Computer Vision: Object Detection using Faster R-CNN, Mask R-CNN, and YOLOv8

University of West Florida

Project Lead: Hector R. Gavilanes
Deep Learning Architect: Jacob Knight
Data Scientist: Kyle Knuth Software Engineer: Sai Devarashetty

2024-02-16

Table of contents

Abstract

The initial objective of this research paper is to evaluate the performance of the Faster R-CNN methodology on a dataset provided by Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun in their paper titled "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," introduced at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) in 2015 [1]. We aim to replicate the results presented in this article, and gain a comprehensive understanding of Faster R-CNN, which will serve as the foundation for our future testing endeavors. Furthermore, we will assess the performance of Mask R-CNN, and YOLOv8 to offer a comparison of various object detection methodologies. A custom class label was applied to detect military tanks using a YOLOv8 model.

RCNN


RCNN
person 0.91
snowboard 0.89
Devarashetty
Gavilanes
Knight
Knuth
2021

Glossary

Term	Definition
CNN	Convolutional Neural Network
COCO	Microsoft Common Objects in Context
CPU	Central Processing Unit
DFL	Distribution Focal Loss
FC	Fully Connected Layer
FCN	Fully Convolutional Network
GPU	Graphics Processing Unit
IoU	Intersection over Union
mAP	mean Average Precision
NMS	Non-Maximum Suppression
ReLU	Rectified Linear Unit
RoI	Region of Interest
RoIAlign	Region of Interest Align
R-CNN	Region-based Convolutional Neural Network
RPN	Region Proposal Network
SPP	Spatial Pyramid Pooling
SVM	Support Vector Machine
YOLOv8	You Only Look Once, version 8

Summary

The faster R-CNN approach significantly enhances object recognition. It meets the requirement for high precision near real-time detection. The Regional Pattern Network (RPN), Region of Interest (RoI) Network, Backbone Convolutional Neural Network (CNN), and Training/Pattern Framework are the primary components of the rapid R-CNN technique.

Proposal Network (RPN) is a significant advancement in Faster R-CNN technology. The convolutional feature map is used to efficiently produce regional dimensions. This leads to a considerably more versatile, end-to-end trainable system that does not require external recommendation techniques like selective search.

Upon obtaining the local parameters, identically sized objects are extracted from the feature map for every local parameter using the ROI pooling layer. Then, bounding box return and classification are performed using these features. By using this method, the model may concentrate on the areas of the picture that are most important, which increases accuracy and speed.

To extract information from the visual input, the spinal cord's Convolutional Neural Network (CNN) is employed. The most widely used options for spinal networks are ResNet and VGG-16. The model gains from transfer learning, and is able to extract well-formed features from the input photos by utilizing a pre-trained CNN as its foundation.

Using a multitasking loss function that combines classification loss with bounding box regression loss, the model is trained from beginning to end. As a result, the model may learn to predict bounding boxes for objects, and classify them concurrently. The computationally trained model is appropriate for low latency applications since it may be utilized for real-time object detection.

The extensively utilized Faster R-CNN technique has served as the foundation for numerous extremely sophisticated object recognition systems. It has positively contributed to different fields including, but limited to autonomous driving, robotics, medical diagnostic, manufacturing, and video surveillance.

1 Introduction

Object detection is a fundamental task in computer vision, playing a crucial role in various applications such as autonomous vehicles, surveillance systems, and medical imaging. Recent advancements in deep learning have led to the creation of sophisticated architectures such as Faster R-CNN, Mask R-CNN, and YOLO (You Only Look Once) [2], all of which have substantially improved object detection performance. These architectures have revolutionized the field by achieving superior accuracy in detecting objects within images. However, despite significant advancements, a critical trade-off between accuracy and processing speed persists. It is essential to carefully consider this balance when selecting the most suitable model for practical applications.

The efficiency of object detection models is essential, particularly in real-time applications where rapid decision-making is vital. Currently, object detection primarily relies on region proposal methods, and region CNNs. For instance, Selective Search, a widely used region proposal method, exhibits a speed of 3.9 seconds per class per image [3] when implemented on the CPU, which is considered slow for many real-world applications. On the other hand, Edge Boxes, another region proposal method, offers improved speed, clocking in at 0.25 seconds per image [4]. However, a notable constraint remains – even with advancements, the region proposal step still demands substantial computational resources when executed on the CPU rather than the GPU.

Moreover, comparing the performance of Faster R-CNN with region proposal methods poses challenges. Since R-CNNs leverage GPU acceleration while region proposal methods primarily rely on CPU computation. A fair comparison between the two becomes challenging. To address this issue, Region Proposal Networks (RPNs) were introduced in 2015 as part of the Faster R-CNN framework [1]. A novel approach that integrates the region proposal step directly into the object detection algorithm itself. This architectural innovation aims to streamline the object detection process and improve overall efficiency by leveraging GPU resources effectively.

Nonetheless, to mitigate this challenge, the introduction of Region Proposal Networks (RPNs) within the Faster R-CNN framework in 2015 marks a pivotal innovation. By integrating the region proposal step directly into the object detection algorithm, RPNs aim to streamline the process, optimizing resource utilization, and enhancing overall efficiency. This architectural advancement underscores the ongoing efforts to strike a balance between accuracy and speed, heralding a promising future for real-time object detection in diverse applications.

2 Literature Review

The evolution from R-CNN to faster R-CNN represents a significant advancement in object detection algorithms, especially in speed and efficiency. A brief history of the development progression of R-CNN, Fast R-CNN, and Faster R-CNN adds valuable context to our study on object detection architectures. It will help to understand the evolution of these models, the motivations behind their development, and the improvements made over time.

2.1 R-CNN (Region-based Convolutional Network)

R-CNN was a breakthrough in object detection. It employed a multi-stage approach that involved a selective search for generating region proposals followed by a convolutional neural network for feature extraction and a support vector machine (SVM) for object classification within each region.

2.1.1 Region Proposal Generation

R-CNN began with generating region proposals using the selective search algorithm [5]. Selective search is a method for identifying potential object regions in an image based on low-level features such as color, texture, and intensity. It produces a set of bounding boxes that mostly have objects.

2.1.2 Feature Extraction

Following the generation of region proposals, R-CNN employed a pre-trained convolutional neural network to independently extract features from each region. Typically, the chosen CNN model was AlexNet, pre-trained on the ImageNet dataset specifically for image classification tasks.

2.1.3 Fine-tuning and Classification

Following feature extraction, the extracted features were input into a distinct classifier to ascertain the presence of objects within the regions. R-CNN utilized a support vector machine (SVM) [5] for this classification task. Each SVM was trained to discern whether the feature corresponded to a particular object or background.

2.1.4 Bounding Box Regression

Following classification, R-CNN conducted bounding box regression to enhance the accuracy of the detected object locations. This process adjusts the bounding boxes produced by the region proposal algorithm to align more precisely with the actual object locations within the regions.

2.1.5 Non-Maximum Suppression

Lastly, R-CNN implemented non-maximum suppression to eliminate redundant detections, ensuring that each object is detected only once.

2.1.6 Drawbacks

An initial drawback of the R-CNN architecture was its computational inefficiency during inference, primarily stemming from its sequential processing of region proposals. Processing each region proposal independently led to redundant computations and prolonged inference times.

2.2 Fast R-CNN

Fast R-CNN, a significant advancement in object detection, introduces several innovations enhancing both training and testing efficiency while improving detection accuracy. Unlike its predecessors, Fast R-CNN leverages the VGG16 network, achieving a remarkable 9x increase in training speed compared to R-CNN [6]. At test-time, Fast R-CNN demonstrates an impressive speed improvement of 213x, making it significantly faster, and more practical for real-world applications. Additionally, Fast R-CNN outperforms previous methods in terms of mean Average Precision (mAP) on benchmark datasets like PASCAL VOC 2012 [6].

Fast R-CNN tackles the computational inefficiencies of R-CNN by introducing a unified architecture that consolidates region proposal generation, feature extraction, and object classification into a single network. This approach dramatically reduces redundant computations, and accelerates the inference process.

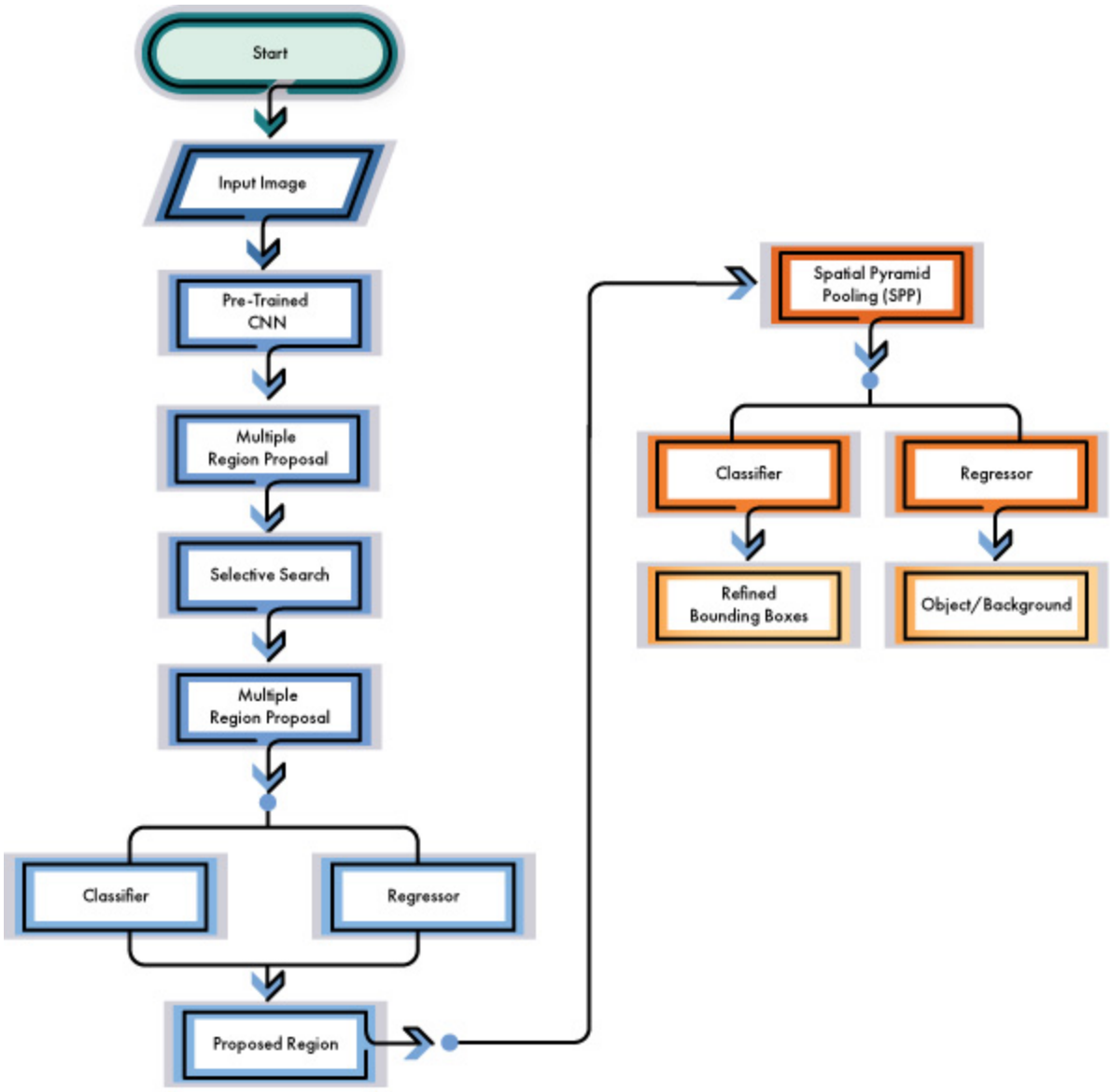

Figure 2.1: Flowchart of R-CNN

2.2.1 Architecture

Expanding upon its architecture, Fast R-CNN incorporates several key components to achieve its performance gains. Fast R-CNN addresses the speed and efficiency limitations of R-CNN by proposing a unified architecture that integrates region proposal generation, feature extraction, and object classification into a single network.

Firstly, it utilizes a Region Proposal Network (RPN) to generate region proposals directly from the convolutional feature maps, eliminating the need for external proposal methods like selective search. The RPN functions by sliding a small grid (essentially a compact CNN) across the evolving feature map to predict the spatial dimensions (bounding boxes) and associated probability scores for objects within each sliding window. This streamlines the detection process and enhances efficiency. Furthermore, Fast R-CNN introduces the Region of Interest (RoI) pooling layer [6], allowing feature extraction from region proposals of varying sizes without the need for expensive resizing operations. This enables precise alignment of features with the corresponding regions of interest, leading to improved localization accuracy. Moreover, Fast R-CNN adopts a unified network architecture, enabling end-to-end training of both the region proposal and object detection tasks. This approach ensures better optimization, and facilitates seamless integration of different components, contributing to the overall efficiency and effectiveness of the model.

2.2.2 RoI Pooling Layer

The RoI pooling layer employs max pooling to transform features within any valid region of interest into a compact feature map with a fixed spatial extent of H $\times$ W (e.g., 7 $\times$ 7) [6], where H and W are layer hyper-parameters independent of any specific RoI. A more efficient training strategy leverages feature sharing throughout the process. During Fast R-CNN training, stochastic gradient descent (SGD) minibatches are hierarchically sampled: N images are initially sampled, followed by R/N RoIs from each image. Notably, both during forward and backward passes, ROIs from the same image share computation and memory. Fixed-size feature maps yielded from RoI pooling are input to fully connected layers for object classification and bounding-box regression. Feature classification aims to detect features and assign class probabilities for each proposed location, while bounding box regression seeks to enhance the localization accuracy of detected features. Fast R-CNN introduces multitasking loss functions that amalgamate classification and bounding box regression losses. This enables joint training of both tasks, ensuring the network learns to predict object location and precise bounding boxes concurrently.

2.2.3 Advantages of Fast R-CNN

In contrast to R-CNN, which comprises multiple independent steps (such as region proposal generation, feature extraction, classification, and bounding box regression), Fast R-CNN inte-

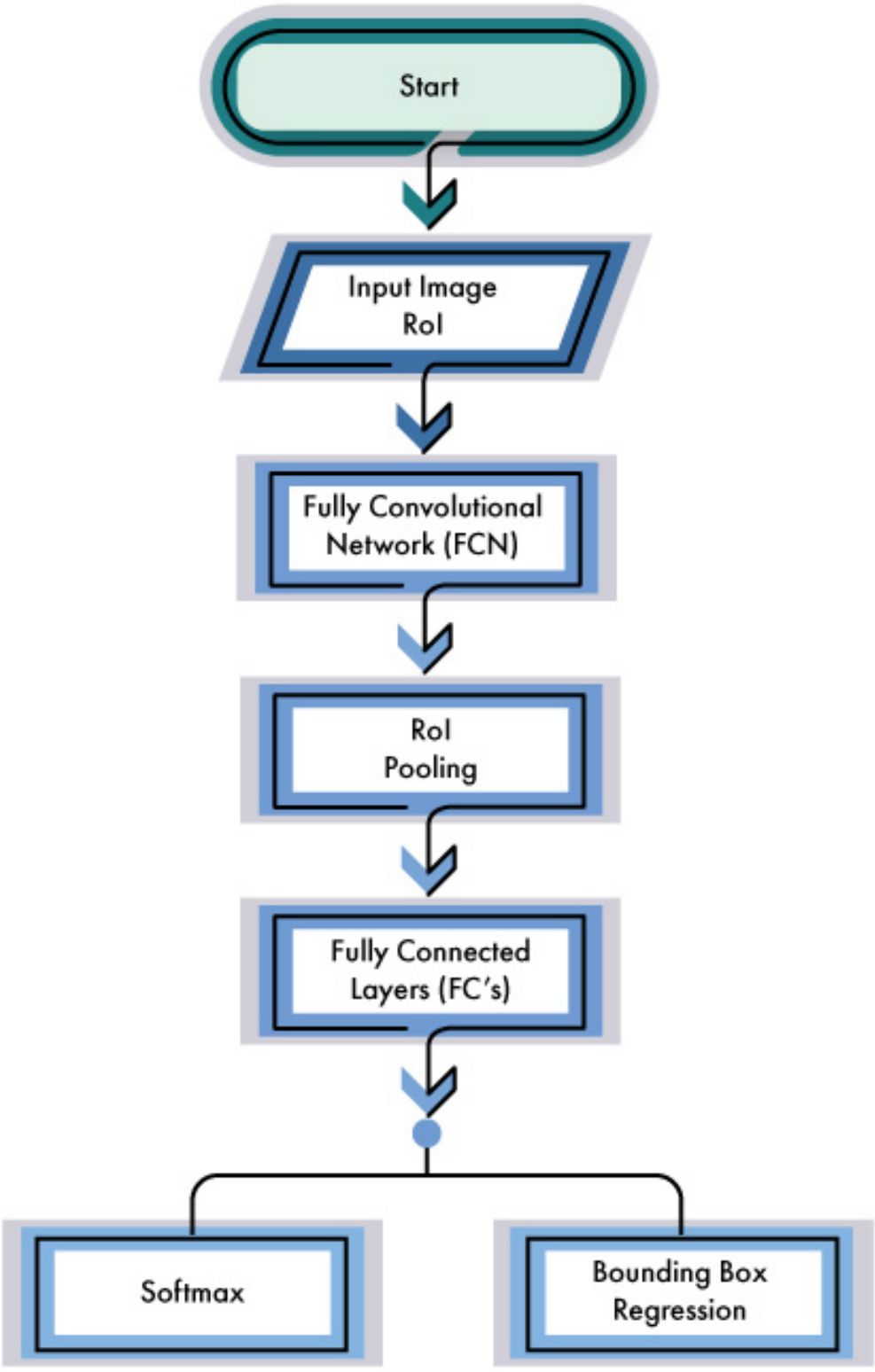

Figure 2.2: Flowchart of Fast R-CNN

grates these processes within a unified network architecture. This enables end-to-end training of the entire pipeline, leading to improved optimization and potentially higher accuracy. Fast R-CNN achieves variable sharing across all spatial dimensions of an image. Unlike R-CNN, which extracts features independently for each image, Fast R-CNN performs feature extraction on the entire image only once. This shared computation across convolutional features significantly minimizes redundant calculations, expediting the inference process.

2.3 Faster R-CNN

Compared to preceding methodologies, Faster R-CNN achieves more efficient object recognition by seamlessly integrating regional proposal steps directly into the network architecture. Subsequent advancements have expanded upon this framework, establishing it as a fundamental component in object recognition. The architecture commences with a feature extraction backbone, typically a convolutional neural network (CNN) pre-trained on extensive datasets such as ImageNet. This backbone extracts pertinent features from the input image.

2.3.1 Region Proposal Network (RPN)

Faster R-CNN further improves the efficiency of object detection by introducing the Region Proposal Network (RPN), which generates region proposals directly from the convolutional feature maps. It eliminates the need for external region proposal methods like Selective Search, resulting in faster and more accurate proposal generation.

The feature maps derived from the backbone network are input to a Region Proposal Network (RPN) [1]. The RPN, a compact fully convolutional network, employs a sliding window approach (typically 3x3) over the feature maps to generate region proposals. The RPN yields a collection of bounding box proposals accompanied by objectness scores, indicating the probability of containing an object. These proposals are generated by leveraging predefined anchor boxes of varying scales and aspect ratios.

2.3.2 Anchor Boxes

The RPN generates region proposals by predicting offsets and scales for a predefined set of anchor boxes at each spatial position in the feature maps. These anchor boxes, varying in size and aspect ratio, serve as reference boxes for the proposal generation process.

2.3.3 Region of Interest (RoI)

Region of Interest (RoI) refers to a specific area or region within an image that is selected for further analysis or processing. In the context of object detection and image segmentation tasks, RoIs typically represent regions where objects of interest are located.

2.3.4 Classifier and Bounding Box Regressor

The RoI-pooled or RoIAlign features are separately input into branches for classification and bounding box regression.

- **Classification:** The features traverse a classifier (e.g., fully connected layers followed by softmax) to predict the probability of each region proposal belonging to different object classes. Bounding Box Regression: Another set of fully connected layers predicts refined bounding box coordinates for each region proposal.
- **Loss Function:** The network undergoes end-to-end training using a multi-task loss function, which combines losses from the RPN (objectness score prediction and bounding box regression) with those from the classification and bounding box regression branches.

2.3.5 Non-Maximum Suppression (NMS)

Following prediction, non-maximum suppression is employed to discard redundant and overlapping detections based on their confidence scores and bounding box coordinates.

2.3.6 Output

The final output comprises a collection of object detections alongside their bounding boxes and corresponding class labels.

2.3.7 Advantages of Faster R-CNN

- **Efficient Region Proposal Generation:** The Region Proposal Network (RPN) significantly reduces the computational load of region proposal methods, enabling nearly cost-free region proposals by sharing convolutional features.
- **End-to-End Training:** RPN and Fast R-CNN can be jointly trained to share convolutional features, leading to a unified network architecture.
- **High-Quality Proposals:** RPN produces high-quality region proposals, enhancing detection accuracy compared to traditional methods like Selective Search and EdgeBoxes.

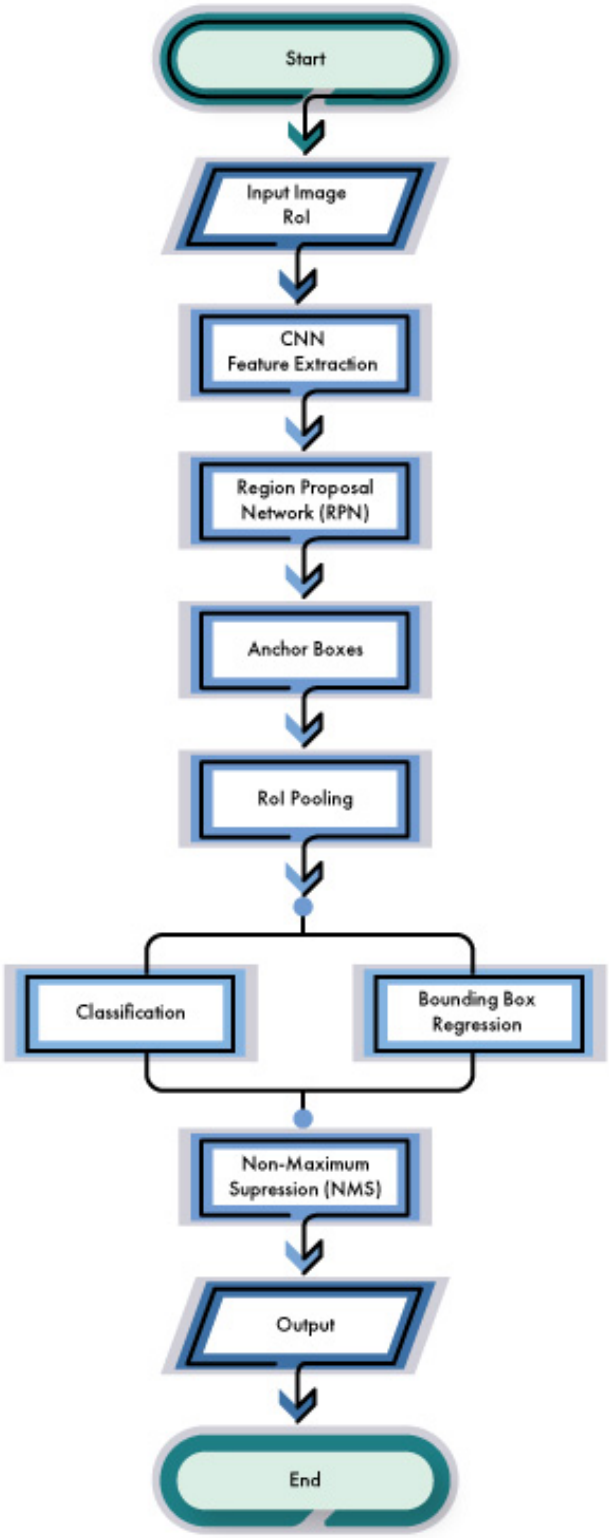

Figure 2.3: Flowchart of Faster R-CNN @ren2015fasterrcnn

- **Practical Implementation:** The proposed method achieves state-of-the-art object detection accuracy on benchmarks like PASCAL VOC [1] while maintaining a practical frame rate of 5fps on a GPU.

2.4 Mask R-CNN

Mask R-CNN extends Faster R-CNN by adding a branch to predict the segmentation mask at each Region of Interest (RoI), alongside existing branches for classification and bounding box regression. This is achieved by introducing a small, fully convolutional network on top of each RoI. Below we briefly describe the architecture components.

2.4.1 Backbone CNN

Similar to Faster R-CNN, Mask R-CNN begins with a backbone that extracts features from the input image. This backbone network is typically pre-trained on a large dataset like ImageNet to learn generic image features.

2.4.2 Region Proposal Network (RPN)

The RPN takes feature maps from the backbone network and generates region proposals using anchor boxes, refined based on their likelihood of containing objects. The region proposals generated by the RPN are subsequently used for both object detection, and instance segmentation in the Mask R-CNN framework.

2.4.3 Region of interest Align (RoIAlign)

For each region proposal generated by the RPN, features are extracted from the feature maps using RoIAlign. RoIAlign is a technique used in CNNs for object detection tasks. RoIAlign improves upon RoIPool, a previous method, by eliminating the quantization step in the pooling operation, resulting in more accurate feature extraction from regions of interest [7]. It precisely aligns features extracted from arbitrary-shaped regions with the spatial layout of the feature map, enabling more accurate object localization and segmentation.

The RoIAlign layer extracts features from each region proposal, preserving spatial information better than previous pooling methods, ensuring accurate alignment of features with the region of interest.

2.4.4 Parallel Branches

Mask R-CNN introduces parallel branches for object detection and instance segmentation.

- Object Detection Branch: Determines the class of each object within the region proposal and refines bounding box coordinates through classification and regression. – Mask Prediction Branch: Predicts segmentation masks for each object within the region proposal, generating pixel-level masks for object instances using a small fully convolutional network applied to each RoI.
- Loss Functions: Mask R-CNN employs a multi-function loss function that combines object detection (classification and bounding box regression) and instance segmentation (mask prediction) losses, weighted by hyperparameters.

$$L = L_{cls} + L_{box} + L_{mask}$$

- Training: The entire Mask R-CNN network is trained end-to-end using backpropagation with stochastic gradient descent (SGD) or other optimization algorithms, typically starting with pre-trained weights and fine-tuning for the specific task.

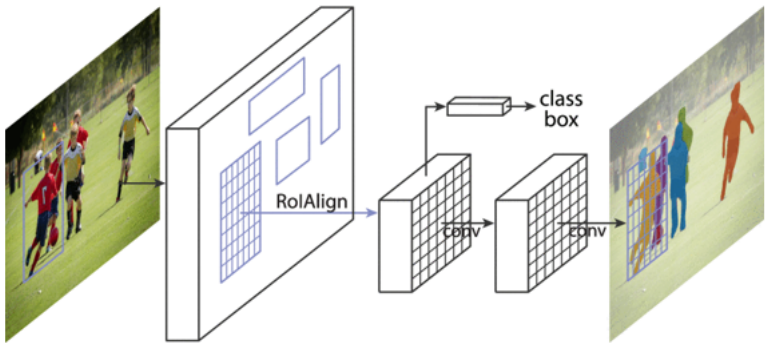

Figure 2.4: Mask R-CNN framework for Instance Segmentation @he2020maskrcnn

2.4.5 Advantages over Faster R-CNN

- Offers pixel-level segmentation alongside object detection and bounding box regression.
- Facilitates instance segmentation, segmenting each object instance in an image separately.
- Enhances understanding of object shapes and boundaries with finer detail.
- Despite increased complexity, maintains comparable speed and efficiency, particularly in scenarios necessitating instance-level segmentation.

Overall, Mask R-CNN represents a substantial advancement over previous methods by integrating object detection and instance segmentation into a unified architecture, rendering it a versatile solution for diverse computer vision applications.

2.5 YOLO (You Only Look Once)

YOLO (You Only Look Once) is a popular object detection algorithm that revolutionized the field of computer vision due to its speed and accuracy [8]. YOLO processes images in a single pass through a neural network, enabling real-time object detection. YOLOv8 is one of the iterations of the YOLO algorithm, incorporating various improvements over its predecessors to enhance performance and efficiency.

2.5.1 Architecture

The architecture of YOLO consists of a backbone network, detection head, and output layers. The backbone network, typically based on Darknet or CSPDarknet, extracts features from the input image [9]. These features are then processed by the detection head, which predicts bounding boxes, confidence scores, and class probabilities for detected objects. Earlier YOLO models utilizes anchor boxes [9] to improve localization accuracy and efficiency. Finally, the output layers produce the final detection and confidence scores.

2.5.2 Anchor-Free Approach

YOLOv8 employs an anchor-free approach [2], eliminating the need for predefined anchor boxes. This simplifies the training process and improves model flexibility.

2.5.3 Efficient Backbone

YOLO utilizes lightweight backbone networks such as CSPDarknet53, which strikes a balance between performance and computational efficiency

2.5.4 Improved Object Detection

YOLO incorporates advanced techniques such as focal loss, which helps to address class imbalance and improve object detection accuracy [10].

2.5.5 Real-Time Performance

YOLOv8 is optimized for real-time object detection applications [2], offering fast inference speeds without compromising accuracy.

2.5.6 Advantages over Mask R-CNN

Both Mask R-CNN and YOLOv8 are instance segmentation models; however, Mask R-CNN is better suited for semantic segmentation while YOLOv8 is not initially designed for that purpose. Both models are pre-trained on the MS-COCO dataset. Mask R-CNN is an improved version of Faster R-CNN and uses ResNet-101 as its backbone and has the addition of the prediction of object masks. YOLOv8 uses CSPDarknet53 as its backbone, which has 53 convolutional layers.

The main difference between YOLOv8 and Mask R-CNN is the region proposal network. YOLOv8 is an anchor free model so it does not need a region proposal network. Another difference between the two models is that Mask R-CNN is a two stage model and YOLOv8 is a one stage model. The first stage of Mask R-CNN is the application of the region proposal network and the second stage executes "bounding box regression, classification, and mask prediction" [11].

On the other hand, YOLOv8 has one stage that consists of "directly [predicting] bounding boxes and class labels for objects in an image" [11]. YOLOv8 and Mask R-CNN also differ in their loss functions. YOLOv8 uses complete intersection over union loss, distribution focal loss, and binary cross entropy. Mask R-CNN uses classification loss, regression loss, and mask loss. In addition YOLOv8 is best used for real time object detection with low latency [11].

3 Methodologies

3.1 Fast R-CNN Training

Fast R-CNN follows a single-stage training process [6]. It begins by processing the image through convolutional and max pooling layers, resulting in a convolutional feature map. Each object proposal then undergoes region of interest pooling, generating fixed-length feature vectors. These feature vectors are fed through fully connected layers, leading to two output layers: one for object class probabilities and the other for real-valued object coordinates. Fast R-CNN's training approach enables higher detection quality, updates all layers during training, and eliminates the need for disk storage for feature caching, contributing to its efficiency and accuracy [6].

3.2 Faster R-CNN Training

On the other hand, Faster R-CNN training is a two-stage process, jointly training a Region Proposal Network (RPN) and a Fast R-CNN detector [1] . In the first stage, the RPN generates region proposals by sliding a small network over the convolutional feature map, refining them, and assigning scores based on their likelihood of containing objects. These proposals serve as input for the second stage, where the convolutional feature map undergoes region of interest (RoI) pooling. This process results in fixed-length feature vectors for each proposal, which are then processed through fully connected layers to predict object class probabilities and refine object bounding box coordinates [6]. This joint training approach allows Faster R-CNN to produce high-quality region proposals and accurately classify and localize objects in images.

3.2.1 Region Proposal Network (RPN)

To produce the object proposals used during training, a Region Proposal Network (RPN) is used. The RPN takes in an image and produces a set of object proposals with an associating score. This region proposal is generated by sliding a small network over the convolutional feature map. At the center of each sliding window is a Translation-Invariant Anchor [1]. This anchor is a reference box that each proposal in that region is relative to. These anchors being translation-invariant means that regardless if the image is translated in any way the prediction of that region should not change.

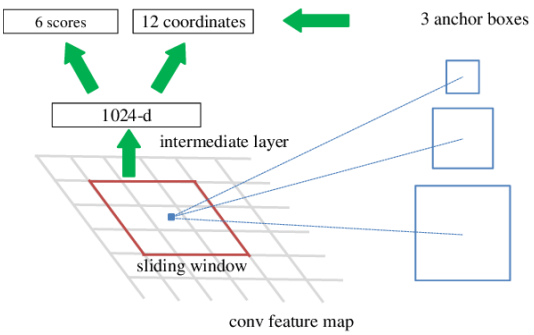

Figure 3.1: Improved anchor boxes in Faster R-CNN @zhou2022visual

Three anchor boxes are depicted within each grid, as illustrated above. The image encompasses numerous points, equivalent to a 600 x 600 image, accommodating a 38 x 38 network overlay. The blue point within the image signifies the center of the grid. Each grid center is associated with three anchor boxes and three squares of varying sizes, representing the original anchor boxes delineated in the image [12].

3.2.2 Multi-task Loss Function

RPNs use positive and negative values to assign anchors. A positive value represents a high Intersection-over-Union overlap between the anchor and the ground truth box [1]. A negative value represents the dissociation of an anchor and a certain prediction. The higher the value, the more associated the anchor is to a certain prediction. These values are used in a multi-task loss function. This function will both train for classification and perform bounding box regression. The loss function is defined as:

$$L = L_{cls} + \lambda * L_{reg}$$

Where the RPN consist of two losses.

$$L = (1/N) * \Sigma[L_{cls}(p_i, p_i*) + \lambda * L_{reg}(t_i, t_i*)]$$

The authors describe the above equation by saying [1], “Here, i is the index of an anchor in a mini-batch and pi is the predicted probability of anchor I being an object. The ground-truth label p $*$ i is 1 if the anchor is positive, and is 0 if the anchor is negative. ti is a vector representing the 4 parameterized coordinates of the predicted bounding box, and t $*$ i is that of the ground-truth box associated with a positive anchor. The classification loss Lcls is log loss over two classes (object vs. not object). For the regression loss, we use Lreg (ti, t$*$ i) = R(ti $-$ t $*$ i) where R is the robust loss function (smooth L1). The term p $*$ i Lreg means the regression loss is activated only for positive anchors (p $*$ i = 1) and is disabled otherwise (p

$*$ i = 0). The outputs of the cls and reg layers consist of {pi} and {ti} respectively. The two terms are normalized with Ncls and Nreg, and a balancing weight ."

For regression, the formula for the parameterized coordinated of the predicted bounding box:

$$\Delta x = (x' - x)/w, \quad \Delta y = (y' - y)/h, \quad \Delta w = log(w'/w), \quad \Delta h = log(h'/h)$$

- (x, y, w, h) represents the coordinates of the default anchor box,
- (x', y', w', h') represents the coordinates of the predicted bounding box,
- Δx, Δy, Δw, and Δh are the parameterized adjustments to the coordinates of the default anchor box to obtain the predicted bounding box coordinates.

3.2.3 Optimization

To optimize the data, backpropagation and stochastic gradient descent are utilized. Start with the initializing from a zero-mean Gaussian distribution. Perform back-propagation through RoI pooling layers and pass partial derivatives of the loss function concerning an activation input "xi" in the RoI pooling layer.

$$\frac{\partial L}{\partial x_i} = \Sigma_r \Sigma_j [i = i * (r, j)] \frac{\partial L}{\partial y_{rj}}$$

The authors describe this backward function as [6], "In words, for each mini-batch RoI r and each pooling output unit yrj, the partial derivative L/ yrj is accumulated if i is the argmax selected for yrj by max pooling. In back-propagation, the partial derivatives L/ yrj are already computed by the backward function of the layer on top of the RoI pooling layer". As iterations continue, weights and biases will be tuned towards values that optimize predictions.

This formula represents the accumulation of partial derivatives with respect to the output units of the RoI pooling layer during backpropagation. It describes how the gradients of the loss function with respect to the outputs of the RoI pooling layer are computed and accumulated during the backward pass of the neural network training process. It helps us understand how the network learns from its mistakes and improves its predictions over time.

3.3 YOLO (You Only Look Once)

YOLOv8 employs a cross-stage partial bottleneck with two convolutions, which is called C2f module. [13] Using YOLOv7's ELAN structure, YOLOv8 uses one standard convolutional layer and maximizes the Bottleneck module to enhance the gradient branch. [14] This keeps YOLOv8 lightweight while also capturing more gradient flow information.

YOLOv8 incorporates an anchor-free model [14] with a decoupled head. This allows for independent processing of objectness, classification, and regression tasks. The anchor-free model [13] design utilizes multiple branches, each of which focus on a single task. This is vital in improving the accuracy of the model.

YOLOv8 uses a sigmoid activation function as well as a softmax function for class probabilities. The loss functions used for bounding box loss and binary cross-entropy are the Complete IoU Loss, and the Distribution Focal Loss. Complete IoU uses Distance-IoU while taking aspect ratios into consideration [13]. Firstly, DIoU is IoU when trying to minimize the normalized distance between central points of two bounding boxes [15].

When taking aspect ratios into account, the CIoU loss fun

tion is defined as:

$$CIoU_{loss} = 1 - IoU + \frac{\rho^2(b, b^{gt})}{c^2} + av$$

Alpha here is defined as:

$$\alpha = \frac{v}{(1 - IoU) + v}$$

The optimization function for CIoU [15] is the following partial derivatives:

$$CIoU_{loss} = IoU_{loss} - v \cdot \left(\frac{d(x,y)}{c}\right) - \arctan\left(\frac{w_a}{h_a} - \frac{w_b}{h_b}\right) + v \cdot \left(\frac{C}{c^2}\right)$$

$$\frac{\partial CIoU_{loss}}{\partial x} = \frac{\partial IoU_{loss}}{\partial x} - v \cdot \frac{\partial}{\partial x}\left(\frac{d(x,y)}{c}\right) - \arctan\left(\frac{w_a}{h_a} - \frac{w_b}{h_b}\right) + v \cdot \frac{\partial}{\partial x}\left(\frac{C}{c^2}\right)$$

Distribution Focal Loss originates from Focal Loss [16], which was initially developed for one-stage object detection tasks. These tasks commonly encounter a significant class imbalance between foreground and background classes throughout the training process.

The Focal Loss function is defined as:

$$FL(p_t) = -(1-p_t)^{\gamma} \cdot \log(p_t)$$

$$p_t = \begin{cases} p & \text{when } y = 1 \\ 1-p & \text{when } y = 0 \end{cases}$$

Li [16] describes Distribution Focal Loss (DFL) as FL that "forces the network to rapidly focus on the values near label y, by explicitly enlarging the probabilities of yi and yi+1 (nearest two to y, yi y yi+1)."

The DFL function between two stages is define below:

$$\text{DFL}(S_i, S_{i+1}) = -(y_{i+1}) \log(S_i) + (y - y_i) \log(S_{i+1})$$

3.4 Metrics

[One data metric utilized in the chosen paper, and that will also be used in the analysis, is the mean average precision (mAP). This evaluation metric is typically used for object detection, and consists of multiple sub-metrics such as the confusion matrix, recall, precision, and the intersection over union. mAP is mathematically defined as, n is the number of classes, and APk is the average precision of the current class (k). Mean average precision basically takes an established ground truth box around the target and compares it to the detected box from the deep learning model, yielding an accuracy score. A higher accuracy score implies that the deep learning model is accurate with its detection.

$$mAP = (1/n) * \sum_{k=1}^{k=n} (AP_k)$$

- AP_k = the Average Precision of class k
- n = the number of classes.

4 Analysis

The demonstration utilized the ResNet backbone network in conjunction with the COCO dataset. The Microsoft Common Objects in Context (COCO) dataset [17] is a large-scale dataset for object detection, segmentation, and captioning tasks. It contains over 200,000 images, each annotated with bounding boxes around objects in various categories such as people, animals, vehicles, and household items. Additionally, each image is annotated with pixel-level segmentation masks for object instances. The COCO dataset is widely used in computer vision research for benchmarking and evaluating object detection and segmentation algorithms.

Faster R-CNN, known for its proficiency in object detection tasks, has showcased remarkable performance [1] . However, challenges persist regarding its robustness to environmental variations such as shadows, occlusions, and perspective distortions. Furthermore, mislabeling of objects remains a concern, potentially leading to incorrect identifications due to similarities with trained categories.

In addition to Faster R-CNN, Mask R-CNN was employed, utilizing a pre-trained model from the COCO dataset. This model underwent testing on both images, and a video dataset. Transitioning to YOLOv8, preparation of the data involved annotating the training set. Twenty images of tanks were manually annotated using CVAT.ai to draw bounding boxes around the targets. Subsequently, specific bounding box coordinates for YOLOv8 were obtained. Notably, all tank images for training and testing were sourced from Flickr.

To configure the dataset appropriately for YOLOv8, a yaml file was created, specifying the dataset's structure. This file included details such as the overall path to the data, the locations of the training, validation, and test sets, and the names of the classes to be identified.

Following data preprocessing, YOLOv8 training commenced using the tank training set. Initially, twenty epochs were employed with the medium model. However, during testing, Google Colab encountered stability issues, concluded in the selection of the small model, which resulted in smooth execution.

The following flowchart illustrates our process details for the custom military tank detector.

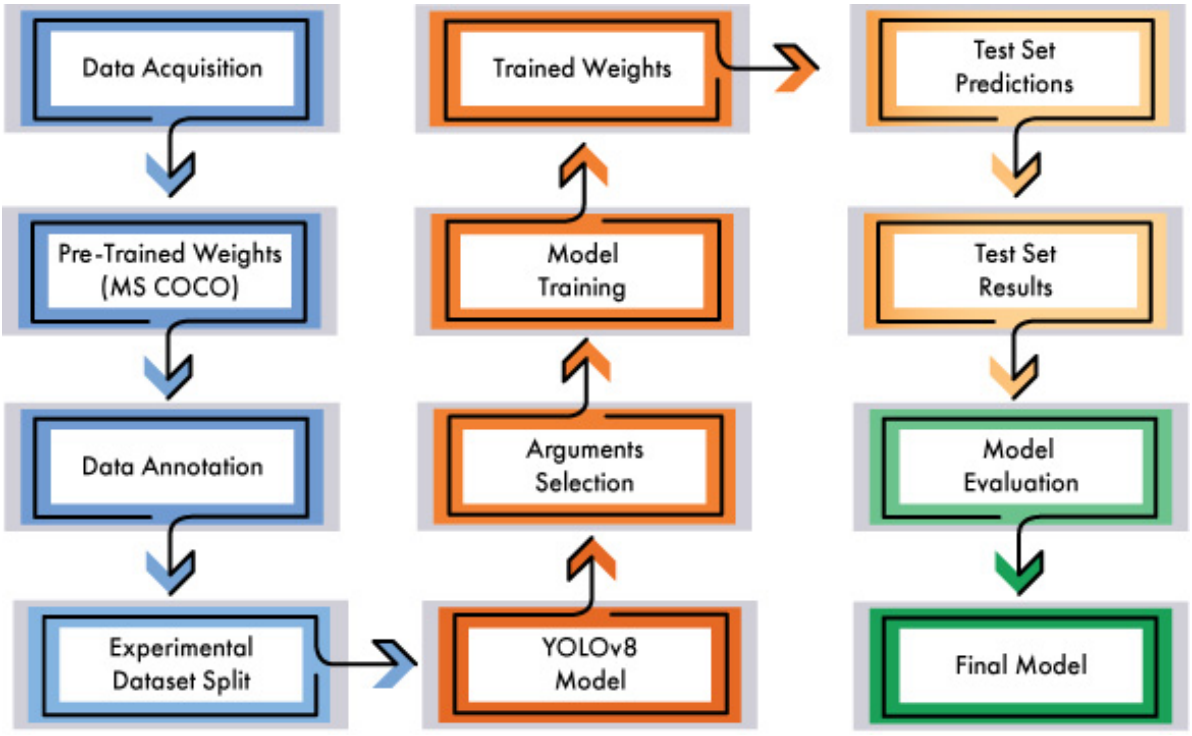

Figure 4.1: Custom Object Detector using YOLOv8

5 Results

5.1 Faster R-CNN

We used the Faster R-CNN model from TensorFlow Hub (Inception-ResNet-V2). According to the authors [18], Inception-ResNet-V2 is a deep convolutional neural network architecture introduced as part of the Inception family.

The Faster R-CNN results can be seen in images 1-4. In the first image, the model was fairly accurate in identifying individual chairs and a couple of the tables.

Figure 5.1: Faster R-CNN result 1

In the second image, Faster R-CNN correctly identified all beetles.

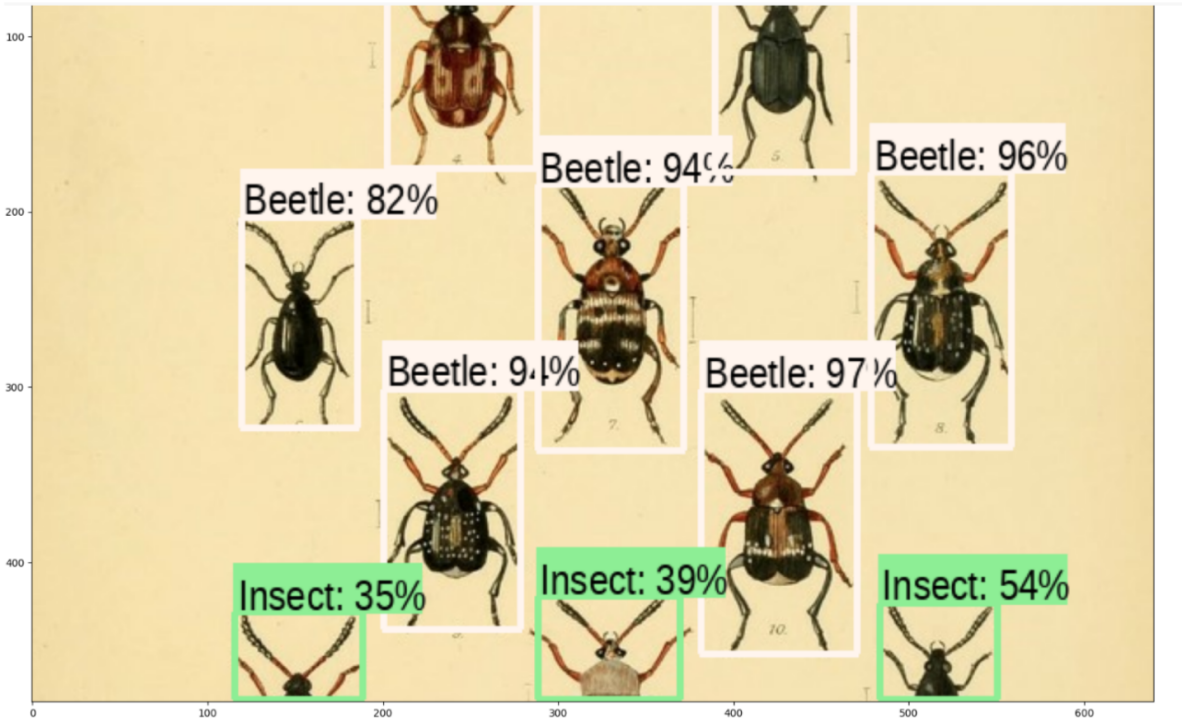

Figure 5.2: Faster R-CNN result 2

In the third image the model was able to distinguish different types of phones.

Figure 5.3: Faster R-CNN result 3

In the fourth image there seemed to be some confusion in the distinction between birds and animals in general.

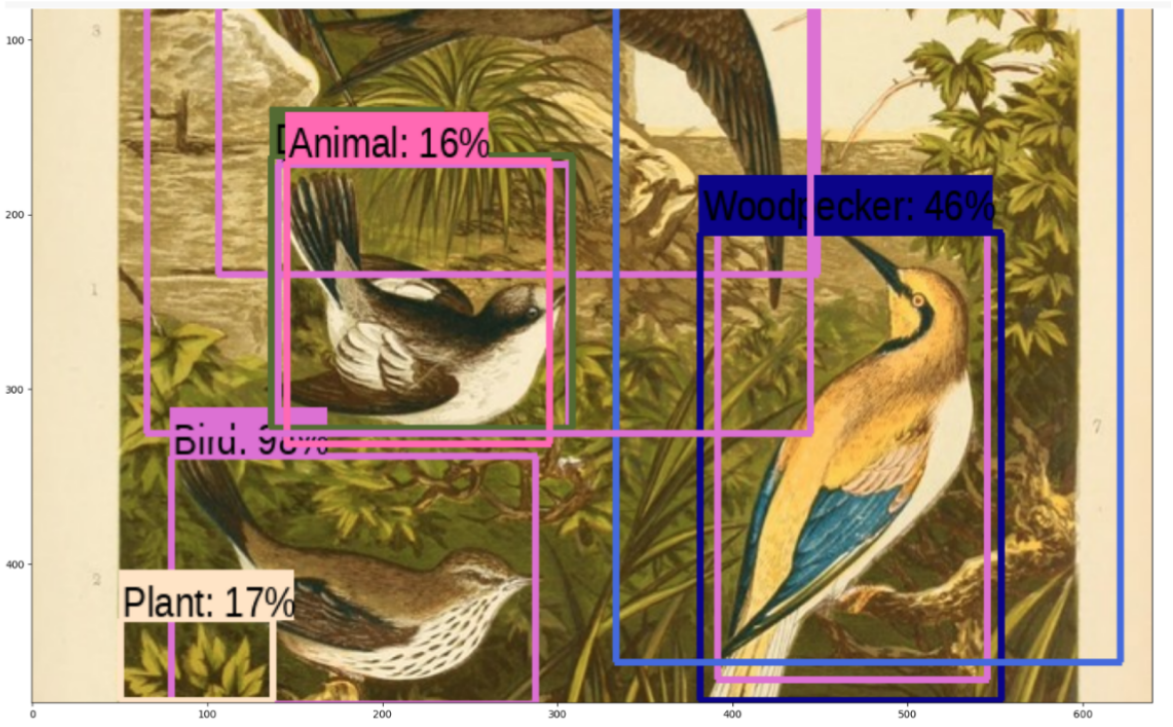

Figure 5.4: Faster R-CNN result 4

5.2 Mask R-CNN

Images 5-7 are the results from the Mask R-CNN model that was trained from the COCO dataset. Based on the results this model seems to perform much better than the Faster R-CNN model.

In image five, Mask R-CNN very accurately detected the giraffe and two zebras.

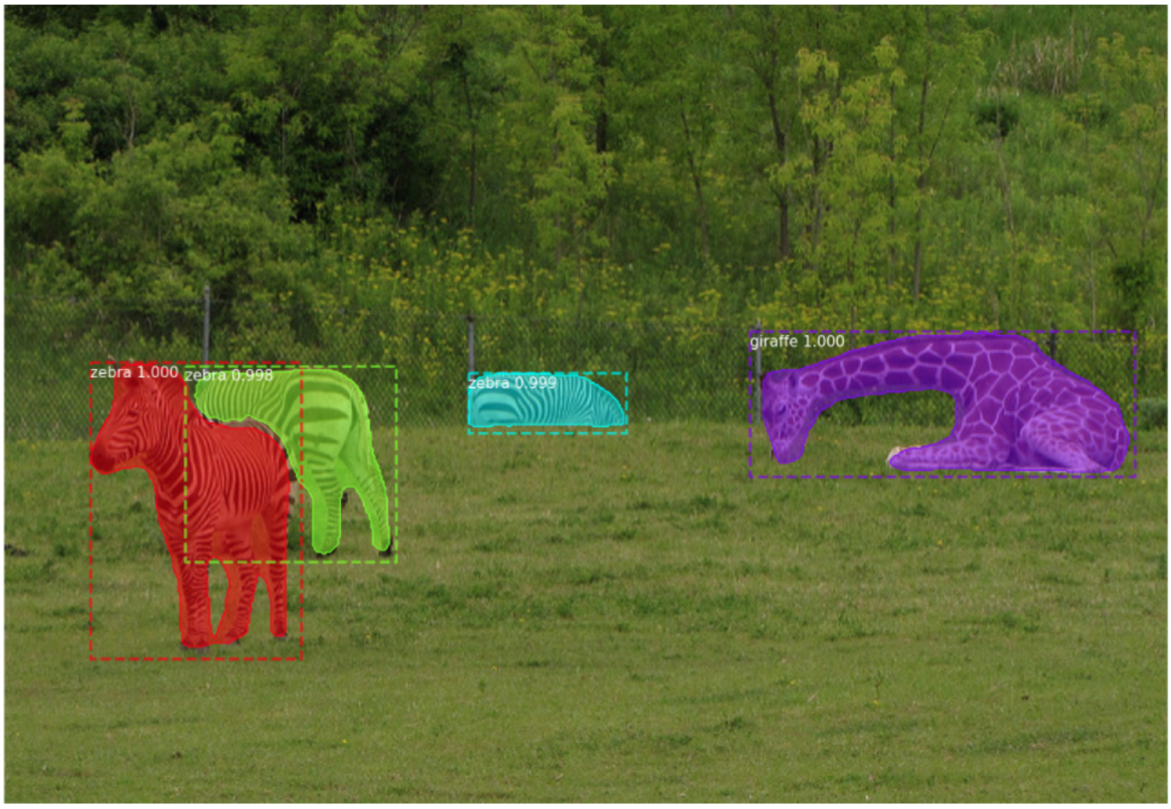

Figure 5.5: Mask R-CNN result 1

In image six, the model was able to clearly detect the objects with a high mean average precision.

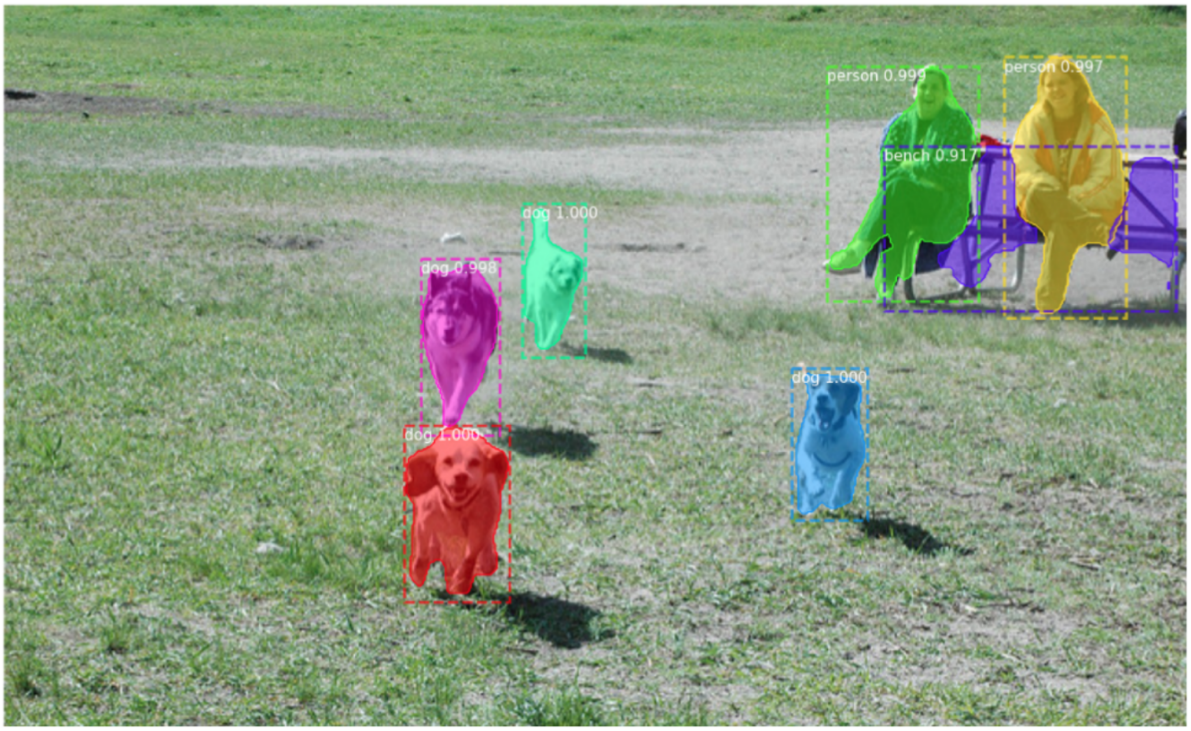

Figure 5.6: Mask R-CNN result 2

Image seven comes from a video that was fed to the Mask R-CNN model. In this frame, the model did have some trouble identifying all the objects.

Figure 5.7: Mask R-CNN result 3

5.3 YOLOv8

The last model used in our analysis was YOLOv8. This model also seemed to perform better than the Faster R-CNN model. Our YOLOv8 small (YOLOv8s) model was trained on a custom dataset that consisted of tanks and was subsequently tested on test photos of tanks. Images 8-10 show the results of the YOLOv8s model. The model accurately identified all of the tanks.

Figure 5.8: YOLOv8 result 1

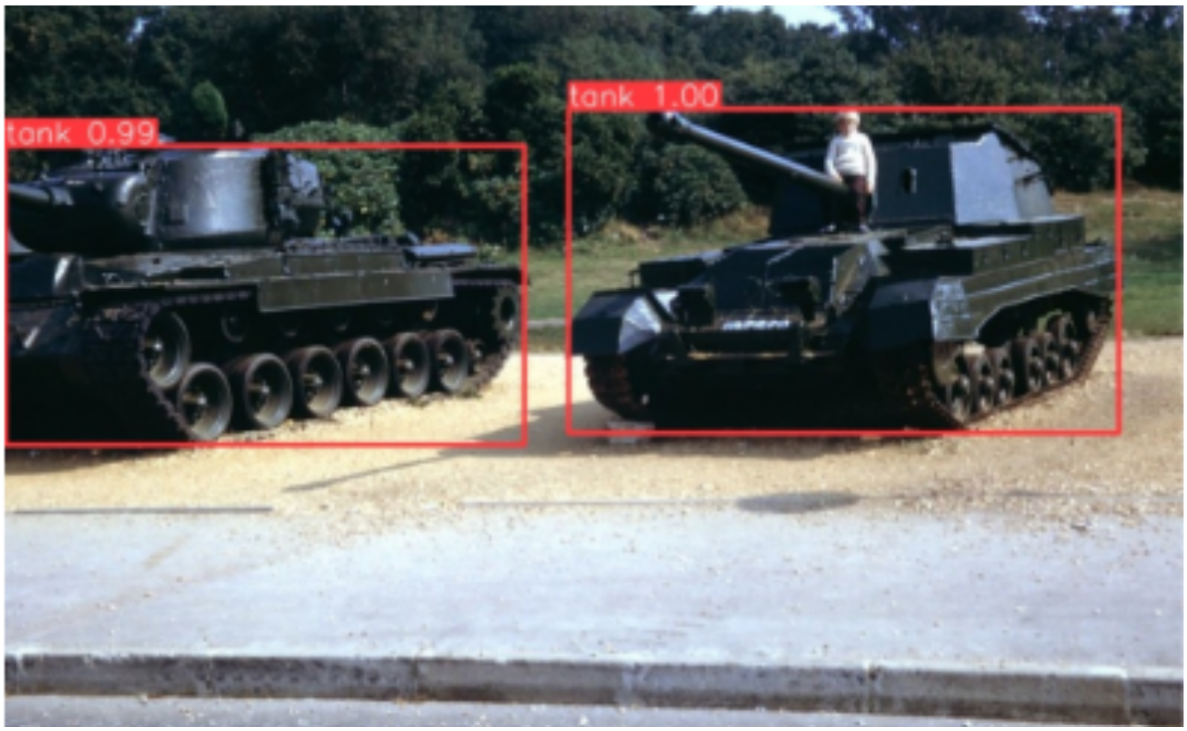

Figure 5.9: YOLOv8 result 2

Figure 5.10: YOLOv8 result 3

The last test was a real-time capture using a webcam using the original YOLOv8 Nano. The pre-trained model was able to detect and label correctly all the class objects in the room.

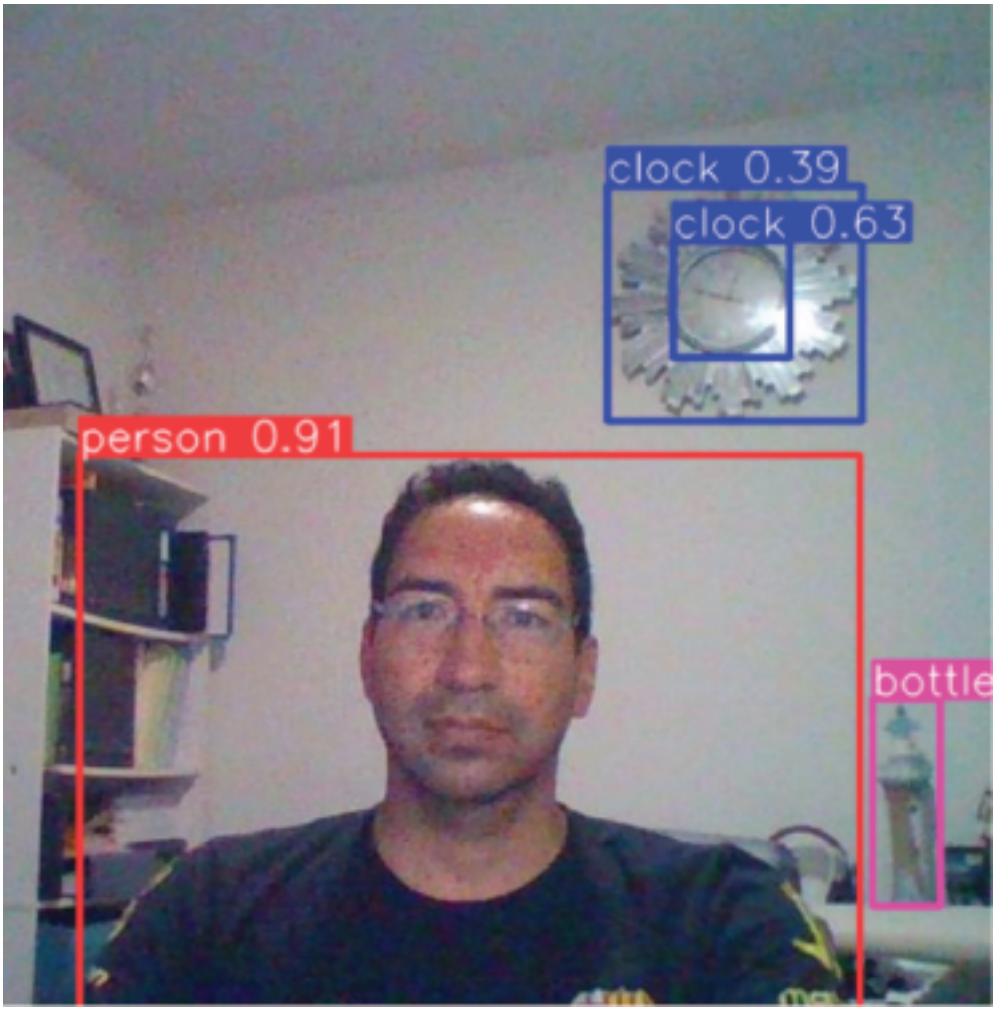

Figure 5.11: Real-Time Object Detection and Classification

6 Conclusion

In conclusion, our aim in this project was to thoroughly examine the Faster R-CNN method, alonside an exploration of alternatives like Mask R-CNN and YOLOv8. Our analysis clearly shows that both Mask R-CNN and YOLOv8 outperform Faster R-CNN in terms of spotting objects accurately, and inference time.

YOLOv8 stands out for its ability to detect objects in real-time situations. On the other hand, Mask R-CNN shines when it comes to accurately identifying objects in images and videos, especially when it requires understanding detailed object shapes and boundaries.

However, it's worth noting that Mask R-CNN's performance might rely heavily on having enough computing power because it needs more processing resources compared to YOLOv8. This highlights the importance of not only considering performance but also practical concerns and limitations when choosing a model for real-world applications.

In conclusion, our research sheds light on the strengths and weaknesses of different object detection methods. By providing practical insights, we believe that YOLOv8 algorithm offers a good trade-off between detection performance, and inference time. As technology progresses, it becomes increasingly important to refine these methodologies to meet the ever-changing demands of various real-world applications.

References

[1] 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*, 2015, pp. 91–99.

[2] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLOv8." 2023. Available: https://github.com/ultralytics/ultralytics

[3] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *International Journal of Computer Vision*, vol. 104, no. 2, pp. 154–171, 2013.

[4] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *European conference on computer vision*, 2014, pp. 391–405.

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *2014 IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587. doi: 10.1109/CVPR.2014.81

[6] R. Girshick, "Fast r-CNN," in *2015 IEEE international conference on computer vision (ICCV)*, 2015, pp. 1440–1448. doi: 10.1109/ICCV.2015.169

[7] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-CNN," in *Proceedings of the IEEE international conference on computer vision (ICCV)*, 2017, pp. 2961–2969. doi: 10.1109/ICCV.2017.322

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

[9] A. Bochkovskiy *et al.*, "YOLOv4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[10] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9627–9636, 2019.

[11] Z. Ameli, S. J. Nesheli, and E. N. Landis, "Deep learning-based steel bridge corrosion segmentation and condition rating using mask RCNN and YOLOv8," *Infrastructures*, vol. 9, no. 1, 2024, doi: 10.3390/infrastructures9010003. Available: https://www.mdpi.com/2412-3811/9/1/3

[12] Y. Zhou, X. Wang, and L. Zhang, "Visual identification and pose estimation algorithm of nut tightening robot system." Feb. 2022. doi: 10.21203/rs.3.rs-1391065/v1

[13] J. Terven, D.-M. Córdova-Esparza, and J.-A. Romero-González, "A comprehensive review of YOLO architectures in computer vision: From YOLOV1 to Yolov8 and yolo-nas," *Machine Learning and Knowledge Extraction*, vol. 5, no. 4, pp. 1680–1716, 2023, doi: 10.3390/make5040083

[14] X. Wang, H. Gao, Z. Jia, and Z. Li, "BL-Yolov8: An improved road defect detection model based on yolov8," *Sensors*, vol. 23, no. 20, p. 8361, 2023, doi: 10.3390/s23208361

[15] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IOU loss: Faster and better learning for bounding box regression," in *Proceedings of the AAAI conference on artificial intelligence*, 2020, pp. 12993–13000. doi: 10.1609/aaai.v34i07.6999

[16] X. Li, W. Wang, X. Hu, J. Li, J. Tang, and J. Yang, "Generalized focal loss V2: Learning reliable localization quality estimation for dense object detection," in *2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2021. doi: 10.1109/cvpr46437.2021.01146

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2016, pp. 770–778.

[18] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," in *Proceedings of the thirty-first AAAI conference on artificial intelligence (AAAI-17)*, 2017, pp. 4278–4284.

A Contributions

A.0.1 Hector Gavilanes – Project Lead

Prepared the GitHub repository and pages for hosting the research paper online. Proofread and edited all chapters for the final version of the research paper. Researched and implemented code for Faster R-CNN, Mask R-CNN, and YOLOv8 methodologies. Analyzed, experimented, and made inferences during training, validation, and test phases. Developed a custom class label for detecting military tanks using YOLOv8. Evaluated the performance between various object detection and segmentation models. Kept communication with the professor, and peers.

A.0.2 Jacob Knight – Deep Learning Architect

Wrote about the different methodologies used, such as Region Proposal Network, Translation-Invariant Anchor, and back-propagation. Conducted peer reviews of others' work.

A.0.3 Kyle Knuth – Data Scientist

Wrote the objectives, introduction, metrics, analysis, and results. Helped gather training and testing images for the YOLOv8 model and assisted with the annotation of the training data. Assisted in the training and implementation of the YOLOv8 model.

A.0.4 Sai Meghana – Software Engineer

Researched and wrote the literature review chapter, and made comparison between different methodologies. Drafted flowcharts to explain the architecture of each model. Assisted in writing methodologies chapter.

B Resources

For your convenience, we have provided the following links accompanying the book report referencing the code notebooks used for the projects, the report in PDF format, and web applications.

B.1 Repositories

- Online Report - GitHub repository
- Project code - Google Colab
- Report - PDF version

B.2 Web Applications

- RCNN - Webcam app
- RCNN - YOLOv8 Models