Improving the Energy Efficiency of Real-time DNN Object Detection via Compression, Transfer Learning, and Scale Prediction

Debojyoti Biswas, M M Mahabubur Rahman, Ziliang Zong, Jelena Tešić

Outline

- Introduction
- Related Work
- Motivation
- Contributions
- Dataset
- Challenges
- Methodology
- Results
- Conclusions and Future Work

Introduction

- Deep Neural Networks (DNNs) have successfully solved many challenging problems in computer vision.
- They have been widely used to support various exciting and powerful applications such as Object Detection, Image Segmentation, Object tracking, etc.
- DNNs are typically computation intensive, memory demanding, and power hungry.
- Lightweight models are required to deploy on resource-constrained systems such as NVIDIA TX2 or Jetson Nano.
- The goal of this project is to reduce power and memory consumption and minimize the GFLOPs of the state-of-the-art object detection model.

Related Work

There have been several works to make object detection from aerial imagery.

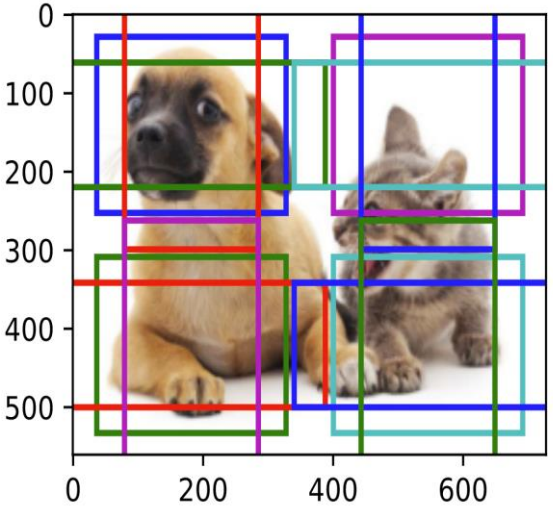
There are mainly two categories of object detection models.

One-stage detection: One-stage detectors propose regions using a different scale and aspect ratio of anchors. They take input and give predictions in one pass, such as YOLOv4, YOLOv5, SSD, etc.

Lightweight one-stage models, such as:

1. Improved YOLOv4
2. Lightweight RetinaNet

Two-stage detection: The two-stage detectors are mainly Region Proposal Network (RPN) based, such as RCNN, Faster-RCNN, etc.



Source: https://blog.csdn.net/qq_56591814/article/details/124916601

Motivation

The incremental applications of UAV-based surveillance systems.

The necessity of low power consumption models suitable for mobile computing.

Analyze the trade-off between detection performance and other critical metrics for object detection in resource-constrained conditions.

Source: https://www.urbanairmobilitynews.com/inspection-and-surveillance/india-national-highways-authority-makes-drone-survey-mandatory-for-all-national-highways-projects/

Contributions

- Application-specific reduced scale prediction for surveillance and rescue systems.
- Backbone compression using energy-aware bottlenecks and the cross-stage partial (CSP) network.
- 1 X1 convolution at task-specific layers for channel shrinkage.
- Faster inference time with the reduced parameters of the models.
- Introduce detection performance, power consumption, and memory consumption metrics to present an in-depth analysis of our proposed models.

Dataset

We have worked with the DIOR dataset for our experiments.

The dataset contains 23,462 images and 192,472 instances.

The DIOR dataset comes with four different angles and characteristics.

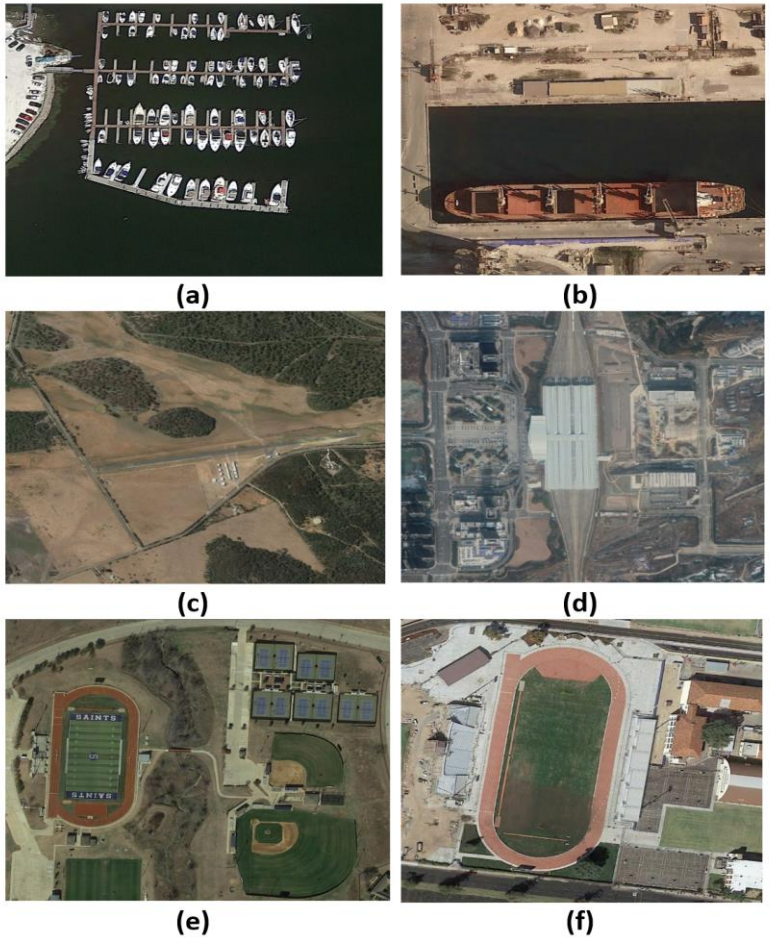
1. Large-scale on the object categories.

2. Large range of object size variations.

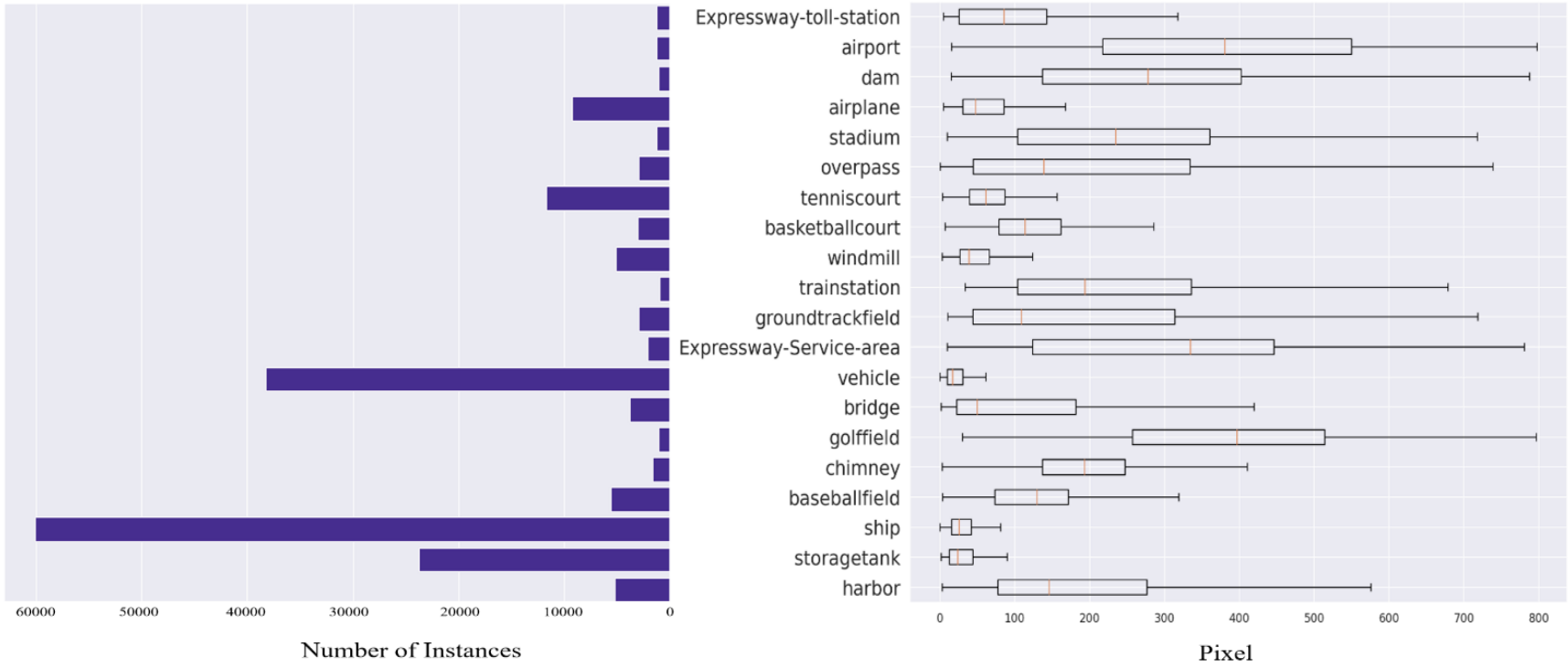
3. Variation in terms of geographical area and weather conditions.

4. High inter-class similarity and intra-class diversity.

The training set contains 22,450 images, and the validation set contains 1012 images.

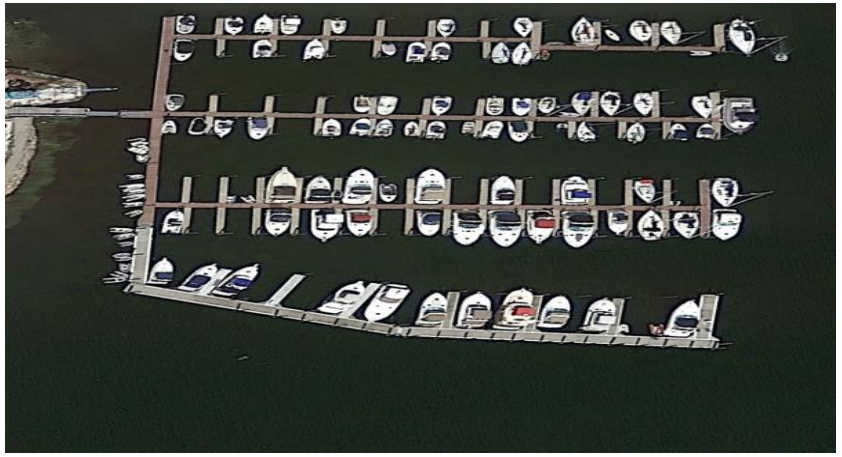


Dataset size and instances distribution



Challenges in aerial object detection

- Object with small size
- Densely packed objects
- Large variety in object orientation
- Imbalance Easy and Hard Examples
- Uniform features across the object



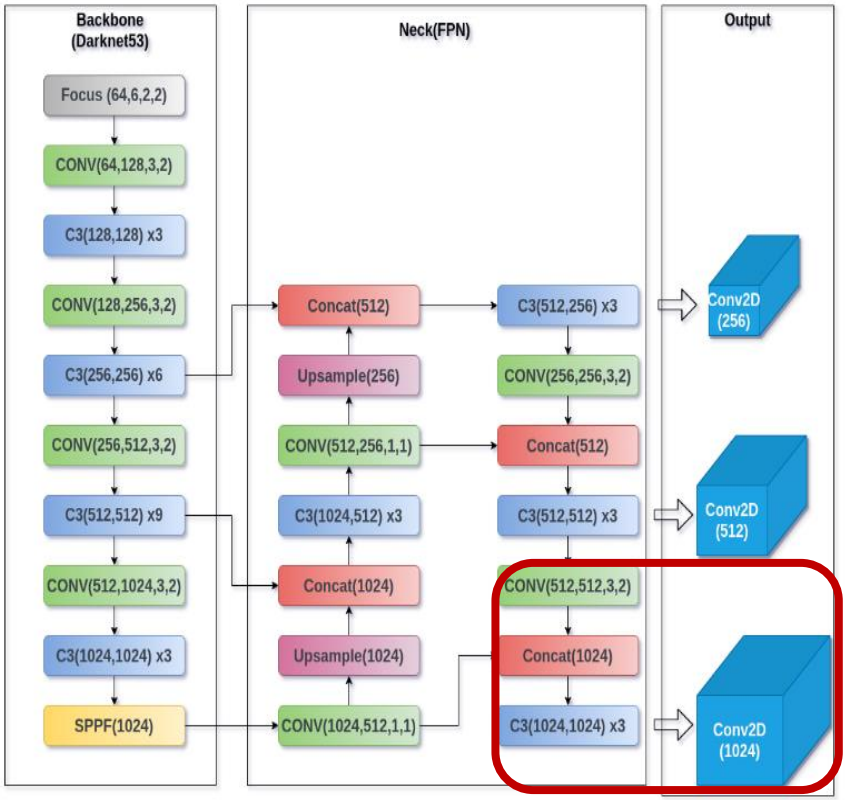
Methodology: YOLOv5 baseline architecture

- There are three major parts in the YOLOv5 model:

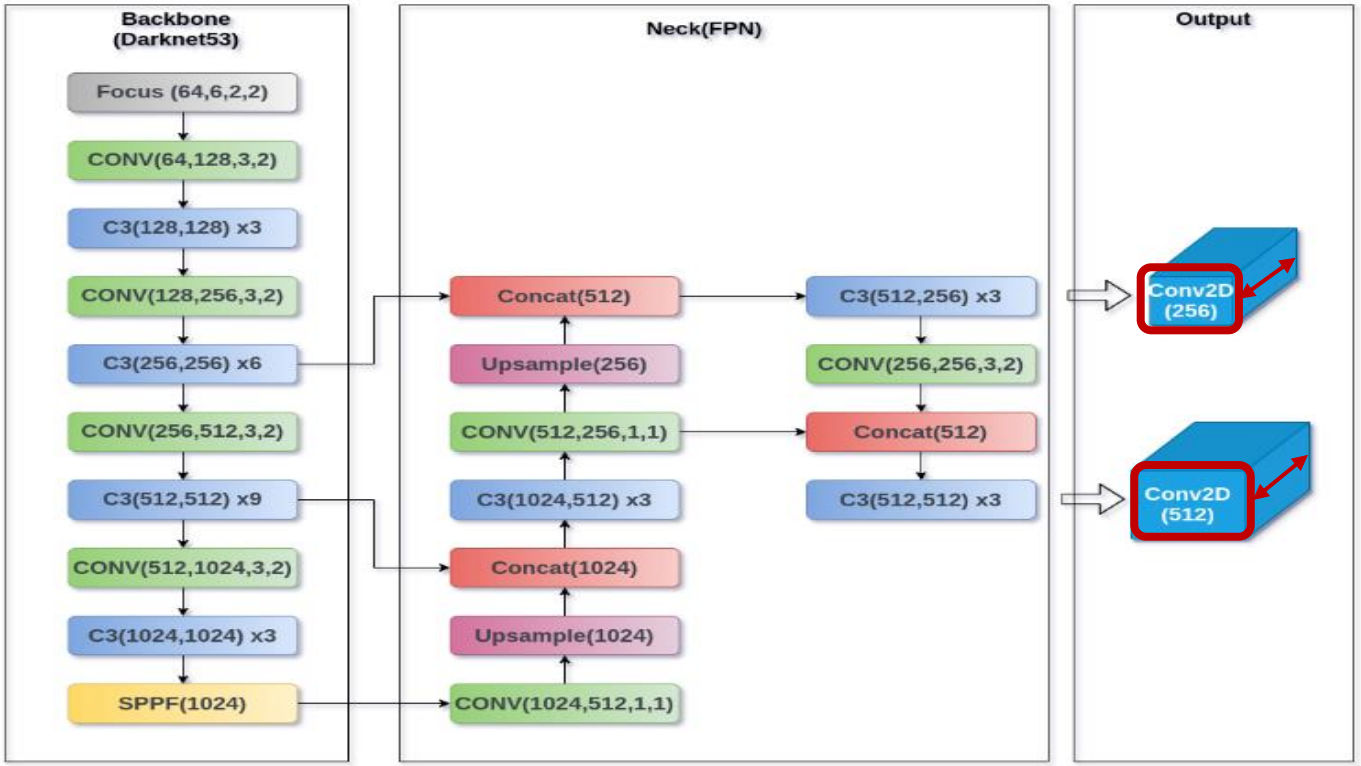
1. Backbone: The Darknet53 consists of Convolution Layers (Conv), CSP Bottleneck Layers(C3), and Spatial Pyramid Pooling Layers (SPP).

2. Neck: The Neck in this architecture is Feature Pyramid Network (FPN).

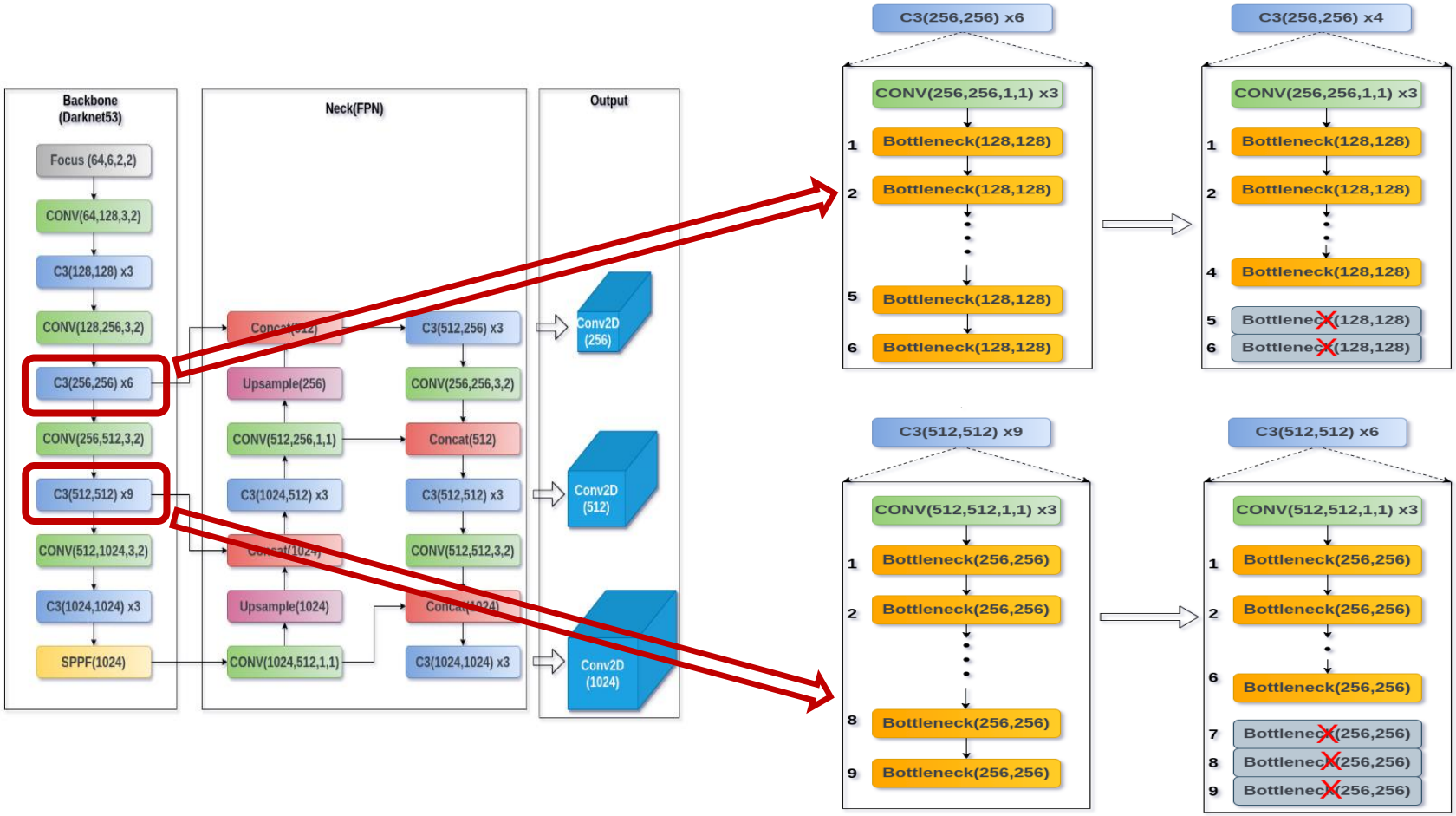
3. Detection Head: The extracted features from the FPN are used for multi-scale detection.



Methodology: YOLOv5 Reduced scale prediction



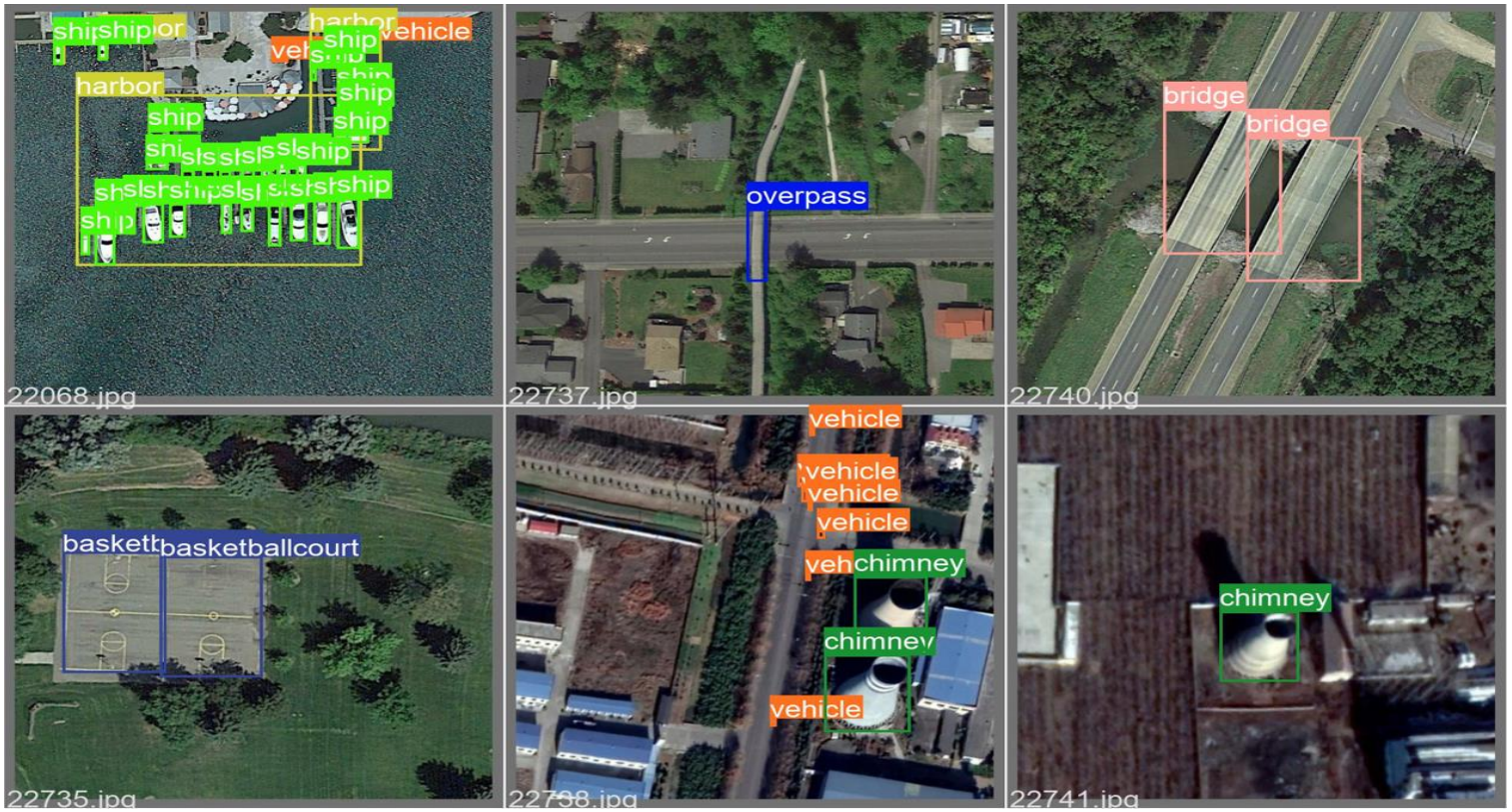
Methodology: Backbone Compression



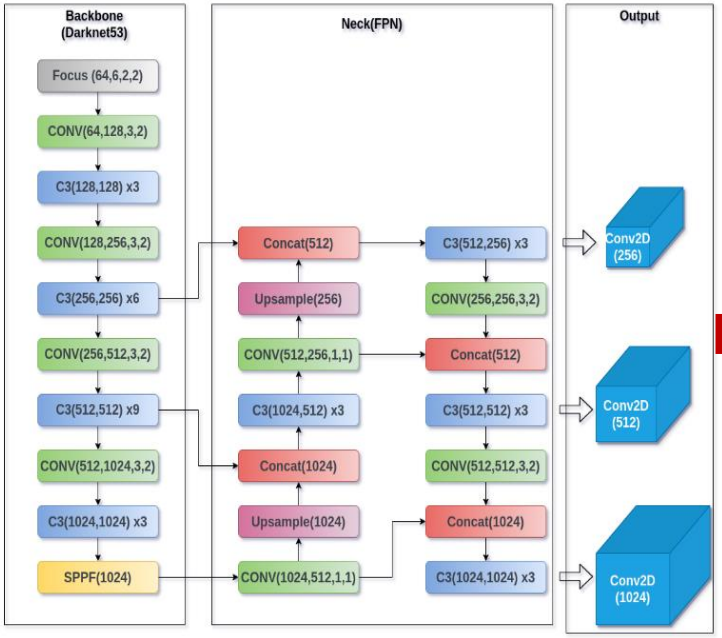
System Specification

System	Configuration
Operating System	18.04
CPU	11th Gen Intel® Core™ i9-11900K @ 3.50GHz × 16
GPU	NVIDIA Corporation GP102 [TITAN Xp]
GPU Memory	12GB
RAM	125GB

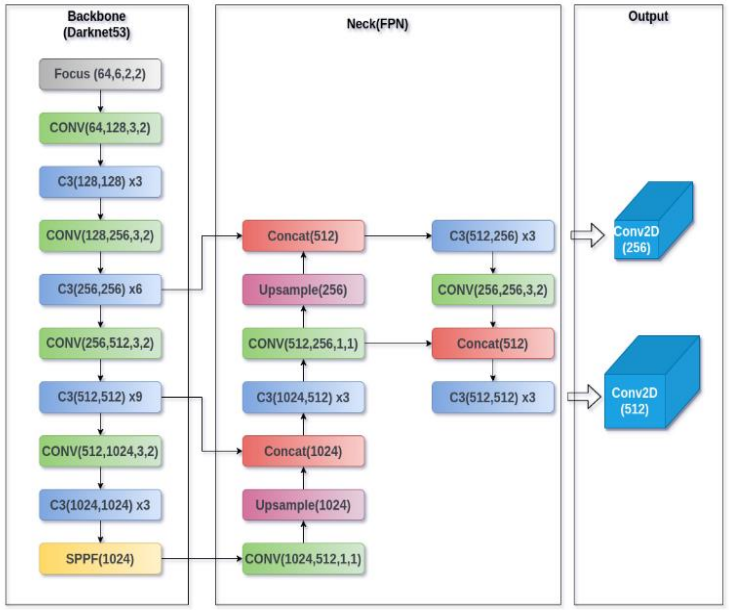
Results: Ground truth DIOR



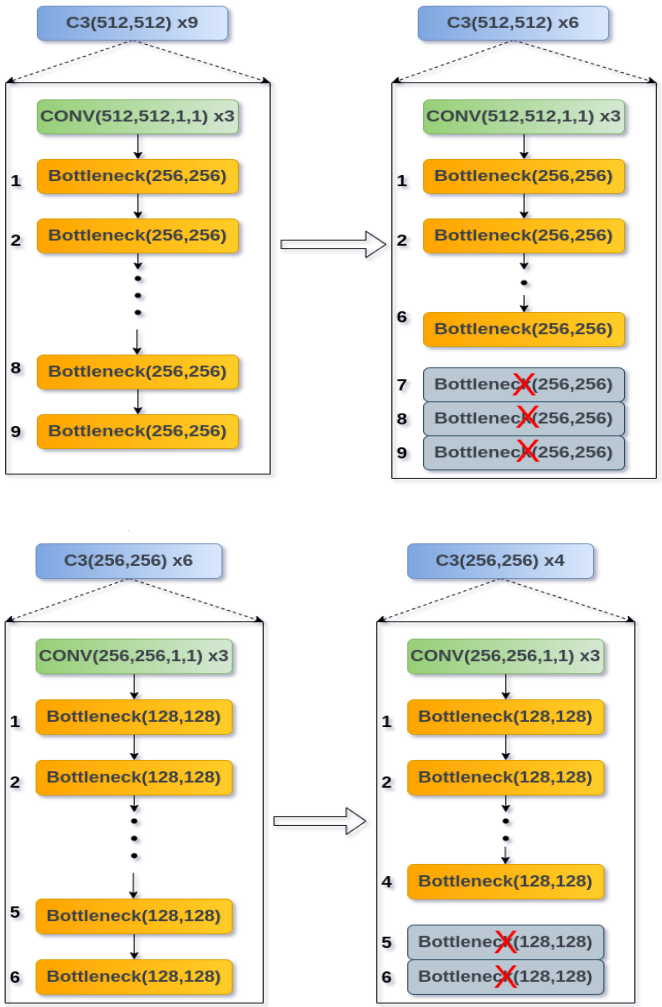
Results: Detection from baseline model



Results: Detection from small and medium-scale prediction

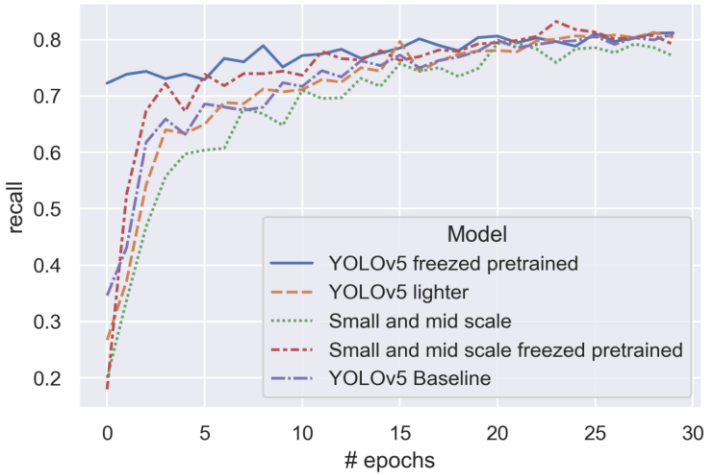


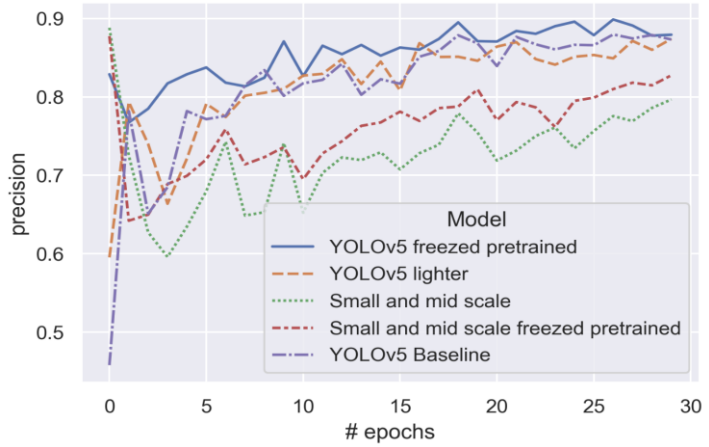
Results: Detection for YOLOv5 lighter model



Results: Summary of performance

Model name	Precision	Recall	mAP_0.5	mAP_0.5:0.95	Inference Time (ms)
YOLOv5 Baseline	0.8731	0.8098	0.8662	0.6381	76.2
Small and mid scale	0.7966	0.7857	0.8188	0.5649	68.21
YOLOv5 lighter	**0.8740**	**0.8041**	**0.8651**	**0.6310**	**69.6**
YOLOv5 freezed pretrained	0.8792	0.8118	0.8706	0.6386	70.3
Small and mid scale freezed pretrained	0.8274	0.7911	0.8403	0.5871	69.13





Power measurement tool

- P4400 p3 Kill-A-Watt meter.
- Using this device, we measured the power consumption from the CPU line.

Results: Summary of power consumption

Model name	GFLOPs	Power (watt) approx.	Number of parameters	GPU memory usage (GB)
YOLOv5 Baseline	108.3	384	46,240,609	3.83
Small-Medium scale	98.3	**374**	**33,83,1702** **~(-26%)**	3.48
YOLOv5 Lighter	**97.7**	380	43,942,753 **~(-5%)**	3.63
YOLOv5 freezed pretrained	**97.7**	377	43,942,753	**2.54**
Small and mid scale freezed pretrained	98.3	**372**	43,942,753	**2.42**

Conclusions and Future work

- Backbone compression, output channel shrinkage, and reduced scale prediction can reduce computational expenses significantly.
- Our proposed model achieved **63.10% mAP** and inference time of **69.6ms/frame.** [15 frames per sec]
- This work reduces the number of GFLOPs, memory consumption, power consumption, and the number of learnable parameters of the YOLOv5 baseline model without dropping the performance significantly.
- In the future, we plan to reduce the GFLOPs by careful selection of anchors and reduced IOU calculations and deploy our models in remote constrained devices such as Nvidia TX2 and Nvidia Jetson Nano with GPU memory of 6GB and 4GB respectively.

Questions?