State of-the-Art Analysis of Multiple Object Detection Techniques using Deep Learning

Kanhaiya Sharma[1], Sandeep Singh Rawat[2], Deepak Parashar[3], Shivam Sharma[4], Shubhangi Roy[5], and Shibani Sahoo[6]
Symbiosis Institute of Technology Pune, Symbiosis International (Deemed University), Pune, India[1, 3, 4, 5, 6]
School of Computer and Information Sciences, IGNOU, New Delhi, India[2]

***Abstract*—Object detection has experienced a surge in interest due to its relevance in video analysis and image interpretation. Traditional object detection approaches relied on handcrafted features and shallow trainable algorithms, which limited their performance. However, the advancement of Deep learning (DL) has provided more powerful tools that can extract semantic, high-level, and deep features, addressing the shortcomings of previous systems. Deep Learning-based object detection models differ regarding network architecture, training techniques, and optimization functions. In this study, common generic designs for object detection and various modifications and tips to enhance detection performance have been investigated. Furthermore, future directions in object detection research, including advancements in Neural Network-based learning systems and the challenges have been discussed. In addition, comparative analysis based on performance parameters of various versions of YOLO approach for multiple object detection has been presented.**

Keywords*—Deep learning; neural networks; object detection; YOLO

I. INTRODUCTION

Object detection involves the process of identifying the location of objects within an image (object localization) and assigning each object to its corresponding class (object classification) [1]. Commonly utilized techniques for object detection include frame difference, background subtraction, optical flow, and Hough transform [2], but they have limitations regarding accurate object detection. On the other hand, object recognition focuses on determining the presence of a specific object in visual data and often involves feature extraction [3].

Nowadays, object detection is applied in various fields such as face detection [4][5], mask detection for COVID-19 compliance [6], railway signal detection [7], and multiple object tracking for counting purposes. Many object detection methods are being developed from many years. Researchers are trying to come up with new methods which will be stable and give accurate results irrespective of the data size. The emergence of popular algorithms for object detection included R-CNN, Fast R-CNN, and Faster R-CNN. R-CNN was initially slow due to its inability to share processing, requiring a ConvNet forward pass for each proposed object [8]. To address this, spatial pyramid pooling networks (SPPNets) were introduced to accelerate R-CNN by enabling computation sharing. Subsequently, Fast R-CNN was developed, training the deep VGG16 network nine times faster than R-CNN, achieving a significantly faster testing speed (213 times faster), and higher Mean Average Precision (MAP) [9] applications requiring fast and accurate object detection [10-13]. YOLO9000, an extension of YOLO, employs joint optimization of detection and classification to identify over 9000 object types in real time. This approach combines data from diverse sources such as ImageNet [14], [15] and COCO [16] using joint optimization techniques and Word Tree [17]. YOLO9000 significantly bridges the dataset size gap between detection and classification. [18]. YOLOv3, a combination of Darknet-19 and residual network technology, features 53 convolutional layers known as Darknet-53. YOLOv3 performs comparable to SSD variations regarding COCO's average mean average precision measure but is three times faster [19].

Real-time object detection functions enable widespread and cost-effective utilization of standard Graphics Processing Units (GPUs). While the most accurate neural networks currently available cannot operate in real-time and require multiple GPUs for training, YOLOv4 addresses these challenges. YOLOv4 is designed to run efficiently on a standard GPU in production systems, optimizing parallel calculations rather than relying solely on low computation volume theoretical indicators [20], [21]. This paper comprehensively reviews various object detection models and their evolutionary advancements. Due to the availability of lots of object detection algorithms, the question arises which is the better and most suitable for handling complex data and giving high accuracy.

The main objective of this study is to presents a detailed analysis of multiple object detection techniques using deep learning. The study is organized as follows: Section I provides introduction; Section II describes the related work. Experimental work is described in Section III. The results are analyzed and discussed in Section IV. Section V presents the conclusion of the article.

II. RELATED WORK

In this section, the existing work based on multiple object detection techniques using deep learning has been reviewed [28]. On making a comparative study of YOLOv5 models' performance based on [31], [33], it has been observed that using YOLOv5 for object detection has gained significant popularity across diverse applications. In a prior study by P. Mishra et al., YOLOv5 was successfully employed to identify objects for agricultural monitoring [34]. The research showcased the proficiency of YOLOv5 in accurately detecting various elements such as crops, weeds, and other objects on large-scale agricultural landscapes. Another study by S. Gupta et al. utilized YOLOv5 for real-time object detection in surveillance videos, showing its efficiency and accuracy in

detecting multiple object classes [35]. These studies highlight the successful application of YOLOv5 in different domains even in detecting the more minor objects findings in the automobiles for which the YOLO5 [36] developed series achieved a very good accuracy, indicating its versatility and performance. The network architecture and overview of YOLOv5 is shown in Fig. 1 and 2, respectively.

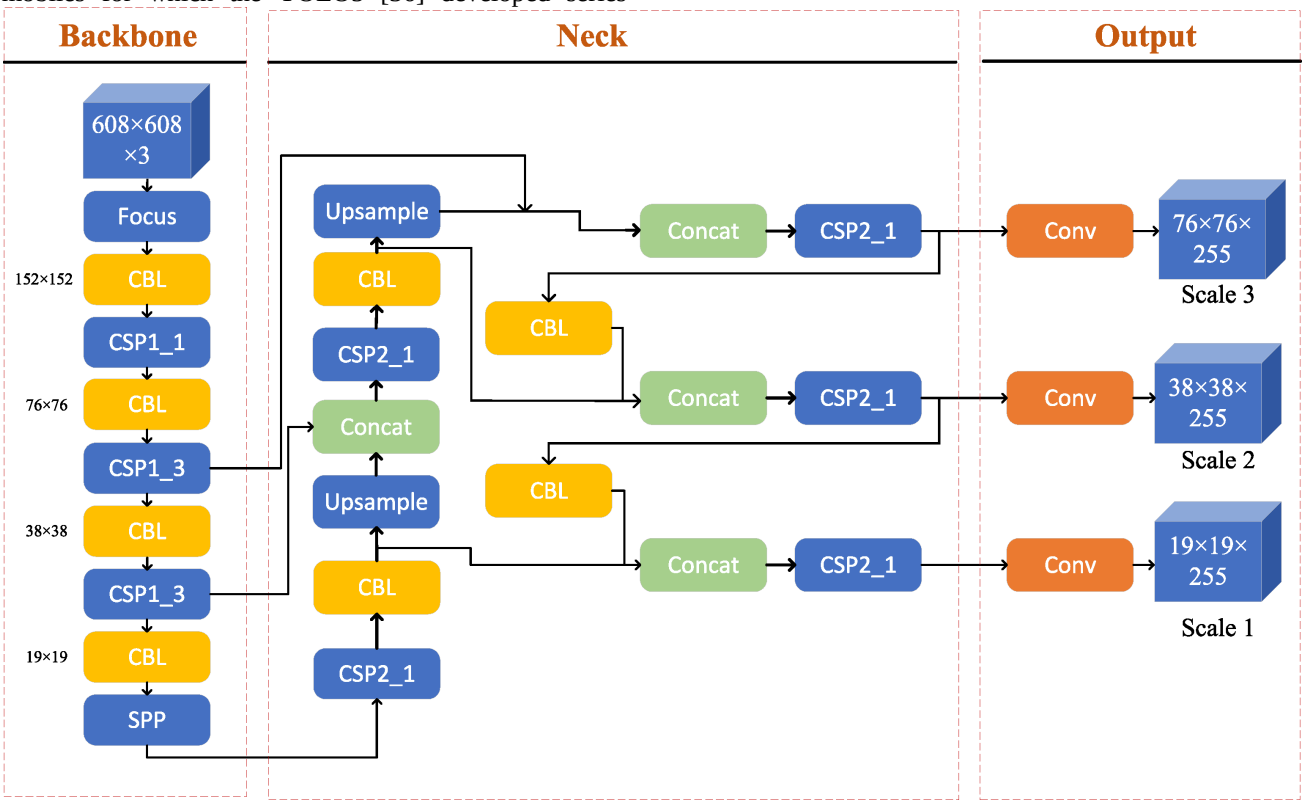

Fig. 1. The network architecture of YOLOv5. [25].

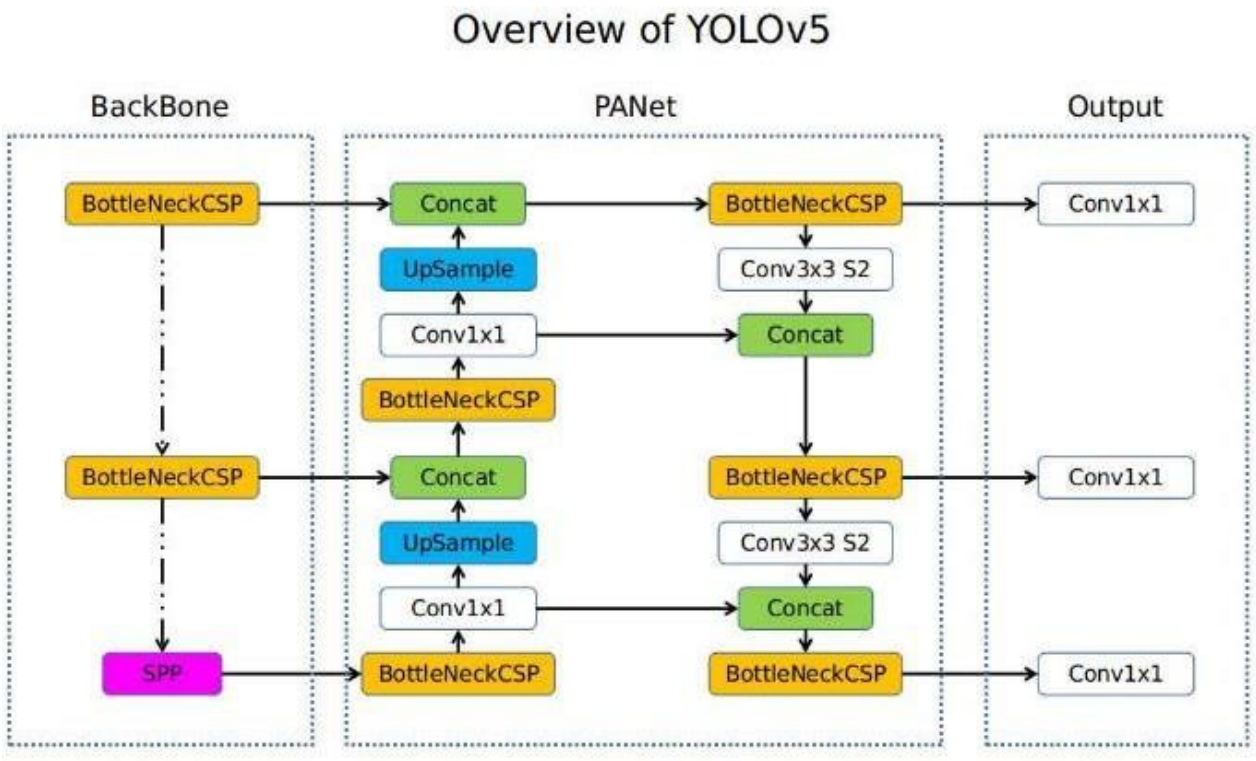

Fig. 2. The overview of YOLOv5. [26].

Data Annotation and Model Training with Roboflow, has been extensively used for data annotation and model training in various computer vision projects. In a study by L. Chen et al., Roboflow was employed to annotate images and train a YOLOv5 model for traffic sign detection [37]. The researchers leveraged the capabilities of Roboflow to generate accurate bounding box annotations, resulting in an exact and efficient traffic sign detection system. Additionally, a study by M. Rodriguez et al. utilized Roboflow for annotating medical images and training a YOLOv5-based model to detect abnormalities in lung X-rays [38]. These examples demonstrate the successful integration of Roboflow in diverse applications, emphasizing its role in facilitating data annotation and model training pipelines and in medical one of its standard applications, is detecting the Lung nodule using YOLO5 [39].

Object Detection in Unconstrained Environments: Object detection in unconstrained environments has been a challenging task. Previous work has leveraged YOLOv5 and Roboflow to address this challenge. In a study, YOLOv5 was used for Outdoor Navigation System for Visually Impaired People [40], [41]. Combining YOLOv5's real-time detection capabilities and Roboflow's efficient data annotation and model training workflow enabled accurate and timely object detection in dynamic environments. This work demonstrates the potential of using YOLOv5 and Roboflow in real-world scenarios with complex backgrounds and varying lighting conditions.

Object Detection for Robotics Applications: YOLOv5 and Roboflow have also found applications in robotics. In a research project by A. Kumar et al., YOLOv5 and Roboflow were employed for object detection in an autonomous drone system [42]. The combination of YOLOv5's fast inference speed and Roboflow's annotation capabilities allowed the drone to detect and track objects in real time, enabling autonomous navigation and interaction with the environment. This study showcases the integration of YOLOv5 and Roboflow in robotics applications, highlighting their potential for enhancing situational awareness and decision-making capabilities. These examples demonstrate the successful utilization of YOLOv5 and Roboflow in various domains, including aerial monitoring, surveillance, unconstrained environments, and robotics. The combination of YOLOv5's real-time object detection capabilities and Roboflow's annotation and model training platform has proven effective in achieving accurate and efficient object detection systems. With the improvement of the YOLOv5 framework, YOLOv6 has been developed for which a customized quantization method is introduced. The latest version of YOLOv6s demonstrates improved mean Average Precision (MAP) compared to all previous iterations of YOLOv5. Additionally, it achieves approximately twice the inference speed [43]. Roboflow for annotation and data management has streamlined the preprocessing stage, ensuring the availability of adequately labeled training data for training object detection system. This has significantly contributed to the development process by expediting the annotation process and allowing more focus on the algorithmic aspects of the system. Furthermore, keeping up- to-date with the latest research papers in the field of object detection, including YOLOv7 [44], [55].

Deep Learning-based Object Detection: Deep learning has revolutionized the field of computer vision, enabling highly accurate object detection. The seminal work by R. Girshick et al. introduced the R-CNN (Region-based Convolutional Neural Networks) framework [45], laying the foundation for subsequent advancements. Numerous variants, such as Fast R-CNN [46], Faster R-CNN [47], and Mask R-CNN [48], have been proposed to improve detection accuracy and processing speed. These methods have significantly influenced the development of stick-based object detection system. Single-Shot Object Detection: Single-shot object detection algorithms have gained popularity due to their real-time processing capabilities. Among them, YOLO family of models [49] has achieved remarkable performance. YOLO models detect objects in a single pass through the Neural Network, making them suitable for resource-constrained environments. It can be observed from literature that YOLO effectively used for object detection.

Mobile object detection aims to enable object detection on mobile devices with limited computational resources. MobileNet [50], is a lightweight deep neural network architecture specifically designed for mobile applications. Its efficient design and small memory footprint make it ideal for real-time object detection on low-power devices. The concepts underlying MobileNet have influenced the development of stick-based object detection system. Contextual object detection methods utilize contextual information to improve detection accuracy. Context R-CNN [51] incorporates context reasoning into the detection pipeline, leveraging the relationship between objects and their surrounding context. Stick-based object detection system also considers contextual cues to enhance object identification and classification.

Focal Loss for Dense Object Detection: They claim that the main barrier stopping one-stage object detectors from outperforming top-performing, two-stage techniques, including Faster R-CNN versions, is class imbalance. They developed the focused loss, which adds a modulating term to the cross-entropy loss, to focus learning on challenging examples and de-weight the many obvious negatives [52] and PointRCNN [53]. These methods leverage multi-scale feature MAPs and anchor-based strategies to improve detection performance in challenging scenarios. Stick-based object detection system integrates similar strategies to handle unconstrained environments effectively.

Sensor-Based Object Detection: Object detection approaches based on sensors use data captured by diverse sensors, including LiDAR, radar, and cameras, to identify and track objects. LiDAR-based methods, such as Point RCNN [54] and PIXOR [55], leverage 3D point cloud data for accurate object localization. Although primarily based on visual information, stick-based object detection systems can benefit from incorporating sensor fusion techniques to enhance detection accuracy.

III. METHODOLOGY

The methodology below explains the object detection process using the base model as YOLOv5. Fig. 3 depicted the block diagram of object detection procedure.

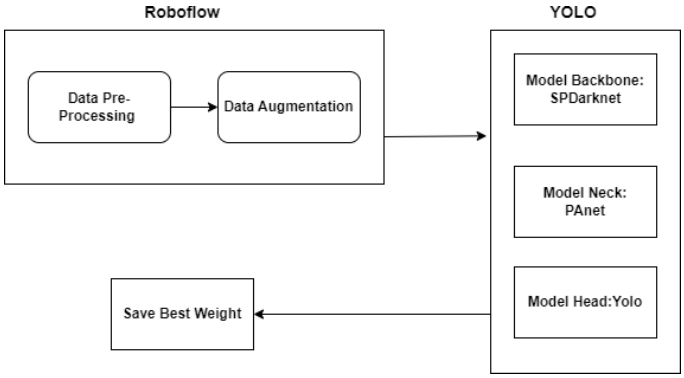

Fig. 3. Block diagram of object detection system.

A. Preprocessing

Roboflow has been used for Data preprocessing and data augmentation.

- It provides a platform for better data collection, pre-processing, and model training techniques.
- It can handle a variety of different annotation formats.
- The pre-processing of data includes resizing, image orientations, contrasting, and data augmentations.
- There are choices for model deployment and visualization, spanning the whole state-of-the-art.

1) Noise: Adding noise to a hazy photograph might help it to stand out. The picture seems to be made up of white and black dots when "salt and pepper noise" is applied.

2) Crop: An area of the image is picked, cropped, and resized to its original size.

3) Flip: Horizontally and vertically; the picture is flipped: The pixels are rearranged while the picture's characteristics are preserved when the image is flipped.

4) Rotation: The picture is rotated by a degree ranging from 0 to 360 degrees. Each rotated image will be different in the model. Each rotated image will be different in the model. The picture's brightness changes, resulting in a darker or brighter image. This technique allows the model to recognize photos in various illumination situations.

5) Blur: The image quality will vary because photographs come from various sources. Some images will be of outstanding quality, while others will undoubtedly be of terrible quality. In these circumstances, blur the original photographs, making model more resistant to the image quality used in the test data.

6) Shear: Shearing is rotating an image along a central axis to add or remove discriminating points. Typically, it is used to magnify images so that computers may understand how different viewpoints affect how things are perceived.

7) Bounding boxes: Bounding boxes are rectangles defining the boundaries of photograph items. Bounding boxes can be annotated in a variety of ways. The bounding box coordinates are represented differently in each format [22].

8) Exposure: Exposure refers to the quantity of light that reaches your camera's sensor over some time, resulting in visual data. It could be a second or several hours.

9) Model Utilized: Three main vital parts of YOLOv5 are as follows:

a) Model Backbone: Model Backbone's primary goal is to extract essential features from an image. In YOLOv5, the CSPNet [23] backbone is used to extract a wealth of valuable characteristics from an input image.

b) Model Neck: Model Neck is preferable while developing feature strategies. Models can generalize their object marking and scaling using feature extractions. Recognizing the same object in various scales, marks, and shapes is helpful. The neck in YOLOv5 uses PANet to build feature pyramids.

c) Model Head: In YOLO, the detection process incorporates the Head component of the model, responsible for the final stage. Following the application of anchor boxes to the extracted features, the Head component further contributes to the detection process. Generated output vectors represent the final results of the detection process. The heads of the YOLOv5 models follow a similar structure to those found in the v3 and v4 versions. YOLOv4 is the superior architecture from this perspective. It's worth mentioning that YOLOv4 is trained in the Ultralytics YOLOv3 repository (rather than the Darknet), which includes most of the training changes in the YOLOv5 repository, resulting in MAP increases.

YOLOv5 has notably impacted by transitioning the Darknet research framework to the PyTorch framework. The Darknet framework, predominantly coded in C, offers meticulous control over the network's operations. Developed in the C language, Darknet grants extensive control over network activities. This low-level control is beneficial for research in several aspects. However, incorporating new research findings becomes more challenging as each addition requires custom gradient computations [24].

B. Data Augmentation Approach

While training the batch YOLOv5 use a data loader that helps add data online with each set. Data loader performs scaling, mosaic augmentation, and color space. Mosaic data augmentation, for example, mixes four photos into four random-ratio tiles. Mosaic augmentation allows the model to learn to deal with "small object problems" in which the smaller items are not detected correctly compared to more significant objects. Thus, it is an effective method for object detection identification benchmark. It's not worth experimenting with the set of augmentations to maximize performance on a specific work that is wrathful. Pre-trained models abound in YOLOv5. The trade-off between model size and inference time separates them. The received picture is first run through the YOLOv5 algorithm. The real-time snapshot in this study is partitioned into matrix grids. The image may be divided into any number of grids as the image complexity changes. After the photos have been divided, each grid holding the item undergoes classification and localization. All of the grids are given a confidence score. Depending on whether the item is spotted or not, the confidence score and the bounding box for each grid will alter. Training techniques are just as crucial as the final performance of an

object detection system while being less talked about. Data augmentation alters the base training data to expose the model to more semantic variance than the training set alone.

IV. EXPERIMENTAL WORK

The procedure for the experiment began with the collection of data, followed by the training of the YOLO network, and finally the testing of the output with test photos. The network was designed to detect 106 class labels and performed annotations [27] of all the images of data. A sample dataset of multi-object detection and category-wise objects in the available dataset are shown in Fig. 4 and 5, respectively.

In the dataset, augmentation is preformed and improved model performance which helped to increase the size and help to generalize the model. A function is lost. IOU is a popular target detection index. It is utilized to assess the positive and negative samples and, in most anchor, -based approaches to calculate the distance between the expected and actual locations.

The research paper introduces a proposed regression positioning loss, which considers multiple factors including overlapping, area, center point distance, and aspect ratio. These factors play a vital role in calculating the loss for regression positioning and are deemed crucial in the proposed approach.

1) Network output analysis: The output must be viewed as a feature MAP or a vector onto which the features are being MAPped. If N = no of bounding boxes and C = no of classes the detector can detect, these N bounding boxes detect various objects. Fig. 6 graphs are the training loss and validation loss graphs auto-generated by Roboflow software. These graphs show the change in the loss function over training epochs or iterations. The training loss graph displays the loss on the training data, while the validation loss graph shows the loss on a separate validation dataset. These graphs help monitor the model's learning progress and identify potential overfitting or under fitting. Fig. 7 shows results obtained. Table I displays the accuracy percentages obtained for various class labels. It is evident that the swivel chair achieves the highest accuracy, while the class label exhibits the lowest accuracy when compared to the other class labels.

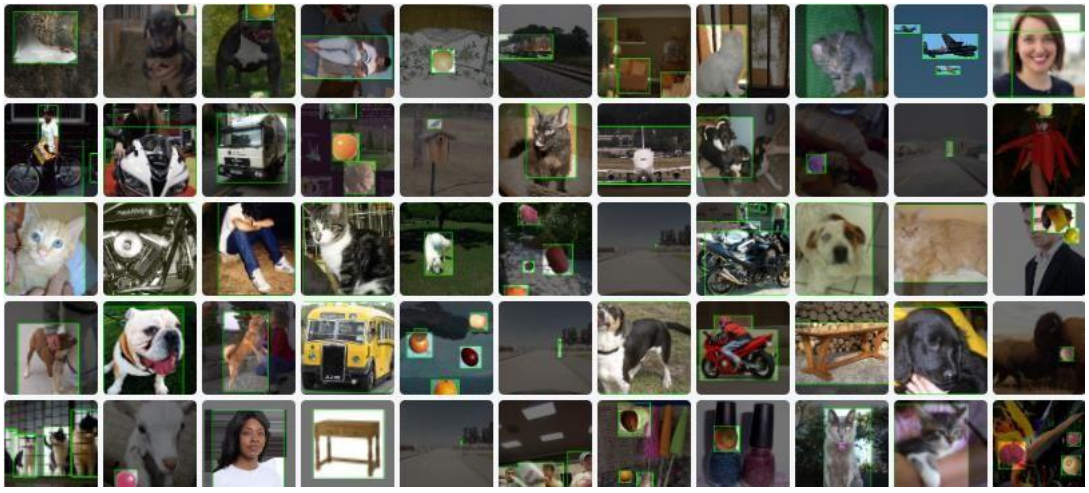

Fig. 4. Sample dataset of multi-object detection.

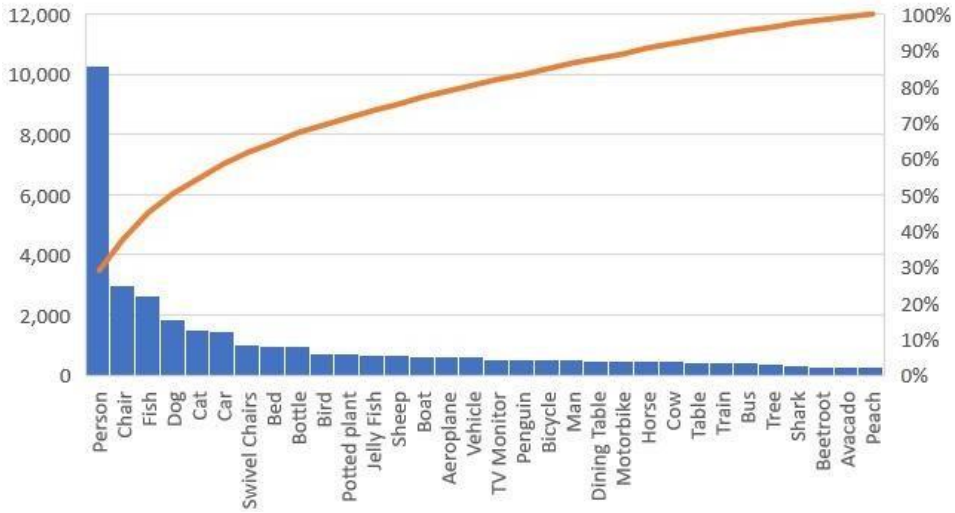

Fig. 5. Category-wise objects in the dataset.

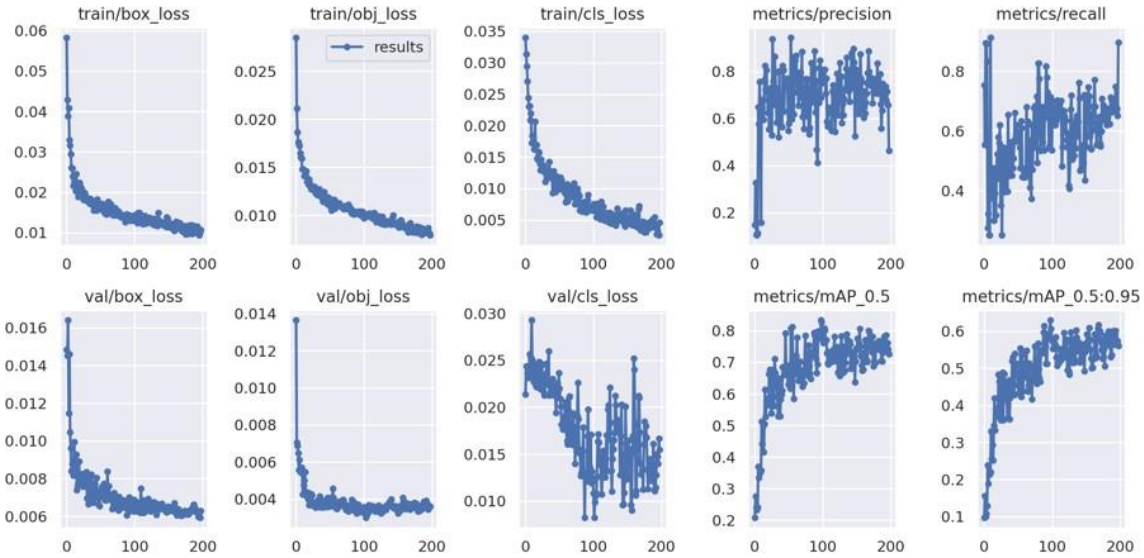

Fig. 6. Graph of performance parameter.

TABLE I. ACCURACIES OBTAINED FOR OUR DATASET CLASS LABELS

Class Labels	*Accuracies*
Swivel Chair	98%
Bed	95.3%
Cat	83.8%
Ambulance	79.2%
Man	76.5%
Dog	76.1%
Bus	74.2%

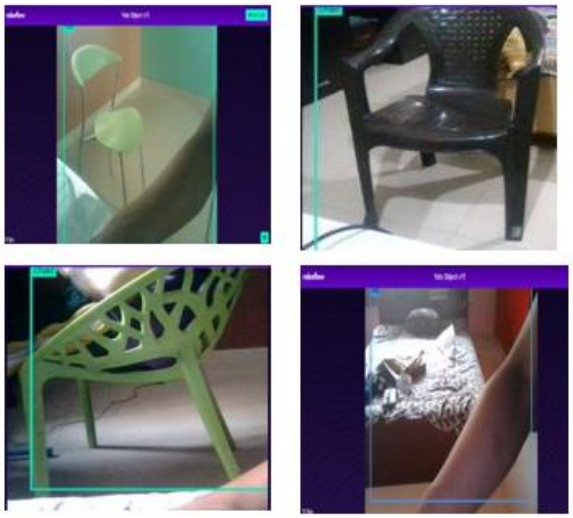

Fig. 7. Obtained results.

The existing research work on the comparisons of the object detection algorithm have been reviewed and presented. The comparison of various YOLO models and their accuracies is presented in Fig. 9. The test time is reduced by 15.6% compared to YOLOv3 due to YOLOv4's various improvements, which lead to the best detection results, as demonstrated above. Despite having fewer training parameters than the YOLOv3-tiny model, the YOLOv4-tiny model's detection results are nonetheless accurate, are subpar. The detection effect is the worst of the models, only achieving 50.06%. The SPP module causes the YOLOv3-SPP3 model's performance to be slightly better than the YOLOv3 but noticeably worse than the YOLOv4. Fig. 8 depicts the YOLO classification loss, the loss compares the predicted class probabilities with the ground truth labels for each object in the image, the loss is being compared between YOLO and Faster R-CNN. Faster R-CNN is better than YOLOv5 in terms of accuracy with approximately 10 times higher inference rate [29]. This optimization occurs by utilizing backpropagation and gradient descent techniques to update the network parameters. To achieve optimal performance, the weights assigned to each loss component can be adjusted to balance their contributions within the overall loss function. Comparative analysis of various YOLO versions for rural road, urban road, and highways image dataset of sample size 120000, 124000, 150000 respectively [30] the Yolov3 has good precession with bad recall and F measure, and with low mAP, FPS. On the other hand, YOLOv4 and YOLOv5 have stable scores in terms of precision and MAP. YOLOv5 outperforms in terms of speed of the algorithm, and precision as compared to YOLOv3, and YOLOv4 [30]. The comparative analysis of YOLO versions for urban road dataset is provided in Fig. 9.

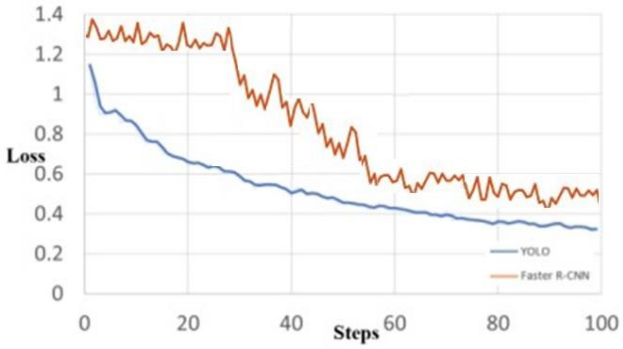

Fig. 8. YOLO classification loss.

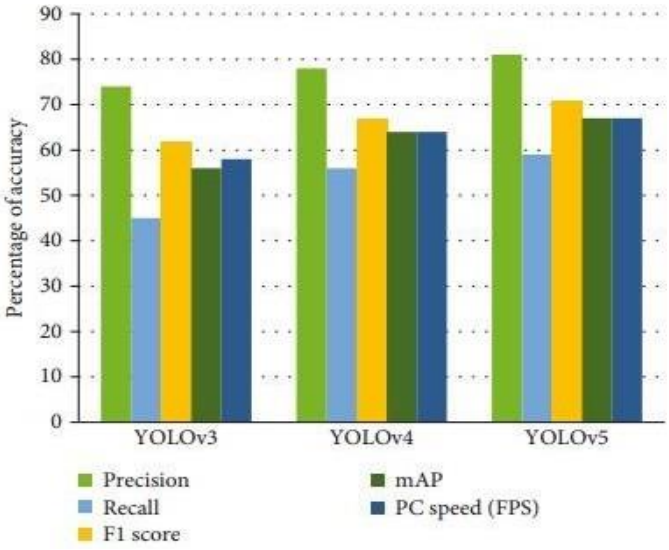

Fig. 9. Comparative analysis of YOLO versions.

In 2021 YOLO series was extended by introducing Yolox [32]. The series continues and the above of the mentioned models were created. Each model in the table above underwent 300 iterations of training. With the help of the ultralytics library's autobatch function, the micro groups' sizes were calculated. The YOLOv5s model's training took the least time. YOLOv5x training took the longest, lasting approximately nine hours, and was completed in under an hour. A comparative study of YOLOv5 is presented in Table II.

TABLE II. COMPARATIVE STUDY OF YOLOV5 VERSIONS DURING THE TESTING

Models	MAP	FPS	Parameters (M)	Test Time(s)
YOLOv3	76.55	35	61.5	122.23
YOLOv3-tiny	62.5	134	8.7	28.14
YOLOv3-SPP3	76.87	40	63.9	128.19
YOLOv4	87.48	72	27.6	103.86
YOLOv4-tiny	50.06	252	7.2	18.41

V. CONCLUSION

In this study object detection algorithms and systems are analyzed based on their accuracy and the speed of detecting objects. It also observed that the accuracy and speed of object detection algorithms are improving daily. Fast R-CNN is an enhanced version of R-CNN that incorporates a selective search for generating Regions of Interest. In contrast, Faster R-CNN utilizes a Regional Proposal Network (RPN), contributing to its superior performance compared to Fast R-CNN. But the Faster R-CNN algorithm required many passes to extract all the objects from the single frame; this is where Single Shot Detector (SSD) came into the picture. Till the time when YOLO was not developed SSD was considered to be the best. YOLOv5 was designed in such a way that it can detect small objects also, especially in autonomous vehicles. Additionally, leveraging Roboflow for annotation and data management has streamlined the preprocessing stage, ensuring the availability of adequately labeled training data for training object detection system. In a nutshell, YOLOv5 is a faster, more scalable, and lighter model compared to other competitors. In future work, it is very useful in IOT, or mobile-based detection, like objecting detecting sticks for blind people, sign language detectors, etc. YOLOv7 is a faster, but heavier model, hence can be used in robotics, satellite imaging, and other related things.

REFERENCES

[1] Xiao, Y.; Wang, X.; Zhang, P.; Meng, F.; Shao, F., “Object Detection Based on Faster R-CNN Algorithm with Skip Pooling and Fusion of Contextual Information.”, Sensors , vol. 20(19),2020.

[2] Daming Shi, Liying Zheng, & Jigang Liu., “Advanced Hough Transform Using A Multilayer Fractional Fourier Method.”, IEEE Transactions on Image Processing, vol.19(6), pp.1558–1566, 2010.

[3] H. Jabnoun, F. Benzarti and H. Amiri, "Object detection and identification for blind people in video scene," *2015 15th International Conference on Intelligent Systems Design and Applications (ISDA)*, Marrakech, Morocco, 2015, pp. 363-367.

[4] D. Garg, P. Goel, S. Pandya, A. Ganatra and K. Kotecha, "A Deep Learning Approach for Face Detection using YOLO," 2018 IEEE Punecon, Pune, India, 2018, pp. 1-4.

[5] Istiak Ahmad et al., “A Novel Deep Learning-based Online Proctoring System using Face Recognition, Eye Blinking, and Object Detection Techniques, IJACSA, vol. 12 (10), , 2021, pp. 847-854.

[6] Guo S, Li L, Guo T, Cao Y, Li Y. Research on Mask-Wearing Detection Algorithm Based on Improved YOLOv5. *Sensors*. 2022; vol.22(13),2022.

[7] Wentao Liu and Zhangyu Wang and Bin Zhou and Songyue Yang and Ziren Gong,” Real-time Signal Light Detection based on Yolov5 for Railway” , IOP Conference Series: Earth and Environmental Science ,vol.769(3), 2021.

[8] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, 2014, pp. 580-587.

[9] M. A. Sarrayrih and M. Ilyas, “Challenges of online exam, performances and problems for online university exam,” International Journal of Computer Science Issues (IJCSI), vol. 10(1), 2013.

[10] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.39, pp. 1137-1149, 2017.

[11] M. F. Haque, H. -Y. Lim and D. -S. Kang, "Object Detection Based on VGG with ResNet Network," *2019 International Conference on Electronics, Information, and Communication (ICEIC)*, Auckland, New Zealand, 2019, pp. 1-3.

[12] Y. Zhang, Y. Huang and L. Wang, "Multi-task Deep Learning for Fast Online Multiple Object Tracking," *2017 4th IAPR Asian Conference on Pattern Recognition (ACPR)*, Nanjing, China, 2017, pp. 138-143.

[13] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 779-788.

[14] A. Krizhevsky, I. Sutskever and G. Hinton, "ImageNet classification with deep convolutional neural networks", Communications of the ACM, vol. 60(6) ,2017, pp. 84-90.

[15] Russakovsky, O., Deng, J., Su, H. et al. ImageNet Large Scale Visual Recognition Challenge. Int J Comput Vis 115, 211–252 (2015).

[16] Lin, TY. et al. (2014). Microsoft COCO: Common Objects in Context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds) Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8693. Springer, Cham. https://doi.org/10.1007/978-3-319-10602-1_48.

[17] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV,

USA, 2016, pp. 779-788, doi: 10.1109/CVPR.2016.91.

[18] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 6517-6525, doi: 10.1109/CVPR.2017.690.

[19] J. Choi, D. Chun, H. Kim and H. -J. Lee, "Gaussian YOLOv3: An Accurate and Fast Object Detector Using Localization Uncertainty for Autonomous Driving," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South), 2019, pp. 502-511.

[20] S. -H. Bae and K. -J. Yoon, "Robust Online Multi-object Tracking Based on Tracklet Confidence and Online Discriminative Appearance Learning," 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 2014, pp. 1218-1225.

[21] X. Yu, T. W. Kuan, Y. Zhang and T. Yan, "YOLO v5 for SDSB Distant Tiny Object Detection," 2022 10th International Conference on Orange Technology (ICOT), Shanghai, China, 2022, pp. 1-4.

[22] M. Taskiran, M. Killioglu and N. Kahraman, "A Real-Time System for Recognition of American Sign Language by using Deep Learning," *2018 41st International Conference on Telecommunications and Signal Processing (TSP)*, Athens, Greece, 2018, pp. 1-5.

[23] R. Yadav et al., "High Speed Single-Stage Face Detector using Depthwise Convolution and Receptive Fields" IJACSA, vol. 12 (2), pp. 738-744, 2021.

[24] Thuan, DoCong. "Do Thuan evolution of yolo algorithm and yolov5: the state-of-the-art object detection algorithm evolution of yolo algorithm and yolov5: the state-of-the-art object detection algorithm." (2021).

[25] Linlin Zhu," Improving YOLOv5 with Attention Mechanism for Detecting Boulders from Planetary Images", *Remote Sens. 13*(18), 2021.

[26] Martinus Grady Naftali, Jason Sebastian Sulistyawan, Kelvin Julian "Comparison of Object Detection Algorithms for Street-level Objects",arXiv:2208.11315,2022.

[27] B. Adhikari and H. Huttunen, "Iterative Bounding Box Annotation for Object Detection," *2020 25th International Conference on Pattern Recognition (ICPR)*, Milan, Italy, 2021, pp. 4040-4046.

[28] N. Adhikari, N. R. Behera, V. R. E, E. S. J. Pimo, V. Chaturvedi and V. Tripathi, "Modeling of Optimal Deep Learning Enabled Object Detection and Classification on Drone Imagery," 2022 International Conference on Augmented Intelligence and Sustainable Systems (ICAISS), Trichy, India, 2022, pp. 303-309.

[29] T. Mahendrakar, J. Cutler, N. Fischer, A. Rivkin, A. Ekblad, K. Watkins, M. Wilde, R. White and B. Kish, "Use of Artificial Intelligence for Feature Recognition and Flightpath Planning Around NonCooperative Resident Space Object," in AIAA, Las Vegas, 2021. Pp. 1-11.

[30] R. Hmida, A. B. Abdelali, and A. Mtibaa, "Speed limit sign detection and recognition system using SVM and MNIST datasets," Neural Computing and Applications, vol. 31(9), pp. 5005–5015, 2019.

[31] Horvat, Marko & Jelečević, Ljudevit & Gledec, Gordan. (2022). A comparative study of YOLOv5 models performance for image localization and classification, CECIIS 2022.

[32] Ge Z, Liu S, Wang F, Li Z, Sun J, "YOLOx: exceeding yolo series in 2021", arXiv:2107.08430,2021

[33] Srivastava, S., Divekar, A.V., Anilkumar, C. et al. Comparative analysis of deep learning image detection algorithms. J Big Data 8, 66 (2021).

[34] Lou, Lijun and Liu, Junya and Yang, Zhen and Zhou, Xin and Yin, Zhijian,"Agricultural Pest Detection based on Improved Yolov5.", Association for Computing Machinery,pp.7-12, 2023.doi:10.1145/3577530.3577532

[35] Jha, S., Seo, C., Yang, E. et al., "Real time object detection and trackingsystem for video surveillance system.", Multimed Tools Appl 80, pp. 3981–3996, 2021.

[36] Benjumea A, Teeti I, Cuzzolin F, Bradley A YOLO-z:improving small object detection in YOLOv5 for autonomous vehicles. arXiv preprint arXiv:2112.11798,2021.

[37] A. J. Lebumfacil and P. A. Abu, "Traffic Sign Detection and Recognition Using YOLOv5 and Its Versions", IEEE 1st International Conference on Cognitive Mobility (CogMob), Budapest, Hungary, 2022, pp. 11-18, 2022.

[38] W. A. K. Adji, A. Amalia, H. Herriyance and E. Elizar, "Abnormal Object Detection In Thoracic X-Ray Using You Only Look Once (YOLO)," 2021 International Conference on Computer System, Information Technology, and Electrical Engineering (COSITE), Banda Aceh, Indonesia, 2021, pp. 118-123

[39] K. Liu, "STBi-YOLO: A Real-Time Object Detection Method for Lung Nodule Recognition," in IEEE Access, vol. 10, pp. 75385-75394,2022.

[40] S. Chandna and A. Singhal, "Towards Outdoor Navigation System for Visually Impaired People using YOLOv5," 2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2022, pp. 617-622.

[41] Rio Arifando: "Improved YOLOv5-Based Lightweight Object Detection Algorithm for People with Visual Impairment to Detect Buses", Appl. Sci. vol. 13(9),2023.

[42] [42]Aydin, Burchan, and Subroto Singh.,"Drone Detection Using YOLOv5" Eng vol.4(1),pp.416-433, 2023.

[43] Li, C., Li, L., Jiang, H., Weng, K., Geng, Y., Li, L., Ke, Z., Li, Q., Cheng, M., Nie, W. and Li, Y., "YOLOv6: A single-stage object detection framework for industrial applications.", arXiv preprint arXiv:2209.02976, 2022.

[44] Chien-Yao Wang,"YOLOv7: Trainable bag-of-freebies sets new state-of- the-art for real-time object detectors",arXiv:2207.02696,July 2022, R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 2014, pp. 580-587.

[45] R. Girshick, "Fast R-CNN," 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 2015, pp. 1440-1448, doi: 10.1109/ICCV.2015.169.

[46] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149, 2017.

[47] K. He, G. Gkioxari, P. Dollár and R. Girshick, "Mask R-CNN," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 2980-2988.

[48] Z. Liu, X. Gu, H. Yang, L. Wang, Y. Chen and D. Wang, "Novel YOLOv3 Model With Structure and Hyperparameter Optimization for Detection of Pavement Concealed Cracks in GPR Images," in IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 11, pp. 22258-22268, 2022.

[49] G. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," arXiv:1704.04861, 2017.

[50] S. Gidaris and N. Komodakis, "Object Detection via a Multi-region and Semantic Segmentation-Aware CNN Model," 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 2015, pp. 1134-1142.

[51] Liu, W. et al. SSD: Single Shot MultiBox Detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds) Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science(), vol 9905. Springer, Cham, 2016.

[52] T. -Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollár, "Focal Loss for Dense Object Detection,",IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 2999-3007.

[53] S. Shi, X. Wang and H. Li, "PointRCNN: 3D Object Proposal Generation and Detection From Point Cloud," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 770-779.

[54] B. Yang, W. Luo and R. Urtasun, "PIXOR: Real-time 3D Object Detection from Point Clouds," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 7652-7660.