MULTI LEVEL REFINEMENT ENRICHED FEATURE PYRAMID NETWORK FOR SCALE AND CLASS IMBALANCE IN OBJECT DETECTION

LUBNA AZIZ

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Doctor of Philosophy

Faculty of Computing
Universiti Teknologi Malaysia

DECEMBER 2022

DEDICATION

Thank you to my supervisor, **Dr. Md Sah bin Hj Salam**, for your patience, guidance, and support. I have benefited greatly from your wealth of knowledge and meticulous editing. I am extremely grateful that you took me on as a student and continued to have faith in me over the years. I am grateful for my parents whose constant love and support keep me motivated and confident. Deepest thanks to my children “Omaisa and Abdullah, who keep me grounded, remind me of what is important in life, and are always supportive of my adventures. Finally, I owe my deepest gratitude to **Aziz ur Rehman Khan**, who is my love. I am forever thankful for the unconditional love and support throughout the entire thesis process and every day. My accomplishments and success are because he believed in me. Immense gratitude as always to Carol for his patience and support. You have always stood behind me, and this was no exception, and for the sacrifices you have made in order for me to pursue a PhD degree.

ACKNOWLEDGEMENT

In preparing this thesis, I was in contact with many people, researchers, academicians, and practitioners. They have contributed towards my understanding and thoughts. In particular, I wish to express my sincere appreciation to my main thesis supervisor, Professor Dr Sah bin Haji Salam, for encouragement, guidance, critics and friendship. I am also very thankful to Professor Dr Usman-Ullah Sheikh from School of electrical UTM, for their guidance, advices and motivation. Without their continued support and interest, this thesis would not have been the same as presented here.

I am also indebted to Higher Education Commission (HEC) Pakistan for funding my Ph.D. study. Librarians at UTM, and Balochistan University of Information Technology, Engineering, and Management Sciences Quetta Pakistan also deserve special thanks for their assistance in supplying the relevant literatures.

My fellow postgraduate student should also be recognised for their support. My sincere appreciation also extends to all my colleagues and others who have provided assistance at various occasions. Their views and tips are useful indeed. Unfortunately, it is not possible to list all of them in this limited space. I am grateful to all my family member.

ABSTRACT

Object detection becomes challenging due to feature unbalancing, less contextual information and class imbalance. The feature pyramid has been used to learn multiscale representation in modern detectors. However, the current version of the feature pyramid failed to integrate useful semantic information across different scales. In addition, many negative anchors are generated during training, resulting in extreme class imbalance. This study proposed a Multi-Level Refinement Enriched Feature Pyramid Network (MREFP-Net) to jointly handle feature-level scale imbalance and class imbalance in object detection. Instead of designing a complex approach, a simple and effective multi-layered feature enrichment scheme was proposed that effectively combines deep, intermediate, and shallow features to obtain important semantic and spatial information for small object detection. In addition, a chained parallel pooling was proposed to capture rich background contextual information. A cascaded anchor refinement scheme was introduced to integrate useful multiscale contextual information into Single Shot MultiBox Detector's prediction layers to improve the multiscale detection's distinctiveness. The ultimate goal of the cascaded anchor refinement scheme was to counteract the class imbalance by refining anchors and enriching contextual features to improve regression and classification. The performance of MREFP-Net was evaluated using two benchmark datasets, MSCOCO and PASCAL VOC 07/ 12. For a 300×300 input on MS-COCO test-dev, MREFP-Net-ResNet101 achieved a state-of-the-art detection accuracy AP of 36.6 with single-scale inference strategy and 39.2 ms on RTX 2060 GPU. For a 512×512 input on MS-COCO test-dev, MREFP-Net obtained an absolute gain of 2.5%. In particular, the results of MREFP-Net-VGG were benchmarked with 800×800 input on MS COCO test-dev: 49.2 AP with a multiscale inference strategy. For 300×300 input, MREFP-Net achieved 82.5% mAP on VOC07+12+COCO, and for 512×512 input, MREFP-Net obtained 84.6% mAP. Finally, feature visualization, object characteristic analysis and false-positive error analysis were performed to highlight the effectiveness of enriched features for small object detection. This study has proven that the proposed MREFP-Net was capable of detecting small objects and learning sensitive features to deal with scale, class imbalances, and appearance complexity across object instances.

ABSTRAK

Pengesanan objek menjadi mencabar disebabkan oleh ketidakseimbangan ciri, kurang maklumat kontekstual dan ketidakseimbangan kelas. Piramid ciri telah digunakan untuk mempelajari perwakilan pelbagai skala dalam pengesan moden. Walau bagaimanapun, versi semasa piramid ciri gagal untuk menyepadukan maklumat semantik yang berguna merentas skala yang berbeza. Di samping itu, sejumlah besar sauh negatif dijana semasa latihan menyebabkan ketidakseimbangan kelas yang melampau antara latar depan dan latar belakang. Kajian ini mencadangkan Rangkaian Piramid Ciri Diperkaya Penapisan Berbilang Tahap (MREFP-Net) untuk pengendalian bersama ketidakseimbangan skala tahap ciri dan ketidakseimbangan kelas dalam pengesanan objek. Daripada mereka bentuk pendekatan yang kompleks, skim pengayaan ciri berbilang lapisan yang mudah dan berkesan telah dicadangkan yang menggabungkan ciri dalam, pertengahan dan cetek secara berkesan untuk mendapatkan maklumat semantik dan ruang yang penting untuk pengesanan objek kecil. Di samping itu pengumpulan selari berantai telah dicadangkan untuk menangkap maklumat kontekstual latar belakang yang kaya. Skim penghalusan sauh bertingkat telah diperkenalkan untuk menyepadukan maklumat kontekstual berbilang skala yang berguna ke dalam lapisan ramalan *Single Shot MultiBox Detector* (SSD) untuk meningkatkan keistimewaan pengesanan berbilang skala. Matlamat utama skim penghalusan sauh bertingkat adalah untuk mengatasi ketidakseimbangan kelas dengan menapis sauh dan menggunakan ciri kontekstual yang diperkaya untuk meningkatkan regrasi dan klasifikasi. Prestasi MREFP-Net dinilai menggunakan dua set data penanda aras MS-COCO dan PASCAL VOC 07/ 12. Untuk saiz input 300×300 pada MS-COCO test-dev, MREFP-Net (Backbone: ResNet101) mencapai state-of- ketepatan pengesanan seni AP 36.6 dengan strategi inferens skala tunggal dan 39.2 ms pada GPU RTX 2060. Untuk input 512×512 pada MS COCO test-dev, MREFP-Net mendapat keuntungan mutlak sebanyak 2.5%. Khususnya, keputusan MREFP-Net-VGG telah ditanda aras dengan input 800×800 pada MS COCO test-dev: 49.2 AP dengan strategi inferens berbilang skala. Untuk input 300×300, MREFP-Net mencapai 82.5% mAP pada VOC07+12+COCO dan untuk input 512×512, MREFP-Net mencapai 84.6% mAP. Akhir sekali, visualisasi ciri, analisis ciri objek dan analisis ralat positif palsu telah dilakukan untuk menyerlahkan keberkesanan ciri yang diperkaya untuk pengesanan objek kecil. Kajian ini telah membuktikan bahawa MREFP-Net yang dicadangkan mampu mengesan objek kecil dan mempelajari ciri sensitif untuk menangani skala, ketidakseimbangan kelas dan kerumitan penampilan merentas kejadian objek.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ANN	-	Artificial Neural Network
AP	-	Average Precision
$AP^{IoU=0.50:0.05:0.95}$	-	AP for IoU threshold 50% to 95% with step size 0.05
$AP^{IoU=0.50}$	-	AP for IoU threshold 50%, PASCAL VOC metric
$AP^{IoU=0.75}$	-	AP for IoU threshold 75%, Static metric
AP^{small}	-	Average Precision for small scale
AP^{medium}	-	Average Precision for medium scale
AP^{large}	-	Average Precision for large scale
AR	-	Aspect Ratio
BB	-	Bounding Box
BN	-	Batch Normalization
CAD	-	Computer Assistance Diagnostic system
CPP	-	Chained Parallel Pooling
CPMC	-	Constrained Parametric Min-Cut
CNN	-	Convolutional Neural Network
DNN	-	Deep Neural Network
DPM	-	Deformable Part Model
Fast R-CNN	-	Fast Region based Convolutional Neural Network
FC	-	fully connected
FCOS	-	Fully Convolutional One-Stage Object Detection
FDRM	-	Feature directed refinement module (FDRM)
FFM	-	Feature Fusion Module
FPS	-	Frame Per Second
FPN	-	Feature Pyramid Network
FSM	-	Feature standardization module
GAN	-	Generative Adversarial Network
GD	-	Gradient Decent
GHM	-	Gradient Harmonizing Mechanism
GPU	-	Graphic Processing Unit (Enhance the graphical performance of the computer.)
GT	-	Ground Truth
ILSVRC	-	ImageNet Large Scale Visual Recognition Challenge

IoU	-	Interaction over Union
mAP	-	Mean Average Precision
MSCF	-	multi-scale contextual feature module
MS-COCO		Microsoft Common Objects in Context
MSCNN	-	Multi-Scale Convolutional Neural Network
MSFC	-	Multi-scale contextual feature
MSCA	-	multi-scale context aggregation
MREFP-Net	-	Multi-level Refinement Enriched Feature Pyramid Network
M2Det	-	A Single-Shot Object Detector based on Multi-Level Feature Pyramid Network
NAS	-	Neural Architecture Search
NMS	-	non-maxima suppression
OHEM	-	Online Hard Example Mining
OID	-	open images detection challenge
OM	-	Objectness Module
PANet	-	Path Aggregation Network
PASCAL VOC	-	PASCAL Visual Object Classes
PSIS	-	Progressive and Selective Instance Switching
PR curve	-	Precision Recall curve
RBM	-	Restricted Boltzmann Machine
R-CNN	-	Region based Convolution Neural Network
ReLU	-	Rectified Linear Unit
ResNet	-	Residual Network
RoI	-	Region of Interest
RPN	-	Region Proposal Network
SPP	-	Spatial Pyramid Pooling
SSD	-	Single Shot Multibox Detector
STDN	-	Scale-Transferrable Detection Network
SVM	-	Support Vector Machine
TPU	-	Tensor Processing Unit. Custom build ASIC to accelerate TensorFlow projects
TUM	-	Thinned U-shape module
XML	-	Extensible Markup Language
YOLO	-	You Only Look Once
YOLSO	-	You Only Look Small Objects

LIST OF SYMBOLS

σ	-	Activation function
μ_B	-	is the empirical mean, evaluated over the whole batch B.
σ_B	-	is the empirical standard deviation, also evaluated over the whole mini batch.
m_b	-	is the number of instances in the batch
$\hat{x}^i$	-	is the zero-centred and normalised input
γ	-	is the scaling parameter for the layer
β	-	is the shifting parameter for the layer
$\in$	-	is a tiny number to avoid division by zero
$w_{i,j}^l$	-	Weights
H	-	Refers image height
W	-	Refers image width
α	-	α-balanced cross entropy loss
ω_i	-	contribution of samples
β	-	controls the normalized rank contribution
$(\Delta cx, \Delta cy)$	-	Centre of default bounding box
Δw	-	Width of default bounding box
Δh	-	Height of default bounding box

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Overview

Object detection is the simultaneous prediction of categories and accurate object localization, that helps to understand the image correctly. It is an fundamental issue in computer vision that encompasses many practical applications for instance self-driving (autonomous) cars (Dai, 2019; Geiger et al., 2012), surveillance (Fan et al., 2016; Fu et al., 2019), medical analysis and decision making (Jaeger et al., 2020; Lee et al., 2018) and, other issues in robotics (Du et al., 2021; Hampali et al., 2020; He et al., 2020; Peng et al., 2019).

In the past time, object detection was considered as a machine learning problem and relied on handcrafted features and max margin, linear classifier. The Deformable Part Model (DPM) (Felzenszwalb et al., 2009) is the best known and most successful method of its time (Liu et al., 2020; Oksuz et al., 2020b). As of 2012, deep neural networks have dominated various computer vision problems after the some extremely influential research was conducted. In the current generation of object detection methods, deep neural networks are used instead of handcrafted features and linear classifiers of early-generation object detection methods. This change has led to significant performance improvements in a extensively used object detection benchmark dataset (MS COCO and PASCAL VOC) (Everingham et al., 2015; T.-Y. Lin et al., 2014). The driving force behind the advancement of object detection over the past half decade has been in the hands of deep networks (Dai et al., 2016a; Girshick et al., 2014; Gkioxari et al., 2015; Law et al., 2018; T.-Y. Lin et al., 2017; Liu et al., 2016; Redmon et al., 2016; Ren et al., 2015c). While imbalance problem at various levels has received a great deal of attention in object detection (Cao et al., 2020; Lin et al., 2017); Ouyang et al. (2016); (Pang et al., 2019;

Shrivastava et al., 2016; Singh and Davis, 2018; Singh et al., 2018b). An imbalance problem related to an input property arises when distribution of that property directly influences the performance of object detector. If this is not corrected in a timely manner, an imbalance problem will adversely affect the final detection performance. The background-to-foreground class imbalance is most common imbalance problem in object recognition, which is an extreme disparity between the number of negative instances and the number of positive ones. Although a given picture usually contains a few positive instances, millions of negatives instances can be extracted. However, this can result in slightly negative-dominated training and extreme imbalance between foreground and background classes. Moreover, another substantial problem for many detectors is the handling of scale diversity, since object examples vary over a wide range. Feature-level scale imbalance cause feature inconsistency across different scales and increase risk of overfitting for each scale. These types of imbalances restrict the effectiveness of well-deigned models from being fully exploited, thus significantly affects the accuracy of the recognition, if not corrected.

The literature on deep learning-base object detection identifies different types of imbalance problems. They are taxonomized into four major groups such as scale imbalance, class imbalance, objective imbalance, and spatial imbalance (Oksuz et al., 2020b). Class imbalance arises when there is large disparity between the number of instances related to various classes such as an imbalance between foreground and background classes, and imbalance between the foreground classes (i.e., positive). An imbalance related to scales encounters when objects have varying scales and different numbers of instances that refer to various scales. some factors related to spatial attributes of bounding boxes cause spatial imbalance such as Intersection over Union (IoU), regression penalty, and location. Any imbalance in spatial attributes directly influences the detection performance and training. Ultimately, an objective imbalance happens when various loss functions have to be penalized, as is frequently the case in object detection such as regression and classification losses.

In general, imbalance problems are prevalent in machine learning, computer vision, and pattern recognition. However, class imbalance and scale imbalance problem that arises in object detection are the only emphasis of this work. This

dissertation discusses the issue of background-foreground class imbalance and feature-level scale imbalance and offers a plan to address these imbalances jointly.

1.2 Problem Background

Object detection is a fascinating and active topic of computer vision research, which seek to identify instances of semantic objects belonging to a specific class in videos and digital images (Felzenszwalb et al., 2009). Which objects are where? Is the foundation of object detection framework's functional premise.

Conventional object detection models consist of three steps: selecting an information region, extracting features, and classifying selected region based on the extracted features (Z.-Q. Zhao et al., 2019). Top-down and bottom-up are the two main approaches to object detection. Despite the fact that both strategies were well-known in the early era of object detection, top-down approaches predominate in current object detection methodologies while bottom-up approaches have only lately been developed. The primary distinction between two approaches is that, the top-down approach generates and process holistic object hypothesis, such as anchors, region of interest (RoIs) /proposals, early in detection pipeline. However, in bottom-up approach, holistic objects emerge by grouping sub-object entities like parts or key-points, later in processing pipeline.

Top-down based methods are categorized into two groups: one-stage detection methods and two-stage detection methods. Two-stage methods (Dai et al., 2016a; Girshick, 2015b; Girshick et al., 2014; Ren et al., 2015a) are less sensitive to class imbalance. It aims to reduce the large number of negative examples generating from predefined dense sliding window (i.e., anchors) to a manageable size by using proposal mechanism (Ren et al., 2015a; Uijlings et al., 2013; Zitnick et al., 2014) that determines the region where the objects most likely appear, known as Region of Interest (RoIs). In two-stage methods, these proposals (RoIs) are generated by a separate network using anchors, hence it is called a two-stage detector. In addition, these RoIs are further processed to generate bounding boxes and class scores of

detecting objects. Finally, duplicate or highly overlapping results are eliminated using the non-maxima suppression (NMS) method. So far, it is a universal step used in all modern detectors for post processing.

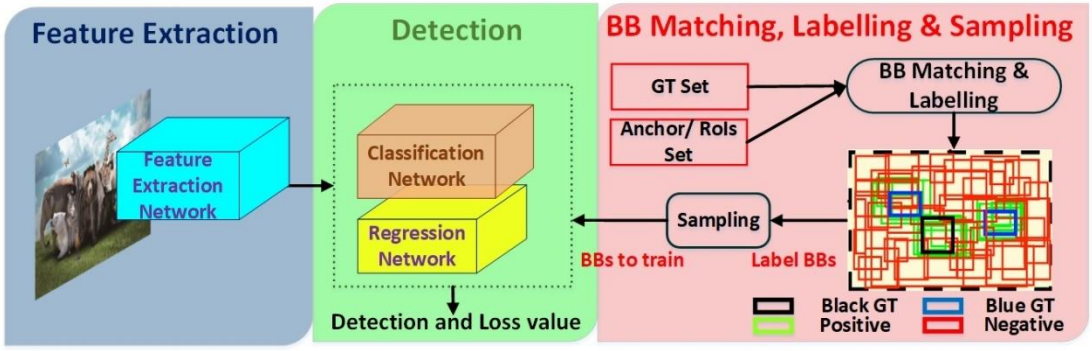
One-stage methods predict the object directly from anchor using feature extracted from input image without any proposal elimination stage. Some well-known approaches that belong to top-down one-stage methods are Single Shot Detector (SSD) variants (Fu et al., 2017; Liu et al., 2016), You Look Only Once (YOLO) variants (Bochkovskiy et al., 2020; Redmon et al., 2016; Redmon and Farhadi, 2017, 2018), and RetinaNet (T.-Y. Lin et al., 2017). To detect small object, a large number of anchors are generated per image, compared to detecting large objects. Only anchors with high Intersection over Union (IoU) to the ground truth are considered as positive examples. Since most anchors have low or no overlap with ground truth bounding boxes, they are considered as negative examples. When densely generated anchors are matched with sparsely located real objects in the images, very small fraction of positive examples are found, resulting in a high-class imbalance.

A typical detection pipeline of one-stage object detector is shown in Figure 1.1(a). The process begins with feature extraction when the input image being fed to feature extraction block i.e., deep network. Then, in bounding box matching, labelling, and sampling phase, a compact set of object hypotheses (i.e., anchors) is generated, that are further labelled and sampled using ground-truth bounding boxes (i.e., Back GT, Blue GT). At last, these labelled anchors or bounding boxes (BB), whose features are generated from feature extraction network, are used for training in classification and regression networks. In addition, as indicated in Figure 1.1(b) various imbalance problems arises at various stages of training pipeline. Whereas, at which phase an imbalance occurs is specified by background color. Scale imbalance arises at feature extraction stage, while class imbalance occurs at bounding box matching and labelling stage.

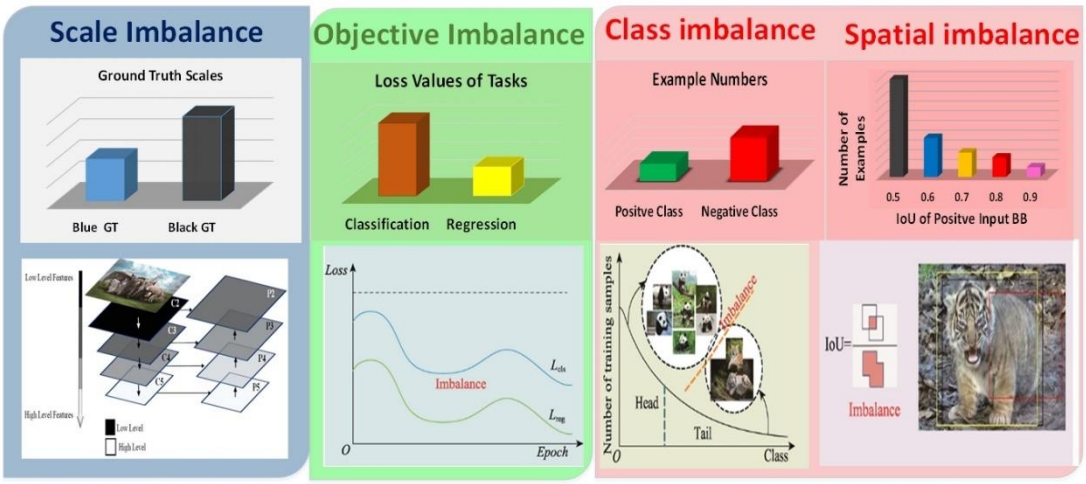
On the contrary, bottom-up object detection approaches (Duan et al., 2019; Law et al., 2018; Zhou et al., 2019a) first predict the contextual key-elements (e.g.,

corners, centres, etc.) on objects and then cluster them to form a complete object using an assemblage method such as brute force search (Zhou et al., 2019a) and associative embedding (Newell et al., 2017).

In general, one-stage top-down methods have advantage that they are more computationally efficient compared to two-stage methods. Contrastingly, two-stage methods (J. Huang et al., 2017) are more accurate but compromise on efficiency. The scope of this work is restricted to investigate the imbalance problem of generic object detection in top-down one-stage method.



(a)



(b)

Figure 1.1 (a) Generic object detection training pipeline. (b) Define the types of imbalance problems for object detection through training pipeline. (Oksuz et al., 2020b).

Scale imbalance is a central and challenging issue for both object detection methods (Lin et al., 2017). The size of object instances varies widely, which hinders the detectors to performed well, in particular gigantic or miniature ones. The multi-scale image pyramid is one of solution to cope with the large scale variation (Adelson et al., 1984), that is most likely used in both traditional methods based on handcrafted features (Dalal et al., 2005; Lowe, 2004), and current deep learning based methods. Previous studies (J. Huang et al., 2017; Liu et al., 2018) show that multi-scale training and testing were beneficial for deep detectors (Dai et al., 2016a; Ren et al., 2016). SNIP (Singh and Davis, 2018; Singh et al., 2018b) proposes a normalization method for scaling that selectively trains the appropriately sized objects at each image scale to avoid training with extreme scales such as big/small objects in larger/ smaller scales. Despite the performance gain, Image pyramid methods are less suitable for everyday applications in real time due to increase of inference time and high memory requirements (Li et al., 2019). Due to memory constraints, another drawback of image pyramid is that detectors are trained at smaller scales but tested on larger scales, leading to inconsistency between training and testing time inference. A large amount of information would be lost if the detector were trained on specific scale, while multiscale training on single detector would induce scale imbalance by preserving data variance.

The feature pyramid (Lin et al., 2017; Liu et al., 2016) is another approach that is used in the network to provide approximately the same performance as an image pyramid with less computational effort and to require significantly less additional memory, that enable the deployment of such network during both the training and test phase in real-time network. In addition, the feature pyramid module can be easily revised and built into modern detectors based on deep neural networks. In (Dollár et al., 2014), a fast feature pyramid for object recognition is constructed by incorporating some feature channels from adjacent scale levels. Whereas in SSD (Liu et al., 2016) multi-scale feature maps from different backbone layers are exploited to detect objects of various scales on each feature layer. Feature pyramid contains multi-level features, whereby the feature uniformity is sacrificed across different scales. This reduces effective training data and increases the risk of overfitting for each scale. Feature pyramid network (FPN) (Lin et al., 2017) integrate the robust semantic information in deep features using lateral connections and top-down

pathway to compensate for the lack of semantics in shallow features. However, contextual features are extracted from objects of different scales from various levels of FPN backbone. Major drawback of FPN is feature unbalancing due to direct integration from backbone network and lose of translation invariance (Li et al., 2019). Therefore, the focus of deep learning designers over the past decade has been to find a more realistic technique to create a more effective and descriptive multi-scale feature in order to resolve the scale variance problem for object detection.

Deep CNN architecture produces hierarchy feature maps due to sampling and pooling operations, resulting different layers of feature maps with different abstractions and spatial resolutions that make it unreliable to predict directly from individual layer. The shallow layer feature maps are of high resolution and contain basic semantic information such as shape, and location, that is used for object localization. On the other hand, feature maps extracted from deeper layers contain rich semantic information that is beneficial for object recognition. Even though high-level features are useful for identifying large objects, they may not be sufficient fo small object detection. Thus, feature fusion at different depths would have a significant positive influence on object detection task. Therefore, these reasons have motivated researchers to work on feature fusion strategies and compare them to the existing state-of-the-art detection methods that have previously shown to be efficient for scale imbalance problem.

Background-to-foreground class imbalance is an over-representation of background and under-representation of foreground classes. This type of problem occurs during training and is unavoidable since most boxes are labels as negative class (i.e., background) by bounding box labelling and matching module. It does not rely on the number of instances per class in the dataset, as they contain no background annotations. Hard sampling is a simple solution for this type of imbalance in object detection. It is based on random sampling and is used in R-CNN detector family (Girshick et al., 2014; Ren et al., 2016). The study shows that other sampling strategies may perform better if the loss value or IoU of input box are taken into account (Cao et al., 2020; Pang et al., 2019; Shrivastava et al., 2016). Hard-example mining methods use the sampled examples instead of random sampling.

Training the model more with hard examples leads to better performance, is the hypothesis behind this method. The origin of this hypothesis is based on bootstrapping, which is used in early research on face detection (Rowley et al., 1995), human detection (Dalal et al., 2005) and object detection (Felzenszwalb et al., 2009). Single shot detector was the first deep model that uses the hard examples (i.e., high loss examples) for training. Online Hard Example Mining (OHEM) (Shrivastava et al., 2016) is another version of hard sampling approach, however it requires additional memory and leads to reduction in training speed.

Some improved mining strategies have been proposed to limit the search space and make mining the hard examples easier. Kong et al. (2017), proposed as an approach to learn objectness (i.e., how likely a box is to cover an object) early in an end-to-end setting to provide direction on where to look for the object. All positive instances with higher objectness than threshold is used in training, whereas negative instances so be chosen to maintain the balance between negative and positive instances. S. Zhang et al. (2018), proposed a module to refine the anchors and determine the confidence scores of anchors. A threshold has also been introduced to remove the simple negative anchors, known as negative anchor filtering. In (Nie et al., 2019a), an SSD-based cascaded-detection scheme were proposed that includes the objectness module (i.e., binary classifier) prior to each prediction module. Classic objectness methods (Alexe et al., 2010; Cheng et al., 2014) are used to decrease the number of proposal windows for quicker detection. Recent state-of-the-art single-stage detectors use the objectness-like mechanism to address the background-to-foreground class imbalance, such as objectness module in RON (Kong et al., 2017), YOLO with objectness (Redmon et al., 2016; Redmon and Farhadi, 2017, 2018), and anchor refinement mechanism in RefineDet (S. Zhang et al., 2018). However, majority of detectors with objectness module still maintain hard example mining or sampling heuristics schemes, which thereby creating the opportunities to propose a newly emerged approach to jointly tackle class imbalance and scale imbalance problems. In the light of the foregoing, this research intends to address the problems highlighted in the preceding discussions.

1.3 Problem Statement

The challenge is to develop a one-stage object detection network to jointly tackle the feature-level scale imbalance and foreground-background class imbalance with minimal computational and memory requirements. Two-stage detectors are more accurate than one-stage detectors, but not suitable for real time detection applications due to high computational cost (Liu et al., 2020). Single Shot Detector (SSD) (Liu et al., 2016) is a good option for real time application due to high efficiency and accuracy. However, there are some obstacles that prevent the model from achieving a significant gain in accuracy. First, it has difficulty in dealing with scale variance due to fixed contextual information. Second, it neglects the bounding box scale diversity as it uses last two layers of backbone network for prediction. It has difficulty in detecting small objects and densely structured complex scenes due to feature unbalancing, less contextual information and class imbalance.

Nature shows that object instances exist at a variety of scales that could prevent the detector to perform well, especially for very large or very tiny ones. A single feature layer does not contain enough information to detect objects at multiple scales. However, existing techniques such as feature pyramid sacrifices the feature consistency across different scales by generating multi-level features. This leads to the higher rick of overfitting for each scale and decrease in effective training data (Li et al., 2019). Although, object detector has shown promising results with feature pyramid, but construction of feature pyramid is intrinsic to the backbone pyramid architecture, that specifically designed for classification purposes rather than detection. This makes the feature map in pyramid less representative for object detection task. Although, each feature map in pyramid constructed from single-level backbone layers that provide only single-level information to detect an object of certain size will yield sub-optimal results. Feature pyramid network has strong representational power of deep model but it not addresses multi-scale problem due feature-level scale imbalance. In general, low-level features in shallow layers are better suited for localization subtask and to describe object with simple appearance, whereas high-level features in deep layers are appropriate for classification subtask and to characterize object with complex appearance. The literature review also

indicates that middle-level feature is not only necessary for shallow features that encode the basic visual geometrical information, but also for higher-level features that encode category level information. The use of middle-level features in object detection is still an open question. Therefore, following research question have been formulated:

1. To what extent will the inclusion of contextual features be effective for small object detection?
2. Does using intermediate features for prediction improve detector performance?

In most cases, both types of detectors rely on an anchoring mechanism to cover shape-diversity and objects with different scale, sampling the dense boxes evenly over the spatial domain. In spite of that, it always causes extreme foreground-background imbalance with increase in anchors (e.g., ~100k), which yield a negative dominated training. In fact, foreground-background imbalance does not equally damage the detection performance in all detectors. In two-stage detectors, such an imbalance due to Region Proposal Network (RPN) is mitigated by filtering out most of the negatives. In contrast, for imbalance sensitive one-stage detectors, number of sampling/ reweighting scheme have been proposed. Despite being effective, these schemes require laborious hyper-parameters tuning and are usually heuristic. Therefore, following research question has been formulated:

3. Instead of complicated, heuristic sampling/reweighting schemes, how much more practical and simple learning-based approach is effective to address the foreground-background class imbalance problem with multi-scale features?

1.4 Aim and Objectives

Based on the formulated research questions, this research aims to propose a simple learning-based scheme to jointly address the problems of scale imbalance, and class imbalance, in order to minimize the hardship of detecting objects with high

shape diversity and scale-variant. This scheme is expected to improve the detection accuracy of modern detectors with less computational cost.

The main objectives of this thesis are:

1. To design a one-stage detector for small object detection that jointly tackle scale imbalance and class imbalance problems.

2. To develop multi layered feature enrichment scheme to tackle feature level scale imbalance and optimized with integration of contextual features using Chained Parallel Pooling (CPP).

3. To develop cascaded anchor refinement scheme to minimize the foreground-background class imbalance by filtering out negative anchors.

4. To benchmark the proposed MREFP-Net in terms of detection performance and efficiency.

1.5 Scope of this Study

The scope of the thesis is limited to the following areas:

1. This thesis focuses solely on the deep learning based one-stage detector for generic object detection. The research considers a Single Shot Detector (SSD) incorporating reformulated feature pyramid network and anchor refinement to evaluate the performance of proposed method.

2. The literature review focuses only on solving the feature-level scale imbalance and background-to-foreground class imbalance in object recognition task.

3. The proposed model (single-stage detector) is tested on two benchmark datasets, such as MS COCO and PASCAL Visual Object Classes (VOC) 07/12.

I. COCO has 1.5 million object instances for 80 object categories. COCO stores annotations in a JSON file. The MS COCO dataset is published by Microsoft Machine Learning and Computer Vision engineers. It is a large object detection, captioning and segmentation dataset, but the COCO object detection dataset is used for this research.

II. The PASCAL VOC dataset contains 20 object categories including vehicles, household, animals, and other: aeroplane, bicycle, boat, bus, car, motorbike, train, bottle, chair, dining table, potted plant, sofa, TV/monitor, bird, cat, cow, dog, horse, sheep, and person. Pascal VOC is an XML file, unlike COCO which has a JSON file

4. The performance of the proposed method is compared with State-of-the-art object detectors.

5. Consequently, the Tensorflow platform is used to implement the scheme in Python language.

1.6 Research Significance

Imbalance problem will adversely affect final detection performance if not managed effectively. Scale variation across object instances prevent the detector from performing well, especially for very large or very small ones. Whereas class imbalance causes negative dominating training. Despite of effective, previous schemes require laborious hyper-parameters tuning and are usually heuristic. Instead of designing complicated, heuristic sampling/ reweighting schemes, it is required to find a feasible and simple learning-based approach to address imbalance problem. The network proposed in this thesis can be used to jointly tackle the scale imbalance and class imbalance problem. It generates more descriptive feature pyramid to handle scale imbalance. In fact, more contextual information is added to improve the detection precision. While using negative anchor filtering mechanism to lessen the

search space and minimize the class imbalances. The proposed network aims to be an effective means of tackling both imbalance problems together.

1.7 Thesis Organization

This thesis is organized into six chapters. **Chapter 1** encompasses an overview, background of the research problem, problem statement, research objectives, the scope and significance of study. The content of the residual chapters is outlined as follows.

Chapter 2 presents a strategic literature review that covers various aspects of the imbalance problems found in generic object detection. It begins with the basic concepts the object detection, their types, details of backbone networks and benchmark datasets used for generic object detection. This is followed by reviews of detection challenges. Afterward, a comprehensive overview of imbalance problem more specifically class imbalance and scale imbalance, their types and proposed solutions is presented.

Chapter 3 explains the methodology employed to accomplish the thesis objectives. It also presents the research framework, which outlines the various stages and activities that led to the actualization of the thesis objectives.

In **chapter 4**, the architecture of the different components which constitute the proposed MREFP-Net (Multi-level Refinement Enriched Feature Pyramid Network) under study are described. This is followed by a brief description of relevant literature. Subsequently, the development of the proposed MREFP-Net for addressing scale imbalance and class imbalance is accomplished using a Single Shot Detector (SSD), then came the formulation of the multi-layered feature enrichment scheme and cascaded anchor refinement scheme. Lastly, the objective loss function is discussed.

Chapter 5 commences with the introduction of the system implementation details and comparison with other modern models on standard benchmark datasets. The

ablation studies and model analysis are discussed in the next section. The variants of Multi-level Enriched Feature module are presented. Consequently, the results of object detection using multi-layered feature enrichment scheme and cascaded anchor refinement scheme are presented. In the last section object detection results using benchmark datasets are presented.

Chapter 6 summarizes the conclusion of the dissertation and highlights the contribution of the proposed work. In addition, possible areas for future research are highlighted.

REFERENCES

Adelson, E. H., Anderson, C. H., Bergen, J. R., Burt, P. J., and Ogden, J. M. (1984). Pyramid methods in image processing. *RCA engineer*, *29*(6), 33-41.

Alexe, B., Deselaers, T., and Ferrari, V. (2010). What is an object? 2010 IEEE computer society conference on computer vision and pattern recognition,

Aziz, L., Salam, M. S. B. H., Sheikh, U. U., and Ayub, S. (2020). Exploring deep learning-based architecture, strategies, applications and current trends in generic object detection: A comprehensive review. *IEEE Access*, *8*, 170461-170495.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, *35*(8), 1798-1828.

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

Bodla, N., Singh, B., Chellappa, R., and Davis, L. S. (2017). Soft-NMS--improving object detection with one line of code. Proceedings of the IEEE international conference on computer vision,

Cai, Z., Fan, Q., Feris, R. S., and Vasconcelos, N. (2016). A unified multi-scale deep convolutional neural network for fast object detection. European conference on computer vision,

Cai, Z., and Vasconcelos, N. (2018). Cascade r-cnn: Delving into high quality object detection. Proceedings of the IEEE conference on computer vision and pattern recognition,

Cao, J., Pang, Y., and Li, X. (2019). Triply supervised decoder networks for joint detection and segmentation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

Cao, Y., Chen, K., Loy, C. C., and Lin, D. (2020). Prime sample attention in object detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

Cao, Z., Simon, T., Wei, S.-E., and Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. Proceedings of the IEEE conference on computer vision and pattern recognition,

Carreira, J., and Sminchisescu, C. (2011). CPMC: Automatic object segmentation using constrained parametric min-cuts. *IEEE transactions on pattern analysis and machine intelligence, 34*(7), 1312-1328.

Chen, E., Peng, X., Chen, J., Luo, B., Xu, T., and Liu, D. (2019). Residual objectness for imbalance reduction. *arXiv preprint arXiv:1908.09075*.

Chen, J., Liu, D., Xu, T., Wu, S., Cheng, Y., and Chen, E. (2021). Is Heuristic Sampling Necessary in Training Deep Object Detectors? *IEEE Transactions on Image Processing, 30*, 8454-8467.

Chen, J., Liu, D., Xu, T., Zhang, S., Wu, S., Luo, B., . . . Chen, E. (2019). Is sampling heuristics necessary in training deep object detectors? *arXiv preprint arXiv:1909.04868*.

Chen, K., Li, J., Lin, W., See, J., Wang, J., Duan, L., . . . Zou, J. (2019). Towards accurate one-stage object detection with ap-loss. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,

Chen, K., Lin, W., Li, J., See, J., Wang, J., and Zou, J. (2020). AP-loss for accurate one-stage object detection. *IEEE transactions on pattern analysis and machine intelligence*, *43*(11), 3782-3798.

Chen, K., Pang, J., Wang, J., Xiong, Y., Li, X., Sun, S., . . . Ouyang, W. (2019). Hybrid task cascade for instance segmentation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

Chen, S., Ma, W., and Zhang, L. (2022). Dual-bottleneck feature pyramid network for multiscale object detection. *Journal of Electronic Imaging*, *31*(1), 013009.

Chen, Y., Yang, T., Zhang, X., Meng, G., Xiao, X., and Sun, J. (2019). DetNAS: Backbone search for object detection. Advances in neural information processing systems,

Cheng, M.-M., Zhang, Z., Lin, W.-Y., and Torr, P. (2014). BING: Binarized normed gradients for objectness estimation at 300fps. Proceedings of the IEEE conference on computer vision and pattern recognition,

Cortes, Corinna, Vapnik, and Vladimir. (1995). Support-vector networks. *Machine learning*, *20*(3), 273-297.

Creswell, J. (2012). Educational research: Planning, conducting, and evaluating quantitative and qualitative research (rd ed.). *Canada: Pearson. Ministry of Higher Education.(2012). Critical agenda projects under the national higher education strategic plan*.

Dahl, G., Ranzato, M. A., Mohamed, A.-r., and Hinton, G. E. (2010). Phone recognition with the mean-covariance restricted Boltzmann machine. Advances in neural information processing systems,

Dai, Jifeng, Qi, Haozhi, Xiong, Y., Li, Y., . . . Wei, Y. (2017). Deformable convolutional networks. Proceedings of the IEEE international conference on computer vision,

Dai, J., Li, Y., He, K., and Sun, J. (2016a). R-fcn: Object detection via region-based fully convolutional networks. Advances in neural information processing systems,

Dai, J., Li, Y., He, K., and Sun, J. (2016b). R-fcn: Object detection via region-based fully convolutional networks. *arXiv preprint arXiv:1605.06409*.

Dai, X. (2019). HybridNet: A fast vehicle detection system for autonomous driving. *Signal Processing: Image Communication*, *70*, 79-88.

Dai, X., Chen, Y., Xiao, B., Chen, D., Liu, M., Yuan, L., and Zhang, L. (2021). Dynamic Head: Unifying Object Detection Heads with Attentions. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

Dalal, Navneet, Triggs, and Bill. (2005). Histograms of oriented gradients for human detection. 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05),

Deng, C., Wang, M., Liu, L., Liu, Y., and Jiang, Y. (2021). Extended feature pyramid network for small object detection. *IEEE Transactions on Multimedia*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. 2009 IEEE conference on computer vision and pattern recognition,

Deng, L., Seltzer, M. L., Yu, D., Acero, A., Mohamed, A.-r., and Hinton, G. (2010). Binary coding of speech spectrograms using a deep auto-encoder. Eleventh Annual Conference of the International Speech Communication Association,

Divvala, S. K., Hoiem, D., Hays, J. H., Efros, A. A., and Hebert, M. (2009). An empirical study of context in object detection. 2009 IEEE Conference on computer vision and Pattern Recognition,

Dollár, P., Appel, R., Belongie, S., and Perona, P. (2014). Fast feature pyramids for object detection. *IEEE transactions on pattern analysis and machine intelligence*, *36*(8), 1532-1545.

Dollar, P., Wojek, C., Schiele, B., and Perona, P. (2011). Pedestrian detection: An evaluation of the state of the art. *IEEE transactions on pattern analysis and machine intelligence*, *34*(4), 743-761.

Du, G., Wang, K., Lian, S., and Zhao, K. (2021). Vision-based robotic grasping from object localization, object pose estimation to grasp estimation for parallel grippers: a review. *Artificial Intelligence Review*, *54*(3), 1677-1734.

Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., and Tian, Q. (2019). Centernet: Keypoint triplets for object detection. Proceedings of the IEEE International Conference on Computer Vision,

Dvornik, N., Mairal, J., and Schmid, C. (2018). Modeling visual context is key to augmenting object detection datasets. Proceedings of the European Conference on Computer Vision (ECCV),

Dwibedi, D., Misra, I., and Hebert, M. (2017). Cut, paste and learn: Surprisingly easy synthesis for instance detection. Proceedings of the IEEE International Conference on Computer Vision,

Erhan, D., Szegedy, C., Toshev, A., and Anguelov, D. (2014). Scalable object detection using deep neural networks. Proceedings of the IEEE conference on computer vision and pattern recognition,

Everingham, M., Eslami, S., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. (2015). The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, *111*(1), 98-136.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International journal of computer vision*, *88*(2), 303-338.

Fadadu, S., Pandey, S., Hegde, D., Shi, Y., Chou, F.-C., Djuric, N., and Vallespi-Gonzalez, C. (2022). Multi-view fusion of sensor data for improved perception and prediction in autonomous driving. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision,

Fan, Q., Brown, L., and Smith, J. (2016). A closer look at Faster R-CNN for vehicle detection. 2016 IEEE intelligent vehicles symposium (IV),

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., and Ramanan, D. (2009). Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, *32*(9), 1627-1645.

Fischler, M. A., and Elschlager, R. A. (1973). The representation and matching of pictorial structures. *IEEE Transactions on computers*, *100*(1), 67-92.

Freund, Yoav Schapire, and E, R. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, *55*(1), 119-139.

Fu, C.-Y., Liu, W., Ranga, A., Tyagi, A., and Berg, A. C. (2017). Dssd: Deconvolutional single shot detector. *arXiv preprint arXiv:1701.06659*.

Fu, Z., Chen, Y., Yong, H., Jiang, R., Zhang, L., and Hua, X.-S. (2019). Foreground gating and background refining network for surveillance object detection. *IEEE Transactions on Image Processing*, *28*(12), 6077-6090.

Gao, Z., Wang, L., Han, B., and Guo, S. (2022). AdaMixer: A Fast-Converging Query-Based Object Detector. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

Ge, Z., Liu, S., Wang, F., Li, Z., and Sun, J. (2021). Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*.

Geiger, A., Lenz, P., and Urtasun, R. (2012). Are we ready for autonomous driving? the kitti vision benchmark suite. 2012 IEEE conference on computer vision and pattern recognition,

Ghiasi, G., Lin, T.-Y., and Le, Q. V. (2019a). Nas-fpn: Learning scalable feature pyramid architecture for object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,

Ghiasi, G., Lin, T.-Y., and Le, Q. V. (2019b). Nas-fpn: Learning scalable feature pyramid architecture for object detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

Girshick, R. (2015a). Fast r-cnn. Proceedings of the IEEE international conference on computer vision,

Girshick, R. (2015b). Fast R-CNN ICCV. 15Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV),

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition,

Gkioxari, G., Girshick, R., and Malik, J. (2015). Contextual action recognition with r* cnn. Proceedings of the IEEE international conference on computer vision,

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Grauman, K., and Leibe, B. (2011). Visual object recognition. *Synthesis lectures on artificial intelligence and machine learning*, *5*(2), 1-181.

Hampali, S., Rad, M., Oberweger, M., and Lepetit, V. (2020). Honnotate: A method for 3d annotation of hand and object poses. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask r-cnn. Proceedings of the IEEE international conference on computer vision,

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, *37*(9), 1904-1916.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition,

He, Y., Sun, W., Huang, H., Liu, J., Fan, H., and Sun, J. (2020). Pvn3d: A deep point-wise 3d keypoints voting network for 6dof pose estimation. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A.-r., Jaitly, N., . . . Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal processing magazine*, *29*.

Hinton, G. E., Krizhevsky, A., and Wang, S. D. (2011). Transforming auto-encoders. International conference on artificial neural networks,

Hinton, G. E., and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, *313*(5786), 504-507.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

Hoiem, D., Chodpathumwan, Y., and Dai, Q. (2012). Diagnosing error in object detectors. European conference on computer vision,

Hong, S., Roh, B., Kim, K.-H., Cheon, Y., and Park, M. (2016). PVANet: Lightweight deep neural networks for real-time object detection. *arXiv preprint arXiv:1611.08588*.

Hosang, J., Benenson, R., and Schiele, B. (2016). A convnet for non-maximum suppression. German conference on pattern recognition,

Hosang, J., Benenson, R., and Schiele, B. (2017). Learning non-maximum suppression. Proceedings of the IEEE conference on computer vision and pattern recognition,

Hou, Z., Yu, B., and Tao, D. (2022). BatchFormer: Learning to Explore Sample Relationships for Robust Representation Learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., . . . Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Hu, H., Gu, J., Zhang, Z., Dai, J., and Wei, Y. (2018). Relation networks for object detection. Proceedings of the IEEE conference on computer vision and pattern recognition,

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

Hu, P., and Ramanan, D. (2017). Finding tiny faces. Proceedings of the IEEE conference on computer vision and pattern recognition,

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition,

Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., . . . Guadarrama, S. (2017). Speed/accuracy trade-offs for modern convolutional object detectors. Proceedings of the IEEE conference on computer vision and pattern recognition,

Ioffe, S., and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

Jaeger, P. F., Kohl, S. A., Bickelhaupt, S., Isensee, F., Kuder, T. A., Schlemmer, H.-P., and Maier-Hein, K. H. (2020). Retina U-Net: Embarrassingly simple exploitation of segmentation supervision for medical object detection. Machine Learning for Health Workshop,

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., . . . Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. Proceedings of the 22nd ACM international conference on Multimedia,

Jiang, N., Zhang, Y., Luo, D., Liu, C., Zhou, Y., and Han, Z. (2019). Feature hourglass network for skeleton detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops,

Kang, J., Shin, J., Shin, J., Lee, D., and Choi, A. (2022). Robust Human Activity Recognition by Integrating Image and Accelerometer Sensor Data Using Deep Fusion Network. *Sensors*, *22*(1), 174.

Kavukcuoglu, K., Ranzato, M. A., Fergus, R., and LeCun, Y. (2009). Learning invariant features through topographic filter maps. 2009 IEEE Conference on Computer Vision and Pattern Recognition,

Kim, S.-W., Kook, H.-K., Sun, J.-Y., Kang, M.-C., and Ko, S.-J. (2018). Parallel feature pyramid network for object detection. Proceedings of the European Conference on Computer Vision (ECCV),

Kirillov, A., Girshick, R., He, K., and Dollár, P. (2019). Panoptic feature pyramid networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,

Kong, T., Sun, F., Tan, C., Liu, H., and Huang, W. (2018). Deep feature pyramid reconfiguration for object detection. Proceedings of the European conference on computer vision (ECCV),

Kong, T., Sun, F., Yao, A., Liu, H., Lu, M., and Chen, Y. (2017). Ron: Reverse connection with objectness prior networks for object detection. Proceedings of the IEEE conference on computer vision and pattern recognition,

Krasin, I., Duerig, T., Alldrin, N., Ferrari, V., Abu-El-Haija, S., Kuznetsova, A., . . . Veit, A. (2017). Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github. com/openimages*, *2*(3), 18.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012a). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*, 1097-1105.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012b). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems,

Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., . . . Duerig, T. (2018). The open images dataset V4: unified image classification, object detection, and visual relationship detection at scale. CoRR ArXiv preprint. *arXiv preprint arXiv:1811.00982*.

Law, Hei, and Deng, J. (2018). Cornernet: Detecting objects as paired keypoints. Proceedings of the European conference on computer vision (ECCV),

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436-444.

Lee, S.-g., Bae, J. S., Kim, H., Kim, J. H., and Yoon, S. (2018). Liver lesion detection from weakly-labeled multi-phase ct volumes with a grouped single shot multibox detector. International Conference on Medical Image Computing and Computer-Assisted Intervention,

Li, Jianan, Liang, X., Shen, S., Xu, T., Feng, J., and Yan, S. (2017). Scale-aware fast R-CNN for pedestrian detection. *IEEE Transactions on Multimedia*, *20*(4), 985-996.

Li, Yanghao, Chen, Y., Wang, N., and Zhang, Z. (2019). Scale-aware trident networks for object detection. Proceedings of the IEEE/CVF International Conference on Computer Vision,

Li, B., Liu, Y., and Wang, X. (2019a). Gradient harmonized single-stage detector. Proceedings of the AAAI Conference on Artificial Intelligence,

Li, B., Liu, Y., and Wang, X. (2019b). Gradient harmonized single-stage detector. Proceedings of the AAAI Conference on Artificial Intelligence,

Li, H., Liu, Y., Ouyang, W., and Wang, X. (2019). Zoom out-and-in network with map attention decision for region proposal and object detection. *International journal of computer vision*, *127*(3), 225-238.

Li, M., Zhang, Z., Yu, H., Chen, X., and Li, D. (2017). S-OHEM: stratified online hard example mining for object detection. CCF Chinese Conference on Computer Vision,

Li, Y., Chen, Y., Wang, N., and Zhang, Z. (2019). Scale-aware trident networks for object detection. Proceedings of the IEEE international conference on computer vision,

Li, Y., Pang, Y., Shen, J., Cao, J., and Shao, L. (2020). NETNet: Neighbor erasing and transferring network for better single shot object detection. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,

Li, Z., Peng, C., Yu, G., Zhang, X., Deng, Y., and Sun, J. (2018). Detnet: A backbone network for object detection. *arXiv preprint arXiv:1804.06215*.

Liang, T., Wang, Y., Tang, Z., Hu, G., and Ling, H. (2021). Opanas: One-shot path aggregation network architecture search for object detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

Lienhart, Rainer, Maydt, and Jochen. (2002). An extended set of haar-like features for rapid object detection. Proceedings. international conference on image processing,

Lin, Tsung-Yi Dollár, Piotr Girshick, Ross He, Kaiming Hariharan, Bharath Belongie, and Serge. (2017). Feature pyramid networks for object detection. Proceedings of the IEEE conference on computer vision and pattern recognition,

Lin, M., Chen, Q., and Yan, S. (2014). Network in Network International Conference on Learning Representations (ICLR'14).

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). Focal loss for dense object detection. Proceedings of the IEEE international conference on computer vision,

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., . . . Zitnick, C. L. (2014). Microsoft coco: Common objects in context. European conference on computer vision,

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

Liu, S., and Huang, D. (2018). Receptive field block net for accurate and fast object detection. Proceedings of the European Conference on Computer Vision (ECCV),

Liu, S., Qi, L., Qin, H., Shi, J., and Jia, J. (2018). Path aggregation network for instance segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition,

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C. (2016). Ssd: Single shot multibox detector. European conference on computer vision,

Liu, W., Rabinovich, A., and Berg, A. C. (2015). Parsenet: Looking wider to see better. *arXiv preprint arXiv:1506.04579*.

Liu, Y., Li, H., Yan, J., Wei, F., Wang, X., and Tang, X. (2017). Recurrent scale approximation for object detection in cnn. Proceedings of the IEEE International Conference on Computer Vision,

Liu, Y., Sun, P., Wergeles, N., and Shang, Y. (2021). A survey and performance evaluation of deep learning methods for small object detection. *Expert Systems with Applications*, *172*, 114602.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, *60*(2), 91-110.

Mateus, A., Ribeiro, D., Miraldo, P., and Nascimento, J. C. (2019). Efficient and robust pedestrian detection using deep learning for human-aware navigation. *Robotics and Autonomous Systems*, *113*, 23-37.

Namysl, M., and Konya, I. (2019). Efficient, lexicon-free OCR using deep learning. 2019 International Conference on Document Analysis and Recognition (ICDAR),

Nataprawira, J., Gu, Y., Goncharenko, I., and Kamijo, S. (2021). Pedestrian detection using multispectral images and a deep neural network. *Sensors*, *21*(7), 2536.

Newell, A., Huang, Z., and Deng, J. (2016). Associative embedding: End-to-end learning for joint detection and grouping. *arXiv preprint arXiv:1611.05424*.

Newell, A., Huang, Z., and Deng, J. (2017). Associative embedding: End-to-end learning for joint detection and grouping. Advances in neural information processing systems,

Newell, A., Yang, K., and Deng, J. (2016). Stacked hourglass networks for human pose estimation. European conference on computer vision,

Nie, J., Anwer, R. M., Cholakkal, H., Khan, F. S., Pang, Y., and Shao, L. (2019a). Enriched feature guided refinement network for object detection. Proceedings of the IEEE/CVF International Conference on Computer Vision,

Nie, J., Anwer, R. M., Cholakkal, H., Khan, F. S., Pang, Y., and Shao, L. (2019b). Enriched feature guided refinement network for object detection. Proceedings of the IEEE International Conference on Computer Vision,

Noh, J., Bae, W., Lee, W., Seo, J., and Kim, G. (2019). Better to follow, follow to be better: Towards precise supervision of feature super-resolution for small object detection. Proceedings of the IEEE/CVF International Conference on Computer Vision,

Oksuz, K., Cam, B. C., Kalkan, S., and Akbas, E. (2020a). Generating positive bounding boxes for balanced training of object detectors. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision,

Oksuz, K., Cam, B. C., Kalkan, S., and Akbas, E. (2020b). Imbalance problems in object detection: A review. *IEEE transactions on pattern analysis and machine intelligence*, *43*(10), 3388-3415.

Oliva, A., and Torralba, A. (2007). The role of context in object recognition. *Trends in cognitive sciences*, *11*(12), 520-527.

Oquab, M., Bottou, L., Laptev, I., and Sivic, J. (2014). Learning and transferring mid-level image representations using convolutional neural networks. Proceedings of the IEEE conference on computer vision and pattern recognition,

Ouyang, W., Wang, K., Zhu, X., and Wang, X. (2017). Chained cascade network for object detection. Proceedings of the IEEE International Conference on Computer Vision,

Ouyang, W., Wang, X., Zhang, C., and Yang, X. (2016). Factors in finetuning deep model for object detection with long-tail distribution. Proceedings of the IEEE conference on computer vision and pattern recognition,

Pan, H., Jiang, J., and Chen, G. (2020). TDFSSD: Top-down feature fusion single shot MultiBox detector. *Signal Processing: Image Communication*, *89*, 115987.

Pang, Chen, K., Shi, J., Feng, H., Ouyang, W., and Lin, D. (2019). Libra r-cnn: Towards balanced learning for object detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

Pang, J., Chen, K., Shi, J., Feng, H., Ouyang, W., and Lin, D. (2019). Libra r-cnn: Towards balanced learning for object detection. Proceedings of the IEEE conference on computer vision and pattern recognition,

Pang, Y., Wang, T., Anwer, R. M., Khan, F. S., and Shao, L. (2019a). Efficient featurized image pyramid network for single shot detector. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

Pang, Y., Wang, T., Anwer, R. M., Khan, F. S., and Shao, L. (2019b). Efficient featurized image pyramid network for single shot detector. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,

Peng, S., Liu, Y., Huang, Q., Zhou, X., and Bao, H. (2019). Pvnet: Pixel-wise voting network for 6dof pose estimation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

Picron, C., and Tuytelaars, T. (2021). Trident Pyramid Networks: The importance of processing at the feature pyramid level for better object detection. *arXiv preprint arXiv:2110.04004*.

Pitts, W., and McCulloch, W. S. (1947). How we know universals the perception of auditory and visual forms. *The Bulletin of mathematical biophysics*, *9*(3), 127-147.

Pont-Tuset, J., Arbelaez, P., Barron, J. T., Marques, F., and Malik, J. (2016). Multiscale combinatorial grouping for image segmentation and object proposal generation. *IEEE transactions on pattern analysis and machine intelligence*, *39*(1), 128-140.

Qian, Q., Chen, L., Li, H., and Jin, R. (2020). DR loss: Improving object detection by distributional ranking. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

Qin, Z., Li, Z., Zhang, Z., Bao, Y., Yu, G., Peng, Y., and Sun, J. (2019). ThunderNet: Towards real-time generic object detection on mobile devices. Proceedings of the IEEE/CVF international conference on computer vision,

Quang, T. N., Lee, S., and Song, B. C. (2021). Object Detection Using Improved Bi-Directional Feature Pyramid Network. *Electronics*, *10*(6), 746.

Real, E., Aggarwal, A., Huang, Y., and Le, Q. V. (2019). Regularized evolution for image classifier architecture search. Proceedings of the aaai conference on artificial intelligence,

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. Proceedings of the IEEE conference on computer vision and pattern recognition,

Redmon, J., and Farhadi, A. (2017). YOLO9000: better, faster, stronger. Proceedings of the IEEE conference on computer vision and pattern recognition,

Redmon, J., and Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

Ren, S., He, K., Girshick, R., and Sun, J. (2015a). Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems,

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

Ren, S., He, K., Girshick, R., and Sun, J. (2015c). Faster r-cnn: Towards real-time object detection with region proposal networks (2015). *Advances in neural information processing systems, 28*, 91-99.

Ren, S., He, K., Girshick, R., and Sun, J. (2016). Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence, 39*(6), 1137-1149.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical image computing and computer-assisted intervention,

Rothe, R., Guillaumin, M., and Gool, L. V. (2014). Non-maximum suppression for object detection by passing messages between windows. Asian conference on computer vision,

Rowley, H. A., Baluja, S., and Kanade, T. (1995). *Human face detection in visual scenes*. Citeseer.

Rowley, H. A., Baluja, S., and Kanade, T. (1996). Human face detection in visual scenes. Advances in neural information processing systems,

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1988). Learning representations by back-propagating errors. *Cognitive modeling, 5*(3), 1.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., . . . Bernstein, M. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision, 115*(3), 211-252.

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., and LeCun, Y. (2013). Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*.

Shang, W., Sohn, K., Almeida, D., and Lee, H. (2016). Understanding and improving convolutional neural networks via concatenated rectified linear units. international conference on machine learning,

Shaoqing, R., Kaiming, H., Ross, G., Xiangyu, Z., and Jian, S. (2016). Object detection networks on convolutional feature maps. *IEEE transactions on pattern analysis and machine intelligence*, *39*(7), 1476-1481.

Shen, Z., Liu, Z., Li, J., Jiang, Y.-G., Chen, Y., and Xue, X. (2017). Dsod: Learning deeply supervised object detectors from scratch. Proceedings of the IEEE international conference on computer vision,

Shrivastava, A., Gupta, A., and Girshick, R. (2016). Training region-based object detectors with online hard example mining. Proceedings of the IEEE conference on computer vision and pattern recognition,

Simonyan, Karen Zisserman, and Andrew. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Singh, B., and Davis, L. S. (2018). An analysis of scale invariance in object detection snip. Proceedings of the IEEE conference on computer vision and pattern recognition,

Singh, B., Najibi, M., and Davis, L. S. (2018a). Sniper: Efficient multi-scale training. Advances in neural information processing systems,

Singh, B., Najibi, M., and Davis, L. S. (2018b). Sniper: Efficient multi-scale training. *arXiv preprint arXiv:1805.09300*.

Song, Haifeng Yang, and Weiwei. (2022). GSCCTL: a general semi-supervised scene classification method for remote sensing images based on clustering and transfer learning. *International Journal of Remote Sensing*, 1-25.

Suissa, O., Elmalech, A., and Zhitomirsky-Geffet, M. (2022). Text analysis using deep neural networks in digital humanities and information science. *Journal of the Association for Information Science and Technology*, *73*(2), 268-287.

Sun, P., Zhang, R., Jiang, Y., Kong, T., Xu, C., Zhan, W., . . . Wang, C. (2021). Sparse r-cnn: End-to-end object detection with learnable proposals. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

Sun, Y., KP, P. S., Shimamura, J., and Sagata, A. (2019). Concatenated feature pyramid network for instance segmentation. 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM),

Sung, K.-K., and Poggio, T. (1998). Example-based learning for view-based human face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(1), 39-51.

Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. Thirty-First AAAI Conference on Artificial Intelligence,

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., . . . Rabinovich, A. (2015). Going deeper with convolutions. Proceedings of the IEEE conference on computer vision and pattern recognition,

Szegedy, C., Reed, S., Erhan, D., Anguelov, D., and Ioffe, S. (2014). Scalable, high-quality object detection. *arXiv preprint arXiv:1412.1441*.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. Proceedings of the IEEE conference on computer vision and pattern recognition,

Tan, M., and Le, Q. V. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*.

Tan, M., Pang, R., and Le, Q. V. (2020). Efficientdet: Scalable and efficient object detection. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,

Tian, Z., Shen, C., Chen, H., and He, T. (2019). Fcos: Fully convolutional one-stage object detection. Proceedings of the IEEE/CVF International Conference on Computer Vision,

Torralba, A., Murphy, K. P., Freeman, W. T., and Rubin, M. A. (2003). Context-based vision system for place and object recognition. Computer Vision, IEEE International Conference on,

Tripathi, S., Chandra, S., Agrawal, A., Tyagi, A., Rehg, J. M., and Chari, V. (2019a). Learning to generate synthetic data via compositing. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

Tripathi, S., Chandra, S., Agrawal, A., Tyagi, A., Rehg, J. M., and Chari, V. (2019b). Learning to generate synthetic data via compositing. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,

Tychsen-Smith, L., and Petersson, L. (2018). Improving object localization with fitness nms and bounded iou loss. Proceedings of the IEEE conference on computer vision and pattern recognition,

Uijlings, J. R., Van De Sande, K. E., Gevers, T., and Smeulders, A. W. (2013). Selective search for object recognition. *International journal of computer vision*, *104*(2), 154-171.

Viola, P., and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001,

Wadley, F. (1952). Probit analysis: a statistical treatment of the sigmoid response curve. In: Oxford University Press Oxford, UK.

Wang, C., and Zhong, C. (2021). Adaptive Feature Pyramid Networks for Object Detection. *IEEE Access*, *9*, 107024-107032.

Wang, H., Wang, Q., Yang, F., Zhang, W., and Zuo, W. (2019). Data augmentation for object detection via progressive and selective instance-switching. *arXiv preprint arXiv:1906.00358*.

Wang, J., Chen, K., Yang, S., Loy, C. C., and Lin, D. (2019). Region proposal by guided anchoring. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

Wang, J., Song, L., Li, Z., Sun, H., Sun, J., and Zheng, N. (2021). End-to-end object detection with fully convolutional network. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

Wang, X., Girshick, R., Gupta, A., and He, K. (2018). Non-local neural networks. Proceedings of the IEEE conference on computer vision and pattern recognition,

Wang, X., Shrivastava, A., and Gupta, A. (2017). A-fast-rcnn: Hard positive generation via adversary for object detection. Proceedings of the IEEE conference on computer vision and pattern recognition,

Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. (2017). Aggregated residual transformations for deep neural networks. Proceedings of the IEEE conference on computer vision and pattern recognition,

Xin, Y., Wang, G., Mao, M., Feng, Y., Dang, Q., Ma, Y., . . . Han, S. (2021). PAFNet: An Efficient Anchor-Free Object Detector Guidance. *arXiv preprint arXiv:2104.13534*.

Yaman, Orhan, and Turker, T. (2022). Exemplar pyramid deep feature extraction based cervical cancer image classification model using pap-smear images. *Biomedical Signal Processing and Control*, *73*, 103428.

Yan, C., Meng, L., Li, L., Zhang, J., Wang, Z., Yin, J., . . . Zheng, B. (2022). Age-Invariant Face Recognition by Multi-Feature Fusionand Decomposition with Self-attention. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, *18*(1s), 1-18.

Yang, F., Choi, W., and Lin, Y. (2016). Exploit all the layers: Fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers. Proceedings of the IEEE conference on computer vision and pattern recognition,

Yu, F., and Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*.

Zeiler, M. D., and Fergus, R. (2014). Visualizing and understanding convolutional networks. European conference on computer vision,

Zeng, X., Ouyang, W., Yan, J., Li, H., Xiao, T., Wang, K., . . . Wang, Z. (2017). Crafting gbd-net for object detection. *IEEE transactions on pattern analysis and machine intelligence*, *40*(9), 2109-2123.

Zhang, J., Zhang, L., Liu, T., and Wang, Y. (2021). YOLSO: You Only Look Small Object. *Journal of Visual Communication and Image Representation*, *81*, 103348.

Zhang, L., Lin, L., Liang, X., and He, K. (2016). Is faster R-CNN doing well for pedestrian detection? European conference on computer vision,

Zhang, S., Wen, L., Bian, X., Lei, Z., and Li, S. Z. (2018). Single-shot refinement neural network for object detection. Proceedings of the IEEE conference on computer vision and pattern recognition,

Zhang, X., Yang, Y.-H., Han, Z., Wang, H., and Gao, C. (2013). Object class detection: A survey. *ACM Computing Surveys (CSUR)*, *46*(1), 1-53.

Zhang, Z., Qiao, S., Xie, C., Shen, W., Wang, B., and Yuille, A. L. (2018). Single-shot object detection with enriched semantics. Proceedings of the IEEE conference on computer vision and pattern recognition,

Zhao, G., Ge, W., and Yu, Y. (2021). GraphFPN: Graph feature pyramid network for object detection. Proceedings of the IEEE/CVF International Conference on Computer Vision,

Zhao, Q., Sheng, T., Wang, Y., Tang, Z., Chen, Y., Cai, L., and Ling, H. (2019a). M2det: A single-shot object detector based on multi-level feature pyramid network. Proceedings of the AAAI Conference on Artificial Intelligence,

Zhao, Q., Sheng, T., Wang, Y., Tang, Z., Chen, Y., Cai, L., and Ling, H. (2019b). M2det: A single-shot object detector based on multi-level feature pyramid network. Proceedings of the AAAI conference on artificial intelligence,

Zhao, T., Zhang, X., and Wang, S. (2021). Graphsmote: Imbalanced node classification on graphs with graph neural networks. Proceedings of the 14th ACM international conference on web search and data mining,

Zhao, Z.-Q., Zheng, P., Xu, S.-t., and Wu, X. (2019). Object detection with deep learning: A review. *IEEE transactions on neural networks and learning systems*.

Zhou, L., Rao, X., Li, Y., Zuo, X., Qiao, B., and Lin, Y. (2022). A Lightweight Object Detection Method in Aerial Images Based on Dense Feature Fusion Path Aggregation Network. *ISPRS International Journal of Geo-Information, 11*(3), 189.

Zhou, P., Ni, B., Geng, C., Hu, J., and Xu, Y. (2018). Scale-transferrable object detection. proceedings of the IEEE conference on computer vision and pattern recognition,

Zhou, X., Zhuo, J., and Krahenbuhl, P. (2019a). Bottom-up object detection by grouping extreme and center points. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,

Zhou, X., Zhuo, J., and Krahenbuhl, P. (2019b). Bottom-up object detection by grouping extreme and center points. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

Zhou, Y., Liu, L., Shao, L., and Mellor, M. (2016). DAVE: A unified framework for fast vehicle detection and annotation. European Conference on Computer Vision,

Zhu, C., He, Y., and Savvides, M. (2019). Feature selective anchor-free module for single-shot object detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

Zhu, L., Deng, Z., Hu, X., Fu, C.-W., Xu, X., Qin, J., and Heng, P.-A. (2018). Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection. Proceedings of the European Conference on Computer Vision (ECCV),

Zhu, L., Lee, F., Cai, J., Yu, H., and Chen, Q. (2022). An Improved Feature Pyramid Network for Object Detection. *Neurocomputing*.

Zhu, X., Pang, J., Yang, C., Shi, J., and Lin, D. (2019). Adapting object detectors via selective cross-domain alignment. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

Zhu, Y., Zhao, C., Wang, J., Zhao, X., Wu, Y., and Lu, H. (2017). Couplenet: Coupling global structure with local parts for object detection. Proceedings of the IEEE international conference on computer vision,

Zitnick, Dollár, Piotr, and Lawrence, C. (2014). Edge boxes: Locating object proposals from edges. European conference on computer vision,

Zoph, B., and Le, Q. V. (2016). Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578*.

Zoph, B., Vasudevan, V., Shlens, J., and Le, Q. V. (2018). Learning transferable architectures for scalable image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition,