EasyChair Preprint

№ 9011

A Review on Real-Time Object Detection Models Using Deep Neural Networks

Lenard Byenkya Nkalubo and Rose Nakibuule

October 6, 2022

A Review on Real-Time Object Detection Models using Deep Neural Networks

Nkalubo Lenard Byenkya[1] and Nakibuule Rose[2]

[1] Kyambogo University, Department of Networks, Data Science and artificial Intelligence, Uganda, lnkalubo@kyu.ac.ug,
[2] Makerere University, Department of Computer Science, Uganda, rnakibuule@cis.mak.ac.ug

Abstract
Object detection is one of the most well-known challenges in computer vision. Many researchers have been employed in numerous application fields, including robotics, autonomous driving, and video surveillance. This paper offers a review of real-time object detection techniques using deep learning approaches. It is aimed at familiarizing the readers with the relevant knowledge, literature, and the latest updates on the state-of-art techniques. This study review records obtained electronically through the leading scientific databases (IEEE, Google Scholar, Scopus, Science Direct, Elsevier, and other journal publications) searched using three sets of keywords: Deep learning, object detection, and convolutional neural networks. Two different categories can be found in the object detection framework, traditional detectors, and deep learning-based detectors. The deep learning object detectors are divided into the two-stage detector and the one-stage detector. One-stage detectors use dense anchor boxes to perform classification and regression without establishing a sparse region of interest collection, while in two-stage detectors, sparse region proposals are created in the first stage of two-stage detectors, after which they are regressed and categorized. Object detection has been applied in crop harvesting, object detector models for blind persons, detection of pedestrians on the road, traffic sign detection and classification, text detection, and remote sensing target detection. In our future work, we propose to develop a one-stage object detection model that may help in guiding blind movements.

Keywords
Deep learning, object detection, convolutional neural networks

1.0 Introduction

Artificial neural networks are the foundation of deep learning techniques (ANNs) (Lecun et al., 2015). When a deep learning-based technique earned an overwhelming victory in a computer vision competition in 2012, it became popular. Deep-learning approaches have improved their accuracy in large-scale visual identification challenges since 2010, and by 2015, they had surpassed human accuracy (Lecun et al., 2015). Traditional feature extraction approaches require human interaction, but deep learning learns image data directly (Ahishakiye et al., 2021). Deep convolutional neural nets (CNNs) have attained state-of-the-art achievements in terms of object detection accuracy and detection speed, thanks to the advancement of deep learning technology in machine vision applications. CNN's key benefit is its capacity to self-learn and extract information from an input image automatically (Junos et al., 2021).

Computer vision is a subfield of computer science that gives computers the ability to perceive, recognize, and analyze things in still images and videos. Numerous computer vision applications, including face detection, face identification, pedestrian counting, security systems, vehicle detection, self-driving automobiles, etc., have been utilized. Object detection processing is connected to some computer vision concepts such as object localization, categorization, and recognition (Kaur & Singh, 2022). In discriminative tasks, Deep Learning models have achieved amazing progress. Deep network designs, sophisticated processing, and access to massive data have all aided this. Because of the development of convolutional neural networks (CNNs), deep neural networks have been effectively applied to Computer Vision applications such as image classification, object identification, and image segmentation (Shorten & Khoshgoftaar, 2019).

Object detection is one of the most well-known challenges in computer vision (Krizhevsky et al., 2012). Handcrafted features and shallow trainable structures are at the heart of traditional object identification systems. However, their performance is easily stagnated by developing complicated ensembles that mix several low-level image features with high-level information from object detectors and scene classifiers. With the rapid advancement of deep learning, more powerful tools that can learn semantic, high-level, and deeper features are being offered to address the issues that traditional architectures have (Zhao et al., 2019). Object detection is defined as the process of determining where objects exist in a given image (object localization) and to which category each object belongs (object classification) (Zhao et al., 2019). We will use the term object recognition to refer to both image classification (a task that requires

an algorithm to determine which object classes are present in an image) and object detection in this study (a task requiring an algorithm to localize all objects present in the image) (Russakovsky et al., 2015). General-purpose object recognition ought to be quick, precise, and capable of identifying a variety of objects. Frameworks for detection have gotten faster and more precise ever since the development of neural networks. Nevertheless, the majority of detection techniques are still limited to a small number of objects (Redmon & Farhadi, 2017). Due to its success in applications such as language processing, object identification, and picture classification, deep learning has emerged as the most discussed technology (Srivastava et al., 2021). A review and history of deep learning and its applications in object detection was one in the study (Zhao et al., 2019).

1.2 Research Objectives and Outline

Motivated by the current developments and many influential studies in the field of real-time object detection models, this study proposes a survey of the studies that have been done on the existing object detection models and their applications, particularly in blind movements.

The rest of the article is organized as follows. In section 2, Materials and methods are discussed; section 3 discusses the results and section 4 discusses the conclusions and recommendations.

2. Main Text

2.1 Materials and Methods

Several review studies contend that it is very important to review articles from high-quality data sources (Xie et al., 2019), (Hsu et al., 2012), (Hwang & Tsai, 2011). During this study, an in-depth keyword-based search was conducted in the leading scientific databases such as Google Scholar, Wiley, Science Direct, Springer, IEEE, Scopus, Nature, Elsevier, and PubMed for publications on object detection approaches. Also, pertinent postgraduate Theses were included in this study.

2.2 Inclusion criteria

Studies that presented object detection approaches using deep learning methods were considered during this study. The PRISMA flow diagram and protocol (Bakator, 2018) were used the identification of the relevant research articles. This approach involves four steps which include; (i) the Identification Phase, this phase involved acquiring articles from various sources; (ii) the screening process. During this phase, article duplicates were excluded, and also inadequate articles were removed. (iii) Eligibility phase. We analyzed articles to determine their eligibility for further review. Ineligible articles were excluded. (iv) The final phase is called the included phase. Articles that were included in this study were analyzed during this phase.

2.3 Exclusion criteria

Studies that involved object detection approaches other than deep learning methods have been excluded from this study.

3.0 Discussion of Results

3.1 Object Detection Models

One of computer vision's most potent applications is object detection, whose primary goal is to identify and categorize the objects in an image (Kaur & Singh, 2022). Two different categories can be found in the object detection framework, traditional detectors, and deep learning-based detectors. The deep learning object detectors are divided into the two-stage detector and the one-stage detector. One-stage detectors (Chen et al., 2019) (Lin et al., 2020) (Tian et al., 2019) (Zhang et al., 2018) (C. Zhu et al., 2019) use dense anchor boxes to perform classification and regression without establishing a sparse region of interest (RoI) collection, while in two-stage detectors (He et al., 2020) (X. Li et al., 2019) (Lu et al., 2019) (Ren et al., 2017), Sparse region proposals are created in the first stage of two-stage detectors, after which they are regressed and categorized (Lu et al., 2020). Because of their simple structures, one-stage detectors are more efficient, yet two-stage detectors still outperform them in terms of accuracy. Despite recent efforts to improve one-stage detectors by mimicking the structural architecture of two-stage detectors, the accuracy gap persists (Lu et al., 2020). Single-stage detectors, on the other hand, approach object detection as a straightforward regression issue that takes the full image as input and generates class probabilities and multiple bounding boxes at the same time. As a result, the model is significantly faster than the two-stage object detectors (Junos et al., 2021). The accuracy offered

by two-stage detectors is sufficient, but the computation time is lengthy. One-stage detectors are therefore suggested to process in less time while managing enough accuracy (Adarsh et al., 2020). SSD and YOLO with its versions are examples of one-stage model algorithms while RCNN, Fast RCNN, and Faster RCNN algorithms are examples of two-stage detector algorithms.

The study (Lu et al., 2020) revealed that two-stage detectors have the following advantages over one-stage ones: 1) One-stage detectors directly face all of the regions on the image and have a problem of class imbalance if no specific design is added. Two-stage detectors, on the other hand, filter away most of the negative recommendations by sampling a sparse group of region proposals. 2) Because two-stage detectors analyze fewer proposals than one-stage detectors, the head of the network (used for proposal classification and regression) can be larger. This allows for the extraction of richer features. 3) Two-stage detectors use the RoIAlign operation to extract the location consistent feature from each sampled proposal, whereas one-stage detectors can allow different region proposals to share the same feature and may result in severe feature misalignment due to the coarse and spatially implicit representation of the proposals. 4) Compared to one-stage approaches, two-stage detectors perform a double regress of the object location (once on each step). One-stage detector performance is substantially hampered by the misalignment between anchor boxes and convolutional features, which is a fundamental problem that affects all one-stage detectors (Chen et al., 2019). According to the study (Lin et al., 2020), the main barrier limiting one-stage object detectors from outperforming top-performing, two-stage algorithms is class imbalance. The study proposed the focal loss as a solution, which modifies the cross-entropy loss and focuses learning on challenging negative examples. The study showed how effective it was by creating a fully convolutional one-stage detector and detailing comprehensive experimental analysis demonstrating that it achieves cutting-edge accuracy and speed. Figure 1 shows object detection techniques.

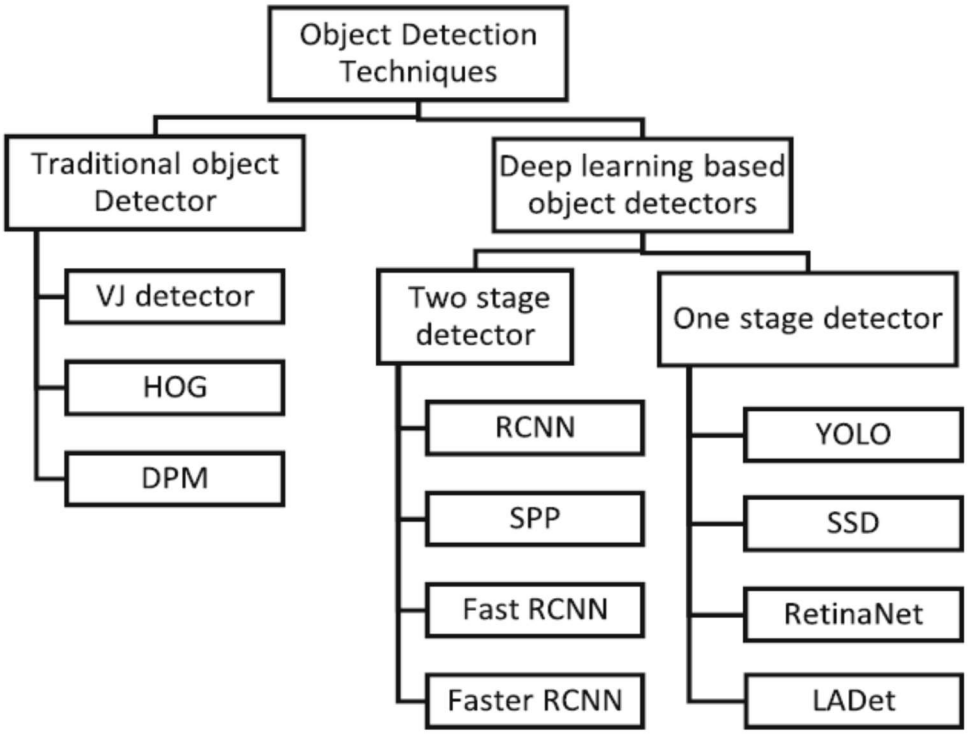

***Figure 1: Object Detection Techniques. Adapted from* (Kaur & Singh, 2022).**

3.2 Review of Object Detection Models

3.2.1 One shot detection models

a) YOLO (You Only Look Once) and its variants

YOLO (Redmon et al., 2016) is a single object detection model. The model is simple to build and can be trained on entire images directly. YOLO is trained on a loss function that directly corresponds to detection performance, unlike classifier-based techniques, and the entire model is learned together. YOLO is the world's fastest general-purpose object detector, and it pushes the boundaries of real-time object detection. YOLO is also adaptable to new domains,

making it excellent for applications that require quick and reliable object detection. When connected to a webcam, YOLO works as a tracking system, recognizing things as they move about and alter their appearance. However, because each grid cell can only predict two boxes and only one class, YOLO places strict spatial limits on bounding box predictions. The number of nearby objects that our model can predict is limited by this spatial constraint. Small items that emerge in groups, such as flocks of birds, are difficult for the model to handle. Also, the model struggles to generalize to objects with new or unusual aspect ratios or configurations since it learns to estimate bounding boxes from data. Because our design contains many downsampling layers from the input image, the model also uses rather coarse characteristics for predicting bounding boxes. Finally, while the model was trained on a loss function that approximates detection performance, our loss function considers errors in small and big bounding boxes the same. A minor error in a large box is usually unnoticed, while a minor fault in a tiny box has a far greater impact on intersection over union (IOU).

YOLOv1 (Redmon et al., 2016) divides the image into S×S grid cells with equal dimensions. If the centroid of the item falls inside a grid cell, that grid cell is responsible for object detection. With a confidence score, each cell may predict a fixed B number of bounding boxes. Five values of x, y, w, h, and confidence score make up each bounding box. A modified YOLOv1-based neural network is proposed for object detection in the study (Ahmad et al., 2020). The study proposed Yolo-LITE (R. Huang et al., 2019), a real-time object detection model designed to operate on mobile devices without a GPU, like a laptop or a smartphone (GPU). YOLOLITE, which was developed to provide a smaller, quicker, and more effective model based on the original object detection algorithm YOLOV2, increases the accessibility of real-time object identification to a variety of devices. YOLOv2 (YOLO9000) (Redmon & Farhadi, 2017), is a cutting-edge, real-time object detection system that can recognize more than 9000 different object categories. In YOLOV2, Convolution layers and batch normalization were combined to increase accuracy and lessen the overfitting issue. It allows for a seamless trade-off between speed and accuracy and can run at a variety of image sizes. The study (R. Li & Yang, 2018) proposed an enhanced YOLOv2 object detection model as a solution to the issues with the YOLOv2 object detection model's excessive number of model parameters and poor performance on small-size objects. First, it enhances the YOLOv2 by substituting the standard convolution used in the YOLOv2 with depth-wise separable convolution. The convolution layer's parameter count is down by 78.83 %. In YOLOv3 (Redmon & Farhadi, 2018), the feature extraction engine of Darknet19, which had trouble recognizing small objects, was upgraded to Darknet 53 to solve the issue. In that study, residual block, skip connections, and up-sampling were introduced, greatly enhancing the algorithm's accuracy. in YOLOv4 (Bochkovskiy et al., 2020), the feature extractors' core was modified to CSPDarknet53 once more, which greatly boosted the algorithm's efficiency and precision. YOLOv5 (X. Zhu et al., 2021) is the most recent and most efficient YOLO algorithm, it employs PyTorch rather than Darknet as its framework. Another variant of YOLO v3 is referred to as YOLO v3-Tiny (Redmon & Farhadi, 2018) since the convolutional layer's depth was reduced. Consequently, while the detection accuracy is decreased, the running speed is substantially faster (around 442% faster than the previous YOLO variations) (Adarsh et al., 2020). YOLO-P (Junos et al., 2021), is an improved YOLOv3 small network, including a lightweight backbone built on a densely linked neural network, a multi-scale detection architecture, and an optimized anchor box size. According to the experimental findings, the proposed YOLO-P model had a satisfactory mean average precision and F1 score of 98.68 % and 0.97 respectively. The study suggested PP-YOLOv2 (X. Huang et al., 2021), which outperforms other well-known detectors like YOLOv4 and YOLOv5 in terms of speed and accuracy.

b) SSD

SSD (Liu et al., 2016) is a detector with a single shot. It achieves a superb balance between results accuracy and speed. The model applies a CNN-based model to the input image just once to compute the feature map. Additionally, it learns the offset rather than identifying the box and uses anchor boxes that are similar to faster RCNNs at different aspect ratios. CNN has many layers, each of which performs processing on a different range of scales and makes use of various feature maps. Consequently, it can detect targets of different sizes. In experiments, SSD outperforms other single-stage approaches in terms of accuracy across a variety of datasets, even when input images are of a small size.

c) Comparison Between YOLO and SSD

SSD does not divide the image into random-sized grids as YOLO does. It forecasts the offset of predefined anchor boxes for each point on the feature map (default boxes). Each box has a fixed size, proportion, and location to the appropriate cell. All of the anchor boxes convolutionally cover the full feature map. The anchors of SSD and YOLO

differ slightly from one another. Because YOLO bases all of its predictions on a single grid and uses anchors that can range in size from a single grid cell to the full image. The SSD's anchors focus on different practical viewpoints and dimensional ratios of its target shapes, but not enough on target size. The anchors of YOLO are computed using k-means clustering on the training data, as opposed to the anchors of SSD, which are calculated using a straightforward algorithm. SSD does not utilize the confidence score, but YOLO determines it to demonstrate confidence in the expected outcomes. SSD does this function by utilizing a distinct background class. A low confidence score in YOLO corresponds to the SSD background class result that is expected. Both show that there is no chance the detector will ever find a target.

d) FCOS

FCOS (Tian et al., 2019) is a fully convolutional one-stage object detector that addresses object detection in a per-pixel prediction manner, similar to how semantic segmentation resolves object detection. Modern object detectors like RetinaNet, SSD, YOLOv3, and Faster R-CNN almost all depend on pre-defined anchor boxes. FCOS, in contrast, is a proposal and anchor box free. FCOS fully avoids the intricate calculations associated with anchor boxes, such as calculating overlaps during training, by doing away with the specified set of anchor boxes.

e) FSAF: Feature Selective Anchor-Free

The FSAF (C. Zhu et al., 2019) module runs more quickly and performs better than its counterparts that use anchors. The FSAF module can consistently outperform the strong baselines across a variety of backbone networks while adding the least amount of computation overhead when working in tandem with anchor-based branches. FSAF surpasses current state-of-the-art single-shot detectors and greatly improves strong baselines with minimal inference overhead.

f) RefineDet

RefineDet (Zhang et al., 2018) maintains efficiency close to one-stage methods while outperforming two-stage methods in terms of accuracy. The anchor refinement module and the object detection module are the two interconnected components that makeup RefineDet. The former specifically seeks to (1) filter out negative anchors to condense the search space for the classifier and (2) coarsely alter the sizes and placements of anchors to improve initialization for the next regressor. To further enhance performance, RefineDet needs an attention mechanism.

3.2.2 Two-shot detection models

a) Region-Based Convolutional Neural Networks (R-CNN)

R-CNN (Girshick et al., 2014) is an abbreviation for region-based convolutional neural networks. The model combines region proposals for object segmentation with powerful CNNs to detect objects. There were numerous problems with this method. The training of the CNN takes a long time because it needs to classify 2000 region proposals. Because it would take about 47 seconds to execute each test image, real-time implementation is unfeasible. Fast R-CNN (Girshick, 2015) is an object detection algorithm that addresses some of the issues with R-CNN. It takes a similar approach to its predecessor, but rather than using region proposals, CNN uses the picture to create a convolutional feature map, which is then used to select and warp region proposals from. The distorted squares are reshaped using an RoI (Region of Interest) pooling layer to a predetermined size so that a fully linked layer can accept them. The RoI vector is then used to forecast the region class with the aid of a SoftMax layer. Because it is not necessary to feed the CNN 2,000 suggestions each execution, Fast R-CNN is faster than its predecessor. Only one convolution operation is performed to produce a feature map per image.

The disadvantage of RCNN is that it uses three different models to detect targets, which increases prediction time because it processes several areas through CNN. The process employed for Fast-RCNN is long and time-consuming. As a result, computation times are still lengthy. The object region proposal is time-consuming for faster RCNN. Systems of several kinds are running sequentially. As a result, the success of the preceding operations is a prerequisite for the successful completion of the complete procedure (Adarsh et al., 2020).

b) Grid R-CNN

Grid R-CNN (Lu et al., 2019) is an innovative framework for object detection that uses a grid-guided localization method to recognize objects accurately. The Grid R-CNN captures spatial information explicitly and benefits from the

position-sensitive feature of a fully convolutional architecture, in contrast to typical regression-based approaches. By using a grid-guided technique for high-quality localization, Grid R-CNN substitutes the conventional box offset regression strategy in object detection. Extensive trials demonstrate that Grid R-CNN achieves state-of-the-art performance, notably on demanding evaluation metrics like AP at IoU=0.8 and IoU=0.9. Grid R-CNN also brings substantial and constant progress.

c) Mask R-CNN

Faster R-CNN is extended by Mask R-CNN (He et al., 2020) by adding a branch for object mask prediction in addition to the existing branch for bounding box recognition. Mask R-CNN runs at 5 frames per second, adds only a little overhead to Faster R-CNN, and is easily trainable. Mask R-CNN is also simple to generalize to other problems, enabling us, for example, to estimate human poses within the same framework.

3.3 Applications of Real-time Object Detection Models

The study (Junos et al., 2021) is an optimized YOLO-based object detection model for crop harvesting. The proposed model was also evaluated for accuracy in identifying fresh fruit bunches of different maturities, and it scored 98.91 %. The extensive experimental findings demonstrate the effectiveness of the proposed YOLO-P model in performing reliable and precise detection at the palm oil farm. The study (Kumar et al., 2019) proposed a deep learning neural network-based object detector model for blind persons to use to detect objects. This approach can be employed to identify items in webcam feeds, films, and even still images. The model's accuracy is greater than 75%. This model requires approximately 5 to 6 hours of training time. The single-shot multi-box detector (SSD) technique was utilized in the model to obtain high accuracy and IOU in real-time for object recognition for a blind person. According to the study (Kaur & Singh, 2022), object detection is crucial for seeing pedestrians on the road. Many researchers have been employed in numerous application fields, including robotics, autonomous driving, and video surveillance. The study further revealed that one of the earliest uses of computer vision is face detection and identification, which has received extensive research. Also, object detection is used in traffic sign detection and classification, text detection, and remote sensing target detection (Kaur & Singh, 2022). To identify items in the images, the study (Kumar et al., 2020) develops an object detection method employing deep learning neural networks. The study employs a multilayer convolutional network with an enhanced SSD method to detect items with excellent accuracy and speed. The study (Kumar et al., 2019) proposed a real-time object detection method for blind persons to use on any device running this model. The proposed model was developed using a convolutional neural network and a single-shot multi-box detection technique. The study (X. Zhu et al., 2021) proposed TPH-YOLOv5, which is particularly effective in object detection in drone-captured circumstances. The study (Potdar et al., 2018) proposed a neural network model for the visually impaired. People who are blind or visually impaired largely rely on their other senses, such as touch and auditory signals, to understand their surroundings.

4. Discussions and Conclusion

Computer vision is a subfield of computer science that gives computers the ability to perceive, recognize, and analyze objects in still images and videos. Numerous computer vision applications, including face detection, face identification, pedestrian counting, security systems, vehicle detection, self-driving automobiles, etc., have been utilized. Two different categories can be found in the object detection framework, traditional detectors, and deep learning-based detectors. The deep learning object detectors are divided into the two-stage detector and the one-stage detector. SSD and YOLO with their versions are examples of one-stage model algorithms while RCNN, Fast RCNN, and Faster RCNN algorithms are examples of two-stage detector algorithms. Object detection has been applied in crop harvesting, object detector models for blind persons, detection of pedestrians on the road, traffic sign detection and classification, text detection, and remote sensing target detection. In our future work, we propose to develop a one-stage object detection model that may help in guiding blind movements.

REFERENCES

Adarsh, P., Rathi, P., & Kumar, M. (2020). YOLO v3-Tiny: Object Detection and Recognition using one stage improved model. *2020 6th International Conference on Advanced Computing and Communication Systems, ICACCS 2020*, 687–694. https://doi.org/10.1109/ICACCS48705.2020.9074315

Ahishakiye, E., Van Gijzen, M. B., Tumwiine, J., Wario, R., & Obungoloch, J. (2021). A survey on deep learning in medical image reconstruction. *Intelligent Medicine*. https://doi.org/10.1016/j.imed.2021.03.003

Ahmad, T., Ma, Y., Yahya, M., Ahmad, B., Nazir, S., Haq, A. U., & Ali, R. (2020). Object Detection through Modified YOLO Neural Network. *Scientific Programming*, *2020*, 1–10. https://doi.org/10.1155/2020/8403262

Bakator, M. (2018). Deep Learning and Medical Diagnosis : A Review of Literature. *Multimodal Technologies and Interaction*. https://doi.org/10.3390/mti2030047

Bochkovskiy, A., Wang, C.-Y., & Liao, H.-Y. M. (2020). *YOLOv4: Optimal Speed and Accuracy of Object Detection*. http://arxiv.org/abs/2004.10934

Chen, Y., Han, C., Wang, N., & Zhang, Z. (2019). *Revisiting Feature Alignment for One-stage Object Detection*. 1–11. http://arxiv.org/abs/1908.01570

Girshick, R. (2015). Fast R-CNN. *Proceedings of the IEEE International Conference on Computer Vision*, *2015 Inter*, 1440–1448. https://doi.org/10.1109/ICCV.2015.169

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 580–587. https://doi.org/10.1109/CVPR.2014.81

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2020). Mask R-CNN. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *42*(2), 386–397. https://doi.org/10.1109/TPAMI.2018.2844175

Hsu, Y. C., Ho, H. N. J., Tsai, C. C., Hwang, G. J., Chu, H. C., Wang, C. Y., & Chen, N. S. (2012). Research trends in technology-based learning from 2000 to 2009: A content analysis of publications in selected journals. *Educational Technology and Society*, *15*(2), 354–370.

Huang, R., Pedoeem, J., & Chen, C. (2019). YOLO-LITE: A Real-Time Object Detection Algorithm Optimized for Non-GPU Computers. *Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018*, 2503–2510. https://doi.org/10.1109/BigData.2018.8621865

Huang, X., Wang, X., Lv, W., Bai, X., Long, X., Deng, K., Dang, Q., Han, S., Liu, Q., Hu, X., Yu, D., Ma, Y., and Yoshie, O. (2021). *PP-YOLOv2: A Practical Object Detector*. 1–7. http://arxiv.org/abs/2104.10419

Hwang, G. J., & Tsai, C. C. (2011). Research trends in mobile and ubiquitous learning: A review of publications in selected journals from 2001 to 2010. *British Journal of Educational Technology*, *42*(4), 65–70. https://doi.org/10.1111/j.1467-8535.2011.01183.x

Junos, M. H., Mohd Khairuddin, A. S., Thannirmalai, S., & Dahari, M. (2021). An optimized YOLO-based object detection model for crop harvesting system. *IET Image Processing*, *15*(9), 2112–2125. https://doi.org/10.1049/ipr2.12181

Kaur, J., & Singh, W. (2022). Tools, techniques, datasets and application areas for object detection in an image: a review. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-022-13153-y

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Neural Information Processing Systems Conference*, *4*(25), 1097–1105. https://doi.org/10.1201/9781420010749

Kumar, A., Reddy, S. S. S. S., & Kulkarni, V. (2019). An Object Detection Technique For Blind People in Real-Time Using Deep Neural Network. *Proceedings of the IEEE International Conference Image Information Processing*, *2019-Novem*, 292–297. https://doi.org/10.1109/ICIIP47207.2019.8985965

Kumar, A., Zhang, Z. J., & Lyu, H. (2020). Object detection in real-time based on an improved single shot multi-box

detector algorithm. *Eurasip Journal on Wireless Communications and Networking*, *2020*(1). https://doi.org/10.1186/s13638-020-01826-x

Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. In *Nature* (Vol. 521, Issue 7553, pp. 436–444). Nature Publishing Group. https://doi.org/10.1038/nature14539

Li, R., & Yang, J. (2018). Improved YOLOv2 Object Detection Model. *International Conference on Multimedia Computing and Systems -Proceedings*, *2018-May*, 1–6. https://doi.org/10.1109/ICMCS.2018.8525895

Li, X., Lai, T., Wang, S., Chen, Q., Yang, C., & Chen, R. (2019). Feature Pyramid Networks for Object Detection. *Proceedings - 2019 IEEE Intl Conf on Parallel and Distributed Processing with Applications, Big Data and Cloud Computing, Sustainable Computing and Communications, Social Computing and Networking, ISPA/BDCloud/SustainCom/SocialCom 2019*, 1500–1504. https://doi.org/10.1109/ISPA-BDCloud-SustainCom-SocialCom48970.2019.00217

Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollar, P. (2020). Focal Loss for Dense Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *42*(2), 318–327. https://doi.org/10.1109/TPAMI.2018.2858826

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016). SSD: Single shot multibox detector. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9905 LNCS*, 21–37. https://doi.org/10.1007/978-3-319-46448-0_2

Lu, X., Li, B., Yue, Y., Li, Q., & Yan, J. (2019). Grid R-CNN. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2019-June*, 7355–7364. https://doi.org/10.1109/CVPR.2019.00754

Lu, X., Li, Q., Li, B., & Yan, J. (2020). MimicDet: Bridging the Gap Between One-Stage and Two-Stage Object Detection. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12359 LNCS*, 541–557. https://doi.org/10.1007/978-3-030-58568-6_32

Potdar, K., Pai, C. D., & Akolkar, S. (2018). *A Convolutional Neural Network-based Live Object Recognition System as Blind Aid*. http://arxiv.org/abs/1811.10399

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2016-Decem*, 779–788. https://doi.org/10.1109/CVPR.2016.91

Redmon, J., & Farhadi, A. (2017). YOLO9000: Better, faster, stronger. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, *2017-Janua*, 6517–6525. https://doi.org/10.1109/CVPR.2017.690

Redmon, J., & Farhadi, A. (2018). *YOLOv3: An Incremental Improvement*. http://arxiv.org/abs/1804.02767

Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(6), 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, *115*(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, *6*(1). https://doi.org/10.1186/s40537-019-0197-0

Srivastava, S., Divekar, A. V., Anilkumar, C., Naik, I., Kulkarni, V., & Pattabiraman, V. (2021). Comparative analysis of deep learning image detection algorithms. *Journal of Big Data*, *8*(1). https://doi.org/10.1186/s40537-021-00434-w

Tian, Z., Shen, C., Chen, H., & He, T. (2019). FCOS: Fully convolutional one-stage object detection. *Proceedings of the IEEE International Conference on Computer Vision*, *2019-Octob*, 9626–9635.

https://doi.org/10.1109/ICCV.2019.00972

Xie, H., Chu, H. C., Hwang, G. J., & Wang, C. C. (2019). Trends and development in technology-enhanced adaptive/personalized learning: A systematic review of journal publications from 2007 to 2017. *Computers and Education*, *140*(June), 103599. https://doi.org/10.1016/j.compedu.2019.103599

Zhang, S., Wen, L., Bian, X., Lei, Z., & Li, S. Z. (2018). Single-Shot Refinement Neural Network for Object Detection. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 4203–4212. https://doi.org/10.1109/CVPR.2018.00442

Zhao, Z. Q., Zheng, P., Xu, S. T., & Wu, X. (2019). Object Detection with Deep Learning: A Review. *IEEE Transactions on Neural Networks and Learning Systems*, *30*(11), 3212–3232. https://doi.org/10.1109/TNNLS.2018.2876865

Zhu, C., He, Y., & Savvides, M. (2019). Feature selective anchor-free module for single-shot object detection. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2019-June*, 840–849. https://doi.org/10.1109/CVPR.2019.00093

Zhu, X., Lyu, S., Wang, X., & Zhao, Q. (2021). TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-captured Scenarios. *Proceedings of the IEEE International Conference on Computer Vision*, *2021-Octob*, 2778–2788. https://doi.org/10.1109/ICCVW54120.2021.00312