Advancements in General Object Detection Architectures

Harshel Malawade KLS Gogte Institute of Technology, Belagavi Email: harshelmalawade21@gmail.com

***Abstract*—A key task in computer vision with many applications is object detection. An overview of recent developments in broad object detection architectures is provided in this work. The Deconvolutional Single Shot Detector (DSSD), Neural Architecture Search Feature Pyramid Network (NAS-FPN), CornerNet, and PeleeNet are four cutting-edge techniques that are explored. The architecture of Deeply Supervised Object Detectors (DSOD) is also investigated. Each of these architectures displays notable advances in accuracy, speed, and model efficiency while addressing particular difficulties in object detection. Experimental findings on benchmark datasets like PASCAL VOC and MS COCO demonstrate how these approaches outperform earlier ones.**

***Index Terms*—Object detection, deep learning, convolutional neural networks, architecture, performance evaluation.**

I. Introduction

In computer vision tasks like autonomous driving, surveillance, and picture analysis, object detection is crucial. Numerous object detection designs have been created over time to handle the difficulties presented by different object sizes, occlusions, and intricate backdrops. The current developments in general object identification architectures are highlighted in this study, with a particular emphasis on four noteworthy methods: DSSD, NAS-FPN, CornerNet, and PeleeNet. Additionally, a solution to the issue of learning object detectors from small datasets—the DSOD architecture—is explored.

Fig. 1: Object Detection

II. Methods

A. Deconvolutional Single Shot Detector (DSSD)

Cheng-Yang Fu et al. [3] propose DSSD, an enhanced large-scale context is incorporated into the object detection algorithm to increase accuracy, especially for small objects. DSSD performs at the cutting edge on the PASCAL VOC and COCO datasets. The performance of earlier approaches, such as R-FCN, was surpassed by the authors' 81.5% mean Average Precision (mAP) on the VOC2007 test, 80.0% mAP on the VOC2012 test, and 33.2% mAP on the COCO. .

B. Neural Architecture Search Feature Pyramid Network (NAS-FPN)

Golnaz Ghiasi et al. [4] introduce NAS-FPN, an object identification architecture that makes use of Neural Architecture Search (NAS) methods. The feature pyramid of NAS-FPN combines top-down and bottom-up connections, increasing the accuracy of mobile object recognition. It has been demonstrated through experimental analysis on the COCO dataset that NAS-FPN outperforms other detectors with a higher detection time efficiency.

C. CornerNet

Heu Law and Jia Deng [6] propose CornerNet, a method for object detection that does not rely on the common anchor box architecture used in single-stage detectors. CornerNet uses a single convolutional neural network and popularises the idea of "corner pooling." CornerNet yields a stunning 42.1% Average Precision (AP), according to experimental results on the MS COCO dataset.

D. PeleeNet

Robert J. Wang et al. [17] present PeleeNet, an effective architecture made to handle the rising demand for Convolutional Neural Network (CNN) models on portable devices with constrained memory and processing power. PeleeNet achieves remarkable accuracy and a model size that is just 66% of MobileNet's size. PeleeNet is able to operate quickly and make predictions in real time on mobile devices by utilising the single-shot multibox detector.

E. Deeply Supervised Object Detectors (DSOD)

Zhiqiang Shen et al. [12] broaden the research on Deeply Supervised Object Detectors (DSOD), a system that allows for the creation of new object detectors. The drawbacks of complicated loss functions and small datasets that were faced in earlier efforts are addressed by DSOD. In terms of detection performance, experimental analysis on PASCAL VOC 2007, 2012, and MS COCO shows that DSOD outperforms state-of-the-art approaches. DSOD has potential in a number of fields, including multi-spectral imaging, depth, and medicine.

F. Single Shot Multibox Detector (SSD)

Wei Liu et al., [10] propose the Single Shot Multibox Detector (SSD), This creates a single deep neural network that combines object detection and categorization. The SSD network calculates scores for several object categories and modifies bounding box predictions to more accurately reflect the form of the object. SSD provides a streamlined and effective method for object detection by doing away with the requirement for a separate proposal generating stage and subsequent resampling. The unified structure of SSD enables effective training and inference with accuracy on par with approaches that also involve an object proposal stage.

G. EfficientDet: Scalable and Efficient Object Detection

Mingxing Tan et al. [14] introduce a compound scaling technique for object detection, It simultaneously scales the box/class prediction networks, feature network, and backbone network's resolution, depth, and width. In terms of effectiveness and resource use, this strategy performs better than earlier techniques. The study offers insightful analysis and suggestions for enhancing the performance of neural network-based object detecting systems. The suggested improvements and the EfficientDet family of object detectors have numerous uses in a variety of industries, such as robotics, surveillance, and self-driving vehicles.

H. Detection with Enriched Semantics (DES)

Zhishuai Zhang et al. [19] propose Detection with Enriched Semantics (DES), a unique single-shot object detection network. DES adds a semantic segmentation branch and a global activation module to a basic deep detector to enhance the semantics of object detection features. There is no need for extra annotation because the segmentation branch is supervised by weak segmentation ground-truth. A self-supervised understanding of the relationship between channels and object classes is also developed by the global activation module. Results from experiments performed on the PASCAL VOC and MS COCO datasets show that DES significantly improved performance.

I. RefineDet: Single-Shot Refinement Neural Network for Object Detection

Shifeng Zhang et al. [18] present RefineDet, a single-shot detector that is more accurate than two-stage methods while still being effective. RefineDet achieves state-of-the-art detection accuracy while performing at a high level of efficiency, as shown by the studies performed on the PASCAL VOC 2007, PASCAL VOC 2012, and MS COCO datasets. The results show that the suggested algorithm offers greater accuracy compared to two-stage methods while maintaining the efficacy of one-stage approaches.

J. RetinaNet: Focal Loss for Dense Object Detection

Tsung-Yi Lin et al., [8] propose RetinaNet, a brand-new one-stage object detector that combines cutting-edge accuracy with the simplicity and speed advantages of one-stage detectors. RetinaNet's fundamental innovation is the use of Focal Loss in training, which enables the detector to function better than two-stage detectors that are currently in use while still operating at the same speed as prior one-stage detectors. RetinaNet greatly improves the field of object detection by its well-balanced accuracy-to-speed trade-off, which has ramifications for numerous applications.

K. CenterNet: Keypoint Triplets for Object Detection

Kaiwen Duan, Song Bai et al., [2] introduce CenterNet, a framework for keypoint-based object detection that deals with the problem of incorrect object bounding boxes. By identifying objects as triplets of keypoints and using unique pooling modules, CenterNet increases precision and recall. On the MS-COCO dataset, CenterNet surpasses conventional one-stage detectors by at least 4.9 percent, achieving an Average Precision (AP) of 47.0 percent. When compared to the top two-stage detectors, it performs similarly while maintaining a higher inference rate. The groundbreaking approach used by CenterNet and its encouraging outcomes have a substantial impact on object detection.

L. Feature Fusion Single Shot Multibox Detector (FSSD)

Zuo-Xin Li, Fu-Qiang Zhou., [7] present FSSD, a more advanced SSD object detection method. In order to overcome the feature pyramid detection constraint of SSD, FSSD introduces a lightweight feature fusion module that dramatically enhances performance without sacrificing speed. A novel feature pyramid is produced by FSSD by combining features from various levels of differing scales. Using a single Nvidia 1080Ti GPU, FSSD completes the Pascal VOC 2007 test with an accuracy of 82.7% at 65.8 frames per second. It performs better than several cutting-edge object detection algorithms and conventional SSD in terms of accuracy and performance.

M. Efficient Multi-category Object Detection

Kye-Hyeon Kim et al., [5] propose a novel strategy for detecting multi-category objects with high accuracy while minimizing computational cost. The network obtains outstanding results on well-known object detection benchmarks by rethinking the feature extraction phase of the pipeline and using methods such concatenated ReLU, Inception, and HyperNet. Its excellent performance is a result of the network's deep and narrow architecture, batch normalisation, residual connections, and learning rate scheduling. Notably, the proposed network is highly efficient and effective for object detection tasks and produces results comparable to ResNet-101 with only 12.3% of the computational cost.

N. RFB Net: Real-Time Object Detection with RFB Modules

Songtao Liu and Yunhong Wang., [9] address the trade-off between accuracy and computational cost in object detection models. To improve feature discriminability and resilience, they suggest a unique RF Block (RFB) module based on the configuration of receptive fields in the human visual

system. The suggested RFB Net detector preserves real-time processing speed while attaining performance on par with sophisticated deep detectors by incorporating RFB modules into the SSD structure. Experiments on important benchmarks show that the RFB Net represents a potential method for striking a balance between accuracy and efficiency in object detection tasks.

O. Sparse R-CNN: Sparse Object Detection via Fixed Object Proposals

Pieze Sun, Rufeng Zhang et al., [13] introduce Sparse R-CNN,a technique for object detection in images that makes use of a predetermined pool of learnt object suggestions. This method differs from others that rely on dense object candidates in that it does not require manually constructed candidates or many-to-one label assignment. On the difficult COCO dataset, the suggested technique performs similarly to established detectors in terms of accuracy, runtime, and training convergence. With a much lower computing cost, sparse R-CNN challenges the usage of dense priors in object detection and advances the field.

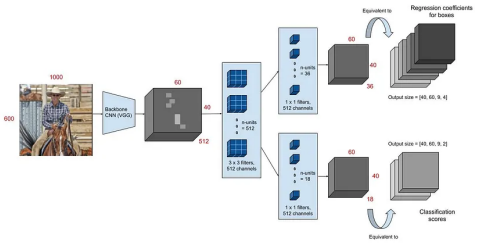

Fig. 2: Faster R-CNN

P. EfficientDet: Efficient Object Detection with BiFPN and Compound Scaling

Mingxing Tan et al., [15] propose optimizations, To increase the efficiency of object detection models, use weighted bi-directional feature pyramid networks (BiFPN) and compound scaling. Based on EfficientNet backbones, the resulting EfficientDet models beat earlier detectors in terms of performance while being substantially smaller and using less processing power. EfficientDet provides compelling evidence of the efficiency of the suggested improvements in enhancing object localization tasks by providing results on the COCO test-dev dataset.

Q. FCOS: Fully Convolutional One-Stage Object Detection

Zhi Tian et al., [16] present FCOS, a fully convolutional one-stage object locator that operates in a per-pixel prediction paradigm.FCOS is an anchor-free and proposition-free detector that works directly on the detection process, in contrast to other detectors that rely on predetermined anchor boxes. FCOS increases detection accuracy by doing away with the requirement for anchor box calculations and hyper-parameter adjustment. With an impressive Average Precision (AP) of 44.7 percent in single-model and single-scale testing with ResNeXt-64x4d-101, it outperforms earlier one-stage detectors. FCOS offers a more straightforward and adaptable detection framework for a variety of instance-level tasks, which represents a considerable improvement in object detection methods.

R. YOLOv3: An Incremental Improvement

Jospeh Redmon and Ali Farhadi., [11] introduce modifications to YOLOv3,resulting in upgrades to the network's training and architecture for greater precision. The revised YOLOv3 is still speedy, operating at 22 milliseconds at 320 x 320 resolution despite a modest size increase. While being many times faster, it achieves precision that is comparable to SSD. YOLOv3 is a potential alternative for object detection tasks because it surpasses RetinaNet in terms of detection inference times for a.5 Intersection over Union (IOU) mean Average Precision (mAP).

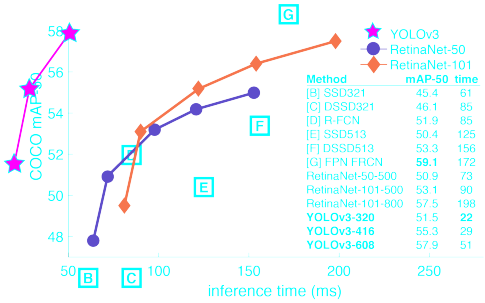

Fig. 3: YOLO v3: Comparison to other detectors

S. Universal Features for Object Detection: Practical Testing and Theoretical Justification

Alexey Bochkovskiy et al., [1] support the theoretical rationale and empirical evaluation of features that improve Convolutional Neural Networks' (CNNs') accuracy. It is suggested that features including self-adversarial training (SAT), weighted residual connections (WRC), cross stage partial connections (CSP), cross mini-batch normalisation (CmBN), and Mish activation are universal features that may be used with a variety of models, tasks, and datasets. On a Tesla V100, the suggested approach achieves a real-time speed of 65 frames per second and an amazing AP of 43.5% (65.7% AP50), illustrative of the value of these attributes in improving object detection performance.

T. TPH-YOLOv5: Transformer Prediction Heads for Drone-based Object Detection

Xingkui Zhu et al., [20] the unique method for object recognition in drone-captured scenes, TPH-YOLOv5. TPH-YOLOv5 uses Transformer Prediction Heads (TPH) with self-attention, a novel prediction head for different-scale objects,

and includes the convolutional block attention model (CBAM) for scenarios involving dense objects. TPH-YOLOv5 surpasses earlier state-of-the-art techniques on the DET-test-challenge dataset by 1.81 percent and obtains results that are comparable on the VisDrone Challenge 2021. TPH-YOLOv5 is a viable alternative for drone-based object detection because to its enhancements over the baseline model.

III. Future Work

There are various directions for object detection research in the future. The effectiveness of detecting algorithms must first be increased in order to attain real-time performance across a variety of hardware platforms. Future study should focus on improving the identification of small items, which are frequently difficult to identify precisely. Furthermore, there is potential in investigating the use of object detection in multi-spectral and depth imaging contexts. For better performance, domain-specific detectors designed for certain applications, such as medicine or robotics, can be created. Future developments in the field must focus on strengthening the object detectors' robustness and generalisation capabilities as well as including interpretability and explainability into the models.

IV. Conclusion

An overview of current developments in broad object detection architectures was provided in this work. The discussion of the DSSD, NAS-FPN, CornerNet, PeleeNet, and DSOD architectures highlighted each one's unique contributions to the area. In terms of accuracy, speed, model size, and the capacity to learn from small datasets, these architectures have shown considerable performance increases. The deployment of more accurate and effective detection models for a variety of computer vision applications is made possible by the ongoing development of object detection architectures.

Acknowledgment

The author would like to express their gratitude to the researchers and authors whose work was reviewed in this paper. Their contributions have significantly advanced the field of object detection and inspired further research in this area.

References

[1] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

[2] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6569–6578, 2019.

[3] Cheng-Yang Fu, Wei Liu, Ananth Ranga, Ambrish Tyagi, and Alexander C Berg. Dssd: Deconvolutional single shot detector. *arXiv preprint arXiv:1701.06659*, 2017.

[4] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. Nas-fpn: Learning scalable feature pyramid architecture for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7036–7045, 2019.

[5] Kye-Hyeon Kim, Sanghoon Hong, Byungseok Roh, Yeongjae Cheon, and Minje Park. Pvanet: Deep but lightweight neural networks for real-time object detection. *arXiv preprint arXiv:1608.08021*, 2016.

[6] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European conference on computer vision (ECCV)*, pages 734–750, 2018.

[7] Zuoxin Li and Fuqiang Zhou. Fssd: feature fusion single shot multibox detector. *arXiv preprint arXiv:1712.00960*, 2017.

[8] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[9] Songtao Liu, Di Huang, et al. Receptive field block net for accurate and fast object detection. In *Proceedings of the European conference on computer vision (ECCV)*, pages 385–400, 2018.

[10] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer, 2016.

[11] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[12] Zhiqiang Shen, Zhuang Liu, Jianguo Li, Yu-Gang Jiang, Yurong Chen, and Xiangyang Xue. Dsod: Learning deeply supervised object detectors from scratch. In *Proceedings of the IEEE international conference on computer vision*, pages 1919–1927, 2017.

[13] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, et al. Sparse r-cnn: End-to-end object detection with learnable proposals. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14454–14463, 2021.

[14] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10781–10790, 2020.

[15] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10781–10790, 2020.

[16] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019.

[17] Robert J Wang, Xiang Li, and Charles X Ling. Pelee: A real-time object detection system on mobile devices. *Advances in neural information processing systems*, 31, 2018.

[18] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z Li. Single-shot refinement neural network for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4203–4212, 2018.

[19] Zhishuai Zhang, Siyuan Qiao, Cihang Xie, Wei Shen, Bo Wang, and Alan L Yuille. Single-shot object detection with enriched semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5813–5821, 2018.

[20] Xingkui Zhu, Shuchang Lyu, Xu Wang, and Qi Zhao. Tph-yolov5: Improved yolov5 based on transformer prediction head for object detection on drone-captured scenarios. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2778–2788, 2021.