EFFICIENT TWO-STAGE TARGET DETECTION FRAMEWORK: ENHANCING OBJECT DETECTION WITH DEEP LEARNING

Abstract

Target detection in computer vision is a vital problem that has been studied in-depth for the past 20 years and used in a variety of applications. The main goal is to locate and recognize multiple objects of categoriesin each image accurately and swiftly. Computer vision algorithm classified into two categories based on their architecture and approach towards problems. In addition to providing a thorough review of the top algorithms in each area, this article compares and analyses a number of typical algorithms employed in this field using both popular public datasets and unique datasets. This essay also forecasts potential difficulties that target detection might encounter in the future.

Keywords: Real-time Detection, Computer Vision, Region-based Convolutional Neural Networks, Feature Extraction

Authors

Nikhil Kumar Gupta
GNIOT Engineering College
Greater Noida
nikhil.ee@gniot .net .in

Naveen Jaiswal
GNIOT Engineering College
Greater Noida
jaiswalnaveen143@gmail.com

Satyam Ray
GNIOT Engineering College
Greater Noida
satyamray14601@gmail.com

Pritam Yadav
GNIOT Engineering College
Greater Noida
pritamyadav2222@gmail.com

Vipin Pandey
GNIOT Engineering College
Greater Noida
pandeyvipin2369@gmail.com

I. INTRODUCTION

The advancement of computer vision requires a number of essential elements, such as object detection, feature extraction, and image processing. Utilizing cameras or other sensors, images are captured, and then they are processed to enhance their quality by lowering noise, boosting contrast, and sharpening edges. Corners, edges, and textures are a few examples of relevant components that feature extraction searches for and extracts from photos.

Object identification, which uses machine learning approaches to detect and categorize items based on their features, brings the process to a successful conclusion. the enormous potential of computer vision to transform a number of industries and offer fresh approaches to challenging issues. Computer vision is projected to alter how people interact with the visual environment and open up new prospects for automation as machine learning and artificial intelligence advance. Computer vision has the potential to totally change a number of industries while also offering creative answers to complex issues. It is anticipated that machine learning and artificial intelligence will change the way we interact with our visual surroundings and open up new possibilities for automation and reasoned decision-making.

II. ENHANCING TARGET DETECTION ACCURACY

1. **R-CNN:** In order to create a list of region recommendations that are likely to contain objects of interest, the R-CNN algorithm uses a selective search strategy. After that, this feature vector is fed into a group of support vector machines (SVMs), which classify objects and improve bounding box predictions. The R-CNN (Region-based Convolutional Neural Network) object detection algorithm was released in 2014 by Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. The enhanced R-CNN model achieved a mean average precision (mAP) of 66%. Figure 1 illustrates the architecture of the model, which begins by generating approximately 2000 region proposals using selective search for each image to be analyzed. These extracted region features are then uniformly transformed into fixed-length feature vectors and fed into an SVM classifier for classification.

   The R-CNN (Region-based Convolutional Neural Network) approach is widely used in computer vision for object detection. However, it suffers from computational inefficiency due to the selective search technique and the individual evaluation of region proposals. To address these limitations, faster and more accurate variants of R-CNN have been developed. For example, Fast R-CNN accelerates computation by utilizing a shared CNN for all region proposals. This allows for end-to-end training and significantly faster processing.

2. **Spatial Pyramid Pooling Network (SPP-Net):** Convolutional neural networks (CNNs) struggle with variable-sized input images, so the Spatial Pyramid Pooling Network (SPP-Net), a deep learning architecture, was developed to solve this problem without the use of resizing or cropping methods. SPP-Net was developed for tasks involving object recognition and was first introduced in 2014 by Kaiming He et al. SPP-Net use a CNN to extract features from the input image. These features are then processed by a spatial pyramid pooling (SPP) layer that divides the feature map into sub-regions of fixed sizes

and applies max pooling operations to each sub-region. SPP-Net employs the SPP layer, which handles input photographs of varying sizes without needing to resize or crop them, making it excellent for real-world applications where input photos may have diverse aspect ratios or sizes needing to resize or crop them.

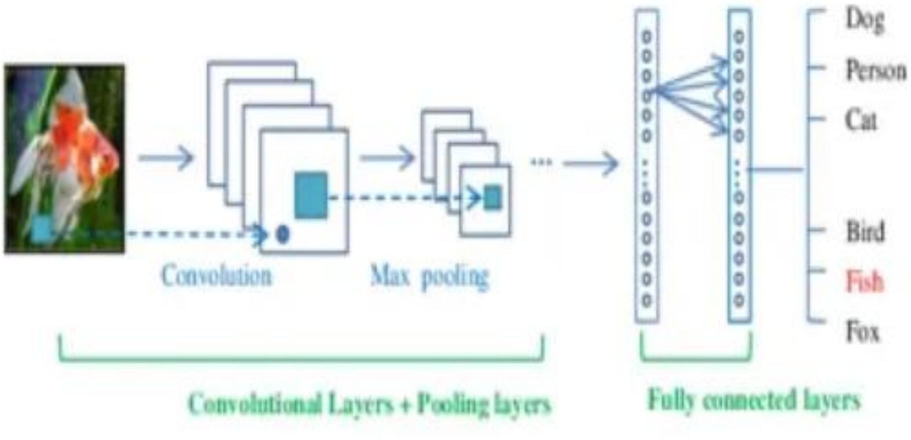


Figure 1: R-CNN architecture

3. **Fast R-CNN:** As a result, the requirement for feature extraction for each region suggestion is eliminated, which significantly reduces the computational efficiency of R-CNN. An area of Interest (RoI) pooling layer is placed after the shared CNN and extracts a fixed-size feature map from each area proposal. This allows the RoI to be fed into a set of fully connected layers for object categorization. As a result, each region recommendation no longer requires feature extraction, which dramatically lowers R-CNN's processing efficiency. After the shared CNN, an area of Interest (RoI) pooling layer is added that extracts a fixed-size feature map from each area proposal. This makes it possible to feed the RoI into a group of completely interconnected layers for object categorization.

   Fast R-CNN is an enhanced iteration of the R-CNN (Region-based Convolutional Neural Network) algorithm, specifically tailored for object detection tasks. Developed by Ross Girshick in 2015, Fast R-CNN aims to address the limitations observed in the original R-CNN approach. One notable improvement is the adoption of a training technique that facilitates the simultaneous optimization of the Convolutional Neural Network (CNN) and the Region of Interest (RoI) pooling layers. This enables an end-to-end learning process, enhancing both efficiency and performance. In contrast, R-CNN, which pre-trains the CNN on a huge dataset, handles the region proposals individually using a collection of support vector machines (SVMs) and bounding box regression. This technique enables Fast R-CNN to recognize objects faster and more precisely than R-CNN.

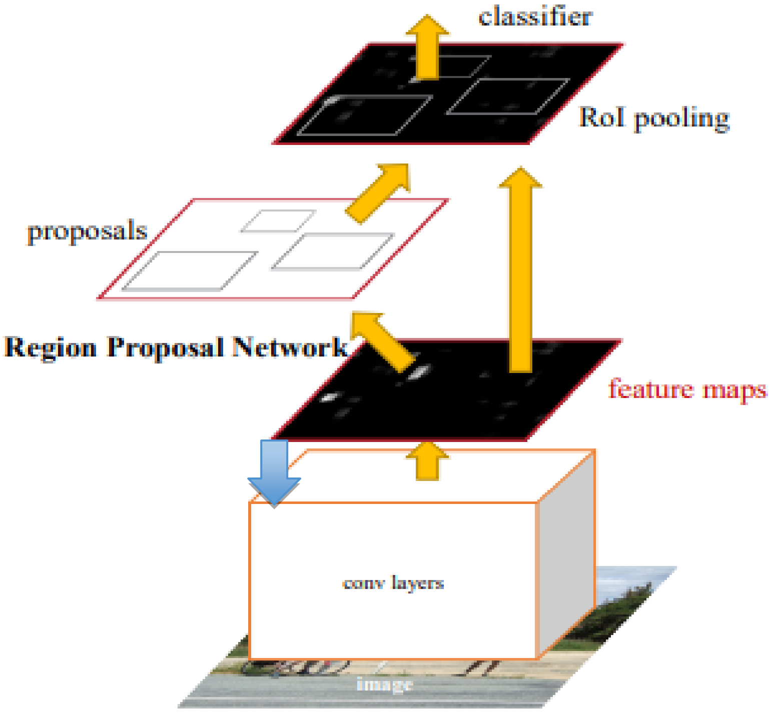


Figure 2

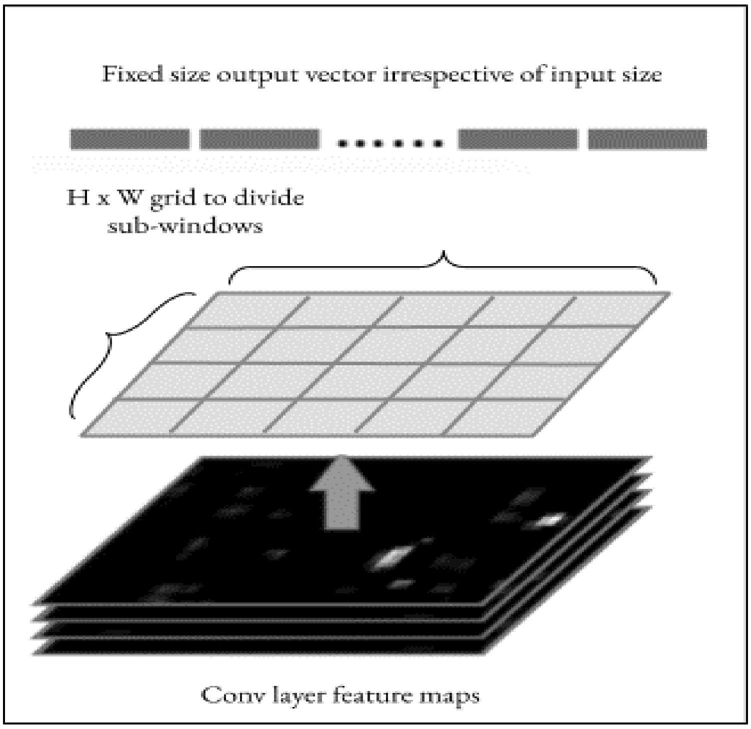


Figure 3: Architecture of Fast R-CNN

4. **Faster R-CNN:** The Faster R-CNN method improves upon Fast R-CNN by incorporating a Region Proposal Network (RPN). This RPN leverages the same convolutional layers as the object recognition network and allows for end-to-end training. This integration

enhances the overall performance and efficiency of the object detection process. The RPN outputs the shared CNN's convolutional feature map and generates region suggestins by sliding a small network over it. Each sliding window predicts many area proposals and the accompanying objectless scores. The region proposals generated by the algorithm are refined through bounding box regression, which helps improve their accuracy. Redundant suggestions are then eliminated using a technique called non-maximum suppression. The refined proposals are then passed through a RoI pooling layer and a set of fully connected layers responsible for bounding box regression and object classification. This method improves the computational efficiency of the object detection system and generates faster and more accurate object detection results than Fast R-CNN by allowing the RPN to share the feature computation with the object detection network.

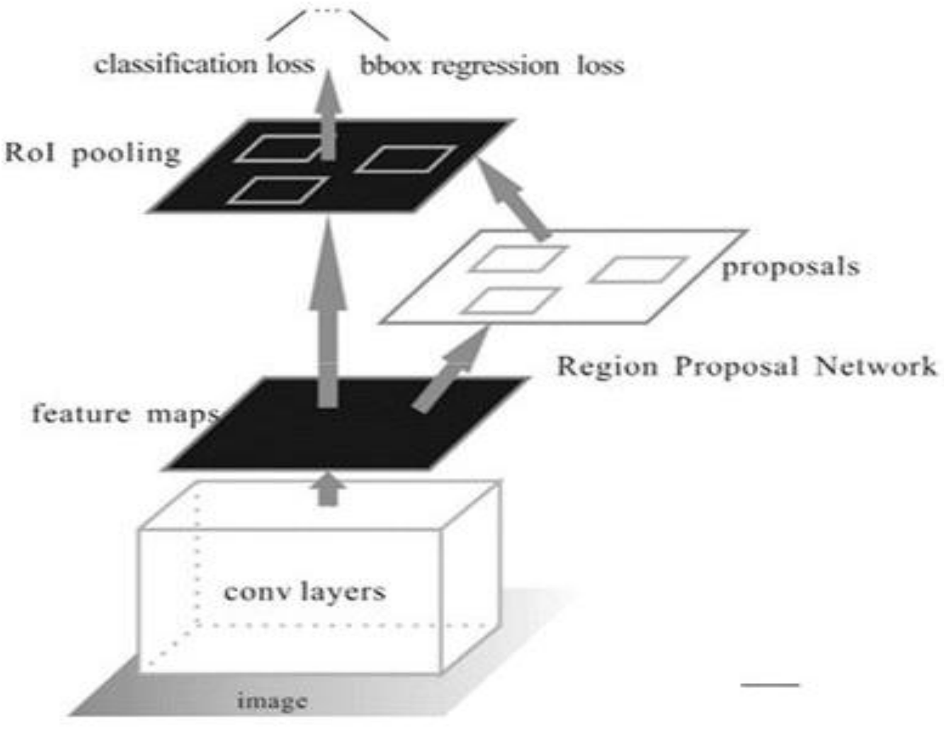


Figure 4: Structure of a Quicker R-CNN

III. SINGLE-STAGE TARGET DETECTION ALGORITHMS

1. **YOLO V1:** Joseph Redmon unveiled YOLOv1 in 2016, an object identification model that does away with the region proposal extraction method in favour of a more traditional one. Using an SS grid to segment the image into cells, each cell in the grid predicts B bounding boxes and confidence scores for a total prediction of B(4+1) values. YOLOv1 enables 100% real-time detection at up to 45 frames per second on a single TitanX. The model performs poorly when attempting to recognize items in groups, which reduces identification accuracy even when there are less background mistakes.

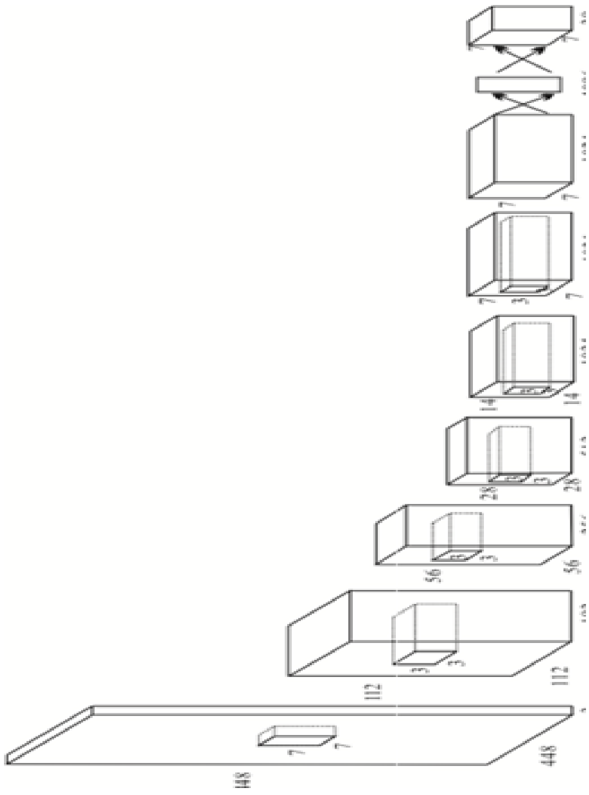


Figure 5: YOLOv1 architecture

2. **YOLOv2:** The YOLO V2 model, which contrasts with YOLO V1 and emphasizes enhanced memory and localization while preserving classification accuracy, was introduced by Joseph Redmon in 2016. YOLO V2 utilizes Darknet-19, a recently developed fully convolutional feature extraction network comprising 19 convolutional layers and up to 5 pooling layers. YOLO V2 incorporates various enhancements, including the anchor box technique, dropout reduction, addition of a batch normalization layer to the convolutional layer, utilization of k-means clustering on the training set of bounding boxes, and the adoption of several other improvements of multi-scale training are all factors that contributed to the model's improved recall and accuracy. The model still finds it challenging to locate targets with significant overlap and small size.

3. **YOLO V3:** YOLO V3, introduced by Joseph Redmon, is a well-balanced object detection model known for its exceptional speed and accuracy. It introduces several key improvements, such as transitioning from single-label classification to multi-label classification for category prediction. To enhance the detection of small targets, the model incorporates three scales and incorporates an up-sampling fusion technique inspired by the Feature Pyramid Network (FPN). The network structure is built upon the deeper feature extraction network known as Darknet-53. While the overall detection accuracy may not have seen significant improvement, YOLOv3 notably enhances the recognition of small targets and further boosts detection speed, particularly with an Intersection over Union (IOU) threshold greater than 0.5.

4. **Single Shot MultiBox Detector (SSD): A Fast and Effective Object Detection Model:** In 2016, Liu introduced the SSD (Single Shot MultiBox Detector) model, which incorporates the anchor box concept from the Faster R-CNN model and regression techniques from the YOLO algorithm. By utilizing both bottom-level and high-level feature maps, SSD enhances the accuracy and efficiency of multi-scale object detection. The final two completely linked layers in the VGG, which is the basic design, are replaced by convolutional layers. Utilising an Nvidia Titan X, SSD achieves 74.3% mAP on VOC2007 at 59 frames per second and is inspired by the RPN network's anchor mechanism. Small targets are harder for the SSD model to classify, and because different scale feature maps are independent, multiple boxes of varying sizes can simultaneously identify the same thing.

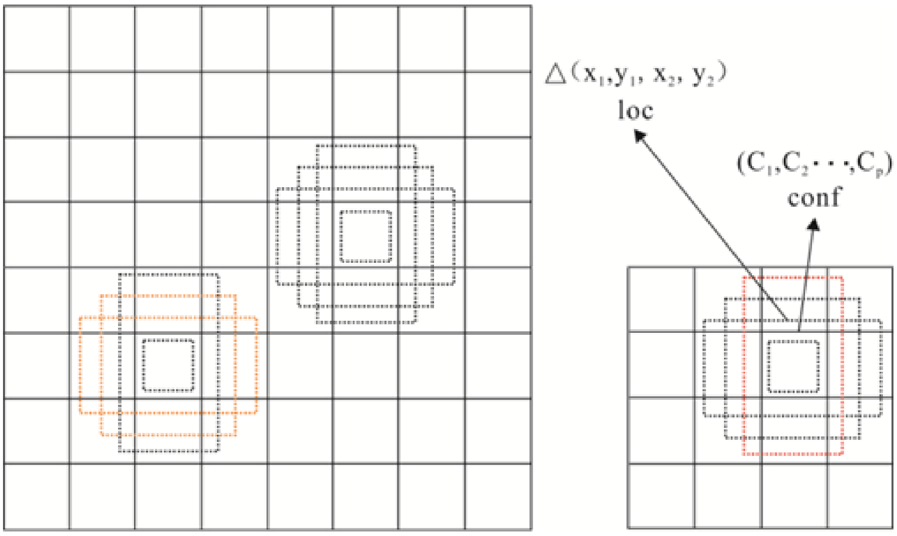


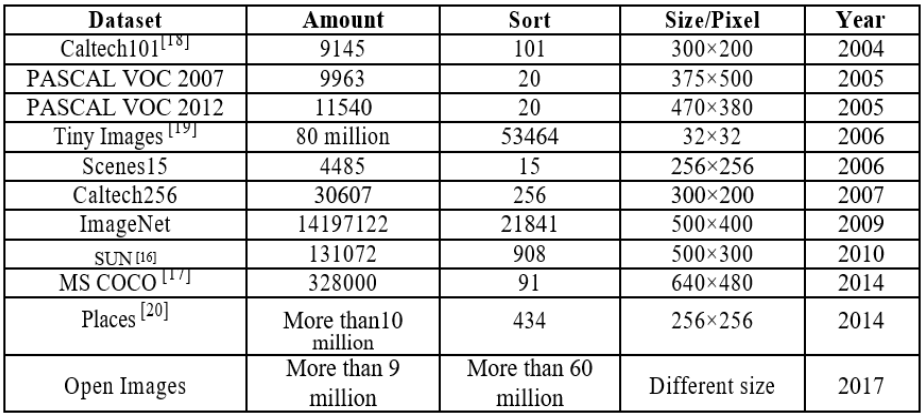
Figure 6: SSD architecture

5. **YOLO v4:** In 2016, Liu introduced the SSD (Single Shot MultiBox Detector) model, which combines the anchor box theory and regression techniques from the Faster R-CNN detection model. SSD leverages both bottom-level and high-level feature maps to improve the accuracy of detecting objects at different scales. The VGG architecture forms the basis of SSD, where the final two fully connected layers are replaced with convolutional layers. Inspired by the anchor mechanism in the RPN network, SSD achieves impressive results, with 74.3% mAP on VOC2007 at a rapid processing rate of 59 frames per second using an Nvidia Titan X. However, the SSD model faces challenges in accurately classifying small targets, and the independent nature of the feature maps at multiple scales allows multiple boxes of different sizes to detect the same object simultaneously.

IV. PERFORMANCE AND DATASETS ACROSS DIFFERENT ALGORITHMS

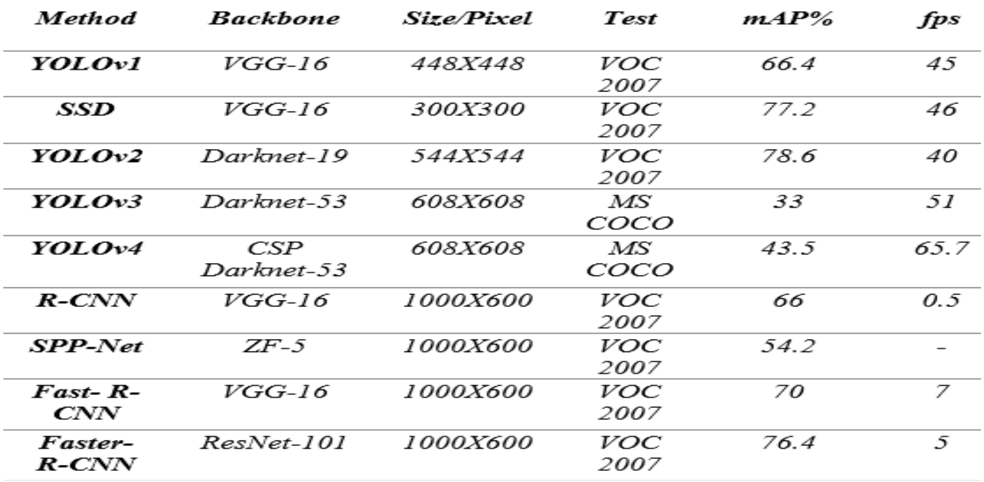
1. **Dataset:** Although the phrase "artificial intelligence" was coined in 1956, it wasn't until 2012 that AI achieved substantial strides, in part because of the expansion of machine learning techniques, computing power, and data volume. Because performance evaluation and algorithm evaluation heavily rely on datasets. In actuality, the use of datasets in research has greatly aided the creation of detecting systems. The parameters for popular public datasets are listed in Table 1.

Table I: Parameters and Characteristics of a Dataset



Dataset	Amount	Sort	Size/Pixel	Year
Caltech101[18]	9145	101	300×200	2004
PASCAL VOC 2007	9963	20	375×500	2005
PASCAL VOC 2012	11540	20	470×380	2005
Tiny Images [19]	80 million	53464	32×32	2006
Scenes15	4485	15	256×256	2006
Caltech256	30607	256	300×200	2007
ImageNet	14197122	21841	500×400	2009
SUN [16]	131072	908	500×300	2010
MS COCO [17]	328000	91	640×480	2014
Places [20]	More than10 million	434	256×256	2014
Open Images	More than 9 million	More than 60 million	Different size	2017

2. **Comparative Analysis of Algorithm Performances:** Statistics and comparisons of single-stage and two-stage detection techniques are provided in Table 2.



Method	*Backbone*	*Size/Pixel*	*Test*	*mAP%*	*fps*
YOLOv1	*VGG-16*	*448X448*	*VOC 2007*	*66.4*	*45*
SSD	*VGG-16*	*300X300*	*VOC 2007*	*77.2*	*46*
YOLOv2	*Darknet-19*	*544X544*	*VOC 2007*	*78.6*	*40*
YOLOv3	*Darknet-53*	*608X608*	*MS COCO*	*33*	*51*
YOLOv4	*CSP Darknet-53*	*608X608*	*MS COCO*	*43.5*	*65.7*
R-CNN	*VGG-16*	*1000X600*	*VOC 2007*	*66*	*0.5*
SPP-Net	*ZF-5*	*1000X600*	*VOC 2007*	*54.2*	*-*
Fast- R-CNN	*VGG-16*	*1000X600*	*VOC 2007*	*70*	*7*
Faster-R-CNN	*ResNet-101*	*1000X600*	*VOC 2007*	*76.4*	*5*

V. CONCLUSION

Object recognition is a complex and significant taskin the area of computer vision, and it has received a great deal of interest recently. Deep learning has emerged as a prominent approach for object detection in various industries. However, despite its advancements, deep learning still faces several challenges, including. Addressing the dependence on large amounts of training data and finding ways to reduce data requirements. Enhancing the reliability of detecting small objects, which can be particularly challenging.

Improving object detection performance across multiple categories, ensuring accurate identification across diverse classes of objects. In summary, while deep learning-based object detection systems have shown great potential, there is ongoing research and development needed to overcome these challenges and further enhance the capabilities of object recognition algorithms.

REFERENCES

[1] Wu, R.B. Research on Application of Intelligent Video Surveillance and Face Recognition Technology in Prison Security. China Security Technology and Application. 2019,6: 16-19.

[2] Tian, J.X., Liu, G.C., Gu, S.S., Ju, Z.J., Liu, J.G., Gu, D.D. Research and Challenge of Deep Learning Methods for Medical Image Analysis. Acta Automatica Sinica,2018, 44: 401-424.

[3] Jiang, S.Z., Bai, X. Research status and development trend of industrial robot target recognition and intelligent detection technology. Guangxi Journal of Light Industry, 2020, 36: 65-66.

[4] Krizhevsky, A., Sutskever, I., Hinton, G. ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems,2012, 25: 1097-1105.

[5] Russakovsky, O., Deng, J., Su, H., et al. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision,2015, 115: 211-252.

[6] Girshick, R., Donahue, J., Darrel, T.,Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In: Computer Vision and Pattern Recognition. Columbus.2014, pp. 580-587.

[7] He, K.M., Zhang, X.Y., Ren, S.Q., Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. IEEE Transactions on Pattern Analysis & Machine Intelligence,2015, 37: 1904-1916.

[8] Girshick, R. Fast R-CNN.In: Proceedings of the IEEE international conference on computer vision. Santiago.2015, pp. 1440-1448.

[9] Ren, S.Q., He, K.M., Girshick, R., Sun, J. Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. Montreal.2016, pp. 91-99.

[10] Redmon, J., Divvala, S., Grishick, R., Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In: Computer Vision and Pattern Recognition. Las Vegas.2016, pp. 779-788. Redmon, J., Farhadi, A. YOLO9000: better, faster, stronger. In: Computer Vision and Pattern Recognition. Hawaii.2017, pp. 7263-7271.

[11] Redmon, J., Farhadi, A. (2018) Yolov3: An incremental improvement. arXiv: Computer Vision and Pattern Recognition.

[12] Liu, W., Anguelov, D., Erhan, D., et al. SSD: Single Shot MultiBox Detector. European Conference on Computer Vision, 2016, pp. 21-37.

[13] Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv: Computer Vision and Pattern Recognition, 2020.

[14] Everingham, M., Eslami, S.M.A., Van Gool, L. The Pascal Visual Object Classes Challenge: A Retrospective. International Journal of Computer Vision,2015, pp.98-136.

[15] Xiao, J.X., Ehinger, K.A., Hays, J.,Torralba, A.,Oliva, A. SUN Database: Exploring a Large Collection of Scene Categories. International Journal of Computer Vision, 2016,pp.3-22.

[16] Lin T Y , Maire M , Belongie S , et al. Microsoft COCO: Common Objects in Context. European Conference on Computer Vision, 2014, pp.740-755.

[17] Li, F.F., Rob, F., Pietro, P. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. Computer Vision and Image Understanding,2007,pp. 59-70.

[18] Torralba, A., Fergus, R., Freeman, W.T. 80 Million Tiny Images: A Large Data Set for Nonparametric Object and Scene Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence,2008, pp.1958-1970.

[19] Zhou, B., Lapedriza, A., Khosla, A., et al. Places: A 10 million Image Database for Scene Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, pp.1452-1464