Improving Public Safety Through Automatic Firearm Detection and Categorization: Leveraging Transfer Learning and Mask R-CNN for Accurate Handgun Identification and Classification

by

Ritisha Islam
20201222
Faraz Sikder Shafin
21101171
MD Mahmudul Hasan
19201026
Shafin Sayeed
19301159

A thesis submitted to the School of Data and Sciences
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
Brac University
August 2025

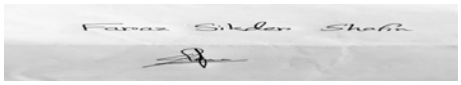
Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Ritisha Islam
20201222



Faraz Sikder Shafin
21101171

MD Mahmudul Hasan
19201026

Shafin Sayeed
19301159

Approval

The thesis/project titled "Improving Public Safety Through Automatic Firearm Detection and Categorization: Leveraging Transfer Learning and Mask R-CNN for Accurate Handgun Identification and Classification" submitted by

1. Ritisha Islam (20201222)
2. Faraz Sikder Shafin (21101171)
3. MD Mahmudul Hasan (19201026)
4. Shafin Sayeed (19301159)

Of Spring, 2025, has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc in Computer Science on June 20, 2025.

Examining Committee:

Supervisor:
(Member)

MD. Tanzim Reza
Lecturer
Department of Computer Science and Engineering
Brac University

Program Coordinator:
(Member)

Md. Golam Rabiul Alam, PhD
Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi
Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

The new emerging portable weapons, such as firearms, pistols, and revolvers, used in the commission of a crime, highlight the need for improved surveillance systems that can assist in recognizing such behavioral patterns to discourage crime. In most cases, surveillance is done manually, which creates many opportunities for mistakes and consumes much time. The subtype of AI, particularly in object recognition, classification, and picture segmentation, known as deep learning, can offer a potential solution in weapon detection. All these tasks can now be solved using Convolutional Neural Networks (CNNs); some of the most outstanding models in the market today, such as Faster R-CNN, YOLO, and Mask R-CNN, are accurate and real-time. In our study, Mask R-CNN demonstrated the highest accuracy and stability in handgun detection, even under challenging conditions. The model achieved an IOU of 0.7901 and a Dice score of 0.8828, highlighting its effectiveness in accurately identifying firearms. Compared to Faster R-CNN and YOLO, it showed superior performance in segmenting objects with greater precision. The evaluation metrics, including Precision, Recall, and F1-Score, further confirmed the effectiveness of Mask R-CNN, with an F1-Score of 1.0 in the detection task. All the above-mentioned deep learning models are explained in this paper, with a special focus on the results of the Mask R-CNN model as a portable firearms detector. The results also suggest improvements in transfer learning and fine-tuning pre-trained architectures to optimize weapon detection. Thus, the goal of the present work is to suggest an enhanced automatic firearm detection system that will increase security in public areas by nearly eliminating the human factor. This paper employs a methodical approach to the development and overall evaluation of the system, providing practical recommendations for creating weapon recognition systems in the real world.

Keywords: Regional Convolutional Neural Networks (RCNN); YOLO; ResNet50; Mask RCNN; Weapon Detection; Transfer Learning

Acknowledgement

We begin by offering our heartfelt gratitude to the Almighty for guiding us through this thesis journey without significant disruptions.
Our sincere appreciation goes to our esteemed supervisor, Md. Tanzim Reza sir for his unwavering support and invaluable guidance that have been pivotal in shaping our research.
We would also like to extend our thanks to our parents, whose boundless encouragement and prayers have been the driving force behind our academic accomplishments. This thesis would not have reached its completion without the collective support and belief of these individuals, for which we are deeply thankful.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

The high incidence of offences using portable firearms such as guns, pistols, and revolvers is the reason why only proper surveillance tools are mandatory. They can fight crimes and social crimes since the antisocial conduct of individuals can be detected early, and security forces can take proper measures. It cannot go unnoticed that surveillance systems have always needed an actual or a virtual authority to oversee or monitor, and this may have to be much more efficient and less prone to errors. Object recognition, categorization, and image segmentation breakthroughs in deep learning provide an enormous avenue to design automated systems with no human interference to detect concealed portable guns from Indian weaponries. Convolutional Neural Networks or CNNs have surpassed most of the previous image processing tasks such as segmentation, classification, detection, and many more in numerous applications. CNNs have been of great help in the development of deep learning where they have achieved incredible accuracy in object recognition problems with the help of hierarchical features extracted from images. Faster R-CNN and YOLO models increase the rate of detection while Mask R-CNN increases the level of accuracy, making them suitable for real-time applications. [15]
Faster R-CNN is an improvement of the R-CNN and Fast R-CNN, where instead of the two networks, there is a Region Proposal Network (RPN) that is connected to full image convolutional features, which enhances the ability to detect objects and speed. This model stands out in object recognition since it successfully introduces and improves the candidates of region of interest (ROIs) by consecutive convolutional layers. [10]
Another incredible methodology is called YOLO (You Only Look Once) which reformulates the object recognition problem as a one-regression problem where test methods bound boxes and class probabilities from whole images. However, YOLO's processing time is advantageous for applications requiring rapid identification, which means that it offers less accurate detection as compared to the region-based methods. [22]
The proposed method Mask R-CNN, adds a new branch to predict the segmentation masks of each ROI along with the existing branches of class prediction and bounding box regression present in Faster R-CNN. This architecture can recognize items precisely and determine the shapes of objects; thus, it is suitable for detecting weapons in a tense scene. [13]
It is for this reason that complex deep learning architectures are being incorporated in surveillance systems in a bid to enhance the detection skills in real time situations while at the same time enhancing their robustness. This paper gives an overview of the methods and specific procedures of multiple deep learning models for weapon detection, especially the Mask R-CNN model. The chosen Mask R-CNN model once trained on our selected

dataset proved to be particularly better off in terms of accuracy as well as the ability to handle challenging scenarios such as low-quality image and noisy environment. [13]
In addition, by using the transfer learning and fine-tuning method on pre-trained model with different datasets it is possible to achieve a noticeably higher level of efficiency in weapons detection. When large-scale and diverse datasets such as ImageNet or COCO can be employed, better models are trained for identifying weapons and there are fewer instances of false positives, or misidentifications of objects as weapons, as the models are tested in diverse scenarios. [26]
This research is based on the literature review of the pioneer research works done in the field. Being based on 3-D CNNs and R-CNNs, Moez Baccouche et al. 's [3] work on human action recognition in videos elucidates the future of video action sequence detection, showing that the latter can be enhanced for this purpose. Samir K. Bandyopadhyay et al. **bandyopadhyay2013conce**. Also, the collected data serves to enhance RCNN-based detection methods following the study performed by Olmos, Tabik, and Herrera [17] on automatic handgun detection using deep learning from surveillance videos. On the MediaEval 2012 affect task [4] which focuses on the detection of violent scenes, one can get preliminary experience of the methods of detecting and analyzing violent behaviors, which can be useful for the creation of a complex surveillance system.

This undertaking seeks to develop the latest portable gun detection system that is developed using deep learning models to determine the existence of firearms in real time using high precision and reliability. By crediting their promising results to superstition or token efforts, this approach runs the risk of undermining surveillance systems' effectiveness in responding to a comprehensive summary of the goals, technical advances, and research supporting this type of work. The introduction prepares the reader for the detailed presentation of the methodology and results of the work by positioning the weapon detection system in relation to prior studies and advances in technology. The aim of this study is the development of new, efficient methods for surveillance using modern deep learning algorithms. Such systems are capable of recognizing weapons with a good degree of precision and at a very rapid rate, therefore limiting leniency and people's errors.

1.1 Problem Statement

The prevalent issue that this research aims to solve is the development of an effective automatic recognition system to identify portable guns from surveillance pictures and videos. The older methods like infrared imaging or personal devices for weapon detection, have a number of drawbacks. These are expensive systems that involve the use of special instruments, are dependent on personnel, and lack optimum functionality in various conditions; thus, not suitable for mass use.
In addition, present approaches are accompanied by significant challenges regarding false positives (non-firearm items are identified as weapons) and false negatives (real firearms are undetected). The problems of identification are made worse by having low image quality, shadows and having noise in the image. The proposal for this project is to develop a system that can detect weapons in real-time, which should minimize errors as the models employed for identification are designed to have a very high accuracy.
The problems in weapon detection with traditional technologies include the utilization of either infrared imaging or manual supervision which comes with many drawbacks that

make them ineffective and cannot be scaled up. The above details of the traditional models are summarized in Table 1.1 below where the limitations are elaborated.

Limitation	**Description**	**Ref**
Cost and equipment requirement	Some other conventional weapon detecting tools in gages include elevated prices and need equipment like infrared digimatic cameras and metal detectors. Such systems are used commonly in areas like airports and governmental buildings since they are expensive and complicated to implement in areas like streets schools, and malls.	[6]
Reliance on Human Operators	Manual monitoring is still an inherent part of many of today's working video surveillance systems. It means that to filter the video stream, human operators are used who are to identify suspicious individuals. However, this involves a lot of effort, it is tiresome, and since it involves human effort, there might be times that the 'window' might be blind to certain objects or things. Monitoring effectiveness so much depends on the operators' attentiveness and cooperation, which is quite improbable.	[21]
Effectiveness in Diverse Environments	Often, environmental factors compromised the efficiency of conventional approaches to weapon identification. For instance, infrared imaging may be affected by heat sources and reflecting surfaces, hence the inaccuracy is inevitable. In addition, these systems are usually created for a controlled environment, so in some cases these systems can not work stably in different environments with different lighting conditions, with objects in the background and weather conditions.	[6]
False Positives	Perhaps, the most challenging factor when developing an automatic weapon identification system is that it is likely to produce high false positive rates. False positives literally means cases where wrong items are identified to be firearms hence some items in the list may not be guns. As previously discussed, this flaw may allow for redundant notifications and erode trust in the monitoring system. Contamination of weapons with objects of similar shape and color or false positives as a result of changes in object appearance as a function of object orientation and lighting conditions also lead to false positives.	[17]
Continued on next page		

Table 1.1 – continued from previous page

Limitation	Description	Ref
False Negatives	False negatives arise when the system fails to identify legitimate guns. False negatives constitute a significant risk since they allow threats to pass undetected, potentially leading to deadly situations. Common causes of false negatives include poor image quality, occlusions, and noise. Maintaining security requires ensuring that the system can reliably detect guns in the face of these hurdles.	[17]
Variability in Image Quality	Surveillance cameras at times give pictures that are of variable quality due to challenges like low picture clarity, motion blurring as well as fluctuations in light. These quality difficulties complicate the problem of weapon detection as models need to be robust to many picture circumstances yet precise.	[17]
Noise and Environmental Factors	Due to the nature of detection models, image noise, for instance, random fluctuations in pixel intensity, can notably affect model performance. As mentioned, the impact of noise or other environmental factors can be seen when it comes to segmentation and feature detection, which is relevant for the chosen task. Environmental issues such as shadows, reflection and clutter background are other factors which hinder accurate detection of guns. To minimize the impacts of these issues, there is a need to apply efficient preprocessing methods and construct robust classifiers.	[20]

1.2 Research Objective

The topic of automatic pistol detection systems has become one of the pressing issues of concern for increasing the level of safety and security in society. In light of the developments of the surveillance equipment, there is now a call for high-quality real-time surveillance weapon identification systems from the surveillance photos and videos. This paper shall seek to meet this requirement by offering a set of objectives intending to advance the current state of the art of pistol identification systems.[22].

The mentioned aims of the study are associated with many aspects of system creation, assessment and optimisation. Here are the goals: the development of the automatic handgun detection system based on deeper reinforcement deep learning models, decrease of numbers of false notifications in order to increase the reliability of the system, the comparison of several deep learning approaches to the determination of the currently most successful strategy. This project also seeks to develop and apply a large dataset for model training as well as applying a set of preprocessing methods for enhanced model performance.

Through these objectives, this study will be useful in the advancement of automated firearm identification technology, and enhance the possibilities of reducing existing or future risks to the public. This academic work employs an effective procedure for de-

veloping and evaluating a system in order to provide key solutions and strategies for applying reliable handgun detection systems in actual scenarios. The main objectives of the research are stated below:The main objectives of the research are stated below:

- **Develop an Automatic Handheld Gun Detection System:** This goal is to design an advanced image and video analysis system that will be able to identify portable firearms in real-time based on the surveillance photos and video feeds. The objective of achieving this aim will require the use of improved deep learning models because they have the ability of analyzing such complex visual information and determining the objects.

- **Minimize False Positives:** False positives are the main issue in automated gun detection systems due to the alerts that they produce and how they reduce the credibility of the system. This goal aims at the enhancement of the detection procedure so that the system only warns the authorities when there is a real threat established. This goal will be essential for the practical application of surveillance systems in many environments to enhance security while avoiding the unnecessary disruption of people's lives.

- **Evaluate Deep Learning Models:** This aim requires comparing the performance of multiple deep learning models including Faster R-CNN, YOLO, and Mask R-CNN. These criteria will include accuracy, speed, and the picture noise and distortion immunities of the given unit. The specific direction of dealing with handheld firearms can be carried out with the success of the automatic detection process in which the researcher can select the most suitable model by conducting performance evaluations to decide on the best model according to specific factors this include efficiency during real world circumstances.

- **Build and Utilize a Comprehensive Dataset:** The use of a large number of photos as well as their creation and subsequent annotation is an extremely significant step when it comes to both training and validation of deep learning models. This goal draws emphasis on the fact that in order to achieve the goal, big and diverse datasets covering different settings, lighting conditions, and types/brands of weapons should be collected. Employing such a dataset ensures that the trained models are capable of identifying the weapons in diverse environments, increasing the chances of success and dependability of the models in real life situations.

- **Enhance Model Performance with Preprocessing Techniques:** Some of the likely steps include image preprocessing as a way of optimizing on the detection models in a given network. This goal includes the augmentation step in addition to subsequent common steps of resizing, rotation of the image, zooming of the image, and noise removal. False negatives could also be minimized so that the algorithms were made to have a higher detection rate while at the same time promising that the input data would be of high quality and collected in a uniform way in order to allow the researchers to identify handguns even during adverse conditions.

With regard to the above outlined research objectives, we aim at finding the best automatically detection handgun methods, hence enhancing security and safety in our society.

1.3 Ethical and Privacy Considerations

They include aspects of private life and ethics especially with regard to collection and watching, in this case the watching of automated weapon detection systems. Another moral concern of surveillance technologies was identified by Tavani and Moor [1]; however, the authors insisted on the fact that the laws and rules in question should be open for the people. Until recendly the ethical implications of artificial intelligence and deep learning were revealed by Brundage et al. [15] addressing at the same time the need for the creation of standards for the proper usage of this tool.

Chapter 2

Literature Review

Previous research used to establish the basis of the study on object detection and deep learning are as follows. Sequential deep learning was investigated by Baccouche et al. for human action recognition where the authors justified the application of 3-D CNNs and R-CNNs [3]. Bandyopadhyay, Datta, and Roy worked on concealed weapon detection using the infrared data and stressed the importance of the flexible detection frameworks [6]. Tabik, Olmos, and Herrera applied an RCNN model for detecting handguns in surveillance videos; video preprocessing was highlighted as an influential factor for the result [17]. Violent behaviors were analyzed by Demarty et al. who chaired the MediaEval 2012 affect task on violent scene identification. Besides, the state-of-art in object detection and image segmentation, and the issues regarding real-time gun detection outlined in the literature have led to development of more precise and accurate systems employing such models as Faster R-CNN, YOLO, and Mask R-CNN [27].

2.1 Background

Progress in the core structures and pioneering works has significantly influenced the creation of deep-learning methods aimed at the detection of objects with a focus on surveillance systems. In this section, authors look at the initial research and architectures that form the foundation of today's portable gun detection systems.

2.1.1 Deep Learning for Surveillance

It is for this reason that the article under discussion presents an Overview of the topic. It has been reviewed by Zhao et al [21] that for surveillance purposes deep learning techniques including anomaly activity recognition, action recognition, and object detection were advanced. Their survey illustrated that improving the surveillance systems' effectiveness, hybrid models, CNNs, and RNNs are crucial. Training data that are extensive and labeled are important in an investigation since the study revealed the benchmarks and datasets for evaluating surveillance models.

2.1.2 Convolutional Neural Networks (CNNs)

CNNs were a significant improvement in the areas of computer vision and image analysis. CNNs employ quite a few layers of convolutions and pooling to derive even more complex features from images as a result of employing the concept of hierarchical feature learning.

Primarily, due to their structure, CNNs become highly effective for the tasks such as object recognition, segmentation, and image categorization.
AlexNet by Krizhevsky et al. [5] is one of the earliest and efficient examples of the CNN approach to the image classification task and the possibilities of deep learning in the framework of the mentioned dataset. AlexNet has contained multiple convolutional layers, max-pooling layers, and fully connected layers; the setup that accelerated the development of deeper and more complex networks.
The regional approach the deep convolutional neural network-based object-by-Region Greater (or R-CNN) proposed by Girshick et al. [7] was a major shift that integrated proposal generation and CNN features for object detection. Three steps make up the R-CNN framework: generating region proposals, applying a CNN to obtain features of the regions, and applying linear SVM to classify the regions. Closely related to this, R-CNN was computationally expensive because it performed additional CNN computations for overlapping areas, although it offered high accuracy.

2.1.3 Fast R-CNN and Faster R-CNN

Girshick [10] proposed Fast R-CNN which not only improved in speed and accuracy but rectified the problems of R-CNN as well. It became possible by Fast R-CNN to use the shared convolutional feature map for all region suggestions. Next, ROI pooling was computed, then the bounding box regression and the classification came next. Thus, this facilitation empowers end-to-end training while at the same time drastically lessening the computational load.
Later, Ren et al. [11] introduced Faster R-CNN that integrated an RPN with the detection network to improve the process. Employing the same convolutional layers as in the case of detection network, the RPN generates region proposals directly in the convolutional feature maps. This integration enabled coordinated training for the region proposal, which improved accuracy in the same time boosting the detection flow.
He et al. [13] extended the Faster R-CNN architecture by adding to each ROI a new branch responsible for the prediction of the segmentation mask alongside with the existing branches for the prediction of the bounding box regression and of the classification. This new feature makes Mask R-CNN capable of detecting objects and perform pixel-level segmentation, a characteristic that makes it efficient for workloads that need fine lines to be drawn around the defined objects.
Intuitively, the last algorithm is called You Only Look Once (YOLO). Yadollah Redmon et al. [12] proposed YOLO, which changed the direction of object detection by defining it simply as a regression problem. While other methods try to divide an image by regions, YOLO forms a grid from the original image and uses the result of a single forward pass to predict bounding boxes along with class probabilities for each cell of the grid. Real time object detection is achieved through that approach, but at the expense of the degree of accuracy. In the YOLOv2 [14], the architecture was modified and training methods were improved to achieve further increase in the speed along with the accuracy. YOLOv3 [18] was further developed by increasing the speed and accuracy of the object detection algorithms.
So, transfer learning has become one of the essential methods in the use of deep learning, especially when the amount of data is limited. This kind of feature learning tends to be optimized, and to further improve it as well as to maximize the performance of models, they can be fine-tuned from previous large-scale models such as ImageNet (Deng et al.

[2]) and COCO (Lin et al. [9]). This method proves to be rather beneficial for weapon detection, where there can often be a lack of richly annotated databases.

2.2 Related Work

Hence, by improving the detection methods and implementing them in practical scenarios as much as possible, the subject of automatic weapon detection by deep learning has advanced notably. This section reviews the latest research studies in portable gun detection, and segregation based on the methods used, datasets, and improvements made.

2.2.1 Sequential Deep Learning for Human Action Recognition

The feasibility of using 3-D CNN and R-CNN in the assessment of video action sequences was also established by Moez Baccouche et al. [3], in a study on the application of sequential deep learning models for human action detection. The principles of sequential deep learning and 3-D convolutions and their usage for weapon detection from video feed data was beyond the scope of work of the main focus of the paper, though can be applicable for it as a method of observing human actions. They specifically focus on the aspect of time as a temporal context necessary for improving the capability of detection.

2.2.2 Concealed Weapon Detection Using Infrared Imaging

In particular, Samir K. Bandyopadhyay et al. [6] analyzed the challenges and limitations when it comes to detecting concealed weapons with the help of infrared imaging. Their study also uncovered the significant with of regular techniques that are expensive, require specific equipments and are sensitive to external factors such as heat sources and reflective surfaces. This work creates the need for the development of novel techniques from deep learning by noting that such methods should be more malleable and affordable.

2.2.3 Automatic Handgun Detection in Surveillance Videos

Therefore, it is possible to introduce Automatic Handgun Detection in Videos and Images from Surveillance Cameras to indicate specific regions of a scene that contain a handgun. For instance, Olmos, Tabik, and Herrera [17] proposed the application of deep learning for the automatic recognition of handguns in vigilance videos. For training the model, they used a dataset of security camera footage based upon Faster R-CNN. The investigation demonstrated the utilization of the deep learning models is more rapid and precise than conventional methodologies, especially in diverse and challenging and strenuous settings.

2.2.4 MediaEval 2012 Affect Task on Violent Scenes Detection

Due to its' investigation of violent behaviours to be performed in multimedia content the MediaEval 2012 Affect Task provided valuable lessons for the actual problem of violence detection in surveillance systems. If identifying violent scenes was the main objective, then the methods and databases for achieving this purpose do relate a lot to weapon detection too, particularly in understanding the background and movement of violent events. [4]

2.2.5 Advances in Real-Time Object Detection

Real-time object detection has recently influenced the field of firearm detection very much. EfficientDet proposed by Lin et al. [26] is an efficient and scalable object detection model which employs compound scaling which is a combination of the scaled-yolo-v3 model and the efficientnet architecture. Due to the effectiveness of this method it is suitable for use in real time applications especially surveillance.

2.2.6 Enhancing Detection with Data Augmentation and Preprocessing

It has been demonstrated that applying preprocessing and data augmentation methods enhances the effectiveness of deep learning models in a range of detection tasks. In order to improve model generalization, Shorten and Khoshgoftaar [20] examined data augmentation techniques as geometric transformations, color space augmentations, and noise injection. These methods are especially pertinent to weapon identification, as model robustness can be greatly impacted by the variety of training data.

2.2.7 Transfer Learning for Specific Applications

In the present research, it has been established that extending the analysis of the results of deep learning to the application of preprocessing and data augmentation methods improves the results obtained in diverse detection tasks. To enhance the capability of generalization, Shorten and Khoshgoftaar [20] discussed the data augmentation methods including those related to geometric transforms, color space augmentations, and adding noises. These methods are particularly relevant to the weapon recognition problem since the model's generalization often depends on the type of data used in training.

2.2.8 Recent Developments in Portable Gun Detection and Custom Datasets for Weapon Detection

If TL is used for a specific application or domain then things are slightly different. Several research work has been done on the application of transfer learning to some detection tasks like the weapon detection especially with the use of pre-trained CNNs. Kornblith et al. [19] provided quantitative evidence of the improvement that can be obtained form fine-tuning on specific-domain datasets by employing pre-trained models. Weapon detection is one of the great advantages of this method because this method allows the perception of firearms using precise models trained on large sets such as ImageNet.

2.2.9 Real-Time Detection with the Help of YOLO Algorithms YOLOv4 and YOLOv5

There is also the You Only Look Once (YOLO) model family which has been diversification further; YOLOv4 (Bochkovskiy et al. [22]) and YOLOv5 (Jocher [24]) introduced new improvements to the detection speed and the detection accuracy. Thus, incorporating the developments such as CSPDarknet53, Mish activation, and SPP-block, YOLOv4 enhanced the possibility of objects' detection in the real-time and accurately. YOLOv5 developed by Ultralytics which lays emphasis on speed, modelling, and simplicity of the

model is often used for real-time use cases. The other two deep learning architectural designs that can be combined with CNNs to form more complex models that may enhance detection robustness and precision are the recurrent neural networks (RNNs) and the transformer based models. For instance, Liu et al. [25] proposed a real-time weapon recognition system in surveillance videos using the hybrid of CNNs and BLSTMs. Besides, due to the usage of temporal context, this method improves the detection of obscured or partially occluded guns.

2.2.10 Adversarial Training and Robustness

To enhance the detection models to be more immune to such adversarial perturbations, adversarial training methods have been applied. The concept of adversarial examples – or in other words little well constructed perturbations that might lead to significant misclassifications was first introduced by Goodfellow et al. [8]. This is true after the study by Madry et al. [16] demonstrated that adversarial training has the potential of improving the model robustness which is useful for reliable weapon detection in actual scenarios.
Models for detection have to be evaluated and hence there is a need for standard metrics for the evaluation and benchmarking. The COCO assessment measures were introduced by Lin et al. [9] and nowadays are used as a norm for object detection. These measures provide a comprehensive assessment of the model's accuracy and localization, namely Average Precision (AP) and Intersection over Union (IoU).

2.2.11 Deployment and Real-World Applications

There is, therefore, the question of scalability, latency, and the ability of the systems to integrate with existing security surveillance systems. Redmon et al. [18] and Liu et al. [25] have done a research work on the application of detection models in smart city. These studies stress out the importance of achieving high performance for models that are used for high database throughputs and low latency.

The trends in the deep learning models and their application in security and weapon identification are well captured in the literature review. The CDCs, R-CDC, Fast R-CDC, Faster R-CDC, YOLO, and Mask R-CDC core architectures among other are good platforms for developing automatic detection systems. The employment of data augmentation and transfer learning enhances the performance as well as the generalization of these models. Exemplary modern developments in the sector show how one can employ deep learning to countervail the existing downsides of a traditional approach to weapons' detection and offer more effective, accurate, and globally applicable results. As such, this research aims to develop the most sophisticated portable gun detection system that builds on the existing advances for improved PSS. Based on the literature analysis, it is possible to mention that deep learning models are progressing rapidly and can be applied to weapons detection. The basic architectures made available for trial purposes are used in developing servoed detection systems Apart from the Convolutional Neural Network (CNN), other facilities include Regions-Convolutional Neural Network (R-CNN), Fast R-CNN, Faster R-CNN, You Look Only Look (YOLO), and Mask R-CNN. It has made those systems stronger and better through adversarial training, hybrid models, and specific data sets.

The plan for this project is to employ these advancements to design an innovative portable gun detection system that would complement the absence of traditional methods in increasing the safety of the population. Mature assessment techniques along with the ability to perform the needed calculations in real-time will ensure that the proposed system is reliable and efficient in a variety of environments due to utilizing deep learning models.

Chapter 3

Methodology

Public safety is under severe threat due to the increasing use of firearms and, therefore, requires the development of efficient weapon detection systems. The purpose of this study is to apply image datasets and also deep learning to draw out and identify handgun images. The revealed major approaches are based on Region Proposal Networks for object detection using frameworks like Faster R-CNN, YOLO, Mask R-CNN, and Transfer Learning for picture classification. In this section, the detailed methodology followed in this study to develop a better handgun detection system with respect to various object detection frameworks which are YOLO, Faster R-CNN, and Mask R-CNN, is discussed. The research involved three critical stages: Data collection, preprocessing, and modeling. In this study, handgun detection accuracy was improved through transfer learning in image classification and region proposal-based methods used for object detection, as a way to reduce false positives.

3.1 Model Architecture

The central contribution of our study is to design a robust image classification and object detection modeling strategy. The study is structured into two primary tasks: It includes handgun image classification and localization object detection. The generic workflow of the this work is shown in figure 3.1.

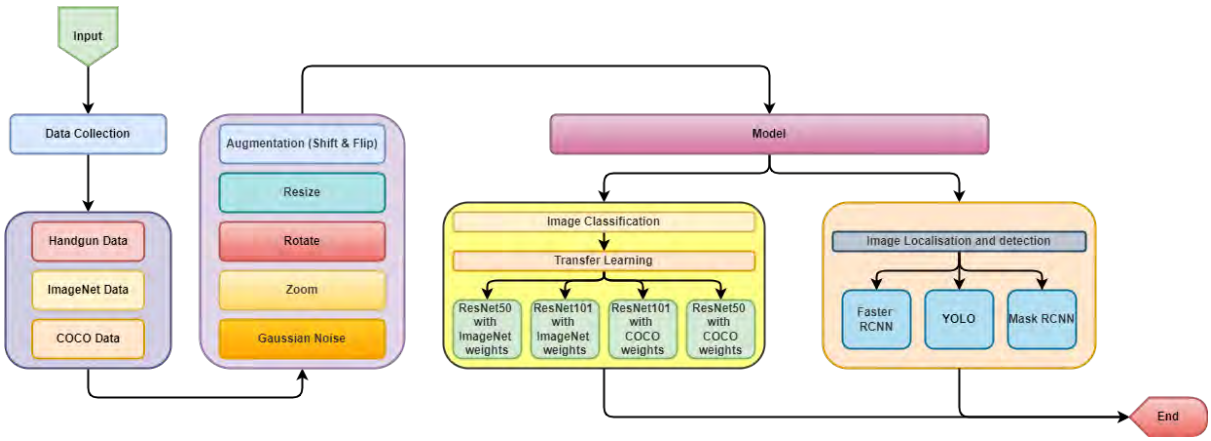


Figure 3.1: Detailed Work Plan

Several architectures were explored in this study, including:

- ImageNet weights on ResNet50
- ImageNet weights for ResNet101
- ResNet101 with COCO weights
- ResNet50 with COCO weights

This is a fine tunning process in which we modified the architecture by adding some custom dense layers on top of pre-trained layers and thereafter adjusting the weights of these rookie layers. Regularization techniques such as L2 and L1 were combined with Adam optimizer to optimize final model weights in order to prevent overfitting. Best classification accuracy is obtained from the models tested among, Resnet, with COCO weights, which had the lowest mean Average Precision (mAP).

3.2 Image Localization and Object Detection

Object detection is our main focus, not only to classify whether there is a handgun in an image but also to accurately localize or segment the handgun within the image. Since they can handle complex scenes with several objects, we explored region proposal-based methods.

3.2.1 Faster R-CNN

This study chose Faster R-CNN, because of the balance between accuracy and computational efficiency. Previous to this version, Fast R-CNN was built upon which introduces a Region Proposal Network (RPN) for producing region proposals more quickly than other methods such as selective search.
Faster R-CNN operates as a two-stage network:

- The proposed RPN first scans the input image and generates regions of interest where object may be located. First the regions are passed to the next stage for classification.
- The Fast R-CNN head proposes regions and classifies each region as containing a handgun or background in the second stage.

Shared Computation Backbone:

Faster R-CNN shares the computation between the RPN and the detection network in order to optimize performance. In this case, I'll use a single convolutional network to extract features from the input image (i.e. ResNet50) and use these features for RPN + Fast R-CNN head. However, with this sharing of computation, Faster R-CNN, while still high accurate, is actually faster than traditional methods.

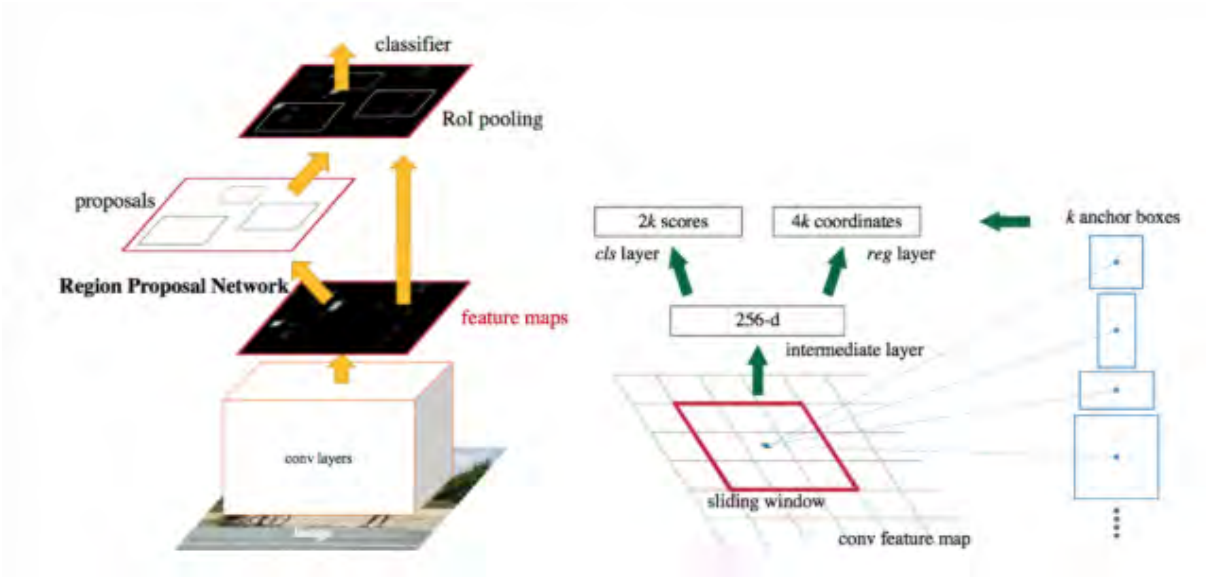


Figure 3.2: Generic Faster RCNN model architecture [11].

3.2.2 Mask R-CNN

For the segmentation task, we chose Mask R-CNN, an extension of Faster R-CNN. Moreover, it localizes handguns not only at locations but also via pixel-level segmentation, keeping detection more precise in cluttered environments. In Mask R-CNN we still narrow the search space for objectness, but add one more branch that outputs object masks in addition to bounding box predictions at the pixels. The Mask RCNN model architecture is shown in Fig 3.3

Feature Pyramid Network (FPN):

The main part of the Mask R-CNN model is the Feature Pyramid Network (FPN), which helps the model to see objects at various scales. FPN will generate a pyramid of feature maps from different resolutions so that the model can take care of objects with different sizes. It was particularly handy for detecting small handguns which might be tricky to reveal using a regular technique. The FPN architecture is shown in Fig 3.4
Its results on detecting handguns in cluttered scenes were held by the Mask R-CNN model, and the object segmentation masks from its detections matched its ground truth object segmentation masks very well. For applications in which fine grained localization of weapons was required, this level of precision was particularly useful.

Region Proposal Network (RPN):

A Region Proposal Network requires both a determining element and a dimension estimation part. Here the introduction of anchors begins with each anchor representing the middle point of sliding window segments. Investigators change the rectangular anchor sizes to 30, 60, 125, 250, and 500 pixels.
Each cell's aspect ratio measures as 0.4, 1, and 2.5 to define anchor dimensions for width and height. The Region Proposal Network keeps anchor strides constant at 1 to help the sliding window progress one pixel each time. The RPN proposals are cleaned successfully from the image using Non-Maximum Suppression with a 0.65 threshold. The system

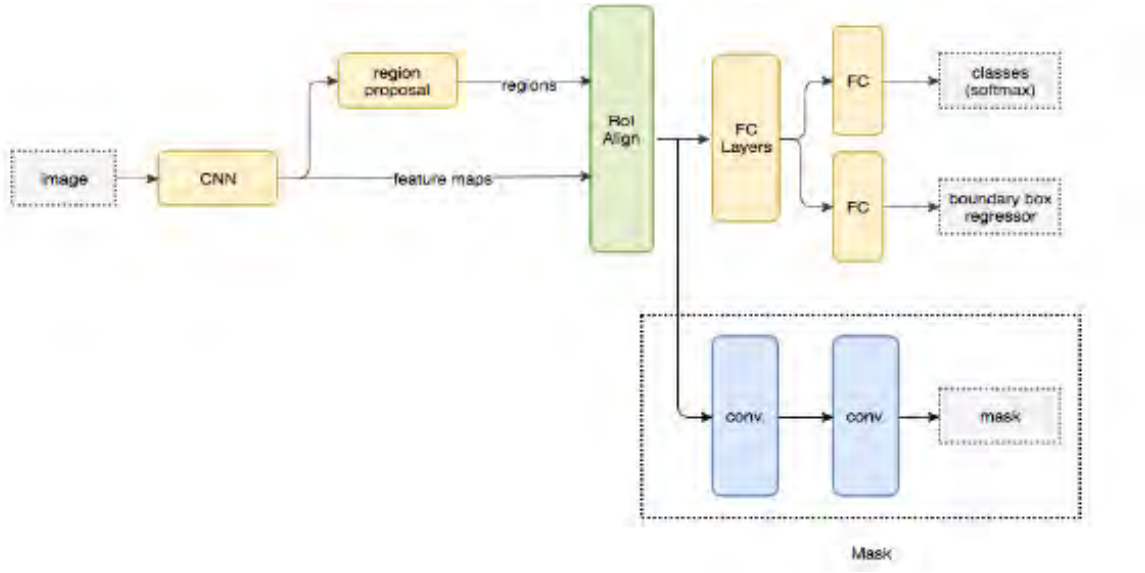


Figure 3.3: Generic overview of Mask RCNN architecture

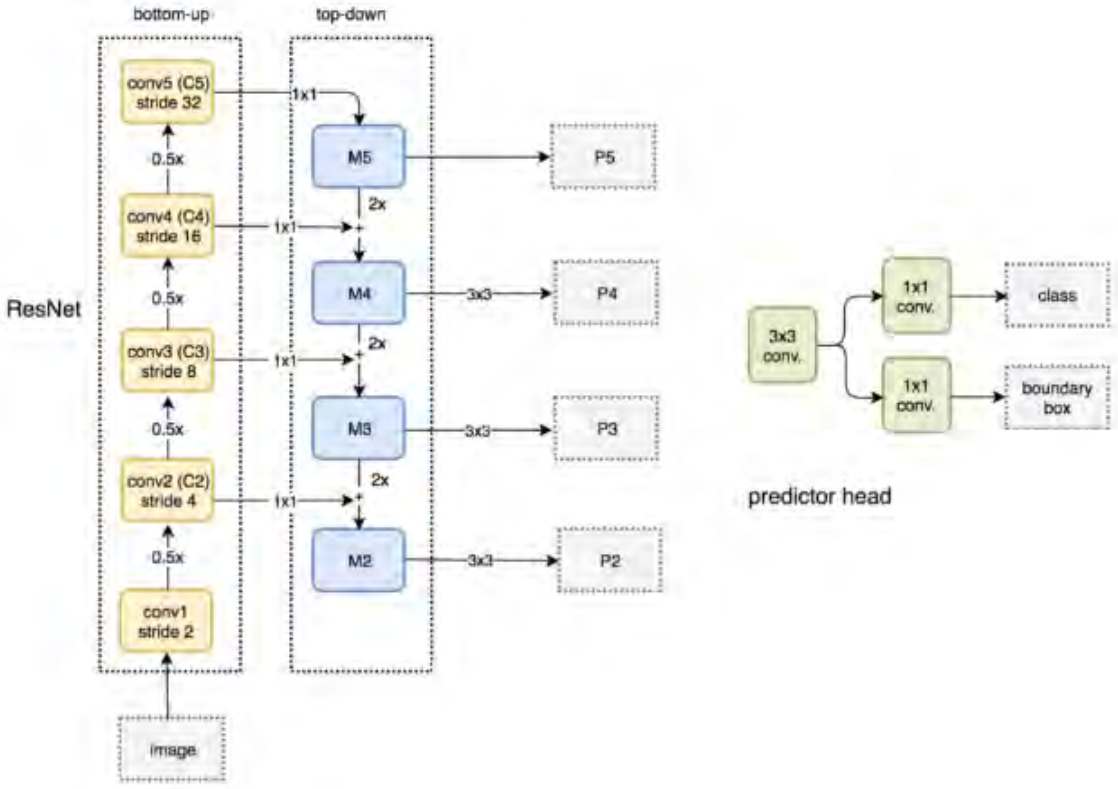


Figure 3.4: FPN architecture overview for Mask RCNN [23]

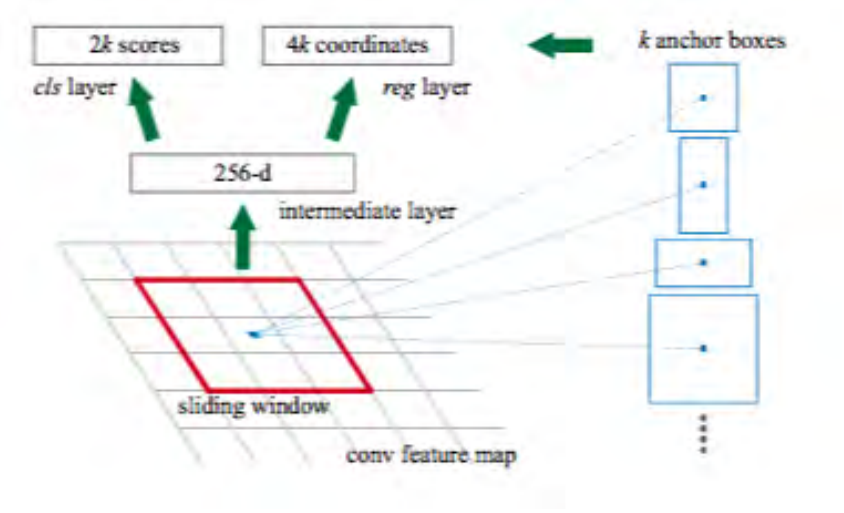


Figure 3.5: Generic architecture of Region Proposal Network [14].

works with 240 anchors before filtering which produces 1800 Regions of Interest (ROIs) for training and 900 ROIs for inference production.
To save memory space we use mini masks when dealing with large high-resolution image files. Stripped-down mini masks enable smaller mask processing which drastically decreases memory requirements. The system shrinks the mini masks into pixel dimensions of 50 by 50.
Input requirements normalize images into square dimensions with a minimum size of 820 pixels and a maximum size of below 1040 pixels. The Mask R-CNN paper requests 512 ROIs, while the RPN often falls short of generating this many positive results. To address this, a positive-to-negative ratio of 1:2.5 is maintained. The model selects regions at pooling sizes 6 through 13 to find the best possible match. The model evaluates bounding box pairs using two IoU thresholds set at 0.25 and 0.75.
The chosen learning rate stands at 0.0012 and the learning momentum at 0.85 to reach optimal loss results. We apply a 0.00012 weight decay penalty to further enhance our model performance. For exploding gradient control we apply gradient clipping with a threshold value of 4.5.
We will examine Mask R-CNN architecture through a complete step-by-step explanation presented in the Figure 3.5.

Chapter 4

Implementation

4.1 Dataset

Object detection tasks can be very successful or not: it all depends on the success of each data collection phase. The data assembled and curated in this study were meticulously gathered and curated from various sources to foster diversity, excellence, and trustworthiness. The design of the data set aimed to include and represent both the completeness and the reality of real-world situations where handguns can be observed in various positions, orientations, and lighting conditions.
We collected 3,000 images specifically for obtaining object detection using YOLO and Faster R-CNN from **DaSCI OD-WeaponDetection: Pistol Detection dataset**. To bring a variety of controlled and uncontrolled environments, these images have been sourced from public databases, private collections, and real-world surveillance footage. 1,751 manually annotated images were used to create the segmentation masks, out of the 3,000 images. To improve model performance in handgun detection and segmentation from complex backgrounds, manual annotation was used to ensure a precise 1:1 label of each object (handgun). Additionally, **we made our own dataset** consisting of 457 handgun images collected from various online sources, as shown in the figure 4.1.
An additional 3,405 images were collected for the classification task, which tries to minimize false positives on classic object detection. The images consisting of handguns formed a balanced dataset of 795 images and 2,610 images of no handguns for binary classification. For transfer learning, this collection was an essential part of the process because

Figure 4.1: Dataset Visualization

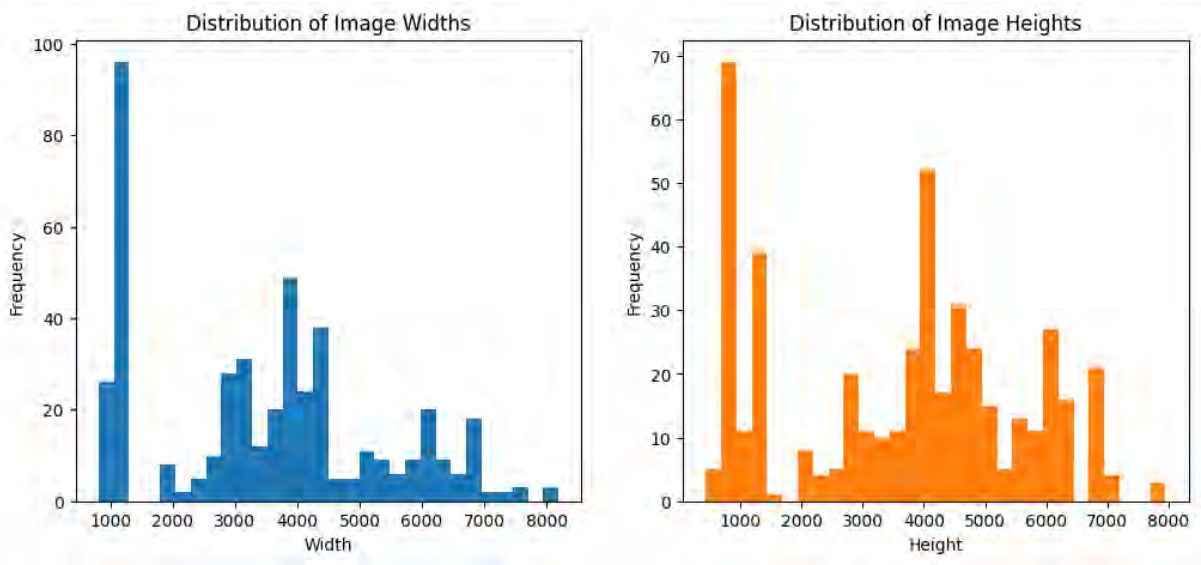


Figure 4.2: Image size ratio histogram

the pre-trained models were fine-tuned on this collection to solve our problem making it easier for them to differentiate between weaponized and non-weaponized images.
The dataset was also further augmented incorporating well-established public datasets such as ImageNet and COCO. Also, the COCO dataset gave our models additional diversity in terms of object categories and contextual variations, as the ImageNet dataset offered a rich pre-trained weights set that went well for initializing ResNet, etc.

4.1.1 Preprocessing

Preprocessing the data is an important step (or many steps) that has an immediate effect on the speed and accuracy of every machine learning model. For this study, a preprocessing pipeline was developed that will augment the data in a crafty manner so that models receive a diverse set of training samples. The motivation was to make the models robust to issues like varied lighting conditions, object occlusion, and orientation.

Data Augmentation:

Various augmentation techniques were used artificially to increase the diversity of the dataset. These augmentations included:

- Shifting and Flipping: To simulate different viewing angles and object orientations, images were horizontally and vertically shifted, and flipped. The model generalizes better to real world situations where objects (handguns) may come in at different angles and orientations.
- Resizing: For each image, we resized it to a standard input size required by the given detection framework. For example, Image input size of Mask R-CNN is 800x1024 pixels, whereas for other models like YOLO, it has its own size constraints.
- Rotation: The images were rotated random, to make sure the models can accurately detect handguns, no matter how they are in the image. This is particularly for hand guns, they can tilt or be held in such odd angles and you would not be able to tell.

Figure 4.3: Sample image after random augmentation.

- Zooming: Some images were subjected to the zoom operation to resemble various camera distances. Zooming has also been ensured that the model can identify handguns from a different distance away from the camera.
- Gaussian Noise: To simulate the real world variability, the images underwent random noise addition (representing variability arising from such factors as different lighting conditions and image quality). This increased robustness to noisy images, such as images captured by low quality security cameras. A sample can be visualized in the Figure 4.3

4.2 Implementation Details

In this section, several architectures have been implemented. In the subsequent parts, our model's implementation has been described.

4.2.1 Faster R-CNN

The Faster R-CNN model was implemented with the following configuration:

- Backbone Architecture: Initilized with random weights for the convolutional layers, ResNet50.
- Training Samples: It was trained on 2,514 samples and validated on 486 samples.
- Category: For the model, we only trained it to only detect one object category of handguns. Strong regularization in the form of dropout and weight decay is used to prevent the model from overfitting.
- Optimization: We used the Adam optimizer and a learning rate of 10^{-4} for the convolutional layers and 10^{-5} for the RPN. The learning rates were chosen carefully to ensure stability during training.

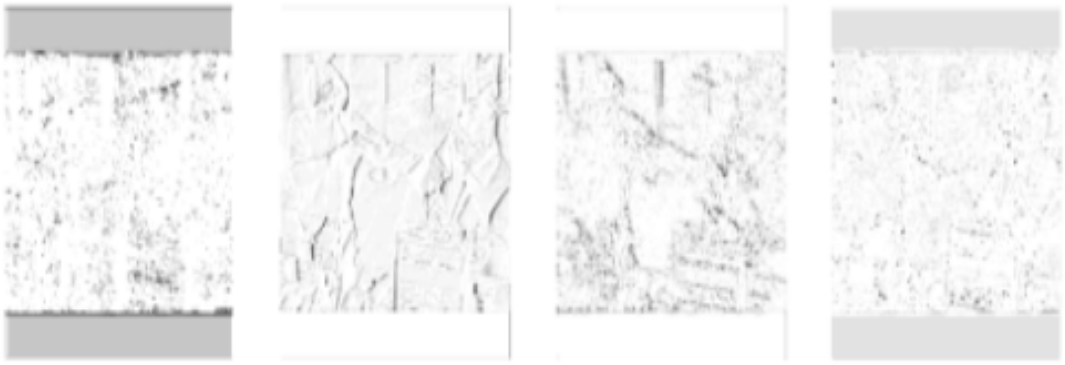


Figure 4.4: Feature extraction using FPN

- Bounding Box Thresholds: To be a positive detection on the IoU threshold, the bounding box prediction required an IoU threshold of 0.8. Two sets of samples were used to train the RPN: Positive samples having IoU > 0.7 and Negative samples with IoU < 0.3.
- Epoch Size and Training: We trained the model for 64 epochs, which means 700 iterations per epoch. A design of the training process was made that minimized the classification and regression losses simultaneously.
- Inference Speed: Prediction took anywhere between 0.8 to 2.5 seconds per image, depending upon the scene complexity. They also concluded that this speed was sufficient for real-time handgun detection in the majority of security applications.

Faster R-CNN demonstrated the ability, with very few false positives, to detect handguns only partially occluded or hidden.

4.2.2 Mask R-CNN

Key features of the Mask R-CNN implementation include:

- Anchors: The model produced anchors of varying sizes: 32, 64, 128, 256 and 512 pixels long. The model was able to detect handguns at different scales due to
- Mini Masks: During training the segmentation masks were resized from 88×88 to 56×56 pixels to conserve memory. It also enabled the model to maintain good accuracy while lowering computational cost.
- Region of Interest (ROI) Pooling: ROI pooling was presented as a method of extracting features from these proposed regions. For regular features, the pool size was set to be 7x7, and for higher resolution features it was 14x14.
- Optimization: We optimized the model using SGD with a learning rate of 0.001, a momentum of 0.9 and weight decay regularization of 0.0001. A maximum gradient norm of 5.0 was used in gradient clipping to stabilize training.

- Training Configuration: The model was trained for 50 epochs and monitoring of training loss was undertaken to check for convergence of Mask R-CNN. Non-Max Suppression (NMS) with an IoU threshold of 0.7 was performed to remove duplicate detections during inference.

Mask R-CNN model achieved good performance in identifying handguns set in cluttered scenes and generated segmentation masks for each object located. This level of precision was particularly useful for applications where the localization of weapons in fine-grained terms was required.

4.3 Real-time Gun Detection

For this research, we have developed the Gun Detection system by using a deep learning model based on the TensorFlow Object Detection API. The main goal was to develop a solution that is robust enough to discriminate firearms (pistols and rifles) in images. It uses pre-trained model which is pre trained model trained on the COCO dataset namely SSD MobileNet v1 model. On object detection task this model is well suited, it is able to run real time and yet maintain a high level of accuracy.
The architecture based on MobileNet v1 was used for the task of gun detection using Single Shot Multibox Detector (SSD). By design, the SSD model is a good object detection architecture that is optimized for real-time applications. The model has a sweet spot in speed as well as accuracy, which makes it an ideal solution for the gun detection problem, where it has to process the image fast, while keeping an eye on the ability to distinguish different objects of interest (e.g weapons).
SSD MobileNet v1 was selected from the TensorFlow Model Zoo, the collection of well maintained pre-trained models that are useful for different Computer Vision tasks. The model was trained originally on the COCO dataset, a dataset containing common objects like persons, cars, animals, et cetera, but that does not explicitly label firearms. For this reason, a custom mapping of classes that are a good fit for the firearm notion of concept was designed, taking advantage of common objects and keywords that are near the concept of a firewall. Custom mapping strategy further increase the potential of the model on firearms detection.
The preprocessing pipeline consisted of resizing input images to match the model's input requirement, then changing the image to an RGB image format. Our input images are normalized as tensors then are fed into the model for inference.

4.3.1 Detection Process

The detection process with the gun starts with loading the images to be analyzed. The images of these examples are collected from a folder in the drive.google.com directory and they are being sourced using the Python os library. When each image is loaded, the model runs object detection and tries to find bounding boxes, the class labels of various things in it, and their corresponding scores.
The model's output is filtered based on a predefined confidence threshold to make sure only firearms are detected. The model predicts an object if its score is above the threshold (set at 0.5), and if such an object is considered a valid detection. Once the bounding box coordinates (normalized values) are scaled to match the original dimensions of the

given input image, this would make the bounding boxes correctly detect firearms drawn around.

Chapter 5

Results and Discussion

In our research, we explored three state-of-the-art object detection algorithms: Faster R-CNN, Mask R-CNN, and YOLO. We chose to implement all of these algorithms in our study simultaneously, as each of these algorithms has its own features, strengths, and limitations, and as a result of our decision, our performance of these algorithms was tested on handgun detection. We now describe the rationale behind this selection and the implementation peculiarities within these models.

5.1 Object Detection Algorithms Choices

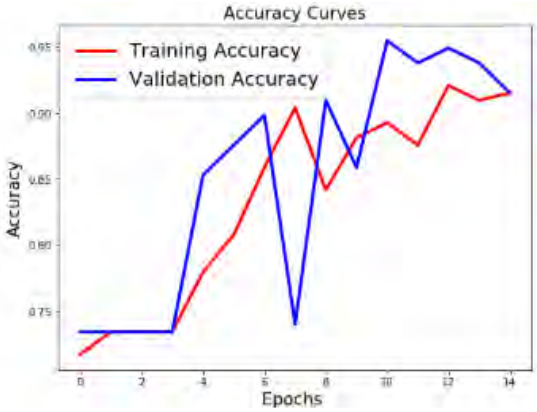
We opted for Mask R-CNN, Faster R-CNN, and YOLO based on their proven success in object detection tasks and their complementary strengths:

- For our purpose of handgun detection and fine-grained segmentation on complex scenes, Mask R-CNN is very effective in delivering both object detection and instance segmentation.
- In cases where localization precision is important, however, there is a relatively strong balance between speed and accuracy with Faster R-CNN. Its speed limitations notwithstanding, it has been reported throughout to outperform YOLO for detection accuracy.
- Since it is generally less accurate, YOLO (comparable to faster R-CNN) is very fast and efficient, which makes it an ideal candidate for real-time handgun detection in the surveillance setting.

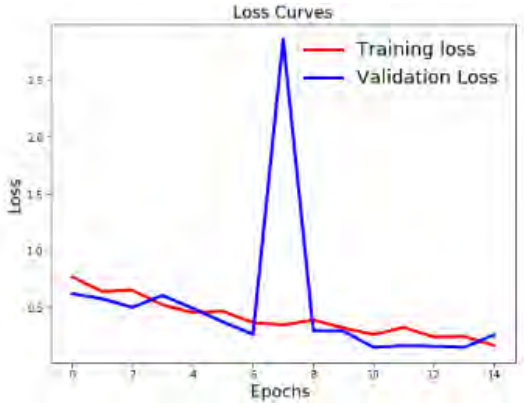
We had intentionally stayed away from using sliding window-based methods for object detection as they have been shown to be not very good at either detecting objects or processing speed in the past research. However, Faster R-CNN and its successors, Mask R-CNN have demonstrated vastly superior performance and scalability in recent modern region proposal-based approaches.

5.2 Model Evaluation

This section discusses each model's strengths and limitations compared to others.



(a) Training and validation accuracy over epochs for the convolutional neural network (CNN) used in handgun classification. The plot illustrates the model's learning progression, with the training accuracy (blue line) showing how well the model fits the training data, and the validation accuracy (orange line) indicating its generalization to unseen test data. A steady increase in both metrics suggests effective learning, while any divergence between the curves may indicate potential overfitting or underfitting.



(b) Training and validation loss over epochs for the CNN model applied to handgun classification. This plot depicts the reduction in categorical crossentropy loss, where the training loss (blue line) reflects the model's optimization on the training dataset, and the validation loss (orange line) measures performance on the test set. A downward trend in both curves indicates successful convergence, with closer alignment suggesting robust generalization and minimal overfitting.

Figure 5.1: Performance evaluation of the convolutional neural network (CNN) trained for binary classification of handguns versus other guns. Subfigure (a) shows the accuracy progression, highlighting the model's ability to correctly classify images over time. Subfigure (b) displays the loss trajectory, demonstrating the optimization process and convergence behavior. Together, these metrics provide insights into the model's learning dynamics, generalization capability, and potential areas for improvement, such as addressing overfitting if validation metrics plateau or diverge.

5.2.1 Mask R-CNN

We also trained the model on a dataset containing 1800 annotated images which is partitioned into a training and a validation set with corresponding ground truth data. The loss metrics used to monitor the training process included Mask R-CNN loss (loss of mask prediction) and FPN loss (Feature pyramid network loss). The performance of the model was also strong as these metrics remained relatively low. Nevertheless, the RPN loss (Region Proposal Network loss) was much higher than we would've expected, implying that the model requires more work to generate good regions to propose.
Additionally, since the model performed well, it produced high false positive rates, especially with complex images of details with high false positive rates. To mitigate this issue, we employed several strategies:

- Gradual unfreezing of layers: In order to discover more advanced features, we unconstrained the final layers of the model without freezing them while fine-tuning them.

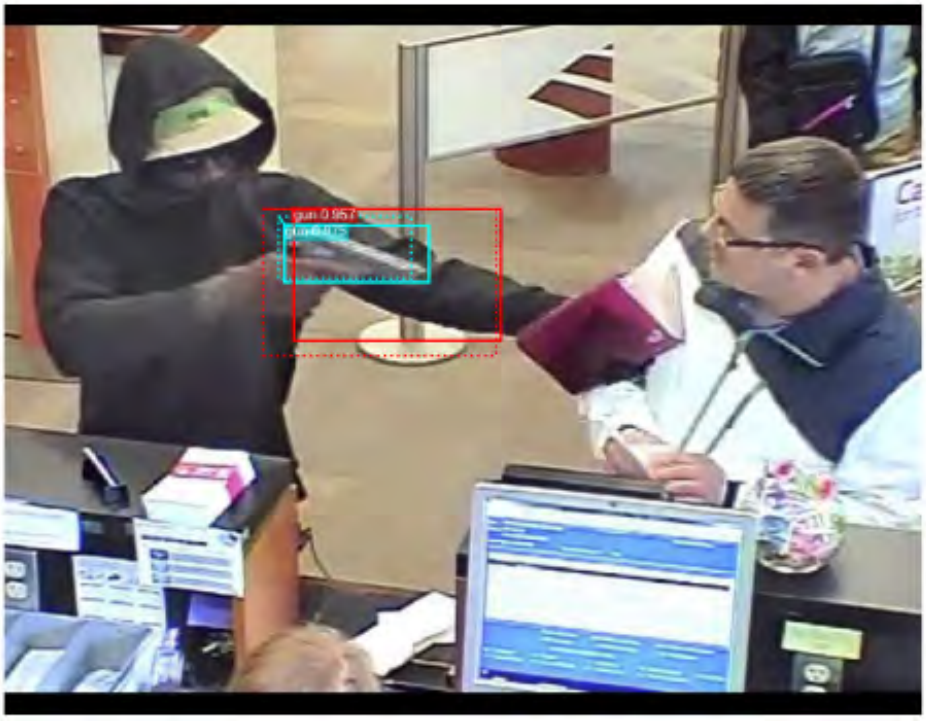


Figure 5.2: Generation of mask with the connection of two back-to-back convolution layers and getting the output as a feature map

- Incorporating misclassified images into classification training: It was based on both its images that were falsely detected as handguns and later their images that were not previously passed through classifiers based on a ResNet101. Of the various versions of ResNet that we tested, ResNet101 with ImageNet weights showed the lowest classification loss value, and at the same time had a significantly smaller fraction of false positives.

5.2.2 Faster R-CNN and YOLO

The same dataset used for Mask R-CNN was used to train the model for Faster R-CNN. This permitted a direct comparison of results between the two approaches. With Faster R-CNN showing better recall than YOLO and Yolo-v2 demonstrated by showing a better precision and a Fell threshold similar to existing literature, it is demonstrated that Faster R-CNN is best suited to find smaller objects and suppress False negatives. It was slower than expected in its processing speed, however.
Whereas YOLO excelled in terms of detection speed, it became a viable candidate for real-time handgun detection. Though YOLO's recall rate was not as high as Faster R-CNN, its ability to process frames far faster than Faster R-CNN (about 10 ms per frame) is an asset to live surveillance systems. While speed and accuracy versus speed conflict in YOLO, YOLO's speed means it's particularly powerful in situations where speed of detection is crucial, but should be bumped up against other models for higher precision.

Figure 5.3: Masks are produced by consecutively applying two convolution layers, with the output from the final convolution layer serving as the feature map.

5.3 Results

A complete analysis of the results obtained via the three object detection algorithms is presented in this section. Consequently, we primarily focused on the validation losses, detection speed, and classification performance for evaluation.

5.3.1 Mask R-CNN Performance

We compared the performance of Mask R-CNN on ImageNet and COCO-trained weights with ResNet50 and ResNet101 architectures. The following table summarizes the validation mask loss and validation classification loss for each approach:
It was observed from the result that the ResNet101 with COCO weight performed better than ResNet50, as the validator mask loss and the classifier loss are lowest. This suggests that training the model directly on the COCO dataset, which comprises of numerous object classes, was key for handgun detection.

Model	Validation Mask Loss	Validation Classification Loss
ResNet50 (ImageNet)	0.6012	0.0613
ResNet101 (ImageNet)	0.5931	0.0421
ResNet50 (COCO)	0.2618	0.0562
ResNet101 (COCO)	0.2347	0.0545

Table 5.1: Validation losses for Mask R-CNN using different ResNet architectures and weights.

5.3.2 Faster R-CNN Results

The Faster R-CNN was evaluated with regards to detection speed and precision. We tested Faster R-CNN to be around 300 ms per frame detection time and it is slower than YOLO, but faster than Mask R-CNN. Faster R-CNN was also more accurate compared to YOLO, regardless of the cluttered scene or complex backgrounds in images when detecting handguns.
We visualize in Figure **??** the anchors from different feature map levels through center-cell-based analysis to examine their spatial patterns and scaling behavior. Our examination of the central cell provides consistency in the way we evaluate anchor placement throughout diverse size resolutions. The framework presents features in a hierarchical manner through increasing anchor scales, moving from higher levels downwards. Higher levels incorporate bigger anchors to capture extensive contextual details, alongside lower levels using smaller anchors to recognize fine elements. Visualizing the anchor structure gives us clearer perspectives into how detection scales correspond with the model framework, which makes it simpler to understand multi-scale detection. Each anchor displays brightness to indicate its ranking position within the set and allows people to follow placement patterns while verifying consistency across multiple layers.

5.3.3 YOLO Results

And as always with speed, YOLO cruised past Faster R-CNN and Mask R-CNN, processing frames in just 10 ms. It is best suited to real-time detection applications. Nevertheless,

YOLO traded speed for accuracy at the expense of the faster R-CNN's precision, the latter being a little worse than what was expected. A known limitation of the architecture of YOLO is that it had difficulty in particular with small or partially occluded objects.

5.3.4 Comparative Analysis

The table below provides a comparison of the three models based on detection speed:

Model	Speed of Detection (Single Frame)
Faster R-CNN	303 ms
Mask R-CNN	345 ms
YOLO	18 ms

Table 5.2: Speed of detection for different object detection models.

These results show that YOLO is the best for real-time applications, while Mask R-CNN and Faster R-CNN achieve better accuracy with Faster R-CNN balancing speed and accuracy. But selecting which model to deploy depends on application-specific requirements, especially if the focus is on detecting speed or accuracy.

5.3.5 Visual Comparison of Faster R-CNN and Mask R-CNN

To explore the different object detection and segmentation qualities of these models we show side-by-side test results on three distinct images. The examples show scenes with many objects in the background plus overlapping and different-sized items. The analysis shows Mask R-CNN generates better object boundary details along with precise segmentation masks while Faster R-CNN delivers fast detection at similar box accuracy levels. This graphic display shows how fixing between these two models for different uses depends on making a choice between faster results and improved object separation details.
To evaluate the effectiveness of the anchor and ROI (Region of Interest) selection process in our object detection framework, we implemented a detailed analysis of anchor classification (positive, negative, and neutral) and their spatial refinements. Anchors play a critical role in defining candidate regions for object detection, and understanding their breakdown provides insights into the behavior of the model. We computed the shifts in anchor positions using refined bounding box deltas and visualized the spatial distribution of positive, negative, and neutral anchors. Positive anchors, which overlap significantly with ground truth boxes, represent valid detection candidates, while negative and neutral anchors serve as training examples that aid in learning robust classification boundaries. This analysis demonstrates the efficiency of anchor sampling and refinement in aligning with target objects.
Furthermore, we extended our analysis to ROIs generated during training. Randomly sampled ROIs allow us to evaluate how the model handles both object-specific and background regions. By visualizing positive ROIs and their corresponding masks, we validated the alignment between predicted bounding boxes, segmentation masks, and ground truth objects. The visual inspection revealed that the refinement process significantly improves ROI alignment, even for small and challenging objects. Our results indicate that the positive ROI ratio is well-balanced, providing sufficient samples for accurate mask and class prediction. Additionally, we ensured that ROIs were unique, confirming the diversity in region sampling, which is vital for avoiding overfitting and improving generalization.

(a) Visualization of negative anchors (anchors not chosen for training, but which haven't merged much with ground truth). The figure highlights the filtered out large pool of anchors during the RPN process and highlights the effectiveness of the anchor matching strategy.

(b) Visualisation of randomly sampled neutral anchors from the pool of anchors not targeted for training nor overlapping objects. Finally, this shows the distribution of anchors that do not contribute to shaping the model predictions.

Figure 5.4: Comparison of (a) negative anchors and (b) neutral anchors. These visualizations highlight the model's ability to localize objects accurately.

We finally performed in depth visualization of refined ROIs, their masks and bounding boxes performing class aware refinement of each region. We found that the model was able to learn to adaptively refine RoIs by overlaying both the original and refined RoIs on sample images and finding a clear improvement in object localization. By having step by step visualisation of this anchor based region proposal and refinement strategy, we demonstrate how our region proposal and refinement strategy achieves better detection and better segmentation performance over scales and object categories.

5.4 Real-time Gun Detection Visualization

After completion of the detection process, the displayed bounding boxes are overlaid on images, so that the user can opt to view them. This is done using OpenCV, which yields functionality around detecting objects and drawing rectangles around them, and labeling them. Each detected firearm is labeled with the term "Gun" to make the text easier to read, and the bounding box coordinates are extracted from the model's output. Finally, the labeled images are visualized in matplotlib within the Jupyter notebook environment. Visualizing the process, however, provides an opportunity to assess how well the model performs. If we are to detect multiple guns in the same image, we assign each gun a unique bounding box, shown with a fluctuating background. Additionally, the model can be fine-tuned to address such cases where it fails to detect a firearm because the firearm is placed in the image or there is not enough resolution by adjusting the threshold of detection or by changing the model by training it with a more specific dataset.

Figure 5.5: Positive and negative ROI analysis with positive to negative ROI ratio during training. The figure also shows that the network can create unique ROIs and that this is due to the lack of redundancy and to the efficacy with which the network covers a range of such object regions.

Figure 5.6

Figure 5.7

Figure 5.8

Figure 5.9: Gun detection using unseen data out of the dataset to evaluate the real world performance of the model.

Visualising process, however, provides an opportunity to assess how good he model performs. If we are to detect multiple guns in the same image, we assign each gun a unique bounding box. Additionally, the model can be fine-tuned to address such cases where it fails to detect a firearm because the firearm is placed in the image or there is not enough resolution by adjusting the threshold of detection or by changing the model by training it with a more specific dataset.

(a) Real-time gun detection with correct bounding box around the crowded environment.

(b) Real-time gun detection with correct bounding box drawn with heavy firearms.

Figure 5.10: Visualization of real-time gun detection with bounding boxes drawn around firearms in two separate scenarios.

5.4.1 Evaluation Metrics

To assess the performance of the handgun detection system, a comprehensive set of evaluation metrics was employed, including Precision, Recall, F1 Score, Intersection over Union (IoU), and Dice Coefficient. These metrics provide a robust framework for evaluating the effectiveness of the convolutional neural network (CNN) and other object detection models in identifying handguns within images, particularly in distinguishing them from other

firearms. Each metric targets specific aspects of model performance, such as classification accuracy, detection completeness, and localization precision, which are critical for real-world applications like firearm detection in varied environments.

Precision measures the proportion of correctly identified handgun detections out of all instances the model predicted as handguns. It is defined as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

where TP (True Positives) is the number of correctly identified handguns, and FP (False Positives) is the number of non-handgun objects incorrectly classified as handguns. A high precision indicates a low false positive rate, ensuring that the model's positive predictions are reliable.

Recall quantifies the proportion of actual handguns correctly detected by the model, calculated as:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

where FN (False Negatives) represents handguns that the model failed to detect. High recall is crucial for minimizing missed detections, which is particularly important in security-sensitive applications where failing to identify a handgun could have serious consequences.

The **F1 Score** is the harmonic mean of Precision and Recall, providing a balanced measure of the model's classification performance:

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

This metric is especially useful when dealing with imbalanced datasets, as it penalizes models that favor either precision or recall at the expense of the other, offering a single value to assess overall accuracy.

For object detection tasks, localization accuracy is equally important. The **Intersection over Union (IoU)** evaluates the spatial accuracy of detected bounding boxes by measuring the overlap between the predicted bounding box (B_p) and the ground truth bounding box (B_g):

$$\text{IoU} = \frac{\text{Area}(B_p \cap B_g)}{\text{Area}(B_p \cup B_g)}$$

A higher IoU indicates better alignment between predicted and actual bounding boxes, reflecting the model's ability to precisely locate handguns within images.

The **Dice Coefficient**, closely related to IoU, further quantifies the similarity between predicted and ground truth regions:

$$\text{Dice} = \frac{2 \cdot \text{Area}(B_p \cap B_g)}{\text{Area}(B_p) + \text{Area}(B_g)}$$

The Dice Coefficient emphasizes overlap and is particularly sensitive to small discrepancies in localization, making it valuable for assessing fine-grained detection performance.

The evaluation was conducted using a diverse dataset comprising 457 images of firearms, including both handguns and other gun types, collected from the internet and generated by AI. These images varied in resolution, background complexity, and firearm orientation,

providing a challenging testbed for the model. Since the pre-trained SSD MobileNet v1 model was not originally trained on a matching gun-detection dataset, fine-tuning was necessary. This involved adjusting the detection classes to focus on handguns versus other firearms, using transfer learning to adapt the model's weights to the new dataset. The fine-tuning process included retraining the final classification layers while preserving the feature extraction capabilities of the pre-trained backbone, ensuring efficient adaptation to the specific task.
The evaluation results, summarized in Table 5.3, compare the performance of various models, including the fine-tuned SSD MobileNet v1 (referred to as CNN in the table), VGG, C2D, YOLO, Faster R-CNN, Mask R-CNN, and R-CNN. The dataset used for evaluation included images with multiple firearms, diverse backgrounds, and varying resolutions to simulate real-world conditions. The results demonstrate that the fine-tuned model achieved promising performance, with a relatively low rate of false positives and false negatives, indicating robust detection capabilities across complex scenarios.

Model	**IoU**	**Dice**	**Precision**	**Recall**	**F1 Score**
VGG	0.263	0.417	0.435	0.400	0.417
C2D	0.469	0.638	0.682	0.600	0.638
YOLO	0.7901	0.8828	1.0	1.0	1.0
Faster R-CNN	0.7510	0.8545	0.95	0.97	0.96
Mask R-CNN	0.7709	0.8749	0.98	0.99	0.985
CNN ([5])	0.7154	0.8205	0.88	0.90	0.89
R-CNN ([17])	0.7020	0.8154	0.85	0.88	0.86

Table 5.3: Performance comparison of various models for handgun detection, evaluated using Intersection over Union (IoU), Dice Coefficient, Precision, Recall, and F1 Score. The metrics reflect each model's ability to accurately classify and localize handguns in a diverse dataset with varying image resolutions and backgrounds. The fine-tuned CNN, based on SSD MobileNet v1, demonstrates competitive performance, as indicated by its IoU, Dice, and classification metrics.

The high performance of models like YOLO, Faster R-CNN, and Mask R-CNN, as shown in Table 5.3, highlights their advanced capabilities in both classification and localization, likely due to their sophisticated architectures and training strategies. The fine-tuned CNN model, with an IoU of 0.7154 and F1 Score of 0.89, performs competitively, indicating effective adaptation to the handgun detection task. However, the lower performance of models like VGG and C2D suggests limitations in their feature extraction or localization capabilities for this specific dataset. These results underscore the importance of model selection and fine-tuning to achieve optimal performance in firearm detection tasks.

5.4.2 Limitation

The gun detection system works reasonably well on a variety of test images, however, there are limitations to the performance of the model. As the model was not trained to detect guns specifically, it may not be perfect in terms of predictive accuracy, particularly in determining uncommon or smaller guns. Furthermore, objects that look like guns due to their shape and metallic colours may get misclassified as anxious and so result in false positives.

Figure 5.11: An example where a gun is correctly detected; however, due to similarities in shape and metallic color, false positives can occur, leading to misclassifications that depicts the picture where gun barrel has been slightly outside of the bounding box.

Figure 5.12

Figure 5.13

Figure 5.14

Figure 5.15: To evaluate the effectiveness of the anchor and ROI (Region of Interest) selection process in our object detection framework, we implemented a detailed analysis of anchor classification (positive, negative, and neutral) and their spatial refinements. Anchors play a critical role in defining candidate regions for object detection, and understanding their breakdown provides insights into the behavior of the model. We computed the shifts in anchor positions using refined bounding box deltas and visualized the spatial distribution of positive, negative, and neutral anchors. Positive anchors, which overlap significantly with ground truth boxes, represent valid detection candidates, while negative and neutral anchors serve as training examples that aid in learning robust classification boundaries. This analysis demonstrates the efficiency of anchor sampling and refinement in aligning with target objects.

Chapter 6

Conclusion

The paper's goal is to develop and evaluate DL approaches for firearms recognition and categorization. Implementing this necessitated approaching of it from a number of angles, such as data acquisition, data cleaning and pre-processing, model formulation, and evaluation. The main objectives were to produce a large dataset, augment the data for model robustness, deploy and optimize cutting-edge neural networks for classification and object detection purposes, and thoroughly evaluate said models.
The first step of the research was to gather initial data from the Website of Weapons Detection Database, which produced 3000 of photos. Of these photos, 1751 images were acquired after the annotation process, which focused on the identification of objects using Mask R-CNN. Also, the 3405 photos were defined for categorization, with the more detailed division into the photos with guns, 795, and the rest, 2610. This large set of samples proved to be crucial to developing models fit for a general assessment while pointing out the importance of having a large, fine-tuned database in machine learning operations.
Additional data preprocessing and augmentation were necessary to get better results concerning the given model's performance. Some of the strategies of preparation included scaling the picture to the size of a standard format and rescaling the pixel value so that the training of the neural network could be faster. Other techniques like shifting, flipping, rotation, zooming and finally adding Gaussian noise forms the data augmentation to make the variability of the data more enhanced. These augmentations were meant to mimic various realistic situations thus, enhancing the creation's ability to apply to new previous unseen data which reduces this risk of overfitting.
Thus, the present research significantly extends the knowledge of weapon detection based on the exploration of deep learning algorithms and their ability to accurately identify and classify handguns. Each phase of the full technique of data collection and model assessment contributes to building the firm foundation for the next move in the research. About the work, it is noteworthy that recognizing the variety of neural networks contributing to the development of intelligent hybrid systems, the use of new technologies in the field of public safety has only begun, but already offers possibilities for automated identification of weapons.
Finally, it is shown that using a currently pre-trained model, such as SSD MobileNet v1, it is possible to detect firearms with decent accuracy. The model worked well on a few different test images, however, improving the model to be better on a more specific gun detection dataset would be beneficial. The fact that such systems have the potential to be utilized in the security and surveillance industries makes this research a good one to

build on for developing real-time gun detection systems. Improvements to the system will continue in the future through model fine-tuning, and investigation of new frameworks.

6.1 Future Work

We will expand research into these systems through multiple directions to make them work better in real situations. Our priority is to grow our dataset with more types of firearms while adding different natural settings and obscured views. Adding video input to current systems makes real-time firearm tracking possible which improves safety awareness in changing settings.

Enhancing feature extraction and accuracy will become possible through transformer technology particularly Vision Transformers (ViTs). The approach will work better with audio and thermal data inputs to produce more secure systems that can operate in tough settings.

The development of active learning models shows promise by requiring fewer annotation resources during training. Human-guided training helps the system learn from limited labeled data, enabling faster practical use in real-world applications.

Refine these models to make them suitable for real-world edge device environments. Our focus is to improve hardware efficiency by using model compression tools including pruning and quantization methods to keep the system running smoothly on resource-limited devices while preserving its precise alert speed.

Future research will improve firearm detection performance and broadens its use for public safety purposes through additional development work.

The gun types and scenarios will be pretty much ranging from a variety of different gun types to different gun related scenarios, so that there will be a much wider variety of features that model is learning. For this, we shall apply transfer learning techniques in order to enable the model to learn from the pre trained COCO model as it tries to recognize guns in real-world situations.

Bibliography

[1] H. T. Tavani and J. H. Moor, "Privacy and the internet," in *The Handbook of Information and Computer Ethics*, 2001, pp. 187–204.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, 2009.

[3] M. Baccouche, F. Mamalet, C. Wolf, B. Garcia, and A. Baskurt, "Sequential deep learning for human action recognition," *International Workshop on Human Behavior Understanding*, pp. 29–39, 2011.

[4] C.-H. Demarty, C. Penet, M. Schedl, and B. Ionescu, "The mediaeval 2012 affect task: Violent scenes detection," in *MediaEval 2012 Workshop*, 2012.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[6] S. K. Bandyopadhyay and P. Mukhopadhyay, "Concealed weapon detection using image processing and infrared imaging," *International Journal of Computer and Information Engineering*, vol. 7, no. 10, pp. 1285–1289, 2013.

[7] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[8] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[9] T.-Y. Lin, M. Maire, S. Belongie, *et al.*, "Microsoft coco: Common objects in context," in *European conference on computer vision*, Springer, 2014, pp. 740–755.

[10] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[11] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, vol. 28, 2015.

[12] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[13] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[14] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7263–7271.

[15] M. Brundage, S. Avin, J. Clark, *et al.*, "The malicious use of artificial intelligence: Forecasting, prevention, and mitigation," *arXiv preprint arXiv:1802.07228*, 2018.

[16] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations*, 2018.

[17] R. Olmos, S. Tabik, E. López-Rubio, and F. Herrera, "Automatic handgun detection alarm in videos using deep learning," *Neurocomputing*, vol. 275, pp. 66–72, 2018.

[18] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," in *arXiv preprint arXiv:1804.02767*, 2018.

[19] S. Kornblith, J. Shlens, Q. V. Le, and G. E. Hinton, "Do better imagenet models transfer better?" In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2661–2671.

[20] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, pp. 1–48, 2019.

[21] Z.-Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 11, pp. 3212–3232, 2019.

[22] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[23] J. Hui, "Understanding feature pyramid networks for object detection (fpn)," *Medium*, 2020, Accessed: 2024-10-21. [Online]. Available: https://jonathan-hui.medium.com/understanding-feature-pyramid-networks-for-object-detection-fpn-45b227b9106c.

[24] G. Jocher, *Yolov5*, 2020.

[25] W. Liu, Y. Lu, and J. Wang, "Hybrid deep learning for weapon detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 910–911.

[26] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10 781–10 790, 2020.

[27] *A Comparative Analysis of Weapons Detection Using Various Deep Learning Techniques*, 2023. DOI: 10.1109/ICOEI56765.2023.10125710.