JOURNAL OF OPTOELECTRONICS LASER ISSN: 1005-0086

Volume 41, Issue 8, 2022

# A Review of One Stage Based Deep Learning Techniques for Object Recognition and Detection

**Anjani Kumar*[1] and Dr. Ram Kumar Karsh[2]**

*[1]National Institute of Technology Silchar, Department of ECE, Ph.D. Scholar,anjani_rs@ece.nits.ac.in.*

*[2]National Institute of Technology Silchar, Department of ECE, Assistant Professor, ram@ece.nits.ac.in.*

## ABSTRACT

With the advent of practical world operations, image detection and labeling have become a significant field of machine learning. The present assessment provides an overview of the major significant developments in image recognition and classification during the last couple of years.Observations from deep learning-dependent 2D entity recognition frameworks using single-stage object recognition methods are the topic of this literature evaluation. This study surveys the most up-to-date techniques for anticipating and recognizing objects. Various strategies, data sets, and promising future developments have been extensively addressed.

**Keywords**: YOLO, One Stage Image Detection, Image Assessment, Object Detection, Over Feat, DSSD,SSD, Deep Learning, Artificial Intelligence, FSSD,Convolutional Neural Network,DSOD.

## I. INTRODUCTION

"Deep learning" (DL), "Artificial intelligence" (AI), and other fields of "Object Recognition" (OR) and "Object Detection" (OD) are crucial topics of research. It is a fundamental prerequisite for complex computerized OR operations, such as image semantic understanding, activity detection, behavior analysis, and monitoring capabilities. It aims to identify the most significant item in a photo, properly define its categories, and offer the object's perimeter area. In running automated vehicles, recovering videos and pictures, conducting intelligent surveillance [1], evaluating medical image data [2], conducting corporate inspections [3], and other fields of target monitoring and identification have been widely applied.

Traditionally employed for specialized recognition tasks, manually characteristic extraction-based recognition techniques typically include 6 phases: "pre-processing", "window sliding", characteristic extracting, characteristic selecting, typical classification, and information "post-processing". Unfortunately, traditional OD has severalsignificant drawbacks, including restricted dataset volume, poor mobility, insufficient relevance, high temporal sophistication, window redundancies, and inadequate efficacy in a few fundamental circumstances.

An AlexNet picture classification approach founded on a "convolutional neural network" (CNN) was discussed by Krizhevsjy [4] in 2012. The winner of the image classification competition held by the "image database Image Net" [5] was krizhevsjy, who outperformed the second-place finisher by 11%. Many investigators have begun utilizing "Deep CNN" (DCNN) to OR issues and have proposed a wide range of improved methods. The "One Stage Detection" (OSD) approach, dependent on region assignment, and the "Two Stage Detection" (TSD) approach depending on the assessment, are the two primary categories into which advanced OR and OD methods may be divided. The functioning example of an OSD system is illustrated in fig 1.

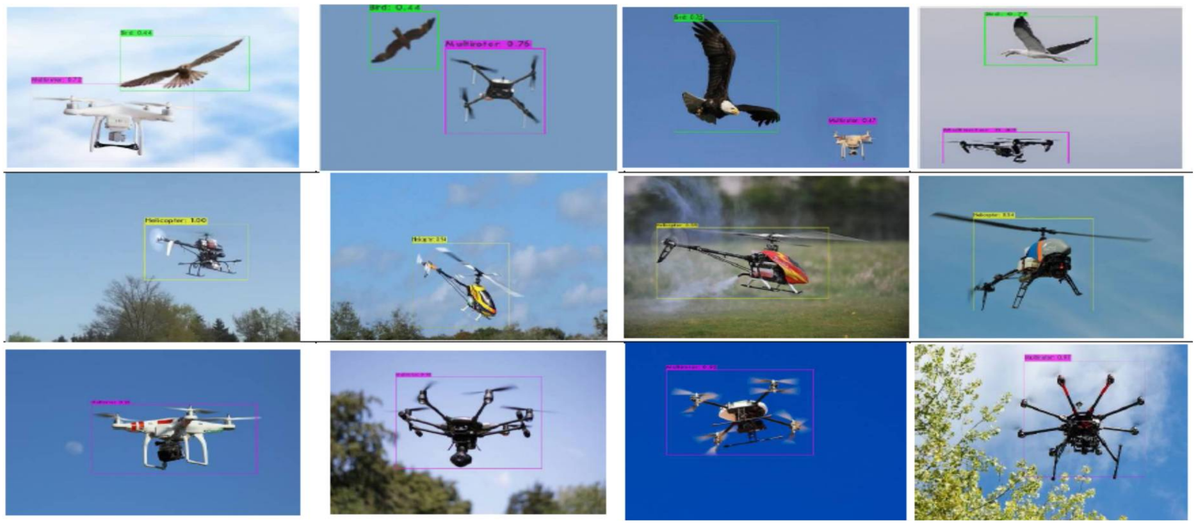

Fig1: OSD system [1].

As seen in Fig. 2, the source images, initial DL infrastructure, infrastructure evolution, and OD are the main components of conventional OD methodologies.

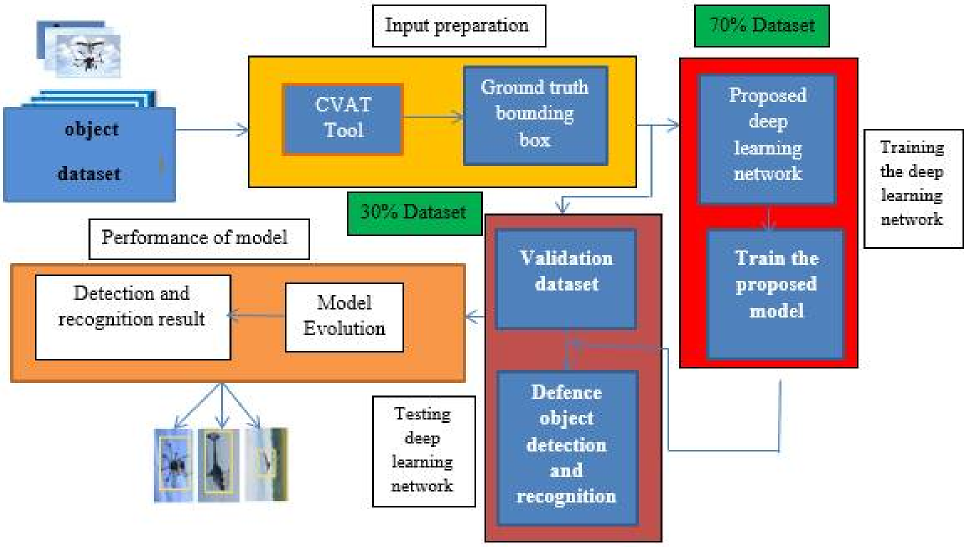
object dataset
Input preparation
CVAT Tool
Ground truth bounding box
70% Dataset
Proposed deep learning network
Training the deep learning network
Train the proposed model
30% Dataset
Validation dataset
Defence object detection and recognition
Performance of model
Detection and recognition result
Model Evolution
Testing deep learning network

Fig 2: Theframework of the traditional OD system[6].

The two types of OR are image location and image classification [7]. Traditional OR is focused on stages separated into components, as illustrated in fig. Three, since the DCNN, has a significant capability for expressing characteristics.

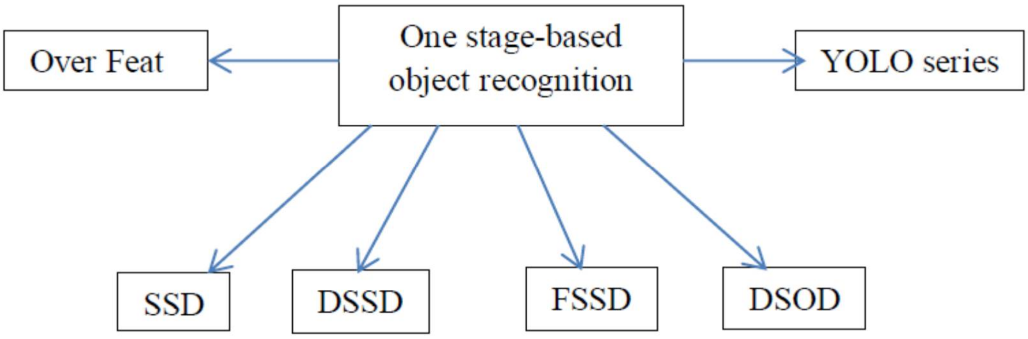

Fig 3:Stage-basedODtechniques.

Instead of the relatively conventional two-stage approach, OSD techniques use DCNN to locate and classify entire entities in one phase. Because the area allocation approach is more accessible to apply than OR, OSD can be employed to calculate an individual's categorization potentials and placement characteristics on a stage right away. Speedy OR is a significant benefit, though TSD approaches often perform better. In one phase of image identification, the "YOLO series", "Over Feat", "SSD", "DSSD", "FSSD", and "DSOD" are all incorporated. This paper reviewed various OSD and item identification techniques and compared them based on their accuracy and usefulness.

## II. STAGE-BASED OR

The most popular OR systems in use today are OSD and TSD approaches. However, with continued efforts to enhance OSD by basing its architectural construction on TSD methods, TSD currently outperforms OSD regarding accuracy, while OSD remains a shortfall [8]. Additional details on OR using OSD methods are provided in the subsequent parts.

### 2.1OSD algorithms

Despite needing additional searching for the region, the DL-based OR framework has an OSD approach that gives the classification likelihood and positioning parameters for the picture right away. Therefore, performing the finished immediate OR in OSD is feasible using the different DL approaches presently accessible [9, 10]. The following parts go into additional information about distinct OSD techniques.

### 2.1.1 over Feat

To combine OD and OR into a single network architecture using DCNN features, "Over Feat" (OF) was proposed by Sermanet [10]. With a multiple-scale fast, moving window, OF attempts to do away with patches for end pooled layers of the DCNN. Instead, the classification score for each patch must be predicted before any patches can be merged. This approach fixed the multiple size problems and the complex image form. Furthermore, OF locates and categorizes items using DCNN's categorization methods. However, while OF outperforms RCNN in speed, it falls short in accuracy.

### 2.1.2 YOLO

Redmon presented YOLO in 2016, a novel methodology that forecasts confidentiality for multiple groups and surrounding locations using the whole outermost functionality mappings [11]. Figure 4 shows the basic framework of YOLO.

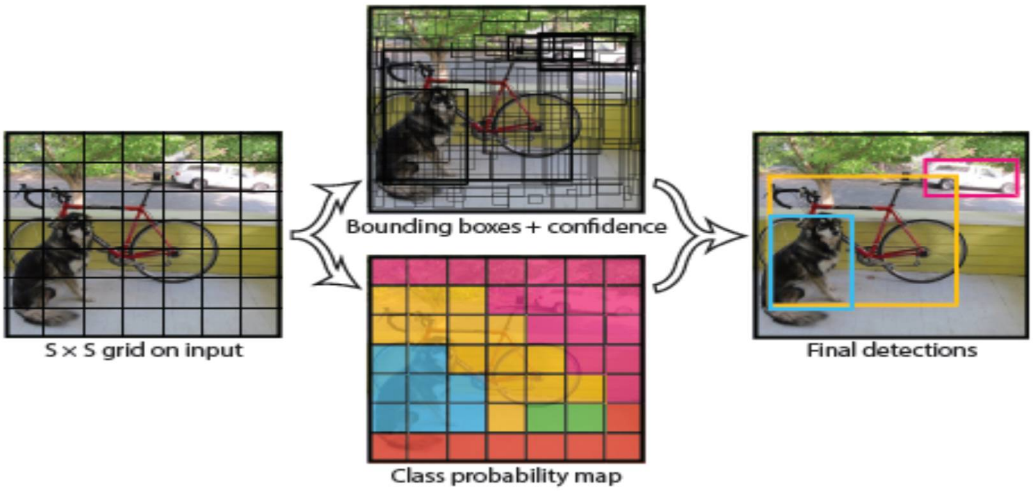

Fig 4:Principalframeworkof YOLO[11].

The complete input picture in the SxS grid cell is converted using YOLO [9, 12]. Images that are in a particular matrix cell must be detected by every matrix cell. The entire image is recorded as an "SxSx (5B+C)" output matrix, with each grid cell forecasting C category potentials, B boundaries squared boxes and probability scores. In Figure 5, the YOLO framework is displayed.

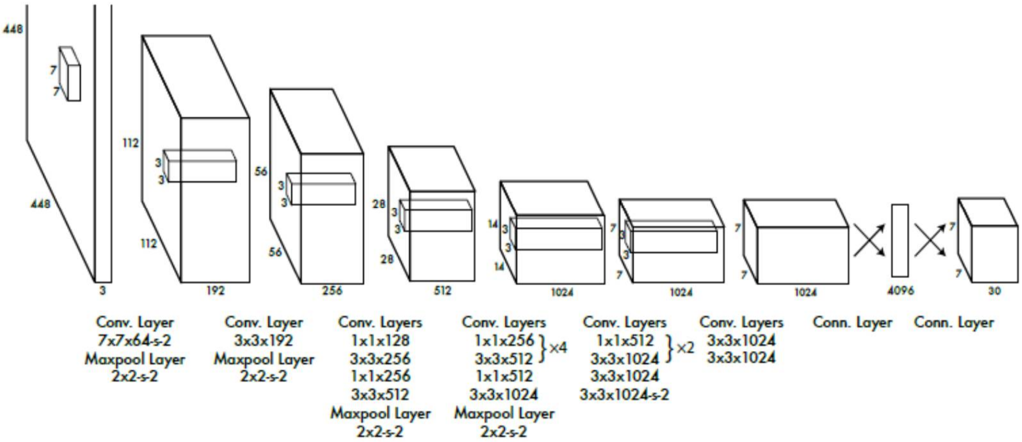

Fig 5: YOLOsimpleconstruction [12]

The YOLO has 24 "Convolution Layers" (CN) and 2 "fully connected" (FC) layers; Some of them group together inception systems that have 3x3 CN after 1x1 reducing phases. The algorithm can analyze photographs in real-time at a rate of 45 "frames per second" (fps). However, a less complex variation called "Quick YOLO" can analyze pictures more effectively than real-time recognition at a rate of 155 fps. Additionally, YOLO produces less misleading results, allowing it to be used in conjunction with Fast R-CNN. As a result, YOLO recognition is quick [13, 14, 15]; however, it has trouble picking up small items and elements that are near. Therefore,Redmon and Farhadi introduced YOLO9000 in 2016 [16], an improved form that includes outstanding techniques, including BN, anchoring boxes, multidimensional clustering, and multi-scale learning.

DOI:10050086.2022.08.68

### 2.1.3 YOLOV2/9000

Around 9000 diverse types of various items may be detected and identified using the real-time OR system YOLOV2/9000. On well-known recognition challenges including "PASCAL VOC" and "COCO," the novel technique, YOLOv2, scores better than earlier iterations, and in VOC 2007, YOLOv2 achieved a 76.8 mAP using 67 frames per second. YOLOv2's 78.6 mAP at 40 fps outperforms faster RCNN with ResNet and SSD, even if they are significantly better than YOLOv2. Redmon and Farhadi [16] have presented a method for simultaneously teaching item recognition and classification. In the following aspects, YOLOV2 [17] is more accurate than YOLOV1 in respect of precision:

(a.) System generalization and standardization are accelerated by batch normalization [18, 19].

(b.) To strengthen YOLO's weak capacity to generalize for different aspect ratios. The anchoring concept is included in YOLOV2's Faster RCNN [20], which also permits forecasting three aspect ratios and three scales per grid cell.

(c). YOLOV2 resolves the instabilities of the architecture by limiting the drifting of the "ground truth" relating to the position of the grid cell region between 0 and 1.

### 2.1.4 YOLOV3

The fundamentals of "YOLOV1" and "YOLOV2/9000" are expanded upon in "YOLOV3 [21], which also strengthens any flaws already present to achieve a balance across performance and accuracy. Utilizing a combination of the "residual block" [22], "Function Pyramid Network" (FPN) [23], and "binary cross-entropy loss" [24], "YOLO" is updated to "YOLOV3". Such modifications make it easier for the classification architecture to identify highly complex items, including multiple groupings and components with different parameters. Consequently, it is a little bigger yet better effective than "YOLOV2/9000". Although it didn't come to YOLOv3 very long to find and identify an item, various studies confirm that YOLOv3 outperforms SSD by a three-fold margin.

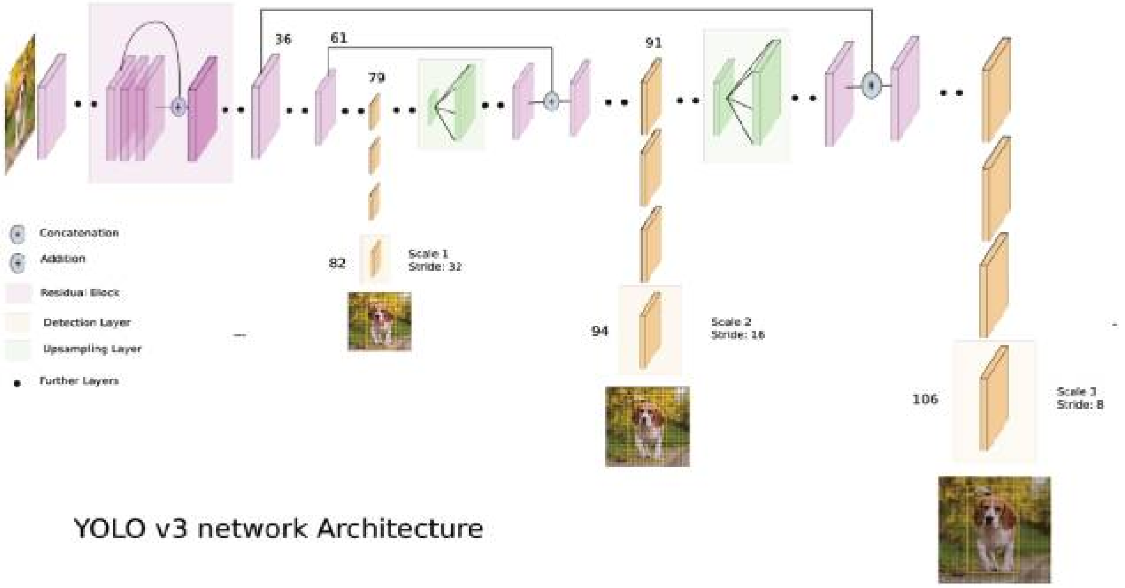

Fig 6: Yolov3 Architecture [24].

YOLO v3 employs a Dark net architecture that was developed on ImageNet. The platform offers an entirely convoluted 106 levels underneath YOLO v3 topology by introducing 53 extra tiers to the existing 53-layer setup first presented on the Darknet. Following an additional 53 layers for identification, the YOLO v3 identification system, shown in Figure 6, comprises 106 CL. Because of this, YOLO v3 operates less well than v2, yet v2 frequently has issues recognizing small entities. There are numerous detection folds employed to overcome this problem. By joining observed layers with earlier levels, fine-grained properties may be retained.

### 2.1.5 YOLOV4

The YOLOV4 platform's main objective is to build an OD system that functions rapidly and maximizes simultaneous computations, not to hypothetically reach a low computing size. As a result, compared to specific other OD techniques that have equal efficacy, "YOLOv4" is twice as fast as its previous YOLO versions. For example, figure 7 shows how it improves "YOLOv3's" working frame per second rate by 10% and AP by 12% [25].

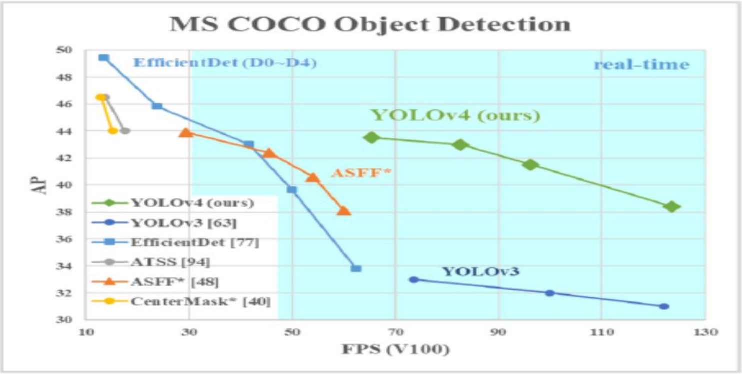

Fig 7: Comparison of “YOLOv4” performance with variousODalgorithms [26].

Only a small number of the parameters proven by YOLOv4 had been employed to improve the ability of categorization and monitoring [26].

### 2.1.6 YOLOV5

Due to the time-consuming recognition,in-time, personalized recognition methods employing current procedures are challenging. The "YOLOv5" algorithms were created to address the limitations of "YOLOv4" in contrast to integrating the "YOLOv4" methodology. The CSP design, concentrated system, and SPP prismatic structure in "YOLOv5" may be leveraged to minimize system characteristics [27, 28, and 29]. The effectiveness chart for "YOLOv5" and its comparable designs is shown in Fig. 8.

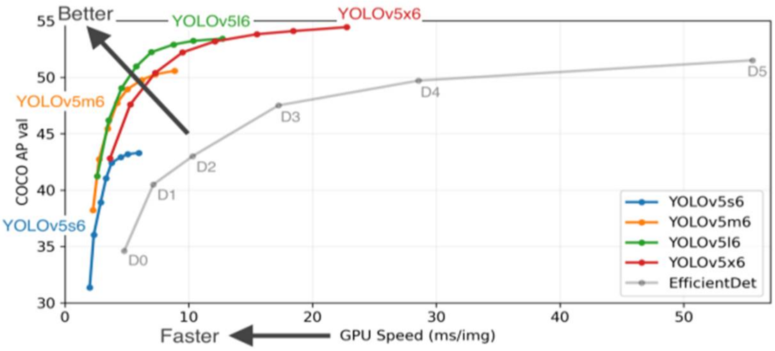

Fig 8:Evaluating the YOLOv5 against similar methodologies [29].

The layout of the DL designs is shown in Figure 9. The "YOLOv3" architecture served as the foundation for the Focused element. The characteristics are initially extracted from the picture using "CSPDarknet53" and are subsequently merged using "PANet".

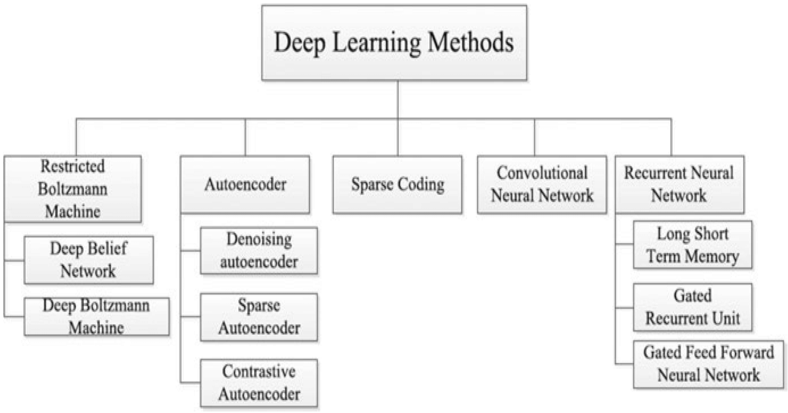

Figure 9: VariousDL designs [2]

The predetermined size limitation of the design is lifted by "Spatial Pyramid Pooling" (SPP). The previous layer is divided into little parts using the "concrete slicing layer" to improve the performance of combinations inside the following terminal.

The most recent iteration for the "YOLO" design is "Yolov5" [30, 31] which performs well in both OR and OD, with a strong detection frequency of upwards to 140 FPS. Furthermore, because "YOLOv5's" OR design framework size of the file is around 90% smaller whenever it refers to actual identification compared to their ancestor "YOLOv4", it is perfect for usage in embedded devices. The "YOLOv5" [31] platform's great identification ability, minimal weight, and rapid identification quickness are, as a consequence, its key advantages.

### 2.1.7YOLO-R

An enhanced "R-YOLO" technique will be presented by Wang in 2021 [32]. "R-YOLO" employs "end-to-end" DL to recognize and categorize the slanted anchoring pixels of an item in a realistic image. In addition, the author's adoption of an extra peripheral identification group resulted in the introduction of the "RDIoU-NMS" approach, an upgrade toward the "box analysis" framework, and a modification of the platform's anomalous property. Fig. 10 depicts the "YOLO-R" topology.

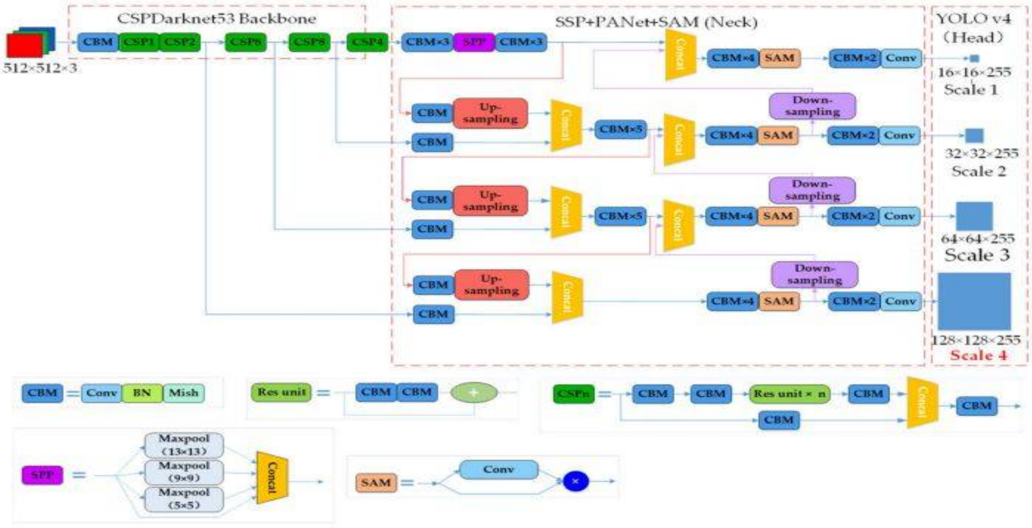

Figure 10: Framework of YOLO-R [32].

### 2.1.8 YOLO-X

OSD YOLOX enhances YOLOv3 by using DarkNet53 architecture. Untangled heads replace the head of YOLO. The Writer initially adopts a 1x1 CL to reduce the signal gateway to 256 while introducing the two simultaneous sections containing two 3 x 3 CL each for classification and forecasting activities. However "YOLOX" technique offers the most outstanding result and accuracy. The "YOLOv3" technique which is still among the foremost widely employed throughout the sector, has had its architecture enhanced to 47.3 % AP on COCO, which is 3.0 times better than best practice [33]. Various researchers, such as Wang et al., 2021, also operated upon the "YOLO-X" design and demonstrated that the results of "Faster R-CNN", "SSD", "Tiny-YOLO", "YOLOv1", "YOLOv2", "Tiny-YOLO", "YOLOv3", and "YOLOv4" with various optimal control strategies were outperformed by X-YOLO [34]. And shows increasing mAP by 96.02% and recall by 98.5 Fig. 11 depicts the YOLO-architectural X's design.

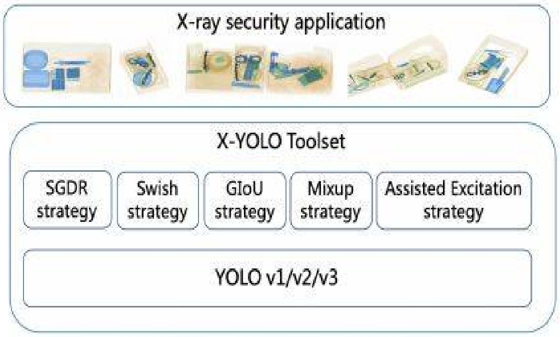

Fig 11:Structure of X-YOLO [34]

### 2.1.9 SSD

SSD is the most recent advancement in OSD ascontrary to past iterations, SSD is both faster and more accurate. A straightforward convolutional screening technique is employed to predict categorization scores and boxed alignment for a specified initial set of anchoring frames [35]. To achieve substantial identification accuracy, the Researcher intentionally divided predictions by aspect ratio and utilized the characteristic layouts of different scales to make predictions for numerous levels. Furthermore, speed vs. accuracy is much improved by the capacity to train rapidly and precisely, even with images of poor resolution.

SSD [36] creates a predetermined dimension array of enclosed blocks using a "feed-forward" CNN, followed by a non-suppressed phase to obtain final detections. In SSD, the shortened core architecture is supplemented with CNN layers. These layers shrink over age, making it possible to anticipate the identification process of varied shapes. A multi-layered identification and forecasting methodology are specific to every characteristic layer. Every subsequent characteristic layer from the central channel could generate a specific set of recognizing predictions by using a combination of convolutional processing. Figure 12 [37] depicts them over the SSD system architecture.

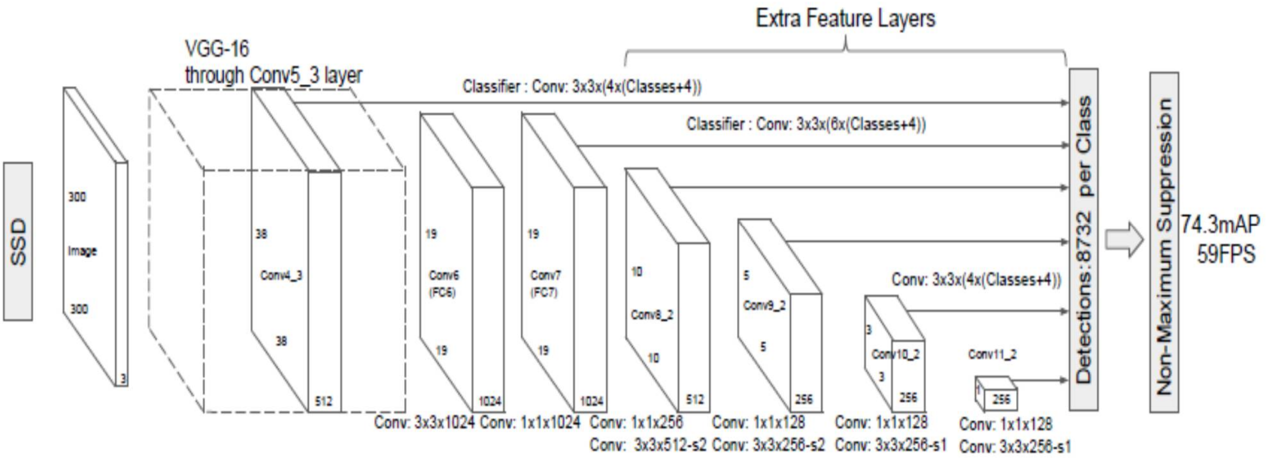

Fig 12: Framework of SSD design [37].

A "33p tiny kernel" is the crucial component for identifying expected characteristics for a collection of features of the mxn matrix with an overall number of "channels" of p. The kernel produces an ultimate response for each of its mxn matrix sections. At each operational arrangement point, the resulting scores of the surrounding rectangular deviations are calculated concerning the conventional framing orientation. The characteristic mappings are tiled using the common borders in a layered fashion, with each box attached to its corresponding unit. The Researcher anticipates anomalies from the typical box configurations for each cell of the characteristic mappings and ranks each category reflecting the presence of a classification frequency in each box [38, 39].

### 2.1.12 DSSD and FSSD

The acronyms "DSSD" and "FSSD" stand for the same approach. SSD [35, 36] used "ResNet101" as based architecture to enhance SSD's capabilities to describe reduced attribute configurations. "Deconvolution" systems and skip-connections could be used to enhance low-level image features [22] significantly and achieve feature uniformity. According to how SSD is used, FSSD transforms low-level characteristics into high-level characteristics, greatly enhancing the user's accuracy.

## III. OBJECT RECOGNITION DATASETS

The OR and OD procedure requires databases. Furthermore, datarecords are necessary to guarantee that comparisons of all techniques are fair. The complexity and wide range of program configurations make it challenging to create a uniform and all-encompassing databases. Several datasets were compiled to evaluate and analyze results based on OSD and assessments. The following section gives an overview of the conventional dataset utilized in the OSD method. Numerous well-known datasets, as well as specific, more recent OR algorithms, have been tested.

### 3.1 Experimental Evaluationof fast age-based datasets:

Furthermore, "PASCAL VOC 2007," [40] "PASCAL VOC 2012," [41] and "Microsoft COCO" [25, 42] have been used to compare OR methods.

There are many methods employed to evaluate the OSD methods and OD, these including "SSD300 [49], SSD500 [46], YOLO [56], YOLOv2 [57], YOLOv3 [23], YOLOv4 [25], YOLOv5 [28], Faster R-CNN [41][45], SPPnet [40], R-CNN [39], R-FCN [64,65], DSOD [61], HyperNet [46], Bayes [43], Mask R-CNN [59], PVANET [55], MR-CNN & S-CNN [47], FPN [58], G-CNN

[52], SD [60], StuffNet [49], and SubCNN [53]". The following analyses of classification and accuracy, in addition to an evaluation of testing use on "PASCAL VOC 2007 vs. 2012", have been made and addressed.

**(A) PASCAL VOC 2007**

Each tag of the "VOC2007" dataset had 20 classifications, including humans, cars, and airplanes. A total of 500,000 Flickr photographs were obtained, and every20 groups received one set of annotations through Flickr. When more than one picture satisfies a search, the latest photograph will be supplied first since Flickr search findings are presented with "recency" and ordered by "recency." The primary areas are identification technique, dataset authentication, and database evaluation. Each subgroup and group's frequency of pictures and object repetitions is shown. [40].

**(B) PASCAL VOC 2012**

Twenty classifications are included in the "PASCAL VOC 2012" sample sources [64]. Every group's "average precision" (AP) and the "mean average precision" (mAP) across all 20 courses were assessed [11]. The current sample collection consists of around 15,000 captioned images that have been divided into 20 groups, which include the subcategories "pet," "dog," and "vehicle." The photographs were scaled to a fixed "aspect ratio of 128x128 pixels" with a range of sizes. Although the Researcher similarly divided the sample 60:40 in support of training configurations, the Researcher first combined the assessment and learning sets at 50:50. Every subgroup also received at least 500 photos for retraining and evaluating the algorithms, with every category having a different amount of images to train and assess. The average sub-object was the only pre-processing technique used on the images.

**(C) Microsoft COCO**

The "Microsoft COCO" database consists of 3 lakh completely segmented images, with 7 item happenings from 80 different categories in every image. Because there are lesser iconic objects of different sizes and a greater demand for object localization, the current data set is much more challenging to deal with than "PASCAL 2012". Therefore, the AP derived at various IoU values and on various image sizes are used to evaluate OD performance. The current investigation results are summarised in Table 1, and the mAP and FPS schematic information analyses are shown in Figures 13 and 14 correspondingly. A characteristic hierarchy for multiresolution visualization may be created using FPN and DSSD's enhanced methodologies, in addition to findings that are the same as those provided by "PASCAL VOC". For accurate object localization, supplementary data from similar prior operations is also valid [65].

Table.1. Presentation ofthe coco dataset

| Algorithm | mAP | Testing | Training | FPS |
|---|---|---|---|---|
| SSD300 | 41.2 | TD | CV | 46 |
| SSD500 | 46.5 | TD | CV | 19 |
| YOLOv2 608×608 | 48.1 | TD | CV | 40 |
| Tiny YOLO | 23.7 | TD | CV | 244 |
| SSD321 | 45.4 | TD | CV | 16 |
| DSSD321 | 46.1 | TD | CV | 12 |
| R-FCN | 51.9 | TD | CV | 12 |
| SSD513 | 50.4 | TD | CV | 8 |
| DSSD513 | 53.3 | TD | CV | 6 |
| FPN FRCN | 59.1 | TD | CV | 6 |
| YOLOv3 320 | 51.5 | TD | CV | 45 |
| YOLOv3 416 | 55.3 | TD | CV | 35 |
| YOLOv3 608 | 57.9 | TD | CV | 20 |
| YOLOv3 tiny | 33.1 | TD | CV | 220 |
| YOLOv3 ssp | 60.6 | TD | CV | 20 |
| YOLOv4 416(without tensorRT) | 41.2 | TD | CV | 30 |
| YOLOv4 512(without tensorRT) | 43.0 | Test-dev | COCO | 30 |
| YOLOv4 608(without tensorRT) | 43.5 | TD | CV | 30 |
| YOLOv5x 640 | 50.4 | TD | CV | - |

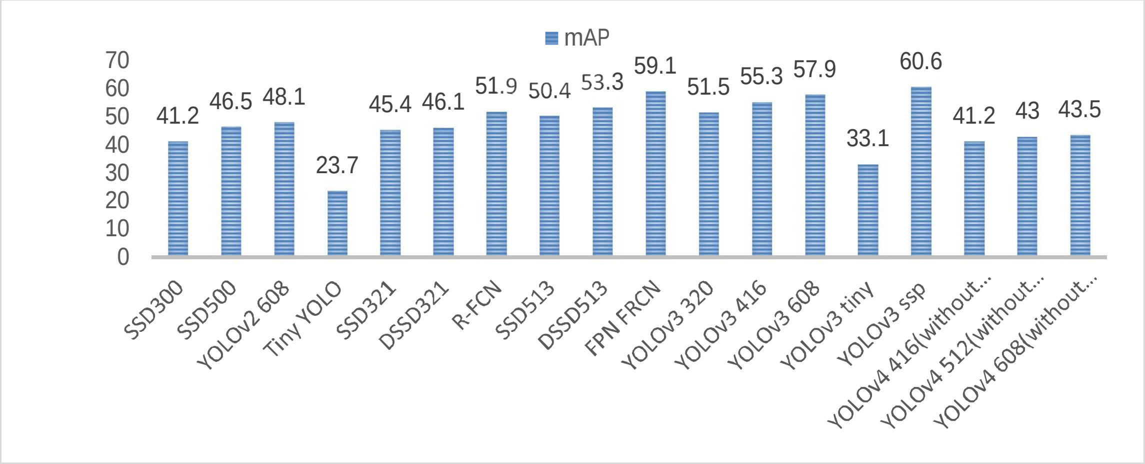

Fig 13:Evaluation of mAPofOSD strategies.

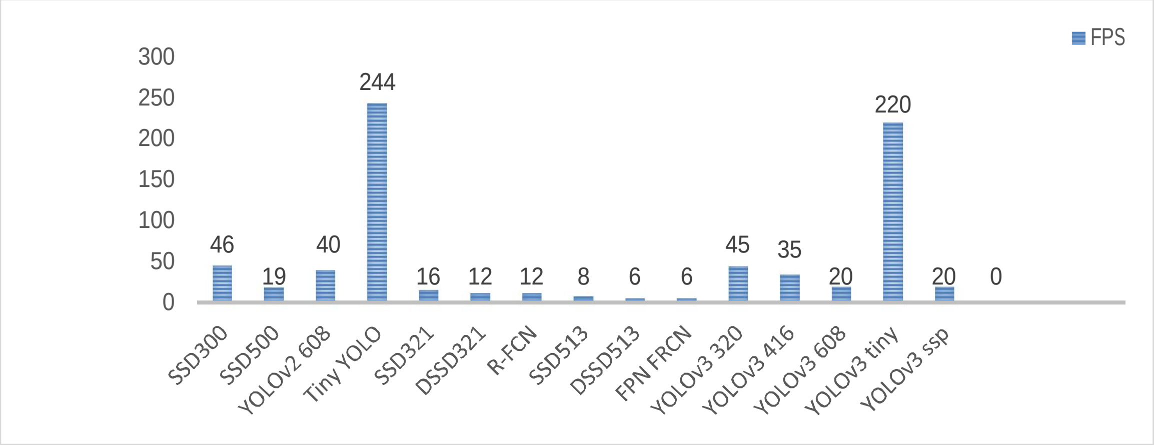

Fig 14: Assessment of FPS for different OSD techniques.

An image database is crucial for OD and its evaluation [51]. The performance of the OD operation is also steadily rising thanks to precise and integrated learning frameworks for the benefits and drawbacks of various OR and OD strategies. A public domain dataset that can be freely accessible in public data and a detailed database may support the enterprise's growth. Our main areas of concern in terms of analysis include location monitoring, OR and OD, and classification methods.

## IV. COMPLEX PROBLEM OF OBJECT DETECTION AND RECOGNITION

In real-world scenario applications, object identification faces several challenges. There include multiple-scale object detection and recognition, tiny object detection, class imbalance, crowded occlusion, and redundant detection. Numerous strategies are put up by researchers as solutions to these problems and concerns. The answers enable the DCNN-based object detection system to advance significantly in real-world applications.

### 4.1 Dense occlusion

In pedestrian identification [66], autonomous driving [67, 68], and other real-world application situations, the problem of dense occlusion frequently arises. Occlusion between things belonging to the same category and occlusion between objects belonging to separate categories are the two circumstances into which it is split. Occlusion can result in the loss of object information, including missing and erroneous detection. Researchers can utilize the extra object data in conventional object detection methods. To get around the problem of thick occlusion, use characteristics like local features, border information, and grey information. This review paper concentrates on DCNN-based approaches to the comprehensive occlusion problem.

### 4.2 Detection and recgnition of small object

One of the challenges in object detection is small object detection. The development of related applications, including automated driving, remote sensing image detection, industrial defect detection, and medical image detection, will be aided by improvements in tiny object detection. Currently, specific traditional detection techniques (such as Faster RCNN, YOLO, and SSD) are not the best for finding small objects.

Object detection, one of the fundamental functions of computer vision, has many practical uses. Depending on the individual duties, object detection technology implementation varies in real-world application settings. This section reviews some significant applications of object detection, such as vehicle detection,face detection [69], salient object detection [70], pedestrian detection [66], remote sensing image detection [71], and medical image detection [72].

## V. CONCLUSION

In-depth information on OSD techniques is provided in the present review paper. Although currently, FPS and mAP have been used, it has demonstrated that the OSD method can surpass competing strategies regarding reliability and quickness with the right amount of study and testing. This study also looks at some of the most significant databases for OD. This study will serve as an insightful analysis of recent research in OD and related DL approaches and will provide precise direction for further development.

## REFERENCES

[1] Tian, Zhi, 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, pp. 9627-9636. 2019.

[2] Fei, Chengwei, Rong Liu, Zihao Li, Tianmin Wang, and Faisal N. Baig. "Machine and deep learning algorithms for wearable health monitoring." In Computational intelligence in healthcare, pp. 105-160. Springer, Cham, 2021.

[3] Wen, Bo, Bo Tao, and Gongfa Li. "Research status and tendency of intelligent industrial robot." In Journal of Physics: Conference Series, vol. 1302, no. 3, p. 032050. IOP Publishing, 2019.

[4] Krizhevsky, Alex, IlyaSutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems 25 (2012).

[5] Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, SanjeevSatheesh, Sean Ma, Zhiheng Huang et al. "Imagenet large scale visual recognition challenge." International journal of computer vision 115, no. 3 (2015): 211-252.

[6] Samadzadegan, Farhad, FarzanehDadrassJavan, FarnazAshtariMahini, and MehrnazGholamshahi. "Detection and Recognition of Drones Based on a Deep Convolutional Neural Network Using Visible Imagery." Aerospace 9, no. 1 (2022): 31.

[7] Yu, Xiangchun, Zhe Zhang, Lei Wu, Wei Pang, Hechang Chen, Zhezhou Yu, and Bin Li. "Deep ensemble learning for human action recognition in still images." Complexity 2020 (2020).

[8] Lu, Xin, Quanquan Li, Buyu Li, and Junjie Yan. "Mimicdet: Bridging the gap between one-stage and two-stage object detection." In European Conference on Computer Vision, pp. 541-557. Springer, Cham, 2020.

[9] Burić, Matija, MiranPobar, and Marina Ivašić-Kos. "Adapting YOLO network for ball and player detection." In Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods, vol. 1, pp. 845-851. 2019.

[10] Sermanet, Pierre, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and YannLeCun. "Overfeat: Integrated recognition, localization and detection using convolutional networks." arXiv preprint arXiv:1312.6229 (2013).

[11] Redmon, Joseph, SantoshDivvala, Ross Girshick, and Ali Farhadi. "You only look once: Unified, real-time object detection." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 779-788. 2016.

[12] Wu, Jiatu. "Complexity and accuracy analysis of common artificial neural networks on pedestrian detection." In MATEC Web of Conferences, vol. 232, p. 01003. EDP Sciences, 2018.

[13] Shafiee, Mohammad Javad, Brendan Chywl, Francis Li, and Alexander Wong. "Fast YOLO: A fast you only look once system for real-time embedded object detection in video." arXiv preprint arXiv:1709.05943 (2017).

[14] Bajestani, Mohammad Farhadi, and Yezhou Yang. "Tkd: Temporal knowledge distillation for active perception." In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 953-962. 2020.

[15] Gupta, Akshat, Milan Desai, Wusheng Liang, and MageshKannan. "Spatiotemporal action recognition in restaurant videos." arXiv preprint arXiv:2008.11149 (2020).

[16] Redmon, Joseph, and Ali Farhadi. "YOLO9000: better, faster, stronger." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7263-7271. 2017.

[17] Liu, Yuanzhong, ZhigangTu, Liyu Lin, Xing Xie, and Qianqing Qin. "Real-time spatio-temporal action localization via learning motion representation." In Proceedings of the Asian Conference on Computer Vision. 2020.

[18] Ioffe, Sergey, and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." In International conference on machine learning, pp. 448-456. PMLR, 2015.

[19] Takahashi, Shinobu, and Kazuhiko Kawamoto. "Object–Action Interaction Region Detection in Egocentric Videos."

[20] Ren, Shaoqing, Kaiming He, Ross Girshick, and Jian Sun. "Faster r-cnn: Towards real-time object detection with region proposal networks." Advances in neural information processing systems 28 (2015).

[21] Redmon, Joseph, and Ali Farhadi. "Yolov3: An incremental improvement." arXiv preprint arXiv:1804.02767 (2018).

[22] He, Kaiming, Xiangyu Zhang, ShaoqingRen, and Jian Sun. "Deep residual learning for image recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778. 2016.

[23] Liu, Guoxu, Joseph Christian Nouaze, Philippe LyonelToukoMbouembe, and Jae Ho Kim. "YOLO-tomato: A robust algorithm for tomato detection based on YOLOv3." Sensors 20, no. 7 (2020): 2145.

[24] Safdar, Muhammad Farhan, ShaymaSaadAlkobaisi, and Fatima Tuz Zahra. "A comparative analysis of data augmentation approaches for magnetic resonance imaging (MRI) scan images of brain tumor." Actainformaticamedica 28, no. 1 (2020): 29.

[25] Bochkovskiy, Alexey, Chien-Yao Wang, and Hong-YuanMark Liao. "Yolov4: Optimal speed and accuracy of object detection." *arXiv preprint arXiv:2004.10934* (2020).

[26] Chung, Chia-Ling, Ding-Bang Chen, and HoomanSamani. "Action detection and anomaly analysis visual system using deep learning for robots in pandemic situation." In 2020 International Automatic Control Conference (CACS), pp. 1-6. IEEE, 2020.

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

[28] Fang, Yiming, XianxinGuo, Kun Chen, Zhu Zhou, and Qing Ye. "Accurate and Automated Detection of Surface Knots on Sawn Timbers Using YOLO-V5 Model." BioResources 16, no. 3 (2021).

[29] Choiński, Mateusz, Mateusz Rogowski, Piotr Tynecki, Dries PJ Kuijper, Marcin Churski, and Jakub W. Bubnicki. "A first step towards automated species recognition from camera trap images of mammals using AI in a European temperate forest." In International Conference on Computer Information Systems and Industrial Management, pp. 299-310. Springer, Cham, 2021.

[30] Liu, Yifan, BingHang Lu, JingyuPeng, and Zihao Zhang. "Research on the use of YOLOv5 object detection algorithm in mask wearing recognition." World Scientific Research Journal 6, no. 11 (2020): 276-284.

[31] ultralytics. yolov5. Available online: https://github.com/ultralytics/yolov5 (accessed on 18 May 2020)

[32] Wang, Xiqi, Shunyi Zheng, Ce Zhang, Rui Li, and Li Gui. "R-YOLO: A real-time text detector for natural scenes with arbitrary rotation" Sensors 21, no. 3 (2021): 888.

[33] Ge, Zheng, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. "Yolox: Exceeding yolo series in 2021." arXiv preprint arXiv:2107.08430 (2021).

[34] Wang, Haoyue, Wei Wang, and Yao Liu. "X-YOLO: A deep learning based toolset with multiple optimization strategies for contraband detection." In Proceedings of the ACM Turing Celebration Conference-China, pp. 127-132. 2020.

[35] Liu, Wei, DragomirAnguelov, DumitruErhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. "Ssd: Single shot multibox detector." In European conference on computer vision, pp. 21-37. Springer, Cham, 2016.

[36] Abebe, Assefa Addis, WenhongTian, and Kingsley NketiaAcheampong. "Extended Single Shoot Multibox Detector for Traffic Signs Detection and Recognition in Real-time." In 2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), pp. 373-379. IEEE, 2020.

[37] Liu, Wei, DragomirAnguelov, DumitruErhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. "Ssd: Single shot multibox detector." In European conference on computer vision, pp. 21-37. Springer, Cham, 2016.

[38] Gong, Meimei, and YimingShu. "Real-time detection and motion recognition of human moving objects based on deep learning and multi-scale feature fusion in video." IEEE Access 8 (2020): 25811-25822.

[39] Bharati, Puja, and AnkitaPramanik. "Deep learning techniques—R-CNN to mask R-CNN: a survey." Computational Intelligence in Pattern Recognition (2020): 657-668.

[40] Everingham, Mark, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. "The pascal visual object classes (voc) challenge." International journal of computer vision 88, no. 2 (2010): 303-338.

[41] Fan, Quanfu, Lisa Brown, and John Smith. "A closer look at Faster R-CNN for vehicle detection." In 2016 IEEE intelligent vehicles symposium (IV), pp. 124-129. IEEE, 2016.

[42] Zhou, Shuren, and JiaQiu. "Enhanced SSD with interactive multi-scale attention features for object detection." Multimedia Tools and Applications 80, no. 8 (2021): 11539-11556.

[43] Wenzel, Florian, Kevin Roth, Bastiaan S. Veeling, Jakub Świątkowski, Linh Tran, Stephan Mandt, Jasper Snoek, Tim Salimans, RodolpheJenatton, and Sebastian Nowozin. "How good is the bayes posterior in deep neural networks really?." arXiv preprint arXiv:2002.02405 (2020).

[44] Kong, Tao, Anbang Yao, Yurong Chen, and Fuchun Sun. "Hypernet: Towards accurate region proposal generation and joint object detection." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 845-853. 2016.

[45] Jiang, Huaizu, and Erik Learned-Miller. "Face detection with the faster R-CNN." In 2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017), pp. 650-657. IEEE, 2017.

[46] Li, Kun, Wei Zhang, Dian Yu, and Xin Tian. "HyperNet: A deep network for hyperspectral, multispectral, and panchromatic image fusion." ISPRS Journal of Photogrammetry and Remote Sensing 188 (2022): 30-44.

[47] Zhai, Junhai, and Dandan Song. "Optimal instance subset selection from big data using genetic algorithm and open source framework." Journal of Big Data 9, no. 1 (2022): 1-18.

[48] He, Kaiming, Xiangyu Zhang, ShaoqingRen, and Jian Sun. "Spatial pyramid pooling in deep convolutional networks for visual recognition." IEEE transactions on pattern analysis and machine intelligence 37, no. 9 (2015): 1904-1916.

[49] Nandimandalam, VenkataDevaraju. "Military and Non-Military Vehicle Detection by Faster R-CNN and SSD300 Models using Transfer Leaning." PhD diss., Dublin, National College of Ireland, 2020.

[50] Dai, Jifeng, Yi Li, Kaiming He, and Jian Sun. "R-fcn: Object detection via region-based fully convolutional networks." Advances in neural information processing systems 29 (2016).

[51] Avramidis, Kleanthis, AgelosKratimenos, Christos Garoufis, AthanasiaZlatintsi, and Petros Maragos. "Deep Convolutional and Recurrent Networks for Polyphonic Instrument Classification from Monophonic Raw Audio Waveforms." In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3010-3014. IEEE, 2021.

[52] Shen, Zhiqiang, Zhuang Liu, Jianguo Li, Yu-Gang Jiang, Yurong Chen, and XiangyangXue. "Dsod: Learning deeply supervised object detectors from scratch." In Proceedings of the IEEE international conference on computer vision, pp. 1919-1927. 2017.

[53] Jordao, Artur, Ricardo Kloss, and William Robson Schwartz. "Latent HyperNet: exploring the layers of convolutional neural networks." In 2018 International Joint Conference on Neural Networks (IJCNN), pp. 1-7. IEEE, 2018.

[54] Zhang, Yuting, KihyukSohn, Ruben Villegas, Gang Pan, and Honglak Lee. "Improving object detection with deep convolutional networks via bayesian optimization and structured prediction." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 249-258. 2015.

[55] Tang, Jiwen, Damien Arvor, Thomas Corpetti, and Ping Tang. "Pvanet-Hough: Detection and location of center pivot irrigation systems from Sentinel-2 images." ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences 3 (2020): 559-564.

[56] Du, Juan. "Understanding of object detection based on CNN family and YOLO." In Journal of Physics: Conference Series, vol. 1004, no. 1, p. 012029. IOP Publishing, 2018.

[57] Boudjit, Kamel, and NaeemRamzan. "Human detection based on deep learning YOLO-v2 for real-time UAV applications." Journal of Experimental & Theoretical Artificial Intelligence 34, no. 3 (2022): 527-544.

[58] Luo, Yihao, Xiang Cao, Juntao Zhang, JingjuanGuo, Haibo Shen, Tianjiang Wang, and Qi Feng. "CE-FPN: enhancing channel information for object detection." Multimedia Tools and Applications (2022): 1-20.

[59] Wu, Minghu, Hanhui Yue, Juan Wang, Yongxi Huang, Min Liu, Yuhan Jiang, Cong Ke, and Cheng Zeng. "Object detection based on RGC mask R□CNN." IET Image Processing 14, no. 8 (2020): 1502-1508.

[60] Girshick, Ross. "Fast r-cnn." In Proceedings of the IEEE international conference on computer vision, pp. 1440-1448. 2015.

[61] Shen, Zhiqiang, Zhuang Liu, Jianguo Li, Yu-Gang Jiang, Yurong Chen, and XiangyangXue. "Dsod: Learning deeply supervised object detectors from scratch." In Proceedings of the IEEE international conference on computer vision, pp. 1919-1927. 2017.

[62] Brahmbhatt, Samarth, Henrik I. Christensen, and James Hays. "StuffNet: Using 'Stuff'toimprove object detection." In 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 934-943. IEEE, 2017.

[63] Xiang, Yu, Wongun Choi, Yuanqing Lin, and Silvio Savarese. "Subcategory-aware convolutional neural networks for object proposals and detection." In 2017 IEEE winter conference on applications of computer vision (WACV), pp. 924-933. IEEE, 2017.

[64] Dai, Jifeng, Yi Li, Kaiming He, and Jian Sun. "R-fcn: Object detection via region-based fully convolutional networks." Advances in neural information processing systems 29 (2016).

[65] Zhang, Yunfeng, and Mingmin Chi. "Mask-R-FCN: A deep fusion network for semantic segmentation." IEEE Access 8 (2020): 155753-155765

[66] Dollar P, Wojek C, Schiele B, Perona P (2009) Pedestrian detection: a benchmark. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 304–311

[67] Chen C, Seff A, Kornhauser A, Xiao J (2015) Deepdriving: learning affordance for direct perception in autonomous driving. In: Proceedings of the IEEE international conference on computer vision (CVPR), pp 2722–2730.

[68] Chen X, Ma H, Wan J, Li B, Xia T (2017) Multi-view 3d object detection network for autonomous driving. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 6526–6534.

[69] Wang WH, Yang J, Xiao JW, Li S, Zhou DX (2015) Face recognition based on deep learning. In: International conference on human-centered computing (HCC), pp 812–820.

[70] Borji A, Cheng MM, Hou Q, Jiang H, Li J (2014) Salient object detection: a survey. Computational Visual Media, pp 1–34.

[71] Ren Y, Zhu C, Xiao S (2018) Small object detection in optical remote sensing images via modified faster R-CNN. ApplSci 8(5):813.

[72] Dollar P, Wojek C, Schiele B, Perona P (2009) Pedestrian detection: a benchmark. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 304–311.