TUGAS AKHIR - EC184801

OPTIMISASI PENDETEKSIAN OBJEK KECIL YOLOv7 UNTUK MENDETEKSI OBJEK-OBJEK *AIRBORNE*

Dion Andreas Solang
NRP 0721 19 4000 0039

Dosen Pembimbing

Reza Fuad Rachmadi, S.T., M.T., Ph.D

NIP 19850403201212 1 001

Dr. I Ketut Eddy Purnama, S.T., M.T.
NIP 19690730199512 1 001

Program Studi Strata 1 (S1) Teknik Komputer
Departemen Teknik Komputer
Fakultas Fakultas Teknologi Elektro dan Informatika Cerdas
Institut Teknologi Sepuluh Nopember

Surabaya
2023

[This page intentionally left blank]

TUGAS AKHIR - EC184801

OPTIMISASI PENDETEKSIAN OBJEK KECIL YOLOv7 UNTUK MENDETEKSI OBJEK-OBJEK *AIRBORNE*

Dion Andreas Solang
NRP 0721 19 4000 0039

Dosen Pembimbing

Reza Fuad Rachmadi, S.T., M.T., Ph.D

NIP 19850403201212 1 001

Dr. I Ketut Eddy Purnama, S.T., M.T.
NIP 19690730199512 1 001

Program Studi Strata 1 (S1) Teknik Komputer
Departemen Teknik Komputer
Fakultas Fakultas Teknologi Elektro dan Informatika Cerdas
Institut Teknologi Sepuluh Nopember
Surabaya
2023

[This page intentionally left blank]

FINAL PROJECT - EC184801

YOLOv7 SMALL OBJECT DETECTION OPTIMIZATION TO DETECT AIRBORNE OBJECTS

Dion Andreas Solang
NRP 0721 19 4000 0039

Advisor

Reza Fuad Rachmadi, S.T., M.T., Ph.D

NIP 19850403201212 1 001

Dr. I Ketut Eddy Purnama, S.T., M.T.
NIP 19690730199512 1 001

Undergraduate Study Program of Computer Engineering

Department of Computer Engineering

Faculty of Faculty of Intelligent Electrical and Informatics Technology

Sepuluh Nopember Institute of Technology

Surabaya

2023

[This page intentionally left blank]

ABSTRAK

Nama Mahasiswa : Dion Andreas Solang

Judul Tugas Akhir : OPTIMISASI PENDETEKSIAN OBJEK KECIL YOLOv7 UNTUK MENDETEKSI OBJEK-OBJEK *AIRBORNE*

Pembimbing : 1. Reza Fuad Rachmadi, S.T., M.T., Ph.D
2. Dr. I Ketut Eddy Purnama, S.T., M.T.

Pada penelitian ini, kami menunjukan percobaan kami untuk meningkatkan kapabilitas YOLOv7 untuk mendeteksi objek airborne. Objek airborne tampak sangat kecil pada kamera karena jarak yang cukup jauh dari kamera. Oleh karena itu, YOLOv7 harus dioptimisasi untuk dapat mendeteksi objek-objek kecil. Beberapa modifikasi diajukan dan diuji pada penelitian ini. Modifikasi-modifikasi ini meliput perubahan arsitektur (Menambah layer deteksi, mengubah sumber feature-map, dan mengganti layer deteksi menjadi anchor-free), dan pada bag-of-freebies (rekalkulasi anchor, dan augmentasi mosaik). Hingga saat ini, kami menemukan bahwa kombinasi augmentasi mosaik, rekalkulasi anchor, dan mengubah sumber feature-map memberikan skor mAP yang paling tinggi 14,09% dibandingkan dengan YOLOv7 yang biasa (mAP=0%).

Kata Kunci: ***Deteksi Objek Kecil, YOLOv7, Modifikasi Arsitektur, Modifikasi Bag-of-Freebies, Objek Airborne***

[This page intentionally left blank]

ABSTRACT

Name : Dion Andreas Solang
Title : *YOLOv7 SMALL OBJECT DETECTION OPTIMIZATION TO DETECT AIRBORNE OBJECTS*
Advisors : 1. Reza Fuad Rachmadi, S.T., M.T., Ph.D
2. Dr. I Ketut Eddy Purnama, S.T., M.T.

In this research, we present an attempt to improve the capability of YOLOv7 to detect airborne objects. Airborne objects appear very small on cameras due to their distance to the camera. For that reason, YOLOv7 needs to be optimized to detect small objects. Several modification proposal was made and tested in this research. These modifications include changes in the architecture (Adding extra detection head, modifying feature-map source, and replacing detection head to a detached anchor-free head), and some bag-of-freebies applications (anchor recalculation, mosaic augmentation). At the current state, we found the combination of mosaic augmentation, anchor recalculation, and modifying feature-map source produces the greatest score in mAP, a 14.09% increase compared to the plain YOLOv7 (mAP=0%).

Keywords: ***Small Object Detection, YOLOv7, Architecture Modification, Bag-of-Freebies Modification, Airborne Object***

[This page intentionally left blank]

Contents

List of Figures

[This page intentionally left blank]

List of Tables

[This page intentionally left blank]

CHAPTER I
INTRODUCTION

1.1 Background

Figure 1.1: An example of airborne object dataset

Autonomous Aerial Vehicle (AAV) have the potential to significantly impact industries, particularly in commercial delivery. One notable example is Amazon Prime Air, which is currently under development. Prime Air aims to deliver goods from Amazon warehouses directly to customers (Amazon, 2022). To accomplish this, AAV require a reliable and efficient autonomy system.

One of the most critical challenges in designing an AAV system is the ability to sense and avoid (SAA) obstacles. While the airspace in which AAVs operate may be relatively sparse, there is still risk of encountering static obstacles or airborne objects such as birds or drones. In commercial applications, such encounters can have consequences not only for the AAV itself but also for the items it carries. Ensuring effective SAA capabilities is essential to mitigate these risks and safeguard both the AAV and the valuable cargo it transports.

Most SAA system of AAVs includes camera as their primary sensor. Cameras provide visual perception to the AAV, real-time, in the form of images. These images must be processed by a computer vision algorithm to identify and localize the obstacles in it so that the AAV can estimate their position and plan actions it needs to do to avoid them. There are other onboard sensing options such as LiDAR or radar, but cameras are more favored due to their lighter weight, cheaper price, and relatively lower energy consumption compared to active sensors.

Airborne objects present a particularly challenging problem in SAA as they can appear unexpectedly and approach rapidly from long distances due to their inherent need for speed in

flight. For this reason, it is important to detect airborne objects while they are still far away. Unfortunately, their far distance cause them to appear very small on images. Figure 1.1 show an example of how airborne objects appear on images. In “Airborne Object Tracking Challenge” (2021), the airborne objects can appear in the range of 4 to 1000 px on a 2048×2448 px image, that is, around 0.00008 - 0.01 % of the image area.

For reasons listed above, the computer vision algorithm to be implemented to detect these airborne objects, must be able to detect small objects accurately. And not only that, it also must be able to do it in real-time. The fact that AAVs operating environments are outdoors, brings more trouble for the computer vision. Outdoor environment brings large complexity and additional variance to the image distribution, making it almost impossible to handcraft feature extractor for it. This is where non-handcrafted feature extractors come in handy, particularly the deep learning approaches. Deep learning models have shown incredible abilities in handling complex data. Given enough training data, a deep learning model can craft itself feature kernels that can extract the objects in an image, despite being subjected to complex outdoor environment. For this reason, deep learning models gained prominence in addressing these challenges.

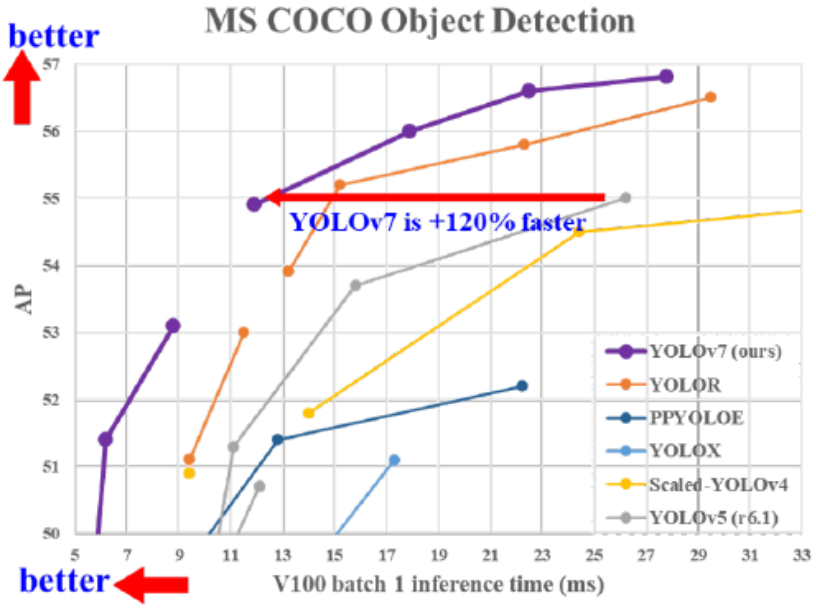


Figure 1.2: YOLOv7 performance on COCO dataset compared to other object detectors.

Introducing YOLOv7, a state-of-the-art deep learning based real-time object detector (Wang, Bochkovskiy, et al., 2022). At the time of the proposal for this research was made (November 2022), YOLOv7 outperform both in speed and accuracy of all known real-time object detectors with inference speed in the range of 5-160 FPS. It also has the highest accuracy (56.8% AP) among object detectors with inference speed greater than 30 FPS on a V100 GPU. The capabilities of this cutting-edge architecture makes it well-suited for AAV computer vision system. However, all the performance metrics of YOLOv7 mentioned before are obtained by training the model using COCO 2017 dataset. A dataset which consist of general objects that people see in their daily life. COCO dataset is going to have very distinct distribution compared to airborne objects. As such, there would be a need for some modification to YOLOv7 so that it could detect airborne objects well.

The topic of this research is about modifying YOLOv7 with objective of optimizing it

to detect small objects, which extends to airborne objects. We experimented with some modification to YOLOv7's bag-of-freebies, bag-of-specials, and network architecture. Then we benchmarked the modified models against the "Airborne Object Tracking Dataset" (2021), and picked the model with the highest AP score that can perform inference in real-time.

1.2 Problem Statement

YOLOv7 is not an object detection model specifically designed to detect small objects or airborne objects. Therefore, this research ask the following question.

- What modification can be done to YOLOv7 to improve its abilities in detecting airborne objects?

1.3 Purpose

The purpose of this research is to find modifications of YOLOv7 that would improve its ability in detecting airborne objects.

1.4 Problem Scope

In this research, we want to formulate the scope of the problem such that the modifications applied on YOLOv7 would not lose its real-time detection ability and is still reasonably computable/trainable. Otherwise, we can just run a Solomonoff induction on airborne object data and call the result a modification of YOLOv7. Thus, we state the following scopes for the problem.

- YOLOv7 is used as the baseline for modifications.
- The result of modification must be able to perform inference in real-time.
- The modified YOLOv7 must be trainable within a reasonable duration using the computational resource available, which is a computer which uses consumer GPU Nvidia RTX 2080 Ti.

[This page intentionally left blank]

CHAPTER II
LITERATURE REVIEW

2.1 Theoretical Basis

2.1.1 YOLO Family Architecture

YOLO is an abbreviation of ”You Only Look Once” which describes what kind of neural network YOLO is, a single stage object detector. It means that this architecture predicts regions and classes both at once. In contrast, two-stage detector predicts objects' regions first and then predicts their classes. Detecting objects in a single-stage manner is what gave YOLO architecture the ability to infer in real-time. This is possible due to how YOLO architecture was designed. YOLO architecture consist of 3 main part, the head, the neck, and the backbone.

Backbone

The backbone is the network that extract features from the inputted image. Typically, the backbone is composed of deep neural network layers that progressively downsample the spatial dimensions of the input while increasing the number of learned features or meaningful abstraction of the data.

The implementation of backbone in YOLO usually varies from one version to another. As example, Redmon and Farhadi (2016)'s YOLOv2 implemented Darknet-19 network as backbone, Redmon and Farhadi (2018)'s YOLOv3 implemented Darknet-53, Bochkovskiy et al. (2020)'s YOLOv4 with their CSP-Darknet-53, or Zhang et al. (2021) with their non-CNN (Transformer) backbone. Each of these network has their own has their own advantages and disadvantage when it comes to accuracy, memory requirement, or latency.

The Neck

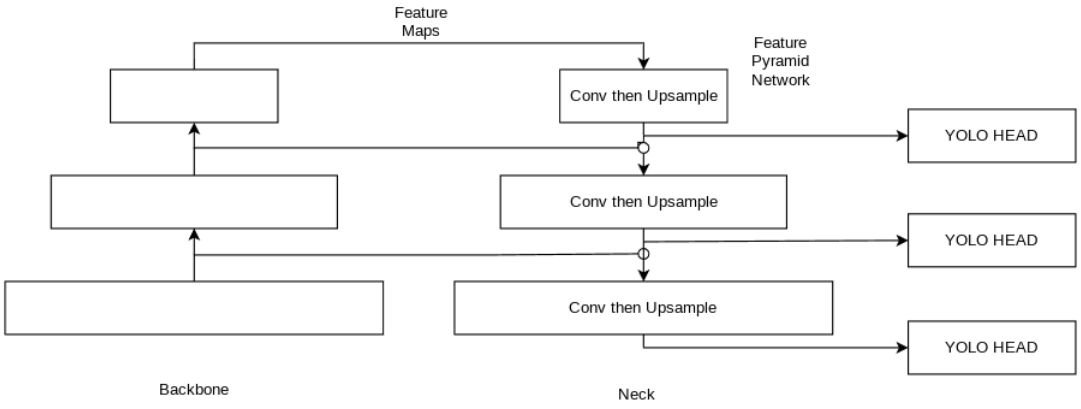


Figure 2.1: Feature Pyramid Network in YOLOv3

The neck is the intermediate network between backbone and head. The main function of the neck is to enhance features extracted by the backbone. Specific implementation of neck for

each YOLO architecture is different one to the other. Some neck implementation try to combine feature maps across different prediction scales of YOLO network.

Redmon and Farhadi (2018) was the first to introduce YOLO prediction in multiple scale in YOLOv3. To enhance the feature maps with using information across scales, YOLOv3 fuses features from multiple parts of the backbone before up sampling them as seen on Figure 2.1. This type of neck network is called Feature Pyramid Network (FPN). A further improvement was made by Bochkovskiy et al. (2020) with their YOLOv4 by introducing Path Aggregation Network (PANet) for the neck. With PANet, feature are fused back to higher scale by adding another FPN-like layer after the original FPN but with reverse direction.

The Head and The Anchors

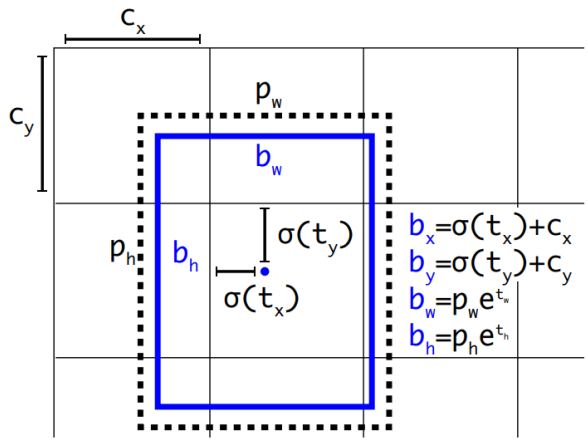


Figure 2.2: Head layer predict anchor box and its offsets from lattice of the feature grid (Redmon & Farhadi, 2018)

The head is where the object detection happens. Extracted and enhanced features of the image is fed to the head in multiple different scales. For each scale, the head will predict a box for each $N \times N$ lattice point on feature map's grid. In total, each layer will output a tensor with size $N_k \times N_k \times [A_k \times (4 + 1 + C)]$ where N_k is the size of feature map grid at the k-th scale, A_k is the number of anchors for that scale, 4 is for the four offsets $[t_x, t_y, t_w, t_h]$ in figure 2.2, 1 is for the objectness score for the grid, and C is for the number of classes it has to predict.

Most of YOLO family architecture head utilizes anchor boxes to assist bounding box prediction. This technique is used in Redmon and Farhadi (2016), Redmon and Farhadi (2018), Bochkovskiy et al. (2020), Wang et al. (2020), Jocher et al. (2022), Wang et al. (2021), and Wang, Bochkovskiy, et al. (2022). Instead of directly predicting the size and position of the bounding box, YOLO head predicts the offsets for each anchor boxes assigned to the head, then it utilizes the objectness score to pick which result of these anchor boxes to be used. Using anchor boxes allows the neural network to converge more quickly because it provides a prior knowledge of the dataset before training.

Loss Function

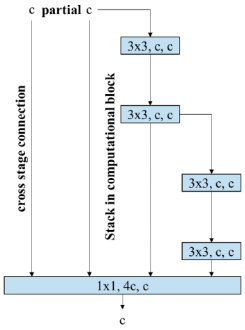
The goal of a YOLO architecture is to (1) predict if an object exist or not, (2) predict the bounding box of such object, and (3) predict the class of the object. These 3 loss functions that correspond to those objectives are called objectness loss, box loss, and class loss respectively.

To update the weights on training, the total loss is calculated as the weighted sum of those 3 losses.

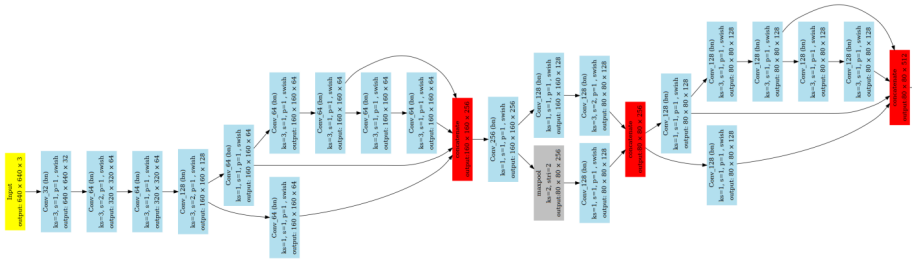
2.1.2 YOLOv7

As said in section 1.1, YOLOv7 is the state-of-the-art real-time object detector. It was made by the authors of YOLOv4, and by the time it was published (july 2022), it surpassed all known real-time object detectors both in speed and accuracy. To achieve this, YOLOv7 implemented some new changes and addition to the neural network.

Backbone



(a) ELAN block

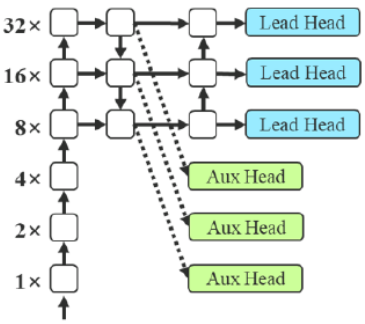


(b) First two ELAN blocks in YOLOv7

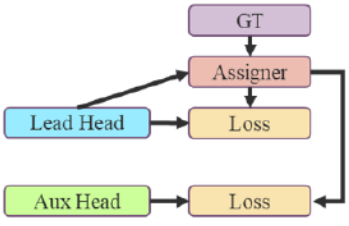
Figure 2.3: ELAN in YOLOv7

YOLOv7 implemented Efficient Layer Aggregation Network (ELAN) and Extended-ELAN (E-ELAN) as the backbone. ELAN is a convolutional neural network that was designed to extract features efficiently by controlling the shortest longest gradient path in the network (Wang, Liao, & Yeh, 2022). This choice of backbone allows YOLOv7 to perform prediction more accurately despite having fewer number of parameters. Figure 2.3b shows how ELAN block in Figure 2.3a used in YOLOv7.

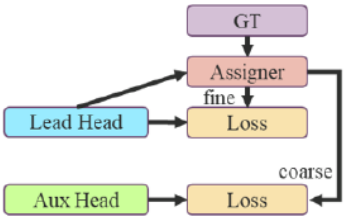
Label Assignment Strategy and Auxilary Head



(a) Auxliary heads attachment in YOLOv7



(b) Lead guided label assignment



(c) Coarse-to-fine lead guided label assignment

Figure 2.4: YOLOv7 Label Assignment Strategy with Auxilary Heads

SimOTA, first introduced in YOLOX, is an algorithm to approximate Optimal Transport Assignment (OTA) in a faster way. Ge et al. (2021) introduced SimOTA in YOLOX because OTA was deemed too slow to compute as it was increasing the training time by 25%. YOLOv7 also implemented SimOTA for its dynamic label assigner.

YOLOv7 deep supervised its training process by attaching auxilary heads to its neural network as seen on Figure 2.4a. These auxilary heads is only used on training, on inference, they are removed from the neural network to improve latency, only the lead head is kept. There is a problem however with assigning labels to the auxilary and lead heads. Most object detection networks that utilizes auxilary heads have 2 independent label assigners, one for auxilary heads and one for lead heads. YOLOv7 done things differently.

YOLOv7 proposed 2 way of assigning labels to auxilary and lead heads. Lead head guided label assignment (Figure 2.4b) and coarse-to-fine lead head guided assignment (Figure 2.4c). For lead head guided label assignment, the assigner gives a copy of lead heads' label assignment to the auxilary heads. For coarse-to-fine lead head guided assignment, the assigner works like lead head guided assigner but gives coarse label assignment to auxilary head. Coarse label assignment is done by relaxing the positive sample constraints of the assigner. Wang, Bochkovskiy, et al. (2022) find that coarse-to-fine label assignment produces the greatest AP scores.

Due to the relaxed constraints, coarse label assignment to auxilary heads assigns more positive labels the auxilary heads' grids. This way, the network will learn more to recall. On inference, this recall ability would be filtered by the lead head to produce accurate prediction.

Reparameterization

YOLOv7 utilized RepConv and YOLOR implicit layers in its network. These 2 layers can be reparameterized after training to simplify the neural network, thereby reducing latency and memory usage but not hurting inference performance.

2.1.3 Anchor Recalculation

Anchor recalculation is a common method of introducing prior distribution of the dataset to an anchor-based object detection networks. Most of the time, anchors provided by pretrained YOLO weights are optimized for the common metrics dataset such as COCO2017 or VOC2012. Thus, recalculating anchor can help the neural network learn faster.

There are multiple ways of recalculating anchors. Some of them can be performed before training or during training. Most of pre-training anchor recalculation method involves with clustering the anchors to the dataset. This is done by using clustering algorithm such as K-means, Gaussian Mixture, and many others. Recalculating anchor on training is a little more complex to do as it will involve some loss function or architectural change on the object detection neural network. Zhong et al. (2018) for example add additional layer on detection part of object detectors that is connected to anchor modifiers such that the anchors will also be updated during training. Wang et al. (2021) mentioned that the implicit multiplication layer of their network can be purposed for anchor refinement.

2.1.4 Mosaic Augmentation

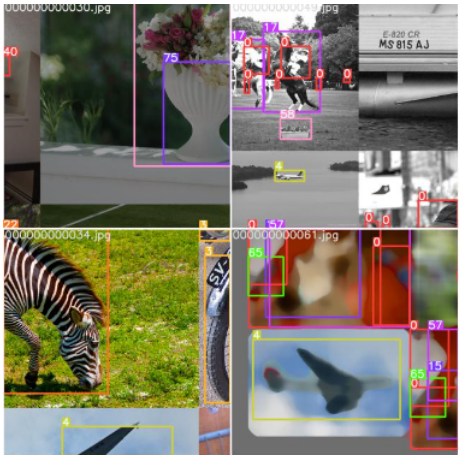


Figure 2.5: Four example of mosaic augmentation (Jocher et al., 2022)

Mosaic augmentation was introduced in Ultralytics' implementation of YOLOv3 (Jocher, 2020). This augmentation technique will randomly pick 4 images of the dataset, and tile them randomly into one image like in 2.5. It's called mosaic due to how the result of the augmented image have mosaic-like appereance. Wang et al. (2019), Bochkovskiy et al. (2020), and Jocher et al. (2022) reported increase in accuracy after applying mosaic augmentation.

2.2 Related Works

2.2.1 YOLO-Z

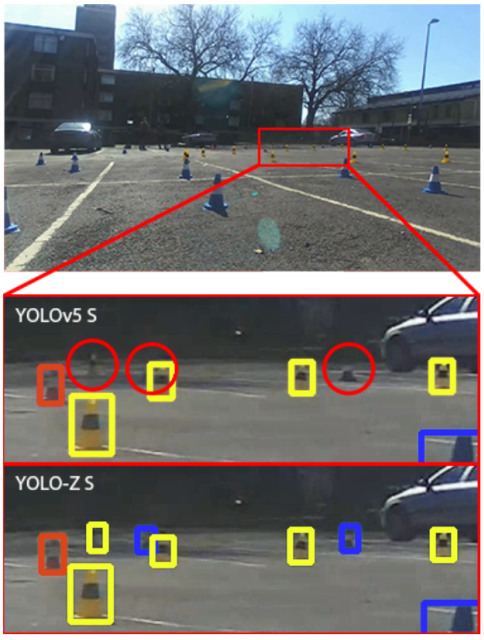


Figure 2.6: Small objects in the image. Comparison of YOLOZ-S and YOLOv5-S. YOLOv5-S was not able to detect the circled objects.

YOLO-Z is a derivative architecture of YOLOv5r5.0. This variant of YOLO modified the backbone, neck, and number of anchors of the original YOLOv5 to enhance its capability of detecting small objects (Benjumea et al., 2021). These changes are backbone change from YOLOv5r5.0 to a downscaled DenseNet, neck change from FPN to biFPN on some YOLO-Z scales, and increasing the number of anchors used at each scale.

YOLO-Z was aimed to be used in autonomous racing car. In this high speed environment, early detection of obstacle is crucial to plan for action. For that reason, the autonomous racing car must detect the cone-shaped obstacles that are far away from it. Since objects that are far away appear small on image captured by camera, YOLOv7 was designed with purpose of small object detection.

2.2.2 exYOLO

exYOLO is a modification of YOLOv3 (Xiao, 2021). exYOLO added a Receptive Field Block before feature-fusion in the neck to one of the feature scale. This change made exYOLO produce a higher mAP score on VOC2007 compared to its baseline YOLOv3.

2.2.3 YOLOv4-tiny with added head

(a) Regular YOLOv4-tiny prediction

(b) YOLOv4-tiny with added head prediction

Figure 2.7: Comparison of regular YOLOv4-tiny and YOLOv4-tiny with added head on Aditya et al. (2022) ASV

Aditya et al. (2022) used YOLOv4-tiny as their Autonomous Surface Vehicle (ASV) object detector due to the computational device constraint. To detect objects that were atleast 30 meters away from the ASV, they applied a modification of YOLOv4-tiny, which was YOLOv4-tiny but with additional head layer. The original YOLOv4-tiny only had 2 head layers, thus was only predicting in 2 scales. An addition of head layer allows it to predict in 3 scales. Using this modification, the network was able to detect small object better and raised the overall mAP score by 4% without significantly reducing latency.

CHAPTER III
METHODS

3.1 Model Searching Method

To find the best model for detecting small objects, trial and error will be performed by adding or changing architecture, bag-of-specials, or bag-of-freebies of YOLOv7. These modifications will be tried out independently and combinatively.

Every modification will be tested on its ability to detect airborne object. The best model will be selected based on $mAP@50$ metric. This metric was chosen instead of $mAP@[50:95]$ because we don't expect for the model to be able to predict a tightly fit bounding boxes for small object and consider a loose 50% coverage IoU is good enough.

The model searching method is comprised of six steps. These steps are dataset preparation, develop training system, create modification model configuration, training the model, analysis, and model selection. Figure 3.1 shows the order these steps will be executed.

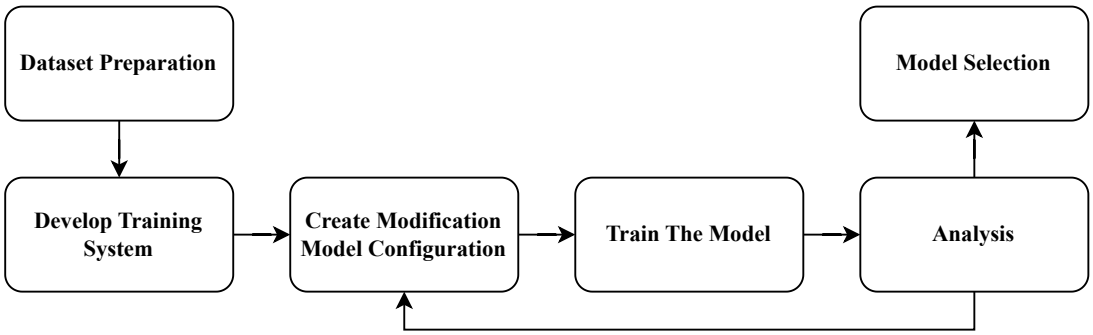


Figure 3.1: Search Steps

To conduct this research, we first have to prepare the dataset. First, the dataset will be downloaded from "Airborne Object Tracking Dataset" (2021). Then, the labels of the dataset will be converted to darknet or COCO format. Next, the dataset will be sampled into training set, validation set, and test set. The sampling will be done in a way that balances the number classes in each set so that there will be no dominating class during training.

The next thing to do is to develop a training system. The purpose of this system is to make the training process easily conducted and monitored. This system will include features such as train fail notification, train queueing, and alert the user when the computer overheats.

Moving on, we have model configuration creation. Here we will create a model configuration file based on the modification that we want to try. The modification configurations will be made according to the modification candidates listed in section 3.2.

The next step is to train the model. The model configurations that was made in previous step will be built and trained from scratch (no pre-trained weights). In this step, the weights of the neural network, training history, and performance metrics will be generated for analysis.

In the analysis step, we will analyze the performance of the modified YOLOv7. This step is done to find other candidate of modification that might work. If such modification candidate was found, we will go back to create the model configuration, and then train the modification.

Finally, in the last step we will select the best model. We will select the model with the best performance among the modified YOLOv7 models. The model selection will be done based on $mAP@50$ metric, with constraint in inference latency. The model that qualify for selection must be able to perform inference with speed of atleast 10 FPS in a consumer GPU Nvidia RTX 2080Ti.

3.2 Modification Candidates

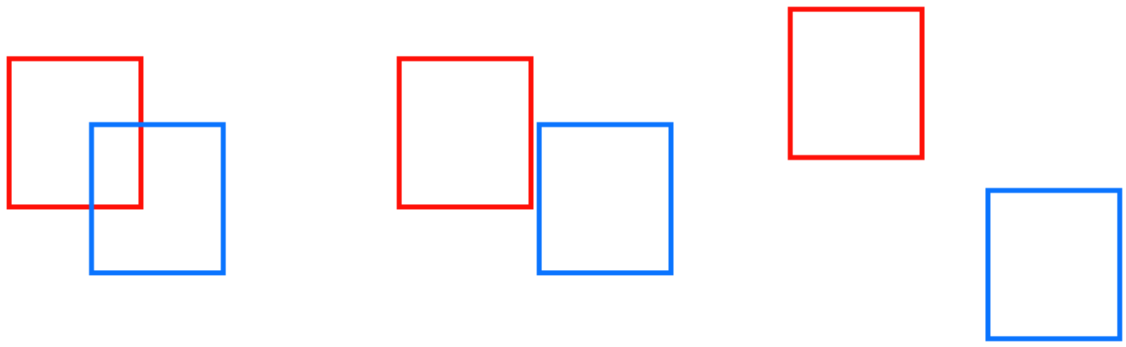
3.2.1 Mosaic Augmentation

As discussed in section 2.1.4, mosaic augmentation was able to increase the detection accuracy of small objects on many object detection neural networks. For this reason, we will experiment by training YOLOv7 with and without mosaic augmentation.

3.2.2 Pre-training Anchor Recalculation

In the implementation code of YOLOv7, the anchors provided was calculated based on COCO2017 dataset. We assume that the dataset distribution of airborne objects to be greatly different from COCO2017. As such, the anchors must be recalculated for faster learning. This recalculation will be conducted using k-means algorithm. One problem however is that k-means might fail to cluster the dataset into the amount of anchors that we needed. To tackle this, we might have to cluster it in logarithmic coordinates.

3.2.3 Replacing Localization Loss to Extended IoU



(a) $IoU > 0$ when 2 boxes intersect

(b) $IoU = 0$ when 2 boxes does not intersect

(c) $IoU = 0$ when 2 boxes does not intersect and far away

Figure 3.2: Cause of IoU vanishing gradient

Extended-IoU (EIoU) is a modification of IoU that are used in neural networks to tackle the problem of vanishing gradient (Peng & Yu, 2021). IoU are known to cause vanishing gradient problem due to its behavior when two bounding boxes are not intersecting. When 2 boxes A and B are not intersecting, the area of intersection ($A \cap B$) would always be 0. This value doesn't give any information whether the boxes are far apart (Figure 3.2c) or near but not intersecting (Figure 3.2b). To solve this, loss involving IoU are usually paired with some regularization

terms. For example Rezatofighi et al. (2019) and Zheng et al. (2019) proposed $GIoU$, $DIoU$, and $CIoU$.

YOLOv7 itself is using $CIoU$ for its localization loss. $CIoU$ is just regular IoU that is paired with distance and box aspect ratio regularization terms. The problem of these kinds of regularized IoU is that when the boxes intersect, it does not behave like IoU anymore. $GIoU$, $DIoU$ and $CIoU$ have residue of their regularization when the boxes intersect. Metrics used to evaluate object detection algorithms depend on IoU (*e.g.* mAP). For this reason Peng and Yu (2021) designed EIoU. The main appeal of EIoU is that it behave exactly like IoU when the boxes intersect and gives non-positive value when the boxes do not intersect. By behaving exactly like IoU, it is hoped that the model would perform better on the metrics.

3.2.4 Utilizing Earlier Feature Map Stage

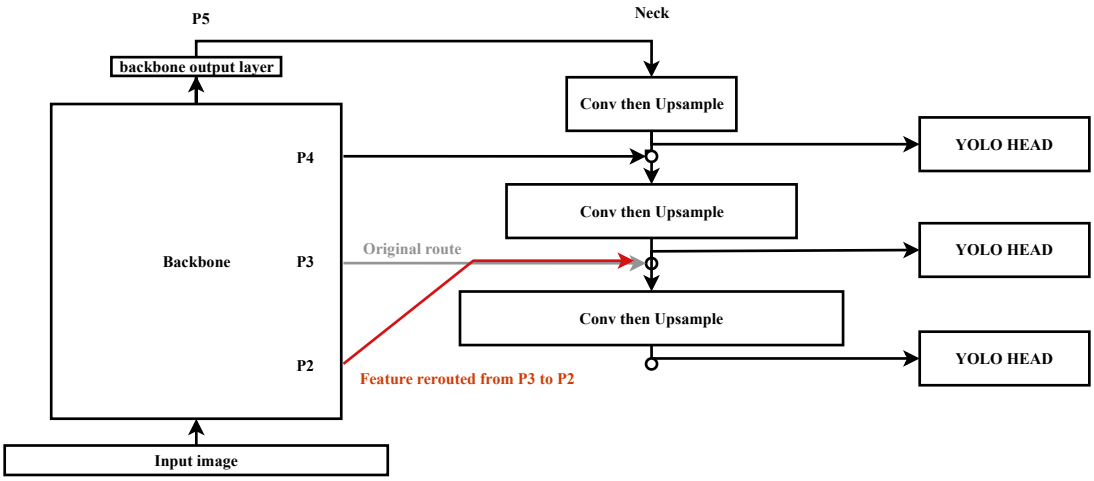


Figure 3.3: Rerouting Neck Connection to Earlier Stage

In a deep neural network, the extracted feature/abstraction of the data become more prominent as the input data passes through deeper layers. However, information loss also increases in the deeper layer. For small object detection, information is crucial. There is a posibility the features of small objects are lost as the data propagates through the network.

If we view it in a data path network design perspective, the greater the length of the path from input to output, the more information will be lost. Therefore, we propose to reroute the orignal connection from backbone to the neck to an earlier stage of inference. For example, YOLOv7 take feature maps from P3, P4, and P5 scales of the backbone, we can reroute the connection from P3 to P2 as shown in Figure 3.3.

3.2.5 Additional YOLO Head

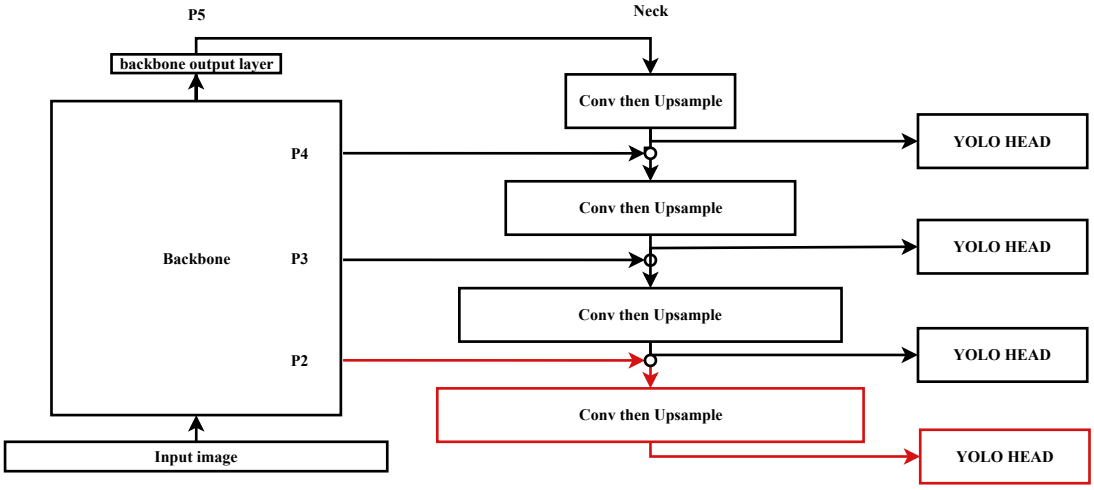


Figure 3.4: Adding an Extra Head Layer

An additional YOLO head means an extra stage of detection. With more stage of detection, the large variance of the objects in the dataset can be learned. This is especially good for "Airborne Object Tracking Dataset" (2021) where the area of the bounding boxes in the dataset can be orders of magnitude apart (see section 3.5).

3.2.6 Replace YOLO Head to YOLOX's Decoupled Anchor-free Head

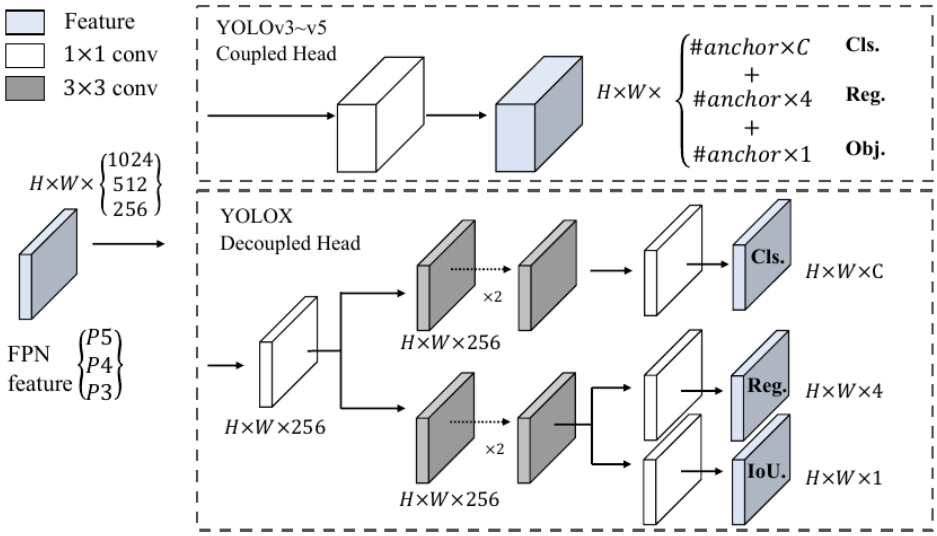


Figure 3.5: Decoupled Anchor-free Head in YOLOX compared to Coupled Head in mainstream YOLO

YOLOv6 and YOLOX used a different kind of head layer compared to usual YOLO (Ge et al., 2021; Li et al., 2022). In mainstream YOLO, prediction for bounding box and classes are both calculated on the same layer. With decoupled anchor-free head, the layers are separated for class prediction and bounding box prediction, and bounding box prediction are done using

4 value without anchor as seen on Figure 3.5. Using decoupled anchor-free head gives us 2 advantages. 1) Reducing the amount of design parameters as we don't have to introduce anchor boxes to the model. 2) Reducing the complexity of interpreting prediction result. Advantage (1) is especially enticing. There is a possibility of us introducing bad priors to the neural network by poorly clustering anchor boxes. By having a model that less dependent on prior, we can reduce such possibility.

A thing to consider when applying decoupled anchor-free head to YOLOv7 is the label assigner. YOLOv7 and YOLOX uses SimOTA as its default label assigner. However, Li et al. (2022) YOLOv6 uses Task-aligned Assigner (TAL) for its label assigner. They reported that TAL performs better than SimOTA. Therefore, in applying decoupled anchor-free head, it might be better to use TAL as label assigner.

3.2.7 Partitioning Image

Figure 3.6: An Image Partitioned to 4 Images

Partitioning the image can significantly improve the accuracy of small object detection. The partitioning process involves cropping the original image into smaller sections, thereby avoiding excessive downscaling of the image before feeding it into the neural network. This approach allows the network to retain finer details and preserve the integrity of small objects, ultimately leading to more accurate detection results. By breaking down the image into manageable segments, the network can focus on individual regions of interest and capture the features of small objects with greater precision.

One challenge associated with partitioning images is the potential increase in inference latency. When an image is partitioned into multiple sections, the neural network needs to perform separate detections on each partition. As a result, the inference time is multiplied by the number of partitions. For instance, if we divide an image into four parts, the network will have to carry out four separate detections, leading to a significant increase in latency.

3.3 Instruments

To conduct this research, we will be using a computer with the following specification:

- CPU — Intel® Core™ i5-9400F CPU @ 2.90GHz
- GPU — Nvidia Geforce RTX 2080 Ti
  - Memory — 11 GB
  - CUDA Compute Capability — 7.5
- RAM — 12 GB
- Hard Drive Available Memory — 1.3 TB
- Operating System — Ubuntu 20.04
- Cuda Toolkit Version — 11.7
- PyTorch Version — 1.13.1

3.4 Pilot Test

To ensure the reliability of the instruments, and to find the instrument-induced constraint, we performed A pilot test. In this pilot test, we trained several YOLOv7 models with varying input sizes with a dummy dataset for 300 epochs. The dummy dataset have the same dimension as the "Airborne Object Tracking Dataset" (2021). The results are presented in Table 3.1.

Table 3.1: Pilot Test Training

No	Model	Input Size	Training Time (300 epochs)	VRAM Usage	Average GPU Temperature
0	`YOLOv7`	640	4.0h	4.5GB	82
1	`YOLOv7 batch-size=2`	640	2.5h	6.5GB	82
2	`YOLOv7`	960	4.7h	6.5GB	84
3	`YOLOv7`	1600	20h	9.5GB	84
4	`YOLOv7`	2000	N/A	Out of Memory	
5	`YOLOv7-d6`	1600		10GB	85
6	`YOLOv7-w6`	1600	N/A	10.5GB	
7	`YOLOv7-e6`	1600	N/A	Out of Memory	
8	`YOLOv7-e6e`	1600	N/A	Out of Memory	

From Table 3.1, we want to pick a baseline model to apply our modification. Since we want to persist the features of the 2048×2448 px images, we pick the largest input size the machine can handle but still able to give enough room for some changes in architecture. For this reason, we picked YOLOv7 with input size 1600 and batch size 1 as the baseline.

3.5 Dataset Preparation

3.5.1 Dataset Source

Dataset containing annotated images of airborne objects can be obtained from "Airborne Object Tracking Dataset" (2021). This dataset is hosted in an AWS S3 Bucket. The dataset consist of several Unmanned Aerial Vehicle (UAV) flights' frame-by-frame camera capture. There are 7 classes in the dataset but most there are 3 dominating classes. The size of the dataset and its class distribution can be seen on Figure 3.2 and 3.3 respectively.

Table 3.2: Original Dataset Splits

Splits	Size (TB)	UAV Flight Sequences	Total Images	Total Labels
Training	11,3	4154	4975765	2891891
Validation + Test	2.1	789	943852	496075
Total	13,4	4943	5919617	3387966

Table 3.3: Dataset' Objects Classes Distribution

Splits	Total Objects	Airplane	Helicopter	Bird	*Other 4 Classes*
Training	2,89 M	0,79 M	1,22 M	0,33 M	0,54 M
Validation + Test	0,50 M	0,13 M	0,17 M	0,06 M	0,14 M
Total	3,39 M	0,92 M	1,39 M	0,39 M	0,69 M

If we look at the area (in px) distribution of the objects' bounding boxes, we will find the histogram in Figure 3.7. Take note that in that figure, the x-axis is using a logarithmic scale, which means the data is very right skewed. The median of the data is $314px$, while the 5th and 95th percentile are $36px$ and $5061px$ respectively.

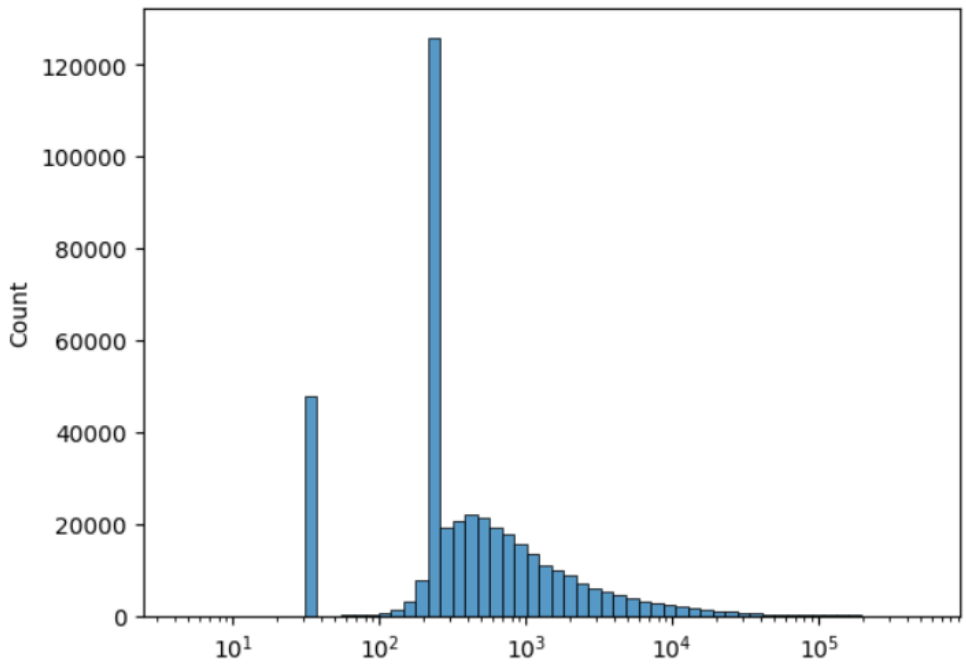


Figure 3.7: Distribution of Bounding Boxes' Area

The distribution of objects position in the camera can be seen on Figure 3.8 Most of the objects are around the center of the image.

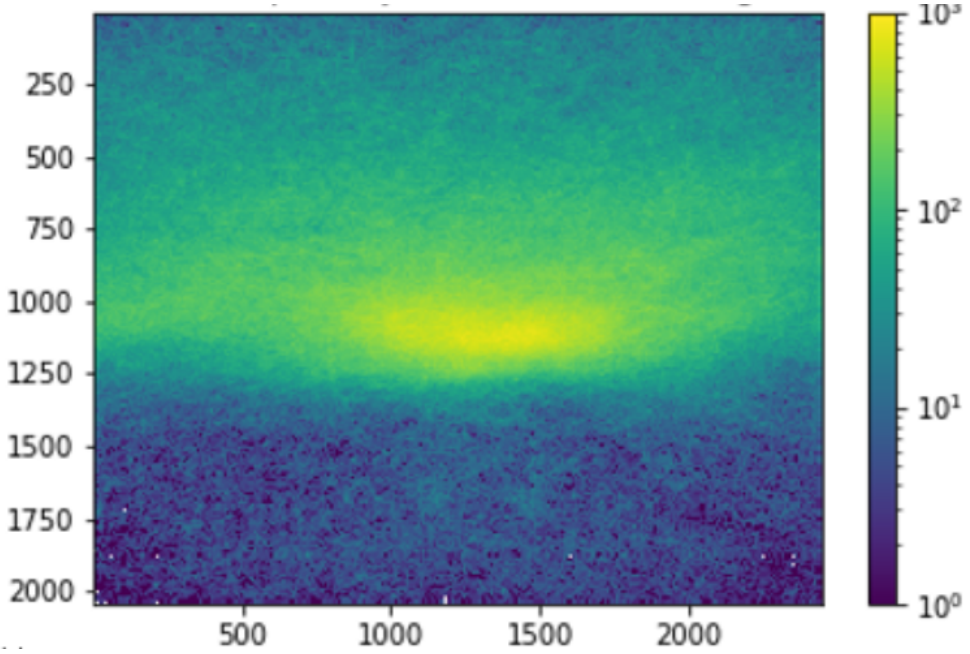
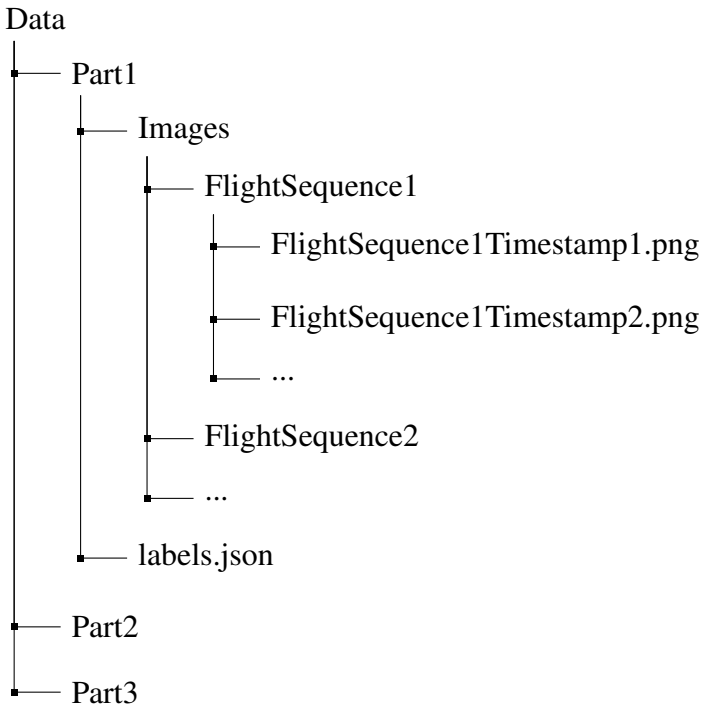
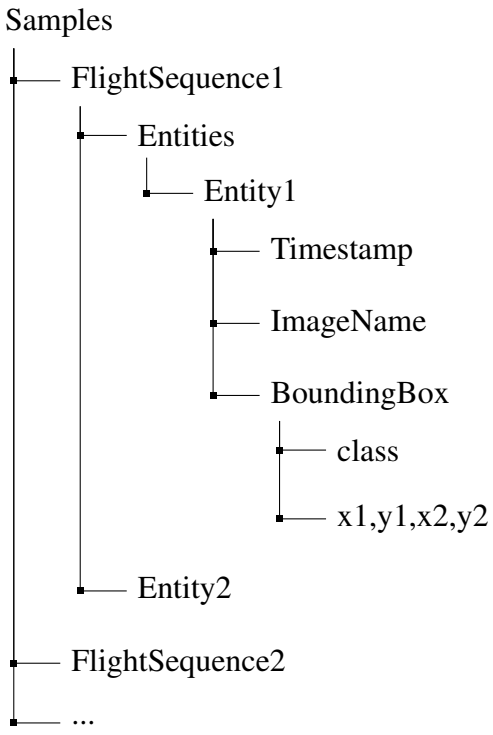


Figure 3.8: Distribution of Bounding Boxes' Area ("Airborne Object Tracking Challenge", 2021)

The dataset is structured in the following manner:



And each labels JSON file is also structured like this:



Notice that the labels are structured entity based. This made the label convertion process more complex. This topic will be discussed more in the next section.

3.5.2 Dataset Sampling

Sampling Distribution

As seen on section 3.5.1, The size of the dataset is massive. For this reason, we only downloaded the planned encounters from the AWS S3 Bucket which in total amounted to 1.1 TB. As this number is still too large due to the limitation of computational resource, we will only sample some of them for training and testing. We will sample in total 700 images from the dataset with splits as shown in Table 3.4.

Table 3.4: Distribusi Sampling Dataset

Splits	Total Images	Classes Percentage				
		Airplane	Helicopter	Bird	Drone	Negative
Training	400	23.75%	23.75%	23.75%	23.75%	5%
Validation	100	20%	20%	20%	20%	20%
Test	200	20%	20%	20%	20%	20%

Sampling and Label Convertion Procedure

Sampling the dataset was not straightforward. Due to how the "Airborne Object Tracking Dataset" (2021) files and labels was structured, we had to make a sampling and converting procedure like the following:

1. Open all `labels.json` file and load all entities into a single list called `entities`

2. Let `Img2Boxes` be a dictionary with image names as key and list of bounding boxes in the image as value.

3. Let `Cls2Img` be a dictionary with 4 class names as key (airplane, bird, etc) and set of image name that contain object of those classes as the value.

4. Populate `Img2Boxes` and `Cls2Img` in the following manner:

```
For entity in entities:
  Img2Boxes[entity.ImageName].append(entity.BoundingBox)
  Cls2Img[entity.BoundingBox.class].append(entity.ImageName)
```

5. Calculate the total image needed for each class according to Table 3.4.
   e.g. Total image needed for airplane class is $\frac{400\times23.75+100\times20+200\times20}{100}$.

6. Distribute the images to train, valid and test.

```
For class in list_of_classes:
  Images[class] = Cls2Img[class].random_choice(TotalImageNeeded[↩
     class])
For class in list_of_classes[-1]: // except negative sample
  addtotrain = images[class][:23.75/63.75*TotalImageNeeded[class]]
  images[class] = images[class][23.75/63.75*TotalImageNeeded[class↩
     ]:]
  train.append(addtotrain)

  do the same to valid and test
process the negative sample independently
```

7. Convert `train`, `valid`, `test` to darknet format.

```
For img in train:
  label = []
  For boxes in Img2Boxes[img]
    // to_darknet converts class xyxy to class x y w h normalized
    label.append(to_darknet(boxes.BoundingBox))
  copy(img, train_folder+img)
  write(label, train_folder+img.with_suffix('.txt'))

repeat for valid and test
```

CHAPTER IV
EXPERIMENTS

4.1 Baseline Performance

For comparison purposes, we first measure the baseline performance of YOLOv7 with all modification candidates from section 3.2 stripped. We call this model as `YOLOv7-plain`. After 300 epochs of training with batch-size 1 like in the experiment setup, we found that this model was unable to detect anything on the test set (a valid detection is detection with $IoU > 0.5$ to groundtruth).

4.2 Mosaic Augmentation and Anchor Recalculation

In this section, we present the effect of mosaic augmentation and anchor recalulation on the $mAP@50$ score when applied independently and combinatively. For comparison purposes, we assigned names the models as `YOLOv7-M`, `YOLOv7-AR`, and `YOLOv7-MAR`. These names correspond to `YOLOv7-plain` with mosaic augmentation, `YOLOv7-plain` with anchor recalculation, and `YOLOv7-plain` with both mosaic augmentation and anchor recalculation, respectively.

The process of applying mosaic augmentation to the dataset is pretty straightforward as YOLOv7's implementation code already provide the mosaic augmentation tool. A mosaic augmentation result example can be seen in Figure 4.1.

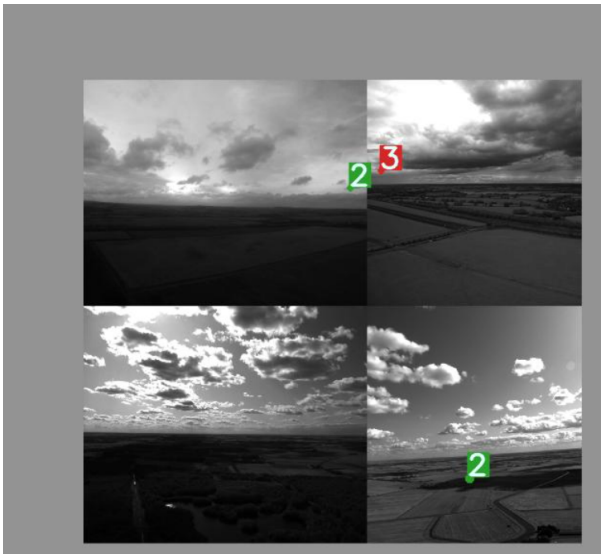


Figure 4.1: An example of mosaic augmented image

Anchor recalculation is done by clustering the training data's widths and heights to 9 centroid using k-means algorithm. However, due to the skewed distribution of the dataset, regular k-means (with L^2 Distance) actually fail to form 9 clusters. To fix this, we used a distance

function

$$L = \sqrt{\ln^2 \frac{w_1}{w_2} + \ln^2 \frac{h_1}{h_2}} \tag{4.1}$$

Distance function in equation 4.1 was actually an L^2 distance but with log-transformed space. With this function, k-means was finally able to cluster the data into 9 centroids. These centroids are used by the head layers as anchors. Each head uses 3. The comparison of the original anchor and recalculated anchor can be seen in Table 4.1 and Figure 4.2.

Table 4.1: Anchor Points Before and After Recalculation

Head	Original Anchors	Recalculated Anchors
1	[12,16], [19,36], [40,28]	[4,4], [14,5], [11,11]
2	[36,75], [76,55], [72,146]	[27,8], [18,17], [41,14]
3	[142,110], [192,243], [459,401]	[43,32], [86,27], [149,70]

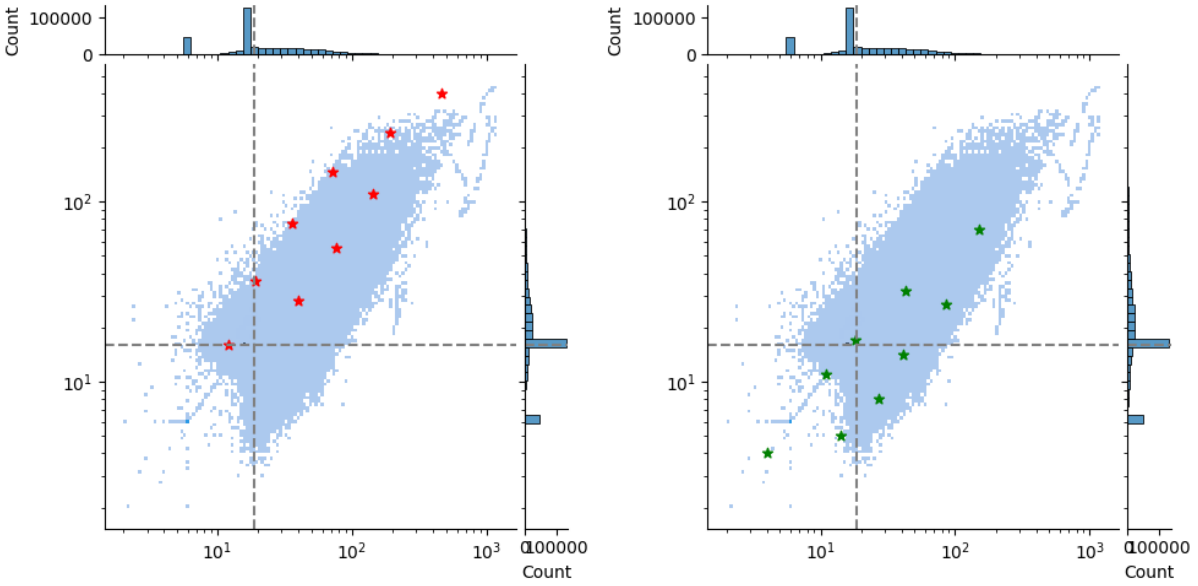


Figure 4.2: Anchor Distribution on Dataset. Left: Original Anchor. Right: Recalulated Anchors

If we observe the distribution of anchors before and after recalculation in Figure 4.2, we can see that indeed the recalculated anchors cover more of the dataset than the original. 8 out of 9 of the original anchors are placed in the first quadrant of the median axis. It means that those 8 anchors are responsible for only 25% of the dataset and leave the rest to 1 anchor. This is indeed very inefficient. In contrast, the recalulated anchors have a more balanced distribution with anchors distributed across each quadrant.

Performance

The performance of the three models, namely `YOLOv7-M`, `YOLOv7-AR`, and `YOLOv7-MAR`, was compared to that of the baseline model YOLOv7-plain. The results are presented in Table 4.2. It can be observed from the table that YOLOv7-plain alone was unable to detect any objects

in the test set. However, after applying both mosaic augmentation and anchor recalculation, the model achieved a notable improvement with a score of 11.2% at mAP@50.

Table 4.2: Mosaic Augmentation and Anchor Recalculation Performance

No	Model	mAP@50
0	`YOLOv7-plain`	0%
1	`YOLOv7-M`	0%
2	`YOLOv7-AR`	0%
3	**`YOLOv7-MAR`**	**11.2%**
	Improvement	**+11.2%**

Since the model `YOLOv7-MAR` was the only model that was capable of detection in the test set, we henceforth establish this model as the baseline for further modification. Meaning, in the subsequent sections, any modifications mentioned should be presumed to be composed of mosaic augmentation and anchor recalculation, in addition to the respective modification that was applied, unless explicitly stated otherwise.

4.3 Replacing Localization Loss to EIoU

In this section, we experimented with $EIoU$. We replaced the original $CIoU$ loss of YOLOv7 to pure $EIoU$ and convexcified version of $EIoU$. The convexciation was done by modifying the $EIoU$ loss from $-EIoU$ to $(1 - EIoU)^2$ for better gradient dynamics during training.

The performance of these modifications can be seen in Table 4.3

Table 4.3: EIoU Localization Loss Performance

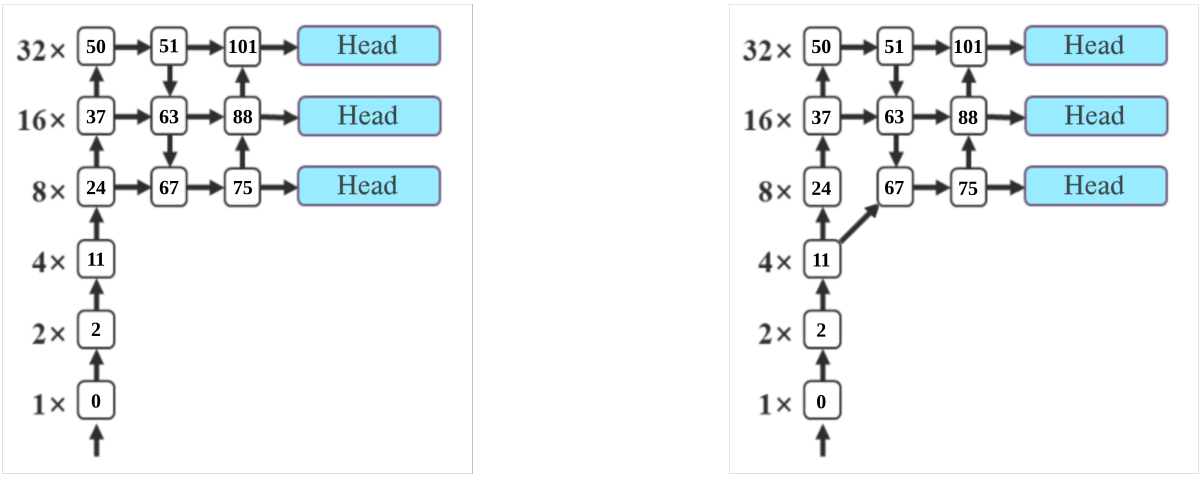
No	Modifikasi	mAP@50
0	**`YOLOv7-MAR +CIoU`** (original)	**11.2%**
1	`YOLOv7-MAR + EIoU`	0%
2	`YOLOv7-MAR + EIoU + Convexication`	4.92%
	Peningkatan	**-6.28%**

Surprisingly, although $EIoU$ outperformed $CIoU$ when applied to networks like Faster-RCNN+ResNet and RetinaNet as Peng and Yu (2021) claimed, it performed worse when applied to YOLOv7 with "Airborne Object Tracking Dataset" (2021). Even after convexciation, $CIoU$ still outperform $EIoU$ by 6.28%.

4.4 Utilizing Earlier Feature Map Stage

For this modification, we reconfigured the source of the feature map by redirecting it from P3 to P2, as shown in Figure 4.3. Specifically, we adjusted the routing on layer 66, changing it from 24 to 11. Since the output of layer 11 is at a different scale compared to 24 (with a scaling factor of 2^{-2} and 2^{-3} respectively), we modified the upsampling factor of layer 55 from 2 to 4

to ensure size compatibility with layer 11. However, this change in upsampling disrupted the subsequent layers. To solve this, we performed downsampling after layer 75 (connected to the head) by setting layer 77's stride to 2 and layer 79's stride to 4, resolving the size mismatch in the subsequent layers. The complete configuration of the layers can be seen in the appendix.



(a) Before Rerouting

(b) After Rerouting

Figure 4.3: Modifying Connection to Earlier Feature Map Stage

As presented in Table 4.4, rerouting the source of feature map to an earlier stage produce an improvement of 2.98% compared to `YOLOv7-MAR`.

Table 4.4: Performance of The Rerouted Model

No	Modifikasi	mAP@50
0	`YOLOv7-MAR`	11.2%
1	**`YOLOv7-MAR + rerouting`**	**14.09%**
	Improvement	**+2.98%**

4.5 Additional YOLO Detection Head

To add additional head layer, we utilized the feature map at the P2 scale of the network. We introduced an upsampling block after layer 75 and fused it with layer 11 at layer 79 by concatenation. To maintain the continuity of the network, we introduced a downsampling block that concludes at layer 87. The subsequent layers in the network retained their original structure, but with layer numbering shifted by 25 as seen on Figure 4.4.

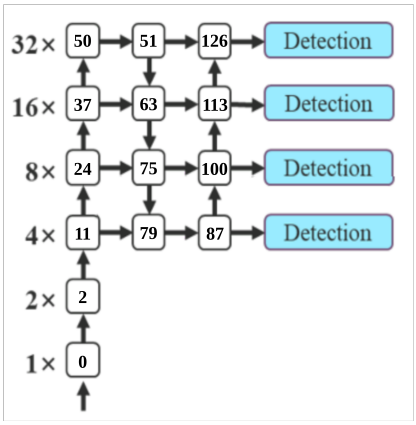


Figure 4.4: Model Architecture After Increasing The Head

Table 4.5: Additional Head Performance

No	Modifikasi	mAP@50
0	`YOLOv7-MAR`	**11.2%**
1	`YOLOv7-MAR + more head`	5.19%
	Improvement	**-6%**

As depicted in Table 4.5, additional head layer hurt the performance of the model. This is counterintuitive. We had seen P2 feature map capable of increasing the mAP in previous section. If we look at the loss at validation set, we can see that `YOLO-MAR + head` is minimizing the localization loss more than `YOLO-MAR + rerouting`. There are several reasons we thought of why this happened:

1. The network was simply producing loose bounding boxes prediction which looks better in loss function but fail when threshold criteria like in $mAP@50$ is introduced.

2. This is an effect of bad choice of hyperparameters.

4.6 Decoupled Anchor-free Head

We experimented with replacing the coupled YOLO head to of a decoupled anchor-free head. Upon replacing the head, we also replaced the label assigner from SimOTA to TAL. The performance of this modification is shown in Table 4.6.

Table 4.6: Anchor-free Head Performance

No	Model	mAP@50
0	`YOLOv7-MAR (anchor head)`	**11.2%**
1	`YOLOv7-MAR + anchor-free head`	0% Bug Fix Data Coming Up

4.7 Partitioning Image

We trained the models using partitioned image of the original dataset. In inference, we expect to partition an image into 4 equal sized images, which will then be fed to the neural network. This way, the network will produce 4 independent output that we can combine to form the final inference.

To generate partitioned data for training, we cropped image with size of $(W/2, H/2)$ based on the location of the objects bounding boxes. Therefore, it can be guaranteed that each image has atleast an object except for negative class images.

Since we are partitioning the image to a smaller size, we can finally use a smaller input size to the neural network. We experimented by using 640 and 960 input size on some promising models.

The performace is shown on Table 4.7. `YOLOv7-MAR + anchor-free` with input size 960 produced the greatest mAP@50 score amongs other model.

Table 4.7: Anchor-free Head Performance

No	Model	Input Size	Partition4 mAP@50	mAP@50
0	`YOLOv7-AR`	960	2.9%	TBA
1	`YOLOv7-MAR`	640	18.46%	TBA
3	`YOLOv7-MAR`	960	37.69%	TBA
4	`YOLOv7-MAR + rerouting`	960	30.04%	TBA
5	`YOLOv7-MAR + more head`	960	10.53%	TBA
6	`YOLOv7-MAR + anchor-free`	640	37.57%	TBA
7	**`YOLOv7-MAR + anchor-free`**	960	**46.18%**	TBA

4.8 Latency

In this section, the latency of the modifications is presented. The data obtained using computer with specification presented in section 3.3

Table 4.8: Inference Speed of Modified Models

No	Model	Input Size	Inference Speed	
			Not Reparameterized	Reparameterized
0	YOLOv7-MAR, plain, M, AR, EIoU	1600	29.6 FPS	TBA
1	YOLOv7-MAR + rerouting	1600	23.28 FPS	TBA
2	YOLOv7-MAR + more head	1600	15.91 FPS	TBA
3	YOLOv7-MAR + anchor-free	1600	TBA	TBA
4	YOLOv7-MAR	Partition 4, 960	TBA	TBA
5	YOLOv7-MAR + rerouting	Partition 4, 960	TBA	TBA
6	YOLOv7-MAR + more head	Partition 4, 960	TBA	TBA
7	YOLOv7-M + anchor-free	Partition 4, 960	TBA	TBA
8	YOLOv7-MAR	Partition 4, 640	TBA	TBA
9	YOLOv7-MAR + rerouting	Partition 4, 640	TBA	TBA
10	YOLOv7-MAR + more head	Partition 4, 640	TBA	TBA
11	YOLOv7-M + anchor-free	Partition 4, 640	TBA	TBA

[This page intentionally left blank]

CHAPTER V
CONCLUSION

5.1 Conclusion

Based on the experiments conducted in previous chapter, we can conclude that: among the modification candidates proposed in this research, we found the combination of mosaic augmentation, anchor recalculation, and rerouting feature map from P3 to P2 produced the greatest mAP@50 score 14.09%. Another improvement we made by partitioning in the input image and perform inference on each of them independently. We find that a YOLOv7 model with mosaic augmentation, and replacement of head layer with decoupled anchor-free head gives out the greatest mAP@50 score of 46.18%.

[This page intentionally left blank]

REFERENCES

Aditya, N., Indaryo, A., Solang, D., Santiung, M., Atmaja, H., Azis, A., Januar, F., Javanica, F., Kautaman, G., Kazaksti, E., Nugraha, D., Permadani, R., Ramadhan, R., Ramadhan, R., Damayanti, Z., Valentia, M., Sundana, P., Shodiq, F., Waisnawa, R., … Dikairono, R. (2022). Roboboat 2022: Technical design report Barunastra ITS roboboat team.

Airborne object tracking challenge. (2021). Retrieved September 1, 2022, from https://www.aicrowd.com/challenges/airborne-object-tracking-challenge

Airborne object tracking dataset. (2021). Retrieved September 1, 2022, from https://registry.opendata.aws/airborne-object-tracking

Amazon. (2022). *Amazon prime air prepares for drone deliveries*. Retrieved September 1, 2022, from https://www.aboutamazon.com/news/transportation/amazon-prime-air-prepares-for-drone-deliveries

Benjumea, A., Teeti, I., Cuzzolin, F., & Bradley, A. (2021). Yolo-z: Improving small object detection in yolov5 for autonomous vehicles.

Bochkovskiy, A., Wang, C.-Y., & Liao, H.-Y. M. (2020). Yolov4: Optimal speed and accuracy of object detection.

Ge, Z., Liu, S., Wang, F., Li, Z., & Sun, J. (2021). Yolox: Exceeding yolo series in 2021.

Jocher, G. (2020). *by Ultralytics* (Version 7.0). https://doi.org/10.5281/zenodo.3908559

Jocher, G., Chaurasia, A., Stoken, A., Borovec, J., NanoCode012, Kwon, Y., Michael, K., TaoXie, Fang, J., imyhxy, & et al. (2022). Ultralytics/yolov5: V7.0 - yolov5 sota realtime instance segmentation. https://doi.org/10.5281/zenodo.7347926

Li, C., Li, L., Jiang, H., Weng, K., Geng, Y., Li, L., Ke, Z., Li, Q., Cheng, M., Nie, W., Li, Y., Zhang, B., Liang, Y., Zhou, L., Xu, X., Chu, X., Wei, X., & Wei, X. (2022). Yolov6: A single-stage object detection framework for industrial applications.

Peng, H., & Yu, S. (2021). A systematic iou-related method: Beyond simplified regression for better localization. *CoRR*, *abs/2112.01793*. https://arxiv.org/abs/2112.01793

Redmon, J., & Farhadi, A. (2016). Yolo9000: Better, faster, stronger.

Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement.

Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., & Savarese, S. (2019). Generalized intersection over union: A metric and a loss for bounding box regression.

Wang, C.-Y., Bochkovskiy, A., & Liao, H.-Y. M. (2020). Scaled-YOLOv4: Scaling cross stage partial network.

Wang, C.-Y., Bochkovskiy, A., & Liao, H.-Y. M. (2022). YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors.

Wang, C.-Y., Liao, H.-Y. M., & Yeh, I.-H. (2022). Designing network design strategies through gradient path analysis.

Wang, C.-Y., Liao, H.-Y. M., Yeh, I.-H., Wu, Y.-H., Chen, P.-Y., & Hsieh, J.-W. (2019). Cspnet: A new backbone that can enhance learning capability of cnn.

Wang, C.-Y., Yeh, I.-H., & Liao, H.-Y. M. (2021). You only learn one representation: Unified network for multiple tasks.

Xiao, J. (2021). exYOLO: A small object detector based on YOLOv3 object detector. *Procedia Computer Science*, *188*, 18–25. https://doi.org/10.1016/j.procs.2021.05.048

Zhang, Z., Lu, X., Cao, G., Yang, Y., Jiao, L., & Liu, F. (2021). Vit-yolo:transformer-based yolo for object detection. *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2799–2808. https://doi.org/10.1109/ICCVW54120.2021.00314

Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., & Ren, D. (2019). Distance-iou loss: Faster and better learning for bounding box regression.

Zhong, Y., Wang, J., Peng, J., & Zhang, L. (2018). Anchor box optimization for object detection.