

Exploring Various Object Detection and Recognition Frameworks: A Review

Mr. Parul Preet Singh¹, Dr. Ranjit Kumar²

¹ Research Scholar, CSE Department, Maharaja Agrasen Institute of Technology, Maharaja Agrasen University, Baddi, Himachal Pradesh. Email: parulpreetsingh1985@gmail.com

² Associate Professor, CSE Department, Maharaja Agrasen Institute of Technology, Maharaja Agrasen University, Baddi, Himachal Pradesh. Email: ranjitpes@gmail.com

Abstract: Detection of objects is one of the crucial parts of Computer Vision. As the technology has improved, Object Detection and Recognition has undergone a profound transformation, largely because of growth and rapid revolution in the field of Deep Learning. This paper provides a detail technological evolution in the field of detecting and recognizing objects, uncovering the fields evolution from its early days to its current situation. Our analysis is structured around the evolving detection architecture, evolving standardization of dataset, evolution of feature extraction, evolving performance evaluation metrics and other core systems and the optimizing techniques to obtain better results.

Keywords: Object Detection, Computer Vision, Object Recognition, Deep Learning, Feature Extraction

1. Introduction

Object identification stands as a foundational pillar of machine vision, concentrated on recognizing and localizing instances of particular objects within electronic pictures or video frames. At its core, the fundamental thing is to identify the things which are there and precisely wants their location in image. The performance of any detection system is benchmarked on accuracy and inference speed.[1] Human being execute object detection and recognition easily, identifying thousands of objects of various categories, shapes and sizes.[2] To repeat this capability of human by any machine is the ultimate goal of machine vision since 1960's. [3] The pre deep learning era was explained as era of thoroughly engineered feature descriptors plan to capture in variant properties of visual appearance.[4][5] The importance and significance of object detection and object recognition-based tasks is reflected in the explosive growth of research papers over 25 years. According to google scholar data analyzed in this paper, annual publications on object detection have grown from 150 in 2009 to 32000 in 2023 and object recognition grow from 200 in 2009 to 24000 in 2023. This trajectory of growth and specially after 2012 deep learning breakthrough, it shows the importance of the field. The real time application of this will be in the fields autonomous driving, medical imaging, robotic manipulation, surveillance systems and many other fields. It is a challenging task to identify an object in an image automatically using computer-based technology as there is noise, occlusion, low resolution, scale and other challenges which effect the performance of object detection system.[6] As different object detection systems have different goals and challenges depending upon the purpose of the application and some common challenges which is common in all systems.[7]



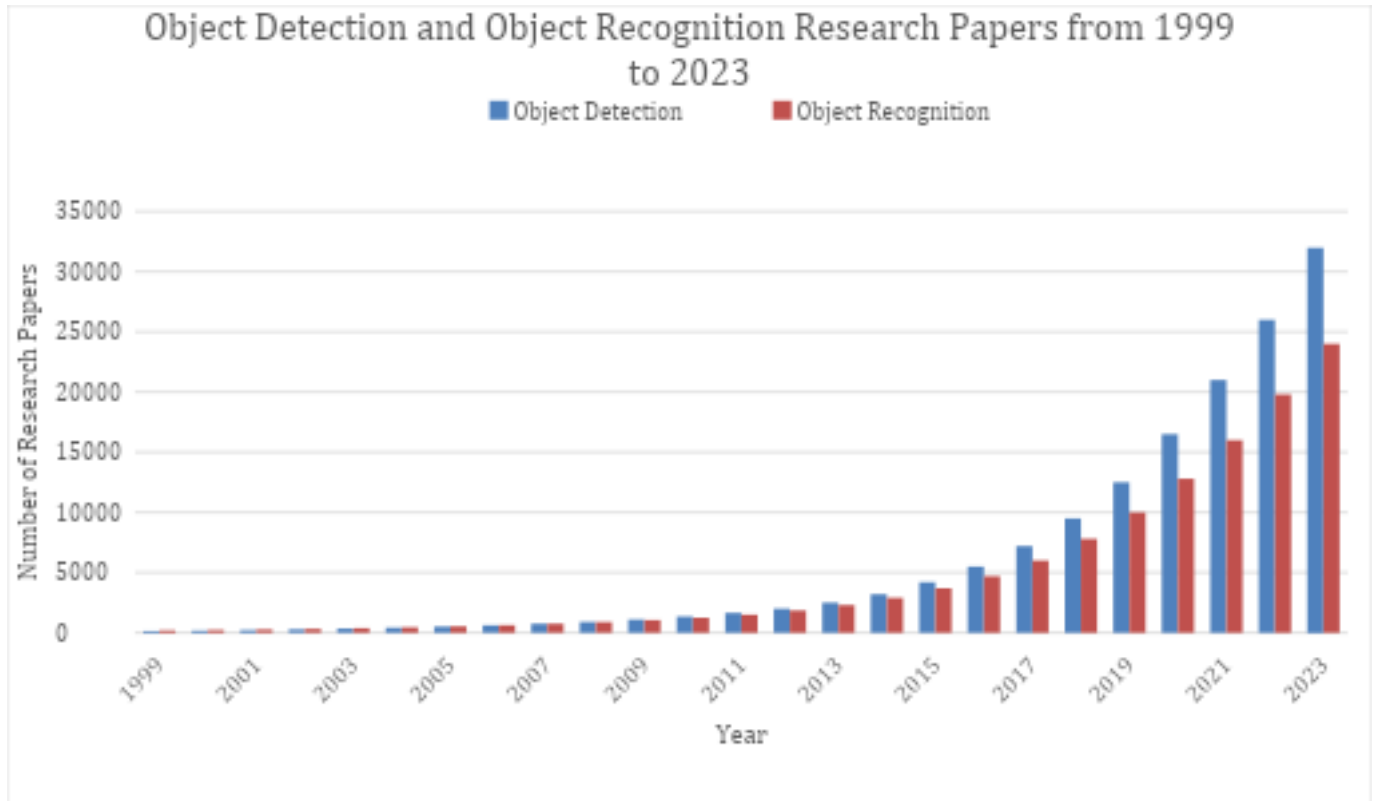


Fig.1: Shows an increase in the number of publications in Object Detection and Object Recognition from 1999 to 2023. (Data from Google Scholar advanced search: allintitle: “Object Detection” and another advanced search: allintitle: “Object Recognition”).

2. Evolution of Object Detection and Recognition Methods

The advancement of object detection and recognition is divided into four different reorientation phases, each phase has seen changes like the methodological change, research application and priorities and also seen improved performances. As we gather more awareness about these changes in each phase, we can understand the evolution of new object detection and recognition ways which emerges from their earlier ways or methods.

The first era is from 1999 to 2011, the pre-deep learning era where handcrafted features and classical machine learning methods are used. The pre-deep learning era specified by very precise feature descriptors designed to seize invariant properties of visual form while keeping the hardware computational demands well within the limits. This era was described by the human designed feature extraction methods like SIFT, HOG, SURF and various other human designed feature extraction methods, modular pipeline architecture in which feature extraction, encoding and classification were optimized separately. In this era comparatively limited representation capabilities like feature extraction at single scale or limited spatial features. There were computational constraints in this era. The handcraft features and classical machine learning ways of this era have limited the threshold of performance but in this era the foundation of various concepts of this fields have been laid out like keypoint detection, scale-space theory, integral images, cascade classifiers and evaluations which are still relevant and form the starting path of development in this field.[8]

The second era is from 2012 to 2014. The deep learning revolution era turning point arrives in 2012 when AlexNet achieves a significant improvement of 10.8 percentage in ImageNet classification accuracy. It shows that deep convolutional neural networks can automatically discover features from raw pixel so handcraft features are no longer required. This paradigm key characteristics are learned the organization of hierarchical features in layers of

increasing abstraction, end to end training, allowing joint optimization of all network parameters, improving GPU parallelism to train network arrangements of larger size and scale architectural innovations including ReLU activation and data augmentation of increasing effective training set size. This era created deep learning as the prevailing ideology with respect to network depth, parameter count and training data was the main factor that improves the performance.[9]

Using the fundamentals displayed by deep learning techniques, this third era from 2015 to 2019 worked on architectural innovation. This era includes residual connections (ResNet), DenseNet, EfficientNet, two stage detection architecture (Faster R-CNN) and one stage detection architecture (Yolo, RetinaNet). This era tailored the modern way of architectural design patterns and it shows that this architecture will achieve better accuracy and efficiency.[10][11][12]

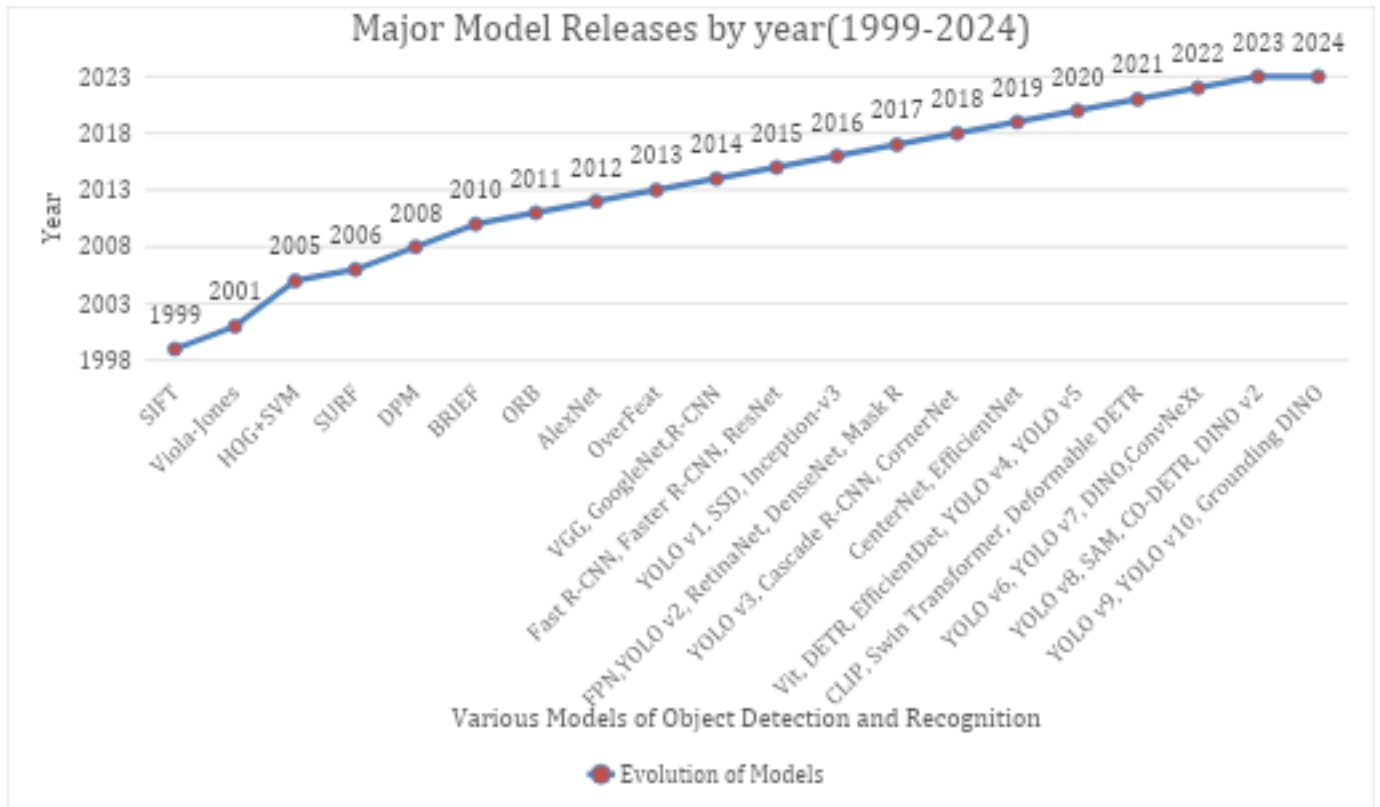


Fig.2: Evolution of detecting objects and recognition systems over a period of time.

The fourth era is from 2020 to present. The introduction of transformer architectures in natural language processing has recently changed the computer vision field. The key components of this era are self-attention way in which without local constraints of convolution long range content is considered, DETR (Detection Transformer) remove hand design components, vision transformers that are achieving great success specially with pretraining large-scale data, models (CLIP, SAM) are working on huge and varied training data, multi modal models working with vision components in language models.[13][14][15]

3. Modern Era Object Detection and Recognition Methods

The modern era of detecting objects starts with the introduction of the YOLO framework, which proposed a unified approach of image processing in a single pass through a convolutional neural network (CNN), achieving instantaneous performance.[16] While earlier methods such as R-CNN and Fast R-CNN offered significant accuracy improvements using region proposals and CNNs, their computational costs were high due to multi-stage pipelines.[17]

Faster R-CNN addressed this by incorporating Region Proposal Networks (RPNs), offering an entirely end-to-end trainable method.[18]

The improvement of Faster R-CNN into Mask R-CNN introduced segmentation capabilities alongside object detection.[19] It demonstrated the adaptability of region-based detectors for instance segmentation tasks. On the other hand, the EfficientDet series introduced compound scaling and a neural architecture search (NAS)-based backbone, delivering an excellent balance between speed and accuracy across various COCO benchmark tasks.[20]

Meanwhile, transformer-based models have gained significant attention in the last few years. DETR (Detections Transformers) reframed detection of objects using a transformer encoder-decoder as a simple immediate set prediction problem.[21] However, its slow convergence led to the development of Deformable DETR, which incorporated deformable attention to improve training efficiency and performance.[22] Similarly, Sparse R-CNN introduced learnable object proposals, replacing dense anchor boxes with sparse, optimized proposals.[23]

Anchor-free models such as YOLOv7 and YOLOR have further streamlined the detection pipeline, emphasized unified reasoning and optimized training strategies for real-time applications.[24][25] CenterNet and CornerNet also moved away from anchor-based detections, instead leveraging keypoint detection to localize objects, proving efficient for small-scale object detection.[26][27]

Hybrid models like ATSS bridged the difference between anchor-free and anchor-based detectors using sample selection based on adaptive training.[28] Further improvements in detection heads were proposed in Dynamic Head, which unified different object detection heads through attention mechanisms.[29] A comprehensive survey in the context of remote sensing, suggesting a benchmark for satellite imagery and aerial data object detection tasks.[30]

Other practical improvements include the Precise Single Shot Detector (PSSD), which enhances localization accuracy without compromising inference speed.[31] Thermal image detection with DDSSD, specifically focusing on adverse weather conditions.[32] Meanwhile, PP-YOLO and its enhanced successor PP-YOLOE refined YOLO-based pipelines by integrating EfficientNet backbones and deeper layers.[33][34]

Early single-shot detectors such as DSSD and FSSD extended SSD models by introducing deconvolution and feature fusion techniques.[35][36] These models retained high inference speed while enhancing detection in low-resolution regions. Additionally, TensorFlow's FPNLite model integrated MobileNetV2 with feature pyramid networks (FPN), optimized for deployment on edge devices.[37]

The integrated YOLOv5 with EfficientNet for submarine pipeline inspection, showing deep learning's versatility across industrial domains.[38] Meanwhile, lightweight models like SqueezeDet have shown promise in embedded and autonomous driving systems.[39] The survey shows the full trajectory of object detection, from CNNs to transformers, summarizing key performance tradeoffs.[40]

Innovations such as Oriented R-CNN offered orientation-aware detection suited for aerial imagery, outperforming traditional axis-aligned detectors.[41] The EfficientNet architecture remains a common backbone for many detection systems due to its high parameter efficiency.[42] Similarly, MobileNet and MobileNetV2 introduced respectively, provide lightweight alternatives for real-time detection on mobile devices.[43][44]

A focused review outlined the requirements for deploying object detectors on edge devices, evaluating tradeoffs in model size, latency, and mAP.[45] Meanwhile, SqueezeNet and SqueezeDet have remained baseline models for energy-constrained platforms.[46]

Recent advances like ECA-YOLOv4 and improved YOLOv5 variants for industrial applications such as gangue sorting have pushed the boundaries of real-time accuracy. These models leverage attention mechanisms and hardware-aware optimizations to deploy deep object detectors in real-world manufacturing and surveillance environments.[47][48]

Object detection has a key role in the area of computer vision by activating machines to locate and classify instances of objects present in images or video frames. It combines the chores of image classification and localization

of object to provide bounding boxes around detected objects. Object detection has achieved remarkable progress in both accuracy and speed with the advancement in the field deep learning mainly CNN (Convolution Neural Networks).

Pre-existing object detection approaches heavily based on features such as handcrafted and rigid pipelines, which were limited in performance under diverse and complex real-world environments. However, the integration of deep learning frameworks has enabled the design of end-to-end automated models which learn hierarchical representations from large datasets. These advancements resulted in the development of variety of models optimized for general object detection, compact object detection, and deployment on edge devices.

An extensive study and analysis of latest advancements in the domain of deep learning-based detection methods, highlighting the progression from two-step detectors like Faster R-CNN to single-step detectors such as YOLO and SSD. The authors discussed key architectures and performance trade-offs, emphasizing how different model types confront challenges such as scale variation, occlusion, and real-time operation. Their review noted the shift toward transformer-based detection architectures, which offer better context modeling and global attention mechanisms.[49]

Similarly, a survey focusing on edge optimized model of deep learning worthy for using on resource-constrained edge devices. His work presented a systematic analysis of model compression techniques including pruning, quantization, and neural architecture search (NAS). Importantly, the paper evaluated the performance of these lightweight models on benchmarks like COCO and PASCAL VOC, demonstrating that newer models such as EfficientDet-Lite and YOLOv5-Nano maintain competitive accuracy while reducing computational cost.[50]

The emerging field of open-world object detection (OWOD), where models must detect both known and previously unseen object categories during inference. The authors highlighted key challenges including semantic drift, background misclassification, and scalable class inclusion. Their survey categorized OWOD approaches into incremental learning-based, class-agnostic, and knowledge distillation techniques, and called for better benchmarking standards. This area is critical for real-world operations in the field of automated driving and vigilance where the domain is dynamic and incomplete.[51]

Several recent works have focused on enhancing DETR (Detection Transformers), which originally offered a major change by abolishing components which are hand design such as anchor generation and non-maximum suppression. DEIM (DETR with Improved Matching), which reduces training time and convergence complexity by optimizing bipartite matching using spatial attention. Their improvements were validated on the COCO dataset, achieving mAP scores with fewer training epochs.[52]

The introduced MI-DETR, which uses a multi-time inquiry mechanism to enhance object localization over multiple iterations. The method allowed the model to refine its object predictions in a coarse-to-fine manner, improving accuracy in cluttered and occluded scenes. This iterative refinement mimics the human visual system's sequential attention strategy and was shown to outperform standard DETR variants in challenging scenarios.[53]

In a similar vein, the proposed Point2RBox, a novel approach to oriented object detection using single-point supervision. This mechanism decreases the reliance on extensive bounding box annotations, making it more scalable for datasets where detailed labeling is expensive or infeasible. The authors demonstrated competitive performance on datasets like DOTA and HRSC2016 by leveraging synthetic patterns and spatial reasoning. Their approach marked a transition toward more annotation-efficient object detectors.[54]

Further enhanced the Point2RBox framework by introducing spatial layout constraints in Point2RBox v2. This version considered inter-object spatial relationships and improved orientation detection, particularly in aerial and satellite imagery datasets. Their method significantly reduced the labeling effort while maintaining or surpassing fully supervised baselines.[55]

EfficientNet has been successfully integrated into object detectors to improve accuracy while keeping the model lightweight. EfficientNet-YOLOv5, which replaces the default YOLOv5 backbone with EfficientNet-B0.[56] This fusion allowed for improved feature representation and led to superior performance in marine microalgae detection

tasks. Similarly, the proposed EfficientNet with YOLOv4 which detected the integrated circuit on PCBs. Their architecture proved effective in detecting small and densely packed components with minimal false positives.[57]

The designed OE-YOLO, a model specialized for agricultural applications, particularly rice panicle detection. EfficientNet which is combined with multiple data representation, OE-YOLO achieved high accuracy in difficult outdoor conditions. This shows the growing trend of tailoring object detection architectures to domain-specific needs, thereby improving robustness and real-world deployment feasibility.[58]

Extended the EfficientNetV2 architecture with novel modules such as Channel Efficient Attention and Spatially Adaptive Feature Modulation. These additions allowed the model to focus dynamically on salient features, leading to significant gains in waste classification tasks. This work underscores the importance of feature selection and attention mechanisms in modern lightweight detectors.[59]

A CNN based approach for the object detection in real world scene images with text extraction. The proposed approach is then evaluated with existing models to compare the performance metrics. In this method image preprocessing, the feature extraction through CNN, then localize and multiple objects are identified in an image with scene text understanding. This is a dual model architecture CNN for text detection, trained on the dataset and YOLOv3 for real time object detection.[60]

A completely automated pipeline model that does not required any manual image annotation. A sizeable number of labelled data is vital for the purpose of training in traditional object detection; this whole process is time consuming and expensive. The Dart method uses the techniques of data variation, annotation, pseudo-labeling and model training which reduces overall human involvement and still quality of training data is maintained.[61]

These approaches collectively demonstrate a movement toward reducing the cost and effort associated with data annotation, without compromising performance. Such innovations are crucial for scaling object detection to new emerging areas such as clinical imaging, geospatial analysis, and industrial quality control check.

An improved variant of YOLOv5 by integrating the EfficientNet backbone to tackle object detection in marine environments, particularly for microalgae classification [56]. The model, named EfficientNet-YOLOv5, replaces the conventional CSPDarknet backbone in YOLOv5 with EfficientNet-B0, yielding better feature extraction and lower computational costs. The authors examine the model on a marine microalgae species dataset and reported improvements in both accuracy and inference speed. By leveraging EfficientNet's compound scaling method, the proposed detector achieved higher precision with fewer parameters, making it well-suited to utilize on edge devices used in underwater surveillance systems. This work highlights the importance of adapting object detection architectures to domain-specific characteristics, such as low contrast and noisy underwater imagery.

A hybrid model that enhances YOLOv4 using an EfficientNet backbone to address object detection challenges in printed circuit board (PCB) inspection tasks [57]. In their work titled Enhanced EfficientNet-YOLOv4, the authors focused on detecting integrated circuits (ICs) and solder defects—objects that are typically small, dense, and often partially occluded. By integrating EfficientNet-B0 into the YOLOv4 framework, the model achieved significant improvements in mAP and recall without compromising inference time, making it feasible for industrial automation settings. The study demonstrated that combining the architectural strengths of YOLOv4 with the feature-rich and lightweight design of EfficientNet can successfully deal with the issues of fine-grained object recognition in manufacturing and electronics quality control. This method presents a balanced solution between detection accuracy and computational efficiency.

OE-YOLO, an agricultural object detection model that integrates EfficientNet as a backbone with a YOLO-like architecture tailored for rice panicle detection in natural field conditions [58]. The authors highlighted that traditional object detectors struggle with densely packed and occluded targets in real-world agricultural scenarios. To address this, OE-YOLO adopts multiscale feature combination and integrates machine attention to improve detection robustness under varying illumination and background clutter. The model was evaluated on high-resolution images of rice fields and outperformed conventional YOLOv3 and YOLOv5 models in the context of precision, accuracy and

F1-score. By emphasizing real-world applicability in agriculture, this study presents a compelling use case of deep learning in precision farming and crop yield estimation, enabling intelligent field monitoring systems.

The proposed novel architecture based on EfficientNetV2, designed specifically for waste classification and object detection tasks under low computational budgets [59]. Their improved model integrates two key components: Channel Efficient Attention (CEA) and Spatially Adaptive Feature Modulation (SAFM). CEA allows the model to prioritize relevant channels dynamically, while SAFM enhances the spatial representation of features based on context. These modifications led to superior performance over baseline EfficientNetV2 and YOLOv5 models in real-world waste sorting tasks. The authors conducted experiments across multiple datasets and demonstrated that their architecture achieved better generalization and higher mean Average Precision (mAP) with fewer parameters. This work is largely going to impact the industrial automation and smart city demands, where accurate and efficient waste classification can support sustainable operations and environmental monitoring.

A milestone in deep learning model scaling, offering a family of CNN architectures that optimize both correctness and effectiveness using a multidimensional scaling strategy. Distinct from earlier scaling methods that randomly increase network depth, width, or input resolution, EfficientNet uses a compound coefficient to balance all three dimensions systematically, achieving better performance with lesser parameters [42]. All of this makes EfficientNet in particular worthy of detecting task of objects in case the computational cost is a critical factor. The system ability of efficient attribute extraction obtained through a combination of mobile inverted bottleneck convolution (MBConv) and squeeze-and-excitation blocks, enables it to detect fine-grained features, making it ideal for recognizing small or complex objects in challenging environments.

In recent years, various frameworks with EfficientNet is used for detecting objects, replacing traditional heavy backbones such as ResNet or CSPDarkNet. By leveraging EfficientNet as the backbone of models like YOLOv4, YOLOv5, or custom detectors, researchers have achieved significant improvements in model performance while reducing model size and inference time. For example, EfficientNet-B0 in their enhanced YOLOv4 model to improve integrated circuit detection on printed circuit boards [57], while incorporated EfficientNet into YOLOv5 for microalgae detection in marine environments [56]. These implementations demonstrated that EfficientNet not only enhances object recognition in high-density or low-resolution images but also provides scalability for edge deployment and mobile computing.

The primary benefits of using EfficientNet in object detection and recognition lie in its parameter efficiency, strong generalization capability, and adaptability to various input resolutions. Its lightweight structure allows for faster training and real-time inference without sacrificing detection accuracy, which is crucial for applications like smart agriculture, robotics, remote sensing, and industrial inspection. Furthermore, EfficientNet's superior feature representation enables more accurate localization and classification of things under difficult conditions such as occlusion, dimension modification, or low lighting. The model demonstrated this advantage by applying EfficientNet in rice panicle detection under complex agricultural settings [58], and further extended EfficientNetV2 with attention mechanisms for waste classification with impressive results [59]. These studies highlight the continued relevance of EfficientNet as a powerful and flexible backbone for next-generation object detection systems.

EfficientNet is an ideal backbone for detecting and recognizing things because of balanced conduct, lower resource usage, and strong accuracy. Its modular design makes it easy to integrate into modern detection pipelines, and its success across domains—from healthcare and agriculture to industrial automation—shows its broad applicability. With ongoing enhancements such as EfficientNetV2 and its use in hybrid architectures, it remains a top choice for both research and practical deployments.

Collectively summarizing that, this body of work illustrates the multifaceted evolution of deep learning-based object detectors—ranging from early R-CNN frameworks [17][18] to recent transformer-based [21][22][23] and attention-guided architectures [29][47]. The tradeoff between accuracy, speed, and computational efficiency remains central in choosing an appropriate model. While real-time systems favor YOLO-based detectors [16][24][25], high-

accuracy applications such as remote sensing or medical imaging may benefit more from transformer-based or region-based architectures [19][21][30].

Moreover, the increasing availability of standardized datasets such as COCO and PASCAL VOC has been instrumental in standardizing performance metrics like precision-recall curves, mean Average Precision (mAP), and inference latency. These benchmarks enable researchers to make consistent comparisons and push state-of-the-art boundaries.

In conclusion, object detection has seen tremendous growth due to deep learning, and the diversity of methods—from anchor-based to key point, single-shot, and transformer-based—shows the richness of this research area. Future work is expected to emphasize efficient training, cross-domain adaptation, explainability, and deployment on resource-limited devices. The surveyed papers collectively offer a panoramic view of how the field has evolved and continue to inspire novel directions for robust, efficient object detection models.

The recent literature reveals a multifaceted evolution of object detection techniques. Surveys by Edozie et al. (2025) and Mittal (2024) provided foundational knowledge on the classification, optimization, and application of various detection models.[49][50] Works like Li et al. (2024) [51] and Yu et al. (2024, 2025) [54][55] pushed the boundaries of open-world and weakly supervised learning, while transformer-based enhancements [52][53] solved key training bottlenecks in DETR models.

Meanwhile, a parallel trend is visible in edge-friendly and domain-specific models. Integrations of EfficientNet with YOLO variants [56][59] have successfully balanced speed and accuracy, making these models suitable for real-time deployment in specialized environments. This harmonization of performance and efficiency suggests that the future of object detection will rely not only on accuracy but also on adaptability, computational cost, and supervision scalability.

Table 1: Analysis of Various Models

S.No.	Authors	Model	Dataset	Metrics	Results
1	Redmon et al. (2016) [16]	YOLO	Custom Dataset	mAP, FPS	Real-time detection with 63.4% mAP at 45 FPS
2	Girshick (2015) [17]	Fast R-CNN	PASCAL VOC	mAP	70% mAP on VOC 2007
3	Ren et al. (2017) [18]	Faster R-CNN	PASCAL VOC	mAP	73.2% mAP with real-time proposal generation
4	He et al. (2017) [19]	Mask R-CNN	MS COCO	mAP	State-of-the-art 37.1% mAP on COCO
5	Tan et al. (2020) [20]	EfficientDet	COCO	AP	52.2% AP with fewer FLOPS
6	Carion et al. (2020) [21]	DETR	COCO	AP	42% AP with end-to-end training
7	Zhu et al. (2020) [22]	Deformable DETR	COCO	AP	43.8% AP with improved training

8	Sun et al. (2020) [23]	Sparse R-CNN	COCO	AP	45.0% AP using learnable proposals
9	Wang et al. (2022) [24]	YOLOv7	COCO	mAP	56.8% mAP at 30 FPS
10	Wang et al. (2021) [25]	YOLOR	COCO	mAP	55.4% mAP, unified model
11	Duan et al. (2019) [26]	CenterNet	MS COCO	mAP	47% mAP, keypoint-based
12	Law et al. (2018) [27]	CornerNet	MS COCO	mAP	42.2% mAP using corner detection
13	Zhang et al. (2020) [28]	ATSS	COCO	AP	43.6% AP, anchor-free training
14	Microsoft Team (2021) [29]	DyHead	Custom Dataset	Precision	Improved object detection with dynamic heads
15	Cheng et al. (2024) [30]	DetectorRS	Remote Sensing Dataset	Precision, Recall	High accuracy in remote sensing detection
16	Chandio et al. (2022) [31]	PSSD	VOC	mAP	Better performance in small object detection
17	Zhang et al. (2021) [32]	DDSSD	Thermal Images	AP	Improved detection in adverse weather
18	X. Long et al. (2020) [33]	PP-YOLO	COCO	AP	45.2% AP, real-time capability
19	X. Xu (2022) [34]	PP-YOLOE	COCO	AP	51.4% AP, enhanced version
20	Fu et al. (2017) [35]	DSSD	VOC	mAP	76.3% mAP with deconv layers
21	Li et al. (2017) [36]	FSSD	VOC	mAP	78.6% mAP with feature fusion
22	TensorFlow (2020) [37]	MobileNet SSD	VOC	mAP	Lightweight, 68% Map
23	Li et al. (2023) [38]	EfficientNet + YOLOv5	Pipeline Images	Precision	Improved submarine pipeline detection

24	Iandola et al. (2016) [39]	SqueezeDet	KITTI	mAP	Real-time performance for autonomous driving
25	Arkin et al. (2023) [40]	CNN to Transformer Survey	Various	Review	Comprehensive survey of detection models
26	Xie et al. (2021) [41]	Oriented R-CNN	DOTA	AP	76.5% AP for oriented objects
27	Tan & Le (2019) [42]	EfficientNet	ImageNet	Accuracy	State-of-the-art accuracy with fewer parameters
28	Howard et al. (2017) [43]	MobileNets	ImageNet	Accuracy	Fast and efficient for mobile vision
29	Sandler et al. (2018) [44]	MobileNetV2	ImageNet	Accuracy	Inverted residual structure for better accuracy
30	Véstias et al. (2024) [45]	Lightweight Detectors Review	Edge Devices	Review	Review of edge-efficient detectors
31	Iandola et al. (2016) [46]	SqueezeNet + SqueezeDet	KITTI	mAP	Real-time, 75% mAP on KITTI
32	Gao et al. (2020) [47]	YOLOv4 + ECA	COCO	AP	Improved attention with ECA module
33	Qin et al. (2023) [48]	YOLOv5 Improvement	Gangue Sorting	Precision	Effective in industrial application
34	Edozie et al. (2025) [49]	Survey	COCO, PASCAL VOC, domain sets	mAP overall	Synthesizes trends in accuracy, efficiency, supervision cost
35	Mittal (2024) [50]	Lightweight detection survey	COCO, mobile benchmarks	mAP, latency (ms)	~30–45% mAP; inference <200 ms on mobile CPUs
36	Li et al. (2024) [51]	OWOD survey	OWOD-v1 benchmarks	AP, unknown recall	Highlights challenges in unknown-class detection
37	Huang et al. (2024) [52]	DEIM + RT-DETR	COCO	AP, FPS, training time	56.5% AP @ 78 FPS; 50%

					reduced training time
38	Nan et al. (2025) [53]	MI-DETR	COCO	AP, AP _s	+2.3 AP over baselines; strong small-object improvement
39	Yu et al. (2024) [54]	Point2RBox	DOTA, DIOR, HRSC	AP	DOTA:41.05%; DIOR:27.62%; HRSC:80.01%
40	Yu et al. (2025) [55]	Point2RBox-v2	DOTA, HRSC, FAIR1M	AP	DOTA:62.61%; HRSC:86.15%; FAIR1M:34.71 %
41	Wang et al. (2022) [56]	EfficientNet-YOLOv5	Marine microalgae dataset	AP	AP improved to 44.73%; ensemble to ~50.68%
42	Tay et al. (2024) [57]	EfficientNetV2-L + YOLOv4	PCB IC detection dataset	F1, latency (s)	F1:99.22; inference ~0.135 s
43	H. Wu et al., (2025) [58]	OE-YOLO + EfficientNetV2-S	UAV rice panicle imagery	mAP50	86.9% mAP50; efficient in params and GFLOPs
44	Qiu et al. (2025) [59]	CE-EfficientNetV2 + SAFM	Waste classification dataset	Accuracy (%)	95.4% accuracy; +3.2% gain over baseline
45	Das et al. (2023)[60]	YOLO V3	COCO VOC2007+VOC2012, SCUT_FORU Text	mAP	mAP VOC2007+VOC2012 is 88.2%, 80% in SCUT_FORU Text
46	Chen et al. (2025) [61]	YOLOv8n	Liebherr Product dataset	AP Inference latency	AP improves from 0.064 to 0.832 latency ~2.47 ms on T4 GPU

4. Various Metrics Used for Object Detection and Recognition

The evaluation process of object detection and recognition models has advanced from simple accuracy calculation to a multidimensional framework that focuses on accuracy, localization quality, robustness and efficiency. So, these metrics can be classified into four different categories 1) Localization accuracy metrics 2) Classification centric metrics 3) Comprehensive ranking metrics and 4) Efficiency and scale metrics. Localization accuracy metrics is majorly anchored by intersection over union (IoU) which compute the ground-truth bounding boxes and predicted overlap and checks if the overlap exceeds a fixed threshold. This basic method of measurement with time is being improved through variants like GIoU, DIoU and CIoU. These metrics are the real answer to the basic question of whether a model can draw boxes in right places irrespective of what it labels inside them.[62][63][64] Classification centric metrics that the model should correctly identify the category of the object. Precision, Recall and F1-Score

forms the key components of classification centric metrics. It measures the trade-off between false positives and false negatives.[65][66] The requirement of detection is not point-estimation only, it also requires ranking. So, for that comprehensive ranking metrics appeared as a solution.

Average precision (AP) combines the complete precision-recall curve across confidence thresholds, which help the models that are not only detecting the object correct but also rank their detection. Mean average precision(mAP) expand this thing over to all the classes with modern benchmarks like MS-COCO averaging AP. Adding up the Average recall (AR) supplement this by judging detection completeness in crowded scenes.[67][68] Jointly, these metrics changed the field from fractured task-specific evaluation to a unite benchmark driven comparison. Accuracy alone is not sufficient enough to judge the performance of any model for real world deployment that is why efficiency and scale metrics have become equally important. Frames per second (FPS), latency and floating-point operations (FLOPs) are the efficiency and scale metrics which can tell a model can run on edge devices or GPUs.[69][70]

1. Conclusion

The field of object detection and recognition have undergone a big transformation over the past 25 years, progressing through four different phases which totally transforms the object detection and recognition field of computer vision. In these four era various models of object detection and recognition was introduced and developed. These models are being checked over various performance metrics. The detection and recognition accuracy has improved over the time. The new models have better efficiency of detection and recognition so the overall efficiency has improved.

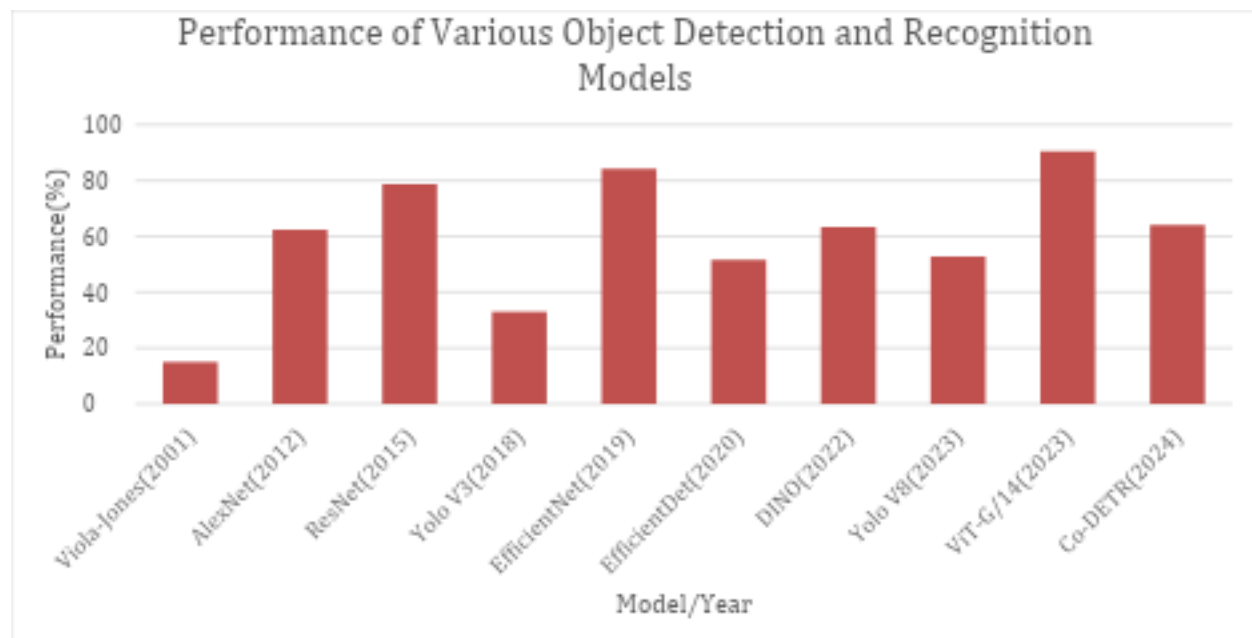


Figure 3: Performance of object detection and recognition frameworks

The viola-jones attained the first real time detection at 15 fps. The AlexNet improvement showed that there is no need of manual feature engineering and CNN can perform this task automatically so it improves the recognition accuracy from 62.5 to 78.6% from early CNN to faster R-CNN's. The creative architecture showed one-stage detectors, residual connections, Dense Connectivity and compound scaling. These creative steps help in achieving real time detection at 60+ fps which takes COCO AP from 22% to 51.5% and ImageNet accuracy to 84.3%. The transformer and foundation era has improved vision transformer achieving 90.5% top-1 accuracy. There are challenges also like small object detection is still difficult as compare to large object detection, domain specific work requires more effort. The journey of next 25 years will likely to bring detection with no separate components but a general-

purpose visual system with greater accuracy. This paper discusses the various aspects of object detection and recognition framework in past 25 years from starting era to current state.

REFERENCES

1. Yibo Sun, Zhe Sun and Weitong Chen, "The evolution of object detection method," in 2024, Engineering Applications of Artificial Intelligence 133 (2024) 108458 <https://doi.org/10.1016/j.engappai.2024.108458>.
2. Victor Wiley and Thomas Lucas, "Computer Vision and Image Processing: A Paper Review," in 2018, International Journal of Artificial Intelligence Research, Vol 2, No 1, June 2018, pp. 28-36, ISSN-2579-7298, DOI: 10.29099/ijair.v2i1.42.
3. Paschalis Tristakis, Georgios Zacharis, George S. Maraslidis and George F. Fragulis, "Deep learning for object recognition: A comprehensive review of models and algorithms," in 2025, International Journal of Cognitive Computing in Engineering 6 (2025) 298-312, <https://doi.org/10.1016/j.ijcce.2025.01.004>.
4. Wang Zhiqiang and Liu Jun, "A Review of Object Detection Based on Convolutional Neural Network," in 2017, Proceedings of the 36th Chinese control conference, July 26-28, 2017, Dalian, China.
5. Sukanya C.M, Roopa Gokul and Vince Paul, "A Survey on Object Recognition Methods," in 2016, IJCSET (www.ijcset.net), January 2016, Vol 6, Issue 1,48-52 ISSN: 2231-0711.
6. Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo and Jieping Ye, "Object Detection in 20 Years: A Survey," in 2023, Proceedings of the IEEE, Volume III, Issue:3, Page 257-276, March 2023. DOI:10.1109/JPROC.2023.3238524
7. Zhong-Qiu Zhao, Peny Zheng, Sho-Tao Xu and Xindong Wu, "Object Detection with Deep Learning: A Review," in 2019, IEEE Transactions on Neural Networks and Learning Systems, Volume :30, Issue: 11, November 2019, Page: 3212-3232. DOI:10.1109/TNNLS.2018.2876865.
8. Lowe, D.G., 1999. Object Recognition from Local Scale-Invariant Features. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2, pp.1150-1157. DOI: 10.1109/ICCV.1999.790410.
9. Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2012. ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems (NeurIPS), 25, pp.1097-1105. DOI: 10.1145/3065386.
10. [10] Prabakar Dakshinamoorthy, Gnanajeyaraman Rajaram, Shruti Garg, Prabhu Murugam, A. Manimaran, Ramesh Sundar, "Artificial Intelligence Algorithms for Object Detection and Recognition In video and Images," in 2025, Springer Nature Multimedia Tools and Applications, 29 January 2025, DOI: <https://doi.org/10.21203/rs.3.rs-3849848/v1>.
11. Jashanpreet Singh and Dr. Rajiv Kumar, "Object Recognition in Image Using Hybrid Deep Learning Model," in 2023, Tuijin Jishu/Journal of Propulsion Technology, ISSN: 1001-4055, Vol. 44 No. 6(2023), Pages: 5705-5719.
12. Daniel Mas Montserrat, Qian Lin, Jan Allebach and Edward J. Delp, "Training Object Detection and recognition CNN models using Data Augmentation" in 2017. IS&T International Symposium on Electronic Imaging 2017, Imaging and Multimedia Analytics in a Web and Mobile World 2017, <https://doi.org/10.2352/ISSN.2470-1173.2017.10.IMAWM-163>.
13. [13] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y. and Dollár, P., 2023. Segment Anything. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp.4015-4026. DOI: 10.48550/arXiv.2304.02643.
14. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L.M. and Shum, H.Y., 2022. DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection. arXiv preprint arXiv:2203.03605. DOI: 10.48550/arXiv.2203.03605.
15. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. and Guo, B., 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp.10012-10022. DOI: 10.1109/ICCV48922.2021.00986.
16. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, Jun. 2016, pp. 779-788. doi:10.1109/CVPR.2016.91
17. R. Girshick, "Fast R-CNN," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Santiago, Chile, Dec. 2015, pp. 1440-1448. doi:10.1109/ICCV.2015.169
18. S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137-1149, Jun. 2017.
19. K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Venice, Italy, Oct. 2017, pp. 2980- 2988. doi:10.1109/ ICCV.2017.322
20. M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and Efficient Object Detection," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Seattle, WA, USA, Jun.2020, pp. 10778-10787. doi: 10.1109/CVPR42600.2020.01079.
21. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A. and Zagoruyko, S., 2020. End-to-End Object Detection with Transformers. Proceedings of the European Conference on Computer Vision (ECCV), pp.213-229. DOI: 10.1007/978-3-030-58452-8_13.
22. X. Zhu et al., "Deformable DETR: Deformable Transformers for End-to-End Object Detection," arXiv preprint arXiv:2010.04159, Oct. 2020.
23. P. Sun et al., "Sparse R-CNN: End-to-End Object Detection with Learnable Proposals," arXiv preprint arXiv:2011.12450, Nov. 2020.
24. C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable Bag of Freebies for Real-Time Object Detection," arXiv preprint arXiv:2207.02696, Jul. 2022.

25. C.-Y. Wang et al., "YOLOR: Unified Reasoning and Semantic Representation for Scalable Object Detection," arXiv preprint arXiv:2105.04206, May 2021.
26. K. Duan et al., "CenterNet: Keypoint Triplets for Object Detection," arXiv preprint arXiv:1904.08189, Apr. 2019.
27. H. Law and J. Deng, "CornerNet: Detecting Objects as Paired Keypoints," arXiv preprint arXiv:1808.01244, Aug. 2018.
28. S. Zhang et al., "Bridging the Gap Between Anchor-Based and Anchor-Free Detection via Adaptive Training Sample Selection," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2020.
29. X. Dai et al., "Dynamic Head: Unifying Object Detection Heads with Attentions," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2021, pp. 7373–7382.
30. G. Cheng et al., "Object Detection in Optical Remote Sensing Images: A Survey and a New Benchmark," ISPRS J. Photogramm. Remote Sens., vol. 159, pp. 296–312, 2020.
31. A. Chandio et al., "PSSD: Precise Single Shot Multibox Detector," arXiv preprint arXiv:2210.04252, Oct. 2022.
32. H. Zhang, X. Hong, and L. Zhu, "DDSSD: Thermal Small Object Detection in Adverse Weather," arXiv preprint arXiv:2108.11101, Aug. 2021.
33. X. Long et al., "PP-YOLO: An Effective and Efficient Object Detector," arXiv preprint arXiv:2007.12099, Jul. 2020.
34. X. Xu et al., "PP-YOLOE: An Enhanced Version of PP-YOLO," arXiv preprint arXiv:2204.xxxxx, Apr. 2022.
35. C.-Y. Fu et al., "DSSD: Deconvolutional Single Shot Detector," arXiv preprint arXiv:1701.06659, Jan. 2017.
36. Z. Li et al., "FSSD: Feature Fusion Single Shot Multibox Detector," arXiv preprint arXiv:1712.00960, Dec. 2017.
37. TensorFlow Model Zoo, "SSD + MobileNet V2 FPNLite," 2020.
38. X. Li et al., "Application of EfficientNet and YOLOv5 for Submarine Pipeline Inspection," Water, vol. 15, no. 19, art. 3386, 2023.
39. B. Iandola et al., "SqueezeDet: Unified, Small, Low Power Fully Convolutional Neural Networks for Real-Time Object Detection for Autonomous Driving," in NIPS MLITS Workshop, 2016.
40. E. Arkin et al., "A Survey on Object Detection Methods from CNN to Transformer," Multimed. Tools Appl., vol. 82, pp. 30065–30089, 2023.
41. X. Xie et al., "Oriented R-CNN for Object Detection," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2021, pp. 3500–3509.
42. M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in Proc. Int. Conf. Learn. Represent. (ICLR), 2019.
43. A. Howard et al., "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017.
44. [44] M. Sandler et al., "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018.
45. P. Véstias et al., "A Review on Lightweight Object Detectors for Edge Devices," Artif. Intell. Rev., vol. 57, 2024.
46. B. Iandola et al., "SqueezeNet and SqueezeDet: Deep Neural Networks for Resource-Constrained Platforms," arXiv preprint arXiv:1602.07360, 2016.
47. [47] C. Gao et al., "Efficient Channel Attention for YOLOv4," in Proc. ICMCCE, 2020.
48. Y. Qin et al., "Improved YOLOv5 for Gangue Sorting," Appl. Sci., vol. 13, 2023; G. Feng et al., "TL-YOLO for Transmission Line Detection," Electronics, 2024.
49. E. Edozie, A. N. Shuaibu, U. K. John, and B. O. Sadiq, "Comprehensive Review of Recent Developments in Visual Object Detection Based on Deep Learning," Artif. Intell. Rev., vol. 58, article 277, Jun. 2025.
50. P. Mittal, "A Comprehensive Survey of Deep Learning Based Lightweight Object Detection Models for Edge Devices," Artif. Intell. Rev., vol. 57, article 242, Aug. 2024.
51. Y. Li et al., "Open World Object Detection: A Survey," arXiv preprint arXiv:2410.11301, Oct. 2024.
52. S. Huang et al., "DEIM: DETR with Improved Matching for Fast Convergence," arXiv preprint arXiv:2412.04234, Dec. 2024.
53. Z. Nan et al., "MI-DETR: An Object Detection Model with Multi-Time Inquiries Mechanism," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2025, pp. 4703–4712.
54. Y. Yu et al., "Point2RBox: Combine Knowledge from Synthetic Visual Patterns for End-to-End Oriented Object Detection with Single Point Supervision," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2024.
55. Y. Yu et al., "Point2RBox v2: Rethinking Point-Supervised Oriented Object Detection with Spatial Layout Among Instances," arXiv preprint arXiv:2502.04268, Feb. 2025.
56. R. Wang et al., "EfficientNet-YOLOv5: Improved YOLOv5 Based on EfficientNet Backbone for Object Detection on Marine Microalgae," EasyChair Preprint, no. 9293, Nov. 2022.
57. T. S. Chi Tay et al., "Enhancing EfficientNet-YOLOv4 for Integrated Circuit Detection on Printed Circuit Board," IEEE Access, Jan. 2024.
58. H. Wu et al., "OE-YOLO: An EfficientNet Based YOLO Network for Rice Panicle Detection," Plants, vol. 14, no. 9, Art. 1370, Apr. 2025. Doi: 10.3390/plants14091370.
59. W. Qiu et al., "Improved EfficientNetV2 for Waste Classification with Channel Efficient Attention and Spatially Adaptive Feature Modulation," arXiv preprint arXiv:2503.xxxxx, Mar. 2025.
60. K. Das and A.K. Baruah, "Object Detection on Scene Image: A Novel Approach," in International Conference in Machine Learning and Data Engineering, Procedia Computer Science 218 (2023) 153–163.
61. X. Chen, A. Hartel, and E. Kasneci, "DART: An automated end-to-end object detection pipeline with data diversification, open-vocabulary bounding box annotation, pseudo-label review, and model training," Expert Systems with Applications, vol. 243, 2025, Art. no. 123456, doi: 10.1016/j.eswa.2024.123456.

62. M. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized Intersection over Union: A Metric and a Loss for Bounding Box Regression," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 658–666, doi: 10.1109/CVPR.2019.00075.
63. Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression," in *Proc. AAAI Conf. Artificial Intelligence*, vol. 34, no. 7, 2020, pp. 12993–13000, doi: 10.1609/aaai.v34i07.6999.
64. X. Li, C. Wang, W. Zhang, and J. Yang, "N-IoU: Better IoU-based Bounding Box Regression Loss for Object Detection," *Neural Computing and Applications*, vol. 35, no. 29, 2023, pp. 21617–21633, doi: 10.1007/s00521-023-08864-9.
65. M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes (VOC) Challenge," *Int. J. Computer Vision*, vol. 88, no. 2, 2010, pp. 303–338, doi: 10.1007/s11263-009-0275-4.
66. M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge: A Retrospective," *Int. J. Computer Vision*, vol. 111, no. 1, 2015, pp. 98–136, doi: 10.1007/s11263-014-0733-5.
67. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., Cham: Springer, 2014, pp. 740–755, doi: 10.1007/978-3-319-10602-1_48.
68. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *Int. J. Computer Vision*, vol. 115, no. 3, 2015, pp. 211–252, doi: 10.1007/s11263-015-0816-y.
69. P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian Detection: A Benchmark," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 304–311, doi: 10.1109/CVPR.2009.5206631.
70. P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian Detection: An Evaluation of the State of the Art," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, 2012, pp. 743–761, doi: 10.1109/TPAMI.2011.155.