagonal direction, whilst we only involve the first two methods, which are more common and effective measures amongst them for data enhancement (Shorten & Khoshgoftaar, 2019)

- Rotation. It is completed via rotating the image on the axis from 1° to 359°. Even though the images of currencies in different angles have been included in the dataset, it is depleted to acquire outperforming results. As a consequence, we further enrich the data by rotating 15° clockwise and counter-clockwise.

- Cropping. It alters the scale of objects such as remain 50 percent pixels of the original pictures. As we exclude the scrutinization of the distances from currencies to the camera, it is indispensable to complete cropping augmentation for the robustness of features. The additional reason to implement cropping is that it is more beneficial than flipping augmentation to promote accuracy (Shorten & Khoshgoftaar, 2019).

- Colour modification. There are various approaches to accomplish colour augmentation containing brightness, contrast, saturation and hue. It plays an important role in the data augmentation of this project since the colour of the transparent window will be various at different angles and lights. The main methods we choose are the alteration of brightness and saturation.

- Noise addition. It refers to the injection of a matrix of random instances, commonly generated from a Gaussian distribution. Due to noises in the data set, the model also effectively generalizes on noisy instances. Considering that noises cover the main feature of objects will elicit the accuracy or even the detection failure, we set the extent of noises ranging from 1 to 5.

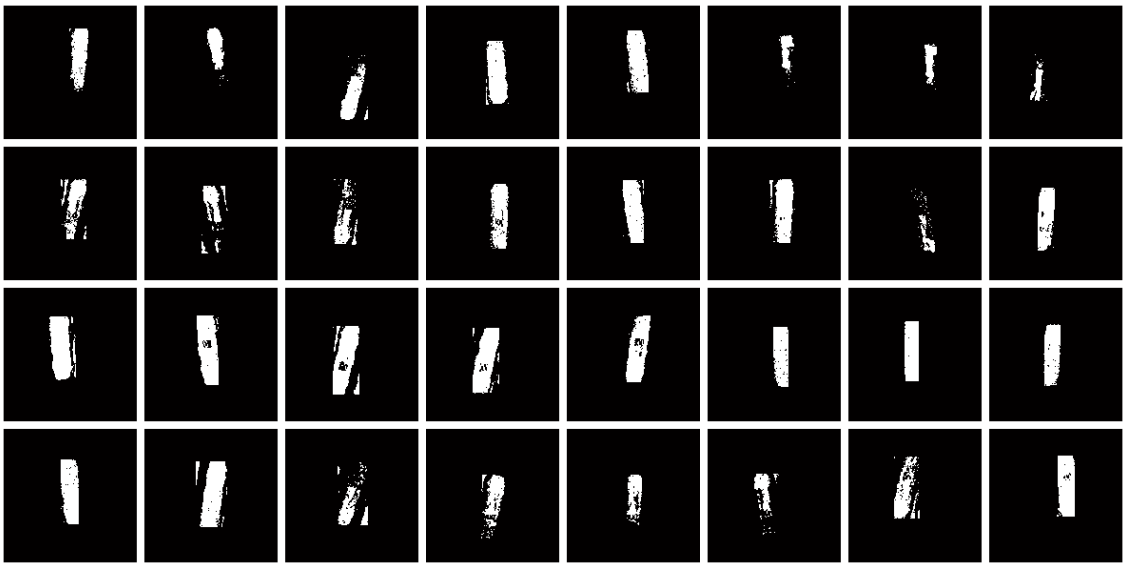

Figure 3.4 The sample of data augmentation

3.3 YOLOv5

The contributing factors for selecting YOLOv5 are the satisfying accuracy, the ability to identify different-scaled objects as well as its extremely outstanding performance. The above strengths will be explained in more detail in the next section. We, however, finally choose YOLOv5l to perform the watermark recognition task as it is less abysmal than YOLOv5x but still has relatively high accuracy and speed for detection objectives. We believe the intermediate size model is sufficient to perform watermark detection.

3.3.1 The Structure of YOLOv5

Differentiate from the previous algorithms such as faster R-CNN, YOLOv5 is a single-stage detector. The frame of YOLOv5 composes of three major components: backbone, neck and output. The only distinction among the architecture of YOLOv5 and YOLOv5-SE is the addition of the SE module. As a result, Figure 3.5, which describes the structure of the former, can convert to YOLOv5-SE ones when adding SE block between *CSP_4 and CBL_5 components.

The input image with the resolution of 640×640×3 passes through the focus module in the backbone. It firstly uses the slicing operation to become the feature map with

320×320×12. Then under 32 convolution kernels, it varies into a 320×320×32 one. Next, incorporating CSP networks into Darknet functions as the main part of the backbone, which is utilized to extract affluent informative features from inputs. It effectively and significantly reduces the duplication of gradient information in the optimization of convolutional neural networks, especially for large-scale backbones. Besides, it conjoins gradient variance and feature map to decrease the parameters and floating-point operations per second, guaranteeing inference speed and precision, whereas shrinking the model size. By altering the width and depth, we obtain four models with different parameters, famous as YOLOv5s, YOLOv5m, YOLOv5l and YOLOv5x. Another important element in the backbone is the SPP module, which can improve receptive fields and obtain different-scaled features.

As for the neck, PANet, which is constructed with a bottom-up path based on FPN structure, is responsible for acquiring feature pyramids. From top to bottom, the FPN layer transmits solid semantic features. In the converse direction, the feature pyramid transfers positional features. This design can enhance the transmission of low-level features and promotes the accuracy of locations for objects.

The head of YOLOv5, which is the same as the fourth generation, engenders feature maps with three different sizes (18×18, 36×36, 72×72) to predict targets in multi-scale including small, medium and oversized objects.

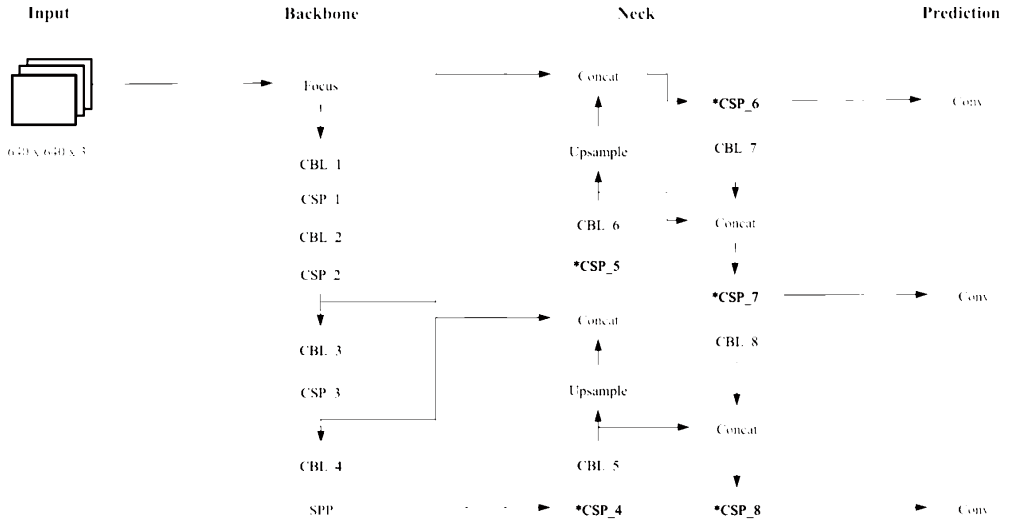

Figure 3.5 The architecture of YOLOv5

3.3.2 Loss Function

To measure loss, YOLO utilizes the sum-squared error between predictions with the highest IoU and ground truth. Its loss function consists of the localization, the confidence (also known as the objectness of boxes) and the classification loss.

As mentioned in the Chapter2, YOLOv5 adopts GIoU to be the localization loss rather than Intersection over Union (IoU). IoU is defined as eq.(3.1):

$$IoU = \frac{|A \cap B|}{|A \cup B|} \tag{3.1}$$

where A and B are two arbitrary convex shapes ($A, B \subseteq S \in R^n$).

IoU refers to the similarity among A and B. It allows the coordinates to be related to each other and has scale invariance, which overcomes the weakness of smooth L1 loss (Rezatofighi, Tsoi, Gwak, Sadeghian, Reid & Savarese, 2019). However, IoU suffers from the optimization problem when A and B have no intersection. Additionally, the IoU value fails to reflect how the prediction and target boxes intersect, assuming that their size intersection is determined and IoU values are identical (Rezatofighi et al., 2019). It is depicted in Figure 3.6:

Figure 3.6 The special case of IoU

Alternatively, GIoU is an effective key to address the above issues. It inherits the advantages of IoU in terms of the invariance of scale and all properties of loss metrics (Rezatofighi et al., 2019). In contrast to IoU, it focuses on overlapping areas and non-overlapping regions, which can better reflect the extent of intersection among A and B. The definition of GIoU is shown in eq.(3.2).

$$GIoU = IoU - \frac{|C(A \cup B)|}{|C|} \tag{3.2}$$

$$GIoU\ loss = 1 - GIoU \tag{3.3}$$

where *C* indicates the minimum encompassing convex object.

GIoU, nonetheless, still has an obvious limitation. When the prediction box (A) is the subset of the target frame (B), it degenerates to IoU and the relationship of its relative position cannot be distinguished. It is described in detail in Figure 3.7.

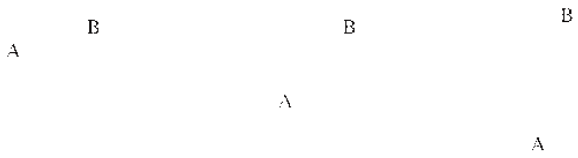

Figure 3.7 The degeneration of GIoU

In summary, the generalization preserves the main properties of IoU but addresses its drawbacks. Thus, GIoU is an appropriate replacement for IoU in computer vision tasks, even though it has some explicit restrictions.

The computation of objectness loss considers the two cases including the presence and absence of the object in the bounding box. If the predictor is responsible for the ground truth box, the object confidence is only penalized by the loss function. If an object is detected in the cell, the value of I_{ij}^{obj} is 1. Otherwise, it is 0. Oppositely, I_{ij}^{noobj} refers to no detected object. $\overline{C}_i$ depicts the confidence score of the box *j* in cell *i*. The factor ω_{noobj} is used to balance the weight with the default value of 0.5.

$$\sum_{i=0}^{S^2}\sum_{j=0}^{B} I_{ij}^{obj}\,(C_i - \overline{C}_i)^2 + \omega_{noobj}\sum_{i=0}^{S^2}\sum_{j=0}^{B} I_{ij}^{noobj}(C_i - \overline{C}_i)^2 \qquad (3.4)$$

The classification loss, which calculates the loss of class probability, is computed by the following equation where $\bar{p}_i(c)$ denotes the conditional class probability for class *c* in cell *i*.

$$\sum_{i=0}^{S^2} I_i^{obj} \sum_{c\epsilon classes}(p_i(c) - \bar{p}_i(c))^2 \qquad (3.5)$$

3.4 YOLOv5-SE

We modified the YOLOv5 model by adding a SE block after the *CSP_4 module, which is depicted in Figure 3.8. As a computational unit, the process of SE can be mainly grouped into squeeze and excitation through the operation of global average pooling and

fully connected layers, respectively. Next, it uses the self-gating mechanism (sigmoid) to limit the output of FC to the range of [0,1], and finally multiply this value as the scale to the C channels to be the input data of the next stage. The principle of this structure is to enhance the important features and cripple the unimportant ones by controlling the size of the scale so as to significantly highlight the extracted features.

Figure 3.8 The architecture of SE block in YOLOv5

Global spatial statistics can be squeezed into a channel descriptor by utilizing global average pooling to produce channel-wise information. The information $z \in R^c$, which is produced by spatial dimensions $H \times W$. U indicates the collection of statistics expressive for the full image ($H \times W \times C$). The c^{th} element of z is computed by:

$$z_c = \frac{1}{H \, x \, W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c \, (i, j) \tag{3.4}$$

Two fully connected layers achieve the squeezing process that aims to aggregate the valuable information, followed by the excitation to catch all channel-wise dependencies. The first layer, followed by ReLU, compresses C channels into $\frac{C}{r}$ (r means the percentage of compression) channels to reduce the spate of computation. According to the research conducted (Hu et al., 2018), the SE module can achieve a superior tradeoff among precision and complication when the value of r is sixteen. Hence, we utilize this value for the experiment. The second full connection is used to restore to C channels, which follows Sigmoid.

3.5 Evaluation criteria

In this research, the model evaluation metrics involve three indicators: precision (*P*), recall rate (*R*) and mAP. In object detection algorithms, both precision and recall are the two fundamental assessment criteria. Precision refers to the percentage of accurately recognized objects amongst all detected samples, whilst recall is defined as the proportion of precisely identified objects among all positive instances detected. The eq.(3.4) and eq.(3.5) are the equations for these two indexes:

$$P = \frac{TP}{FP+TP} \tag{3.4}$$

$$R = \frac{TP}{FN+TP} \tag{3.5}$$

The mAP is a comprehensive indicator that takes precision and recall rate into consideration. It is the computation of average precision (*AP*) over the number of classes (*M*). The *mAP* indicates the performance throughout all classes while *AP* demonstrates the performance on a certain category. The calculation of mAP is exhibited as eq.(3.6):

$$mAP = \frac{1}{M}\sum_{i=1}^{M} AP_i \tag{3.6}$$

where AP is defined as the region under the precision and recall curve, which is instantiated as eq. (3.7):

$$AP = \int_0^1 P(R)dR \tag{3.7}$$

Chapter 4
Results

The main purpose of this chapter is to illustrate the experimental outcomes in comparison. This part, at the terminal, will argue the restrictions of the project, as well.

4.1 Data Collection and Experimental Environment

The objective of this project is to recognize the transparent window of NZD involving the front and back sides before a camera. To guarantee the variability of the dataset, we glean 10 NZD, 50 NZD and 100NZD as templates. In the Chapter3, we mustered data through shooting a video, labelled images via Labelimg and augmented them by traditional methods.

This research is relatively complicated. At first, the background of watermarks is transparent, resulting in the light exposure on the objects, which enhances the predicament of the detection and even affects the experimental accuracy. Another cause is that the colour will be varied with diverse lights and at different angles. Envisaging these influencing aspects, when extracting source images from the video, they must be consistent with high-resolution requirements, entire and central display of transparent windows, different angles and distances from the camera. To beef up the input images in different environments, traditional augmentation approaches, especially colour modification, must be contemplated.

The augment runs on Pycharm with the version of 2019.3.3 installed on MacBook Pro with Dual-Core Intel Core i5 and the memory of 8GB, whereas the experiment is implemented on Google Colab due to its convenience to get free GPU.

Besides, the parameters for training the models are set as Table 4.1. Other parameters not presented in the table are used their default values. The learning rate contributes to the precision in deep learning algorithms. A too-large learning rate will cause large fluctuations in the precision curve when the model converges, which can engender lower accuracy. Conversely, A too-small one can immensely affect the converging speed, which means it will take a longer time on convergence. For 4530 samples, we experiment with the batch size of eight and the learning rate of 0.001 experiencing 100 epochs. In the following part, we will illustrate the specific experimental outcomes.

Table 4.1 The training parameters

Parameters	**Values**	**Applying Models**
Learning Rate	0.01	All
Epochs	100	All
Batch Size	8	All
Pixel Size	416	All
r	16	YOLOv5l-SE

4.2 Watermark Recognition

This experiment uses the original model YOLOv5l as the detector whose results are offered in Table 4.2. By utilizing three measures involving precision, recall and mAP when IoU is 0.5, we assess the performance of YOLOv5l. Each value in the evaluation metrics, fortunately, is relatively high across six classes.

Table 4.2 Experimental outcomes across all classes for YOLOv5l

Classes	**Precision**	**Recall**	**mAP@0.5**
10F	0.999	1	0.996
10B	0.999	1	0.996
50F	1	1	0.996
50B	1	1	0.996
100F	0.999	1	0.996
100B	0.998	1	0.996

The overall measure of YOLOv5l is shown in Figure 4.1. The dataset distribution for training, validation and testing are 70%, 20% and 10%, respectively. A tendency of convergence occurs in training and validating for all measures after training for 100 epochs. The terminal score converges to almost zero on GIoU loss, Objectiveness and Classification, whereas the assessment criteria related to precision approximately reaches 1.

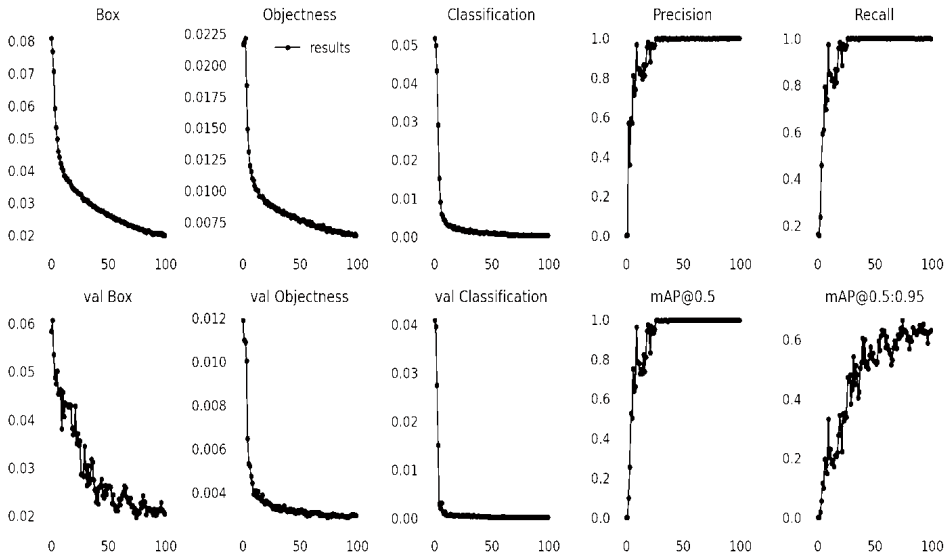

Figure 4.1 The measure of loss and precision for YOLOv5l

The SE module, in this experiment, was added into YOLOv5l, namely YOLOv5l-SE. Similarly, we execute the new algorithm based on the same evaluation metrics and identical parameters apart from the additional reduction ratio. As provided in Table 4.3, all the values almost reach the maximum value, which is significantly satisfying.

Table 4.3 Experimental outcomes across all classes for YOLOv5l-SE

Classes	**Precision**	**Recall**	**mAP@0.5**
10F	0.999	1	0.996
10B	0.999	1	0.996
50F	1	1	0.996
50B	1	1	0.996
100F	0.999	1	0.996
100B	0.999	1	0.996

In this case, we also assess the overall performance of YOLOV5l-SE by utilizing the same evaluation metrics. Its trend is quite similar to the former. The details are depicted in Figure 4.2.

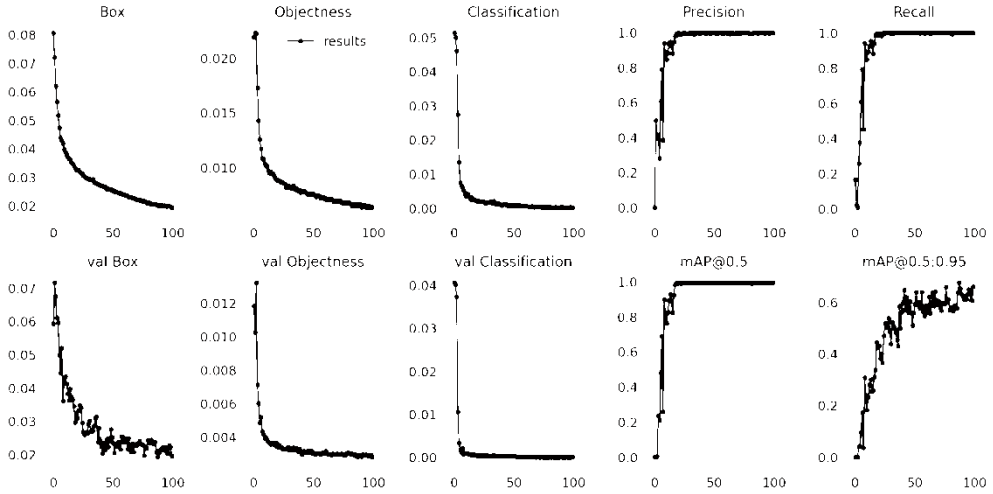

Figure 4.2 The measure of loss and precision for YOLOv5l-SE

We design the model that connects a SE layer at the end of the backbone of YOLOv5l. Consequently, it is indispensable to opt for the original YOLOv5l as the baseline in this case to complete a comparative analysis. In upcoming section, we will compare between the proposed algorithm (Yolov5l-SE) and the baseline (YOLOv5l) in three dimensions in terms of precision, recall and GIoU loss.

The trend of precision, as shown in Figure 4.3, is similar for YOLOv5l and YOLOv5-SE, which initially climb radically and then achieve steady status. First and foremost, there is not much space to improve for the novel and the original algorithms as they perform a near 100% value in the 20th and 28th epoch, separately. Additionally, as a whole, YOLOv5l-SE outperforms YOLOv5l on precision. Approximately from the eleventh epoch, the precision of the latter is higher than the original model. In the 28th epoch, they intersect and reach the perch value, which is stable within the following epochs. Finally, it is apparent that YOLOv5l-SE converges earlier than YOLOv5l since the value gets stable in advance.

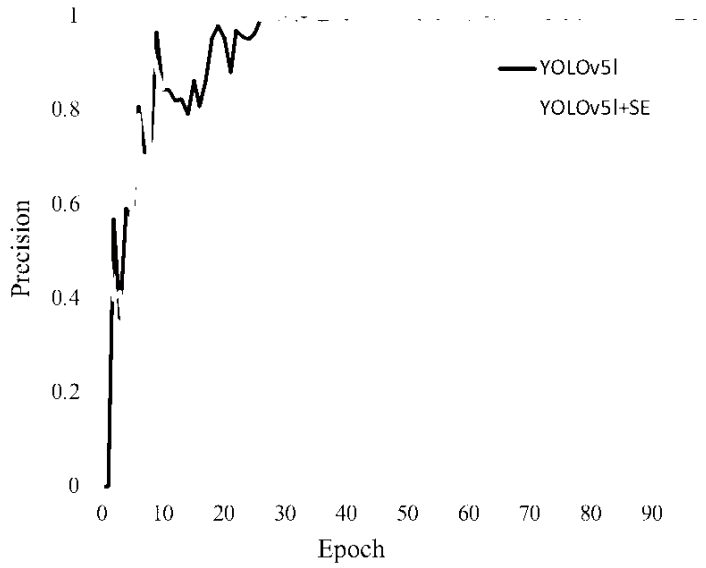

Figure 4.3 The precision in different epochs

In order to evaluate the superior model, we also compare the recall amongst them. As Figure 4.4 depicted, the gap of recall between YOLOv5l and YOLOv5l-SE is marginal. Besides, their recall values are almost identical after converging, which are close to 1. However, the convergence of YOLOv5l-SE arrives earlier than near ten epochs, which is one of the most obvious advantages.

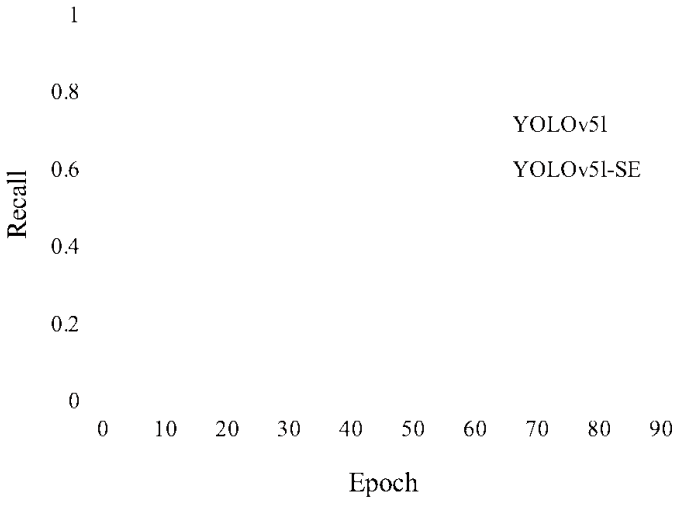

Figure 4.4 The recall in different epochs

In Figure 4.5, as a whole, both YOLOv5l and YOLOv5l-SE present a downward tendency in the loss function. More specifically, they plunge during the previous epochs and then diminish gradually until reaching their nadirs. In the terminal epoch,

the loss value performs superlatively. Last but not least, their loss rate is basically coincident, while YOLOv5l-SE marginally surpasses YOLOv5l within the thorough process.

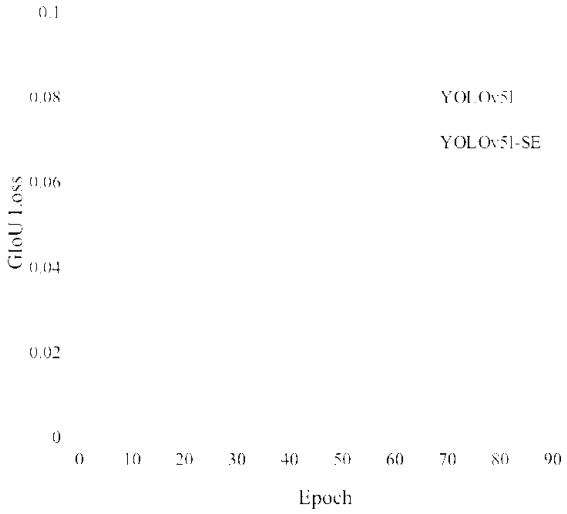

Figure 4.5 The GIoU loss in different epochs

The outcomes of watermark identification are shown in Figure 4.6. It contains six classes represented by bounding boxes with different colours. Specifically, the blue, light green, purple, pink, yellow and green ones describe 10F, 10B, 50F, 50B, 100F and 100B. For instance, 10F means the front side of transparent windows for 10NZD. Instead, 10B depicts the backside of the corresponding currencies. The confidence values behind the class names, are more or less than 0.9, ranging from 0.88 to 0.92.

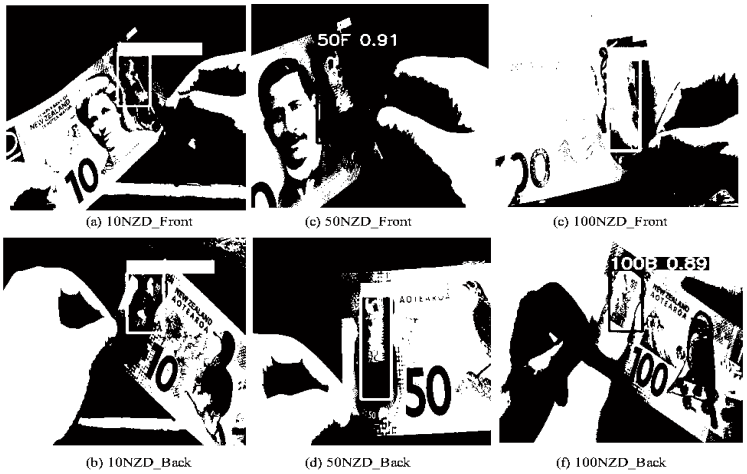

Figure 4.6 The results of currency watermark detection

Figure 4.7 illustrates the effect of noises on object detection. We test the original images and modified instances by drawing orange spots on the surface of transparent windows. On the left, it is the detecting outcome of the sample without any noise, while the right pictures indicate the results of object recognition with stains of different sizes and locations. First, we know from the figures that all samples can be correctly detected. Second, through comparing pictures (b), (c) and (d), we notice that the location of small-size noises has no impact on the confidence value of objects. Third, from the results of (e), it is apparent that the size and the number of noises lightly affect the confidence.

Figure 4.7 The internal comparison (YOLOv5-SE) of

Currency watermark detection with and without noises

As we all know, the only variance among different versions of YOLOv5 is the depth of networks. Therefore, we implemented an experiment, which is shown in Table 4.4, about YOLOv5s, YOLOv5m, YOLOv5l and their updated versions that amalgamate the SE module. The table indicates that YOLOv5l-SE obtains the outperforming outcomes on GIoU loss, accuracy, and mAP with the toll of training time. In other words, with the enhancing depth of networks, the model performs better except for training time. YOLOv5s takes the shortest time to train data compared to other models. With the

addition of SE block, each model costs a longer time on training while other evaluating results such as GIoU and accuracy become superior.

Table 4.4. The effect of the depth of networks

Model	Training Time (s)	GIoU Loss	Precision	mAP@0.5	Recall
YOLOv5s	**439.56**	0.02282	0.9972	0.9959	1
YOLOv5s-SE	461.16	0.02227	0.998	0.996	1
YOLOv5m	666.72	0.02066	0.9989	0.9959	1
YOLOv5m-SE	673.56	0.02047	0.999	0.996	1
YOLOv5l	1004.4	0.01995	0.9988	0.9958	1
YOLOv5l-SE	1100.72	**0.0195**	**0.999**	**0.9961**	1

4.3 Limitations of the Project

As for watermark recognition of New Zealand currencies, we deploy an advanced YOLOv5l by adding a SE block to the terminal of the backbone. In spite of the excellent performance with the accuracy of 99.9% and shorter training time, this experiment has several constraints that we need to promote in the future. They are discussed in details as follows:

(1) The selected domination is limited. We only involve 10NZD, 50NZD and 100NZD, while paper money in New Zealand has five types. It is necessary to collect all of the genres to enrich the dataset.

(2) We just include color modification such as lightness and saturation whereas excluding the effect of different lights in terms of ultraviolet rays and infrared lights. Watermarks will be varied in different lights environment. In the aftermath, we should extend the experiment by considering it to achieve higher generalization.

(3) At present, we restrict the location of the SE layer at the end of the backbone. Hence we have no evidence about the influence of various positions on the accuracy or other evaluation criterion. We will alter the spot of attention blocks in future experiments to obtain the optimal outcomes by comparative analysis.

(4) The attention block SE only contemplates the channel aspect but no spatial dimensions, which is a significant factor in computer vision since the input images

are envisaged spatial structures. We, in the future, promote the proposed algorithm by incorporating those attention blocks that implement both channel and spatial dimensions such as CBAM.

Chapter 5
Analysis and Discussions

We aim to analyze and compare the experimental outcomes in this chapter deeply. The comparison we accomplished among the algorithms applied for watermark recognition. We additionally discuss the optimal model and further assess their strength and weakness.

5.1 Analysis

In conclusion, the attention-based YOLOv5l exceeds the original YOLOv5l according to the evaluation criterion in precision, mAP and GIoU loss. More specifically, the precision of the algorithm with attention reaches 99.99%, while the outcome of YOLOv5l is 99.88%. Then, the mAP of the proposed algorithm achieves the maximum value of 0.999, which surpasses the old model with an elevation of 0.03%. The absolute superiority of attention is that its GIoU loss consistently outperforms YOLOv5ls'.

However, the modified model spends a longer time on computation. Particularly, with the increase of the depth of networks, the training time will be longer. Under the same circumstance, the proposed model will be more outperforming because of the deeper networks. Noticeably, due to the limited space of promotion with the accuracy of nearly the maximum value, we only experiment with the versions from YOLOv5s to YOLOv5l. Notwithstanding that we add more layers into the models, performance promotion is slight and even zero.

5.2 Discussions

In this experiment, we proposed a modified algorithm named YOLOv5-SE, which adds a SE block in YOLOv5. We further analysed the performance of these two models by comparing them. Apart from the support of the experimental outcomes, the opinions that we state are sustained by some references.

Firstly, we revealed that the addition of the SE module impels a more excellent precision, mAP and GIoU loss, even if it takes a longer time on training. For the current state-of-the-art architectures, the SE module can generate crucial performance promotion with minimal extra expenditure on computation, consistent with our experimental results that more satisfying accuracy and loss value are acquired with the sacrifice of the executing speed (Hu et al., 2018). Throughout utilizing global information, the SE attention mechanism can explicitly model dynamic and non-linear dependencies among channels (Hu et al., 2018). Thereby the new model can ease the process of learning and

dramatically hike the representative ability of the network. These are the main reasons for the superiority of the proposed algorithm. However, due to the high accuracy rate of the basic model, the attention-based algorithm has little space to improve the performance. Hence, the betterment of each evaluation criteria is not apparent.

Moreover, to authenticate the effectiveness of detecting noisy data, we contrast the testing images with diverse locations and different amounts of noise. Generally, the site cannot influence the confidence value while both the size and number of noises slightly affect the testing outcomes. This phenomenon, which is possibly caused by the noise addition in data augmentation, further endorses the generalization ability of the model. The supplementary noises are not covered the significant features and the model has learned the pattern of images with certain noises. Hence the new noises have no apparent impacts on detection. Nevertheless, the slight effect of the size and number of noises on currency recognition arises from the distinct extent of covering parts.

Finally, from the experiment of different versions of the YOLOv5 with and without the SE module, we corroborate the effect of the depth of networks and the additional SE module by keeping other parameters invariant. With the enhancing number of network layers, the precision and loss value is getting meliorated, but the speed is becoming slower. With the addition of SE, each version of YOLO acquires higher performance with the sacrifice of training time, attributing to that the deeper networks can extract more features and SE can promote the overall performance but engender more computational costs.

Chapter 6
Conclusion and Future Work

Throughout the previous sections, we have investigated some object detection algorithms and selected the optimal one to perform watermark recognition. We have also accomplished the resultant comparative analysis to answer the research questions proposed at the beginning. We will summarize the content of this research in Chapter 6. In the end, we will introduce the future direction related to watermark recognition based on the experimental outcomes and its insufficiency, as well.

6.1 Conclusion

The main objective of this research is to detect the watermark of NZD based on deep learning techniques since we integrate the SE attention module in YOLOv5 as the detector. Fortunately, the proposed model presents satisfactory outcomes with precision, recall, mAP@0.5 and GIoU loss of 0.999, 1, 0.9961 and 0.0195 in the final training epoch, respectively.

In order to further assess the performance of the modified model and its effectiveness of promotion, we compare the attention-based algorithm (YOLOv5l-SE) and the baseline (YOLOv5l) according to precision, recall, GIoU loss and training time. Supplementarily, to verify the effect of the depth of networks, in this case, we make a contrast among the different versions including YOLOv5s, YOLOv5m, YOLOv5l and their variants with attention block. Eventually, we confirm the influence of noises of different sizes and amounts in the task of watermark detection.

Throughout the corresponding operations, we conclude that the attention-based model outperforms the unmodified with a rise of 0.11%, 0.03% on precision and mAP, separately. Besides, the GIoU loss score of YOLOv5l-SE is always lower than YOLOv5l across the plenary training process. We further summarize that the confidence of watermark recognition is slightly different despite the alteration of size, amount and location of noises. Hence the extra noises have no obvious impact on the transparent window detection. The final epilogue is that the overall performance will be promoted by enhancing network layers while training time will be prolonged.

6.2 Future Work

To compensate for the limitations of this experiment, future work involves the following points:

(1) From the dataset aspect, we will collect more comprehensive currencies in terms of 5NZD and 20NZD under light with different colours.

(2) From the experimental design level, we will compare other classic object

detection models, including CNN, RNN and faster R-CNN, to further illustrate the advantages and disadvantages of the proposed algorithm.

(3) From the detector perspective, we can utilize two-stages algorithms such as faster R-CNN combing with Resnet158 or attention mechanism in terms of CBAM module.

(4) From the detecting target prospect, we can also extend the topic from denomination recognition to authenticity detection.

References

Agasti, T., Burand, G., Wade, P., & Chitra, P. (2017). Fake currency detection using image processing. *IOP Conference Series: Materials Science and Engineering*, *263*, 052047. https://doi.org/10.1088/1757-899X/263/5/052047

Alekhya, D., Prabha, G. D. S., & Rao, G. V. D. (2014). Fake currency detection using image processing and other standard methods. *International Journal of Research in Computer and Communication Technology*, *3*(1), 128-131.

Al-Sarayreh, M., Reis, M., Yan, W., Klette, R.. (2018) A deep learning approach for detecting the adulteration in red-meat products by hyperspectral imaging. In *IEEE Annual Workshop on Smart Sensors, Measurements and Instrumentation.*

Al-Sarayreh, M., Reis, M., Yan, W., Klette, R. (2019) A sequential CNN approach for foreign object detection in hyperspectral images. In *International Conference on Information, Communications and Signal Control*.

Al-Sarayreh, M., Reis, M., Yan, W., Klette, R. (2019) Deep spectral-spatial features of snapshot hyperspectral images for red-meat classification. In *International Conference on Image and Vision Computing New Zealand*.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bello, I., Zoph, B., Vaswani, A., Shlens, J., & Le, Q. V. (2019). Attention augmented convolutional networks. *In IEEE/CVF International Conference on Computer Vision* (pp. 3286-3295).

Bharati, P., & Pramanik, A. (2020). Deep Learning Techniques—R-CNN to Mask R-CNN: A Survey. *In Computational Intelligence in Pattern Recognition* (pp. 657-668).

Springer, Singapore.

Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). YOLOv4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In *European Conference on Computer Vision* (pp. 213-229). Springer.

Chambers, J., Yan, W., Garhwal, A., Kankanhalli, M. (2014) Currency security and forensics: A survey. Multimedia Tools and Applications, 74(11), 4013-4043.

Chandran, R., Yan, W. (2014) Attack graph analysis for network anti-forensics. International Journal of Digital Crime and Forensics (IJDCF) 6 (1), 28-50.

Cui, W., Yan, W. (2016) A scheme for face recognition in complex environments. International Journal of Digital Crime and Forensics (IJDCF) 8 (1), 26-36.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An image is worth 16×16 words: Transformers for image recognition at scale. *ArXiv:2010.11929 [Cs]*. http://arxiv.org/abs/2010.11929

Feng, H., Ling, H., Zou, F., Yan, W., Lu, Z. (2010) Optimal collusion attack for digital fingerprinting. In *ACM International Conference on Multimedia*, 767-770.

Feng, H., Ling, H., Zou, F., Yan, W., Lu, Z. (2012) A collusion attack optimization strategy for digital fingerprinting. ACM Transactions on Multimedia Computing, Communications, and Applications.

Feng, H., Ling, H., Zou, F., Yan, W., Sarem, M., Lu, Z. (2013) A collusion attack optimization framework toward spread-spectrum fingerprinting. Applied Soft Computing 13 (8), 3482-3493.

Frosini, A., Gori, M., & Priami, P. (1996). A neural network-based model for paper

currency recognition and verification. *IEEE Transactions on Neural Networks*, *7*(6), 1482–1490. https://doi.org/10.1109/72.548175

Garhwal, A., Yan, W. (2015) Evaluations of image degradation from multiple scan-print. International Journal of Digital Crime and Forensics (IJDCF) 7 (4), 55-65.

Garhwal, A., Yan, W., Narayanan, A. (2017) Image phylogeny for simulating multiple print-scan. In *International Conference on Image and Vision Computing New Zealand* (IVCNZ)

Garhwal, A., Yan, W. (2018) BIIGA: Bioinformatics inspired image grouping approach. Multimedia Tools and Applications.

Garhwal, A., Yan, W. (2018) BIIIA: a bioinformatics-inspired image identification approach. Multimedia Tools and Applications.

Gautam, K. (2020). Indian currency detection using image recognition technique. In *International Conference on Computer Science, Engineering and Applications (ICCSEA)*, 1–5. https://doi.org/10.1109/ICCSEA49143.2020.9132955

Gu, D., Nguyen, M., Yan, W. (2016) Cross models for twin recognition. International Journal of Digital Crime and Forensics 8 (4), 26-36.

Hassanpour, H., & Farahabadi, P. M. (2009). Using hidden Markov models for paper currency recognition. *Expert Systems with Applications, 36*(6), 10105-10111.

Hu, H., Zhang, Z., Xie, Z., & Lin, S. (2019). Local relation networks for image recognition. *In IEEE/CVF International Conference on Computer Vision* (pp. 3464-3473).

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7132-7141).

Hu, R., Yan, W. (2020) Design and implementation of visual blockchain with Merkle tree. Handbook of Research on Multimedia Cyber Security, 282-295.

Iwana, B. K., & Uchida, S. (2021). An empirical survey of data augmentation for time series classification with neural networks. http://arxiv.org/abs/2007.15951

Ji, H., Liu, Z., Yan, W., Klette, R. (2019) Early diagnosis of Alzheimer's disease based on selective kernel network with spatial attention. In *Asian Conference on Pattern Recognition* 2 (1), 503-515.

Ji, H., Yan, W., Klette, R. (2019) Early diagnosis of Alzheimer's disease using deep learning. In *ICCCV 2019*.

Jiang, H., & Learned-Miller, E. (2017). Face detection with Faster R-CNN. In *IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)* (pp. 650-657).

Jiao, Y., Weir, J., Yan, W. (2011) Flame detection in surveillance. Journal of Multimedia 6 (1).

Jiao, M., He, J., & Zhang, B. (2018). Folding paper currency recognition and research based on convolution neural network. In *International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 18–23.

Joshi, R. C., Yadav, S., & Dutta, M. K. (2020). YOLOv3 based currency detection and recognition system for visually impaired persons. In *International Conference on Contemporary Computing and Applications (IC3A)*, 280–285.

Kamal, S., Chawla, S. S., Goel, N., & Raman, B. (2015). Feature extraction and identification of Indian currency notes. In *National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG)* (pp. 1-4). IEEE.

Kamble, K., Bhansali, A., Satalgaonkar, P., & Alagundgi, S. (2019). Counterfeit currency detection using deep convolutional neural network. In *IEEE Pune Section International Conference (PuneCon)* (pp. 1-4). IEEE.

Laadjel, M., Bouridane, A., Kurugollu, F., Nibouche, O., Yan, W. (2010) Partial palmprint matching using invariant local minutiae descriptors. Transactions on Data Hiding and Multimedia Security V, 1-17

Laadjel, M., Kurugollu, F., Bouridane, A., Yan, W. (2010) Palmprint recognition based on subspace analysis of Gabor filter bank. In *Pacific-Rim Conference on Multimedia*, 719-730.

Lan, W., Dang, J., Wang, Y., & Wang, S. (2018). Pedestrian detection based on YOLO network model. In *IEEE International Conference on Mechatronics and Automation (ICMA)* (pp. 1547-1551). IEEE.

Le, R., Nguyen, M., Yan, W. (2020) Machine learning with synthetic data – A new way to learn and classify the pictorial augmented reality markers in real-time. In *International Conference on Image and Vision Computing New Zealand.*

Lee, Y., Im, D., & Shim, J. (2019). Data labeling research for deep learning based fire detection system. In *International Conference on Systems of Collaboration Big Data, Internet of Things & Security (SysCoBIoTS)*, 1–4.

Li, P., Nguyen, M., Yan, W. (2018) Rotation correction for license plate recognition. In *International Conference on Control, Automation and Robotics* (ICCAR).

Ling, H., Feng, H., Zou, F., Yan, W., Lu, Z. (2010) A novel collusion attack strategy for digital fingerprinting. In *International Workshop on Digital Watermarking*, 224-238.

Ling, H., Zou, F., Yan, W., Ma, Q., Cheng, H. (2011) Efficient image copy detection using multi-scale fingerprints. IEEE Multimedia.

Ling, H., Wang, L., Zou, F., Yan, W., (2011) Fine-search for image copy detection based on local affine-invariant descriptor and spatial dependent matching. Multimedia Tools and Applications 52 (2), 551-568.

Liu, C., Yan, W. (2020) Gait recognition using deep learning. Handbook of Research on Multimedia Cyber Security, 214-226.

Liu, J., Ling, H., Zou, F., Yan, W., Lu, Z. (2012) Digital image forensics using multi-resolution histograms. Crime Prevention Technologies and Applications for Advancing Criminal Investigation.

Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., & Pietikäinen, M. (2020). Deep learning for generic object detection: A survey. *International Journal of Computer Vision*, *128*(2), 261-318.

Liu, X., Yan, W. (2020) Vehicle-related scene segmentation using CapsNets. In *International Conference on Image and Vision Computing New Zealand.*

Liu, X., Nguyen, M., Yan, W. (2019) Vehicle-related scene understanding using deep learning. In *Asian Conference on Pattern Recognition.*

Liu, Z., Yan, W., Yang, B. (2018) Image denoising based on a CNN model. In *International Conference on Control, Automation and Robotics* (ICCAR).

Liu, Z., Wang, H., Weng, L., & Yang, Y. (2016). Ship rotated bounding box space for ship extraction From high-resolution optical satellite images with complex backgrounds. *IEEE Geoscience and Remote Sensing Letters*, *13*(8), 1074–1078.

Lu, J., Nguyen, M., Yan, W. (2020) Comparative evaluations of human behaviour recognition using deep learning. Handbook of Research on Multimedia Cyber Security, 176-189

Lu, J., Nguyen, M., Yan, W. (2020) Human behaviour recognition using deep learning. In *International Conference on Image and Vision Computing New Zealand*.

Menaka, R., Archana, N., Dhanagopal, R., & Ramesh, R. (2020). Enhanced missing object detection system using YOLO. In *International Conference on Advanced Computing and Communication Systems (ICACCS)* (pp. 1407-1411). IEEE.

Mikolajczyk, A., & Grochowski, M. (2018). Data augmentation for improving deep learning in image classification problem. In *International Interdisciplinary PhD Workshop (IIPhDW)*, 117–122.

Mittal, S. & Mittal, S. (2018). Indian banknote recognition using convolutional neural network. In *2018 3rd International Conference on Internet of Things: Smart Innovation and Usages (IoT-SIU)* (pp. 1–6).

Nguyen, M., Le, H., Yan, W., Dawda, A (2018) A vision aid for the visually impaired using commodity dual-rear-camera smartphones. In *International Conference on Mechatronics and Machine Vision*.

Niitsuma, M., Tomita, Y., Yan, W., Bell, D. (2011) Classifying Bach's handwritten C-Clefs. In *International Society for Music Information Retrieval*.

Niitsuma, M., Tomita, Y., Yan, W., Bell, D. (2018) Towards musicologist-driven mining of handwritten scores. IEEE Intelligent Systems.

Onyango, L. A. (2018). Convolutional neural network to enhance stock taking (Doctoral dissertation), University of Nairobi, Kenya.

Pan, C., Yan, W. (2018) A learning-based positive feedback in salient object detection. In *International Conference on Image and Vision Computing New Zealand*.

Pan, C., Yan, W. (2020) Object detection based on saturation of visual perception. Multimedia Tools and Applications 79 (27-28), 19925-19944.

Park, J., Woo, S., Lee, J.-Y., & Kweon, I. S. (2018). BAM: Bottleneck attention module. *ArXiv:1807.06514 [Cs]*. http://arxiv.org/abs/1807.06514

Parmar, N., Vaswani, A., Uszkoreit, J., Ukasz, K., Shazeer, N., and Ku, A. (2018) Image transformer. *arXiv:1802.05751, 2018*.

Perez, L., & Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. http://arxiv.org/abs/1712.04621

Rajan, G. V., Panicker, D. M., Chacko, N. E., Mohan, J., & Kavitha, V. K. (2018). An extensive study on currency recognition system using image processing. In *Conference on Emerging Devices and Smart Systems (ICEDSS)*, 228–230.

Ramachandran, P., Parmar, N., Vaswani, A., Bello, I., Levskaya, A., & Shlens, J. (2019). Stand-alone self-attention in vision models. *arXiv preprint arXiv:1906.05909*.

Ranftl, R., Bochkovskiy, A., & Koltun, V. (2021). Vision transformers for dense prediction. http://arxiv.org/abs/2103.13413

Redmon, J., & Farhadi, A. (2017). YOLO9000: better, faster, stronger. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 7263-7271).

Redmon, J., & Farhadi, A. (2018). YOLOv3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You Only Look Once: Unified, real-time object detection. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 779-788).

Reisa, M., Beersd, R., Al-Sarayreh, R., Shortenb, R., Yan, W., Saeysd, W. (2018) Chemometrics and hyperspectral imaging applied to assessment of chemical, textural and structural characteristics of meat. Meat Science.

Ren, Y., Nguyen, M., Yan, W. (2018) Real-time recognition of series seven New Zealand banknotes. International Journal of Digital Crime and Forensics (IJDCF) 10 (3), 50-66.

Ren, Y. (2018) Banknote Recognition in Real Time Using ANN. Masters Thesis, Auckland University of Technology, New Zealand.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*.

Rey-Area, M., Guirado, E., Tabik, S., & Ruiz-Hidalgo, J. (2020). FuCiTNet: Improving

the generalization of deep learning networks by the fusion of learned class-inherent transformations. *Information Fusion*, *63*, 188–195.

Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., & Savarese, S. (2019). Generalized intersection over union: A metric and a loss for bounding box regression. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 658-666).

Saeed, A., Li, Y., Ozcelebi, T., & Lukkien, J. (2020). Multi-sensor data augmentation for robust sensing. In *International Conference on Omni-Layer Intelligent Systems (COINS)*, 1–7.

Sarfraz, M. (2015). *An intelligent paper currency recognition system. Procedia Computer Science*, *65*, 538–545. https://doi.org/10.1016/j.procs.2015.09.128

Shah, A., Vora, K., & Mehta, J. (2015). A review paper on currency recognition system. *International Journal of Computer Applications*, *115*(20).

Shen, D., Xin, C., Nguyen, M., Yan, W. (2018) Flame detection using deep learning. In *International Conference on Control, Automation and Robotics* (ICCAR).

Shen, T., Zhou, T., Long, G., Jiang, J., & Zhang, C. (2018). Bi-Directional Block Self-Attention for Fast and Memory-Efficient Sequence Modeling. http://arxiv.org/abs/1804.00857

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, *6*(1), 60.

Singh, S., Tiwari, A., Shukla, S., & Pateriya, S. (2010). Currency recognition system using image processing.

Song, C., He, L., Yan, W., Nand, P. (2019) An improved selective facial extraction model for age estimation. In *International Conference on Image and Vision Computing New Zealand* (IVCNZ).

Taylor, L., & Nitschke, G. (2017). Improving deep learning using generic data augmentation. http://arxiv.org/abs/1708.06020

Thuan, D. (2021). Evolution of YOLO algorithm and YOLOv5: The state-of-the-art object detection algorithm. Oulu University of Applied Sciences.

Trinh, H. C., Vo, H. T., Pham, V. H., Nath, B., & Hoang, V. D. (2020). Currency recognition based on deep feature selection and classification. *In Asian Conference on Intelligent Information and Database Systems* (pp. 273-281). Springer, Singapore.

Upadhyaya, A., Shokeen, V., & Srivastava, G. (2018). Analysis of counterfeit currency detection techniques for classification model. In *International Conference on Computing Communication and Automation (ICCCA)* (pp. 1-6). IEEE.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. http://arxiv.org/abs/1706.03762

Wang, G., Wu, X., Yan, W. (2017) The state-of-the-art technology of currency identification: A comparative study. International Journal of Digital Crime and Forensics 9 (3), 58-72.

Wang, J., Yan, W.(2016) BP-neural network for plate number recognition. International Journal of Digital Crime and Forensics (IJDCF) 8 (3), 34-45.

Wang, J., Bacic, B., Yan, W. (2018) An effective method for plate number recognition. Multimedia Tools and Applications 77 (2), 1679-1692.

Wang, L., & Yan, W. Q. (2021). Tree leaves detection based on deep learning. In *International symposium on Geometry and Vision*, *CCS 1386*, Springer.

Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., & Hu, Q. (2020). ECA-Net: Efficient channel attention for deep convolutional neural networks. In *IEEE/CVF Conference on*

Computer Vision and Pattern Recognition (CVPR), 11531–11539.

Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., & Shao, L. (2021). Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. http://arxiv.org/abs/2102.12122.

Wang, X., Yan, W. (2019) Multi-perspective gait recognition based on ensemble learning. Springer Neural Computing and Applications.

Wang, X., Yan, W. (2019) Gait recognition using multichannel convolutional neural networks. Neural Computing and Applications, 1-13.

Wang, X., Girshick, R., Gupta, A., & He, K. (2018). Non-local neural networks. In *IEEE Conference on Computer Vsion and Pattern Recognition* (pp. 7794-7803).

Wang, X., Yan, W. (2019) Cross-view gait recognition through ensemble learning. Neural Computing and Applications.

Wang, X., Yan, W. (2020) Non-local gait feature extraction and human identification. Multimedia Tools and Applications.

Wang, X., Yan, W (2020) Cross-view gait recognition through ensemble learning. Neural Computing and Applications, 32 (11), 7275-7287.

Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018). CBAM: Convolutional block attention module. In *ECCV* (pp. 3-19).

Xin, C., Nguyen, M., Yan, W. (2020) Multiple flames recognition using deep learning. Handbook of Research on Multimedia Cyber Security, 296-307.

Yan, W., Chambers, J. (2012) An empirical approach for digital currency forensics. In *IEEE International Symposium on Circuits and Systems* (ISCAS), 2988-2991.

Yan, W., Chambers, J., Garhwal, A. (2014) An empirical approach for currency identification. Multimedia Tools and Applications 74 (7).

Yan, W., Kankanhalli, M. (2015) Face search in encrypted domain. In *Pacific-Rim Symposium on Image and Video Technology*, 775-790.

Yan, W. (2021) Computational Methods for Deep Learning Theoretic, Practice and Applications. Springer Nature London.

Yan, W. (2019) Introduction to Intelligent Surveillance: Data Capture, Transmission, and Analytics. Springer Nature London.

Yang, G., Feng, W., Jin, J., Lei, Q., Li, X., Gui, G., & Wang, W. (2020). Face mask recognition system with YOLOv5 based on image recognition. In *IEEE International Conference on Computer and Communications (ICCC)* (pp. 1398-1404). IEEE.

Yu, Z., Yan, W. (2020) Human action recognition using deep learning methods. In *International Conference on Image and Vision Computing New Zealand*

Zhang, L., Yan, W. (2020) Deep learning methods for virus identification from digital images. In *International Conference on Image and Vision Computing New Zealand*.

Zhang, Q., & Yan, W. Q. (2018). Currency detection and recognition based on deep learning. In *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)* (pp. 1-6).

Zhang, Q. (2018) Currency Recognition Using Deep Learning. Masters Thesis, Auckland University of Technology, New Zealand.

Zhang, Q., Yan, W. (2018) Currency detection and recognition based on deep learning. In IEEE AVSS.

Zhang, Q., Yan, W., Kankanhalli, M. (2019) Overview of currency recognition using deep learning. Journal of Banking and Financial Technology 3 (1), 59–69.

Zhang, W., Kinoshita, Y., & Kiya, H. (2020). Image-enhancement-based data

augmentation for improving deep learning in image classification problem. In *IEEE International Conference on Consumer Electronics - Taiwan (ICCE-Taiwan)*, 1–2.

Zhu, X., Liu, Y., Qin, Z., & Li, J. (2017). Data augmentation in emotion classification using generative adversarial networks. http://arxiv.org/abs/1711.00648

Zou, Z., Shi, Z., Guo, Y., & Ye, J. (2019). Object detection in 20 years: A survey. *arXiv preprint arXiv:1905.05055*.