IJETIR

ISSN (Online) 2583-0554

International Journal of Emerging Technologies and Innovative Research (IJETIR)

Volume 4, Issue 2, February 2024

Impact Factor: 5.731

An Analysis of Deep Learning-Driven Object Detection Methodologies, Highlighting Advancements in Single-Line Research

Swapnil Nilkanth Patil[1] and Dr. Mukesh Kumar Rana[2]

Research Scholar, Department of Computer Science and Engineering[1]

Research Guide, Department of Computer Science and Engineering[2]

NIILM University, Kaithal, Haryana, India

Abstract: *Target identification is a crucial issue in computer vision, and in the last 20 years, it has gained significant attention as a research hotspot and been used extensively. Its goal is to find and identify a large number of items belonging to specified categories in a given picture in a timely and accurate manner. The algorithms may be categorised into two groups based on the model training method: single-stage detection algorithms and two-stage detection algorithms. The representative algorithms for every level are thoroughly presented in this work. Subsequently, the public and unique datasets often used for target identification are presented, and several sample techniques in this domain are examined and contrasted. Lastly, some possible difficulties with target identification are discussed.*

Keywords: Deep Learning, Convolutional Neural Networks

I. INTRODUCTION

A fundamental area of study in computer vision, deep learning, artificial intelligence, etc. is object detection. More difficult computer vision tasks, such target tracking, event detection, behaviour analysis, and scene semantic comprehension, need it as a necessary precursor. It seeks to identify each target's bounding box, precisely identify the category, and find the target of interest inside the picture. Automatic driving of vehicles, video and image retrieval, intelligent video surveillance, medical image analysis, industrial inspection, and other sectors have all made extensive use of it.

Six phases make up traditional detection techniques for manually extracting features: pre-processing, window-sliding, feature extraction, feature selection, feature classification, and post-processing. These steps are often used for certain identification tasks. Small data size, low portability, lack of pertinence, high temporal complexity, window redundancy, lack of resilience against diversity changes, and strong performance limited to certain basic conditions are its key drawbacks.

Krizhevsjy[4] and others introduced the AlexNet image categorisation model in 2012. It was built on a convolutional neural network (CNN).The picture collection ImageNet[5] hosted an image classification competition, which they won with a significant 11% accuracy edge over the runner-up using conventional techniques. Numerous academics have started using deep convolutional neural networks for target identification problems and have put out a number of top-notch techniques. It may be broadly classified into two categories: area proposal-based single-stage detection algorithms and regression-based two-stage detection algorithms.

II. TWO-STAGE TARGET DETECTION FRAMEWORK

R-CNN

Girshick presented the R-CNN[6] technique in 2014, marking the first practical target identification model based on convolutional neural networks. The enhanced R-CNN model achieves a mAP of 66%. As shown in figure 1, the model extracts around 2000 area recommendations from each picture to be recognised using the Selective Search method

IJETIR
ISSN (Online) 2583-0554
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Volume 4, Issue 2, February 2024
Impact Factor: 5.731

initially. Subsequently, the retrieved proposals' sizes undergo uniform scaling to provide a fixed-length feature vector. These features are then fed into the SVM classifier to perform classification. Ultimately, the regression operation of the bounding box is carried out by training a linear regression model. Although the R-CNN's accuracy is much higher than that of the conventional detection approach, its computation efficiency is too poor and its computation volume is quite enormous. Second, object distortion might result from scaling the region suggestion straight to a fixed-length feature vector.

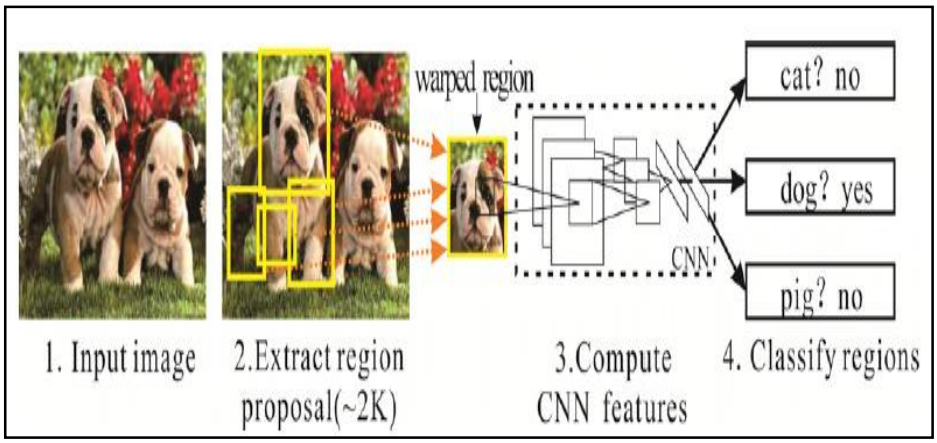

Figure 1. R-CNN architecture

SPP-Net

2015 saw the introduction of the Spatial Pyramid Pooling (SPP) model by He[7], which addresses the issues of R-CNN's fixed input size picture blocks need and poor detection efficiency. Once the original picture has gone through the convolution layer, this approach does a single convolution computation to extract the features of the areas specified on the feature map. In order to extract the feature vector of a defined size, the feature of the area proposal is passed via the spatial pyramid pooling layer, which is introduced concurrently with the final convolutional layer. Spp-Net avoids repeating computations by performing feature extraction on the full picture only once, in contrast to the R-CNN. It still has the same drawbacks as R-CNN, though: 1) Training procedures with several phases are difficult. 2) Additional regressors are needed, and separate SVM classifiers must be trained.

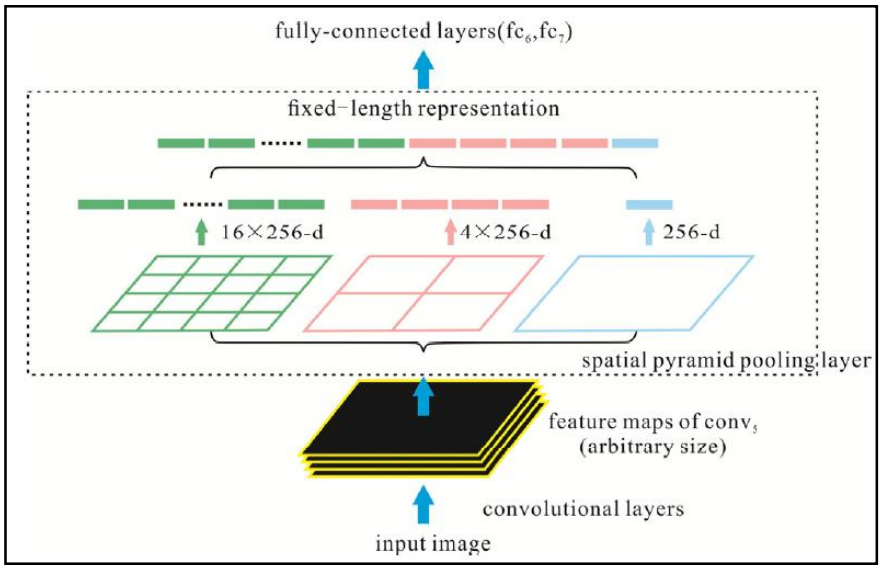

Figure 2. SPP-Net architecture

IJETIR
ISSN (Online) 2583-0554
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Volume 4, Issue 2, February 2024
Impact Factor: 5.731

Fast R-CNN

Fast R-CNN[8] was the model that Girshick suggested in 2015. The mean absolute percentage (mAP) in the combined VOC2007 and VOC2012 dataset [15] is 70.0%. Figure 2 depicts its structure. Fast R-CNN features three modifications over R-CNN. Initially, it employed the softmax function for classification in lieu of the SVM used in R-CNN. Second, to convert the feature of the candidate box into a feature map with a fixed size for access to the full connection layer, the model leverages the pyramid pooling layer in SPP-Net and substitutes the region of interest pooling layer for the final pooling layer in the convolutional layer. Ultimately, two parallel fully linked layers take the role of the CNN network's final softmax classification layer. Still, it is unable to satisfy the demands of real-time detection.

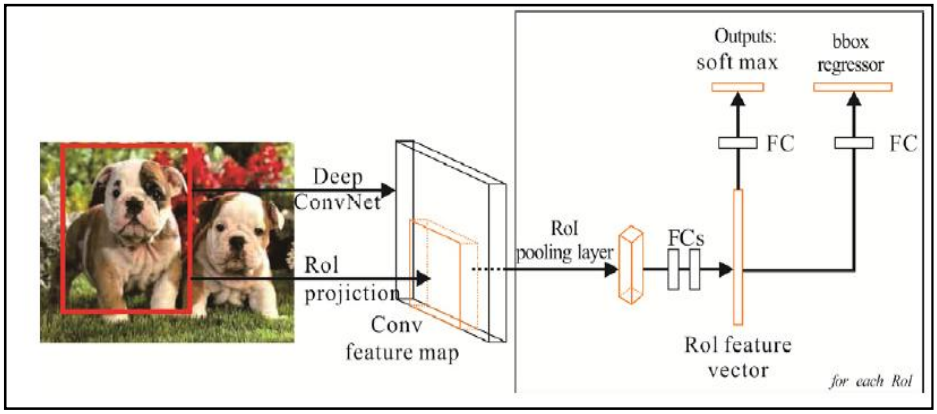

Figure 3. Fast R-CNN architecture

Faster R-CNN

Region proposal networks are used by Ren's Faster R-CNN[9] model to create region proposals instead of the earlier Selective Search technique. The model is split into two modules: the Fast R-CNN detection method and a fully convolutional neural network module that generates all region proposals. These two modules share the same set of convolutional layers. The input picture is carried forward through the CNN network to the final Shared convolutional layer. The picture is carried forward to the particular convolutional layer to create a higher-dimensional feature map; on the one hand, the feature map for the RPN network's input is produced. Even with its exceptional detection accuracy, Faster R-CNN is still unable to accomplish real-time detection.

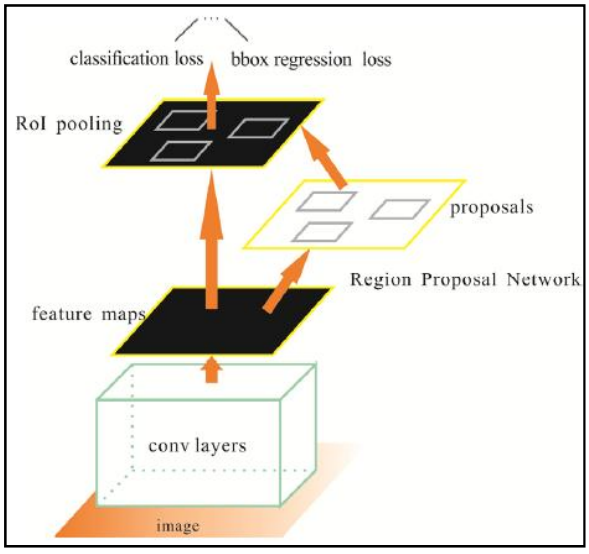

Figure 4. Faster R-CNN architecture

IJETIR
ISSN (Online) 2583-0554
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Volume 4, Issue 2, February 2024
Impact Factor: 5.731

III. ONE-STAGE TARGET DETECTION ALGORITHM

YOLOv1

Joseph Redmon presented the YOLOv1[10] object detection model in 2016. The extraction method of region proposal is not necessary for the YOLOv1 detection model. All of the detection model is nothing more complex than a CNN network structure. Its main concept is to utilise the whole graph as the network's input and to immediately provide the bounding box's category and position at the output layer. An picture is first split into S*S grids, where each grid cell predicts a B bounding box and the confidence scores associated with it. That is, the entire B*(4+1) values are predicted by each cell. Its complete real-time detection speed on a single TitanX is 45 frames per second. When dealing with items in group form, YOLO performs poorly in recognition but generates less background mistakes.

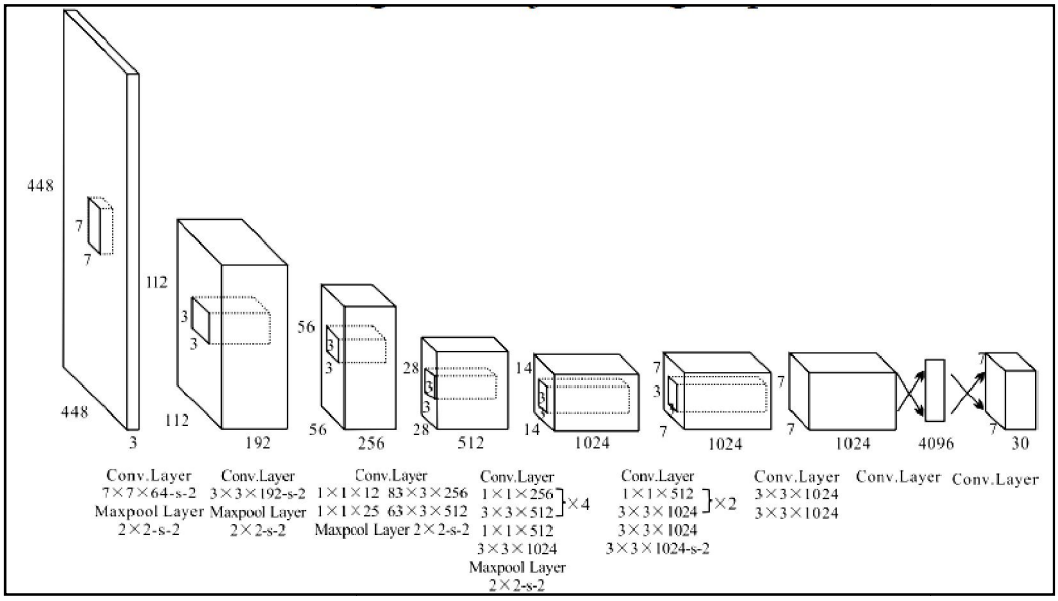

Figure 5. YOLOv1 architecture

YOLOv2

Redmon presented the YOLOv2[11] model in 2016. Improving recall and localisation while preserving classification accuracy is the major objective. The new fully convolution feature extraction network Darknet-19, which has five maximum pooling layers and a total of 19 convolutional layers, is used by YOLOv2. Recall and accuracy are greatly increased by using multi-scale training, k-means clustering on the training set bounding box, anchor box mechanism, batch normalisation layer added to the convolutional layer, and dropout eliminated. Nonetheless, there is still room for improvement in the identification of tiny and high overlap targets.

YOLOv3

By far the best balanced object identification model in terms of both detection accuracy and speed is Redmon's YOLOv3[12]. In terms of category prediction, YOLOv3 primarily consists of converting the initial single-label multi-classification into multi-label multi-classification and substituting a logistic regression layer for the original softmax layer that was used for single-label multi-classification.

The model makes predictions simultaneously by combining many different scales. It uses an upsampling fusion approach that is comparable to FPN before merging three scales, which greatly enhances the identification of tiny objects. This model's network structure uses the Darknet-53 deeper feature extraction network. The detection accuracy has not increased considerably, particularly when IOU>0.5, despite the fact that the YOLOv3 model further increases the detection speed and the detection impact of tiny objects.

SSD

The SSD[13] model was put out by Liu in 2016. The anchor box notion from the Faster R-CNN detection model and the regression idea from the YOLO method are both included into the model. The SSD model suggests using both the

IJETIR
ISSN (Online) 2583-0554
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Volume 4, Issue 2, February 2024
Impact Factor: 5.731

bottom and top level feature maps for detection in order to enhance the effectiveness of multi-scale object detection. The last two fully linked layers of the VGG basic architecture are swapped out for convolutional layers. SSD makes use of the RPN network's anchor mechanism. SSD on an Nvidia Titan X scores 74.3% mAP on VOC2007 at 59 frames per second. Nevertheless, the SSD's poor classification performance for tiny targets and the independence of feature maps at different scales allow boxes of various sizes to simultaneously detect the same item.

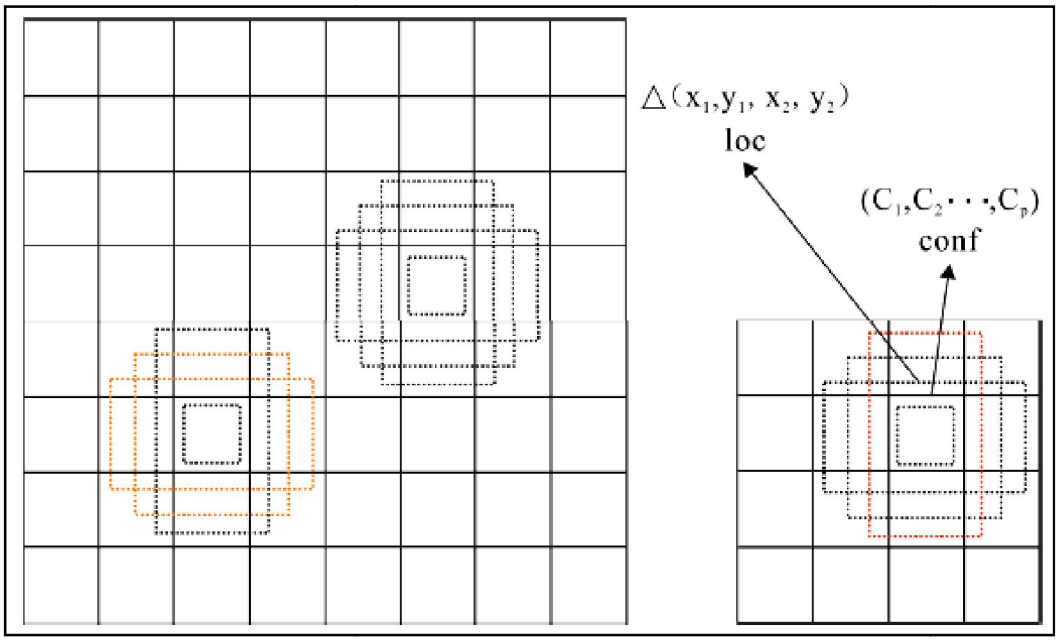

Figure 6. SSD architecture

YOLOv4

Alexey Bochkovskiy suggested the YOLOv4[14] in 2020, and it sets a new standard with the optimal speed-accuracy ratio. Theoretically, YOLOv4 is not that novel. Based on the original YOLO detection framework, it includes Weighted Residual Connection, Cross Stage Partial connection, Cross small Batch Normalisation, Self adversarial training, Mish activation, Mosaic data augmentation, DropBlock, and CIou. Based on the selection of CSP Darknet53 as the backbone network, an SPP module was added in order to extend the receptive field and distinguish the most significant context elements. In the meanwhile, YOLOv4 adheres to the head structure of YOLOv3 and employs PANet as the path aggregation mechanism rather than FPN, as used in YOLOv3. The YOLOv4 is faster and more accurate than the YOLOv3, increasing both by 20% and 10%, respectively.

IV. DATASETS AND PERFORMANCE COMPARISON OF VARIOUS ALGORITHMS

Dataset

The term "artificial intelligence" was first used in 1956. However, the apex of artificial intelligence did not arrive until 2012. The major causes of this are the growth in processing power, the amount of data, and the development of machine learning techniques. The expansion in data volume has a direct bearing on the development of detecting systems. This is due to the fact that datasets are required for both algorithm assessment and performance tests, and they also serve as a strong catalyst for the advancement of the detection techniques study area. Table 1 displays the parameters of popular public data sets.

TABLE I. PUBLIC DATA SET AND ITS PARAMETERS

Dataset	Amount	Sort	Size/Pixel	Year
Caltech101[18]	9145	101	300×200	2004

IJETIR
ISSN (Online) 2583-0554
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Volume 4, Issue 2, February 2024
Impact Factor: 5.731

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

Performance comparison of various algorithms

Single-stage and two-stage detection method comparisons and statistics are shown in Table 2.

TABLE II. COMPARISON OF OBJECT DETECTION ALGORITHMS

Method	Backbone	Size/Pixel	Test	mAP/%	fps
YOLOv1	VGG16	448×448	VOC 2007	66.4	45
SSD	VGG16	300×300	VOC 2007	77.2	46
YOLOv2	Darknet-19	544×544	VOC 2007	78.6	40
YOLOv3	Darknet-53	608×608	MS COCO	33	51
YOLOv4	CSP Darknet-53	608×608	MS COCO	43.5	65.7
R-CNN	VGG16	1000×600	VOC2007	66	0.5
SPP-Net	ZF-5	1000×600	VOC2007	54.2	-
Fast R-CNN	VGG16	1000×600	VOC2007	70.0	7
Faster R-CNN	ResNet-101	1000×600	VOC2007	76.4	5

V. CONCLUSION

Object identification is one of the most fundamental and difficult issues in computer vision, and it has attracted a lot of attention lately. Although deep learning-based detection techniques have been extensively used in several sectors, there are still some issues that need to be investigated:

1) Reduce the dependence on data.
2) To achieve efficient detection of small objects.
3) Realization of multi-category object detection.

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.
[11] Redmon, J., Farhadi, A. YOLO9000: better, faster, stronger. In: Computer Vision and Pattern Recognition. Hawaii.2017, pp. 7263-7271.
[12] Redmon, J., Farhadi, A. (2018) Yolov3: An incremental improvement. arXiv: Computer Vision and Pattern Recognition.
[13] Liu, W., Anguelov, D., Erhan, D., et al. SSD: Single Shot MultiBox Detector. European Conference on Computer Vision, 2016, pp. 21-37.
[14] Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv: Computer Vision and Pattern Recognition, 2020.
[15] 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.
[16] 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.
[17] Lin T Y , Maire M , Belongie S , et al. Microsoft COCO: Common Objects in Context. European Conference on Computer Vision, 2014, pp.740-755.
[18] 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.
[19] 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.
[20] 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.