Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2023.038173

Article

A Progressive Approach to Generic Object Detection: A Two-Stage Framework for Image Recognition

Muhammad Aamir[1], Ziaur Rahman[1,*], Waheed Ahmed Abro[2], Uzair Aslam Bhatti[3], Zaheer Ahmed Dayo[1] and Muhammad Ishfaq[1]

[1]College of Computer Science, Huanggang Normal University, Huanggang, Hubei, 438000, China
[2]FAST School of Computing, National University of Computer & Emerging Sciences (NUCES), Karachi Campus, Karachi, 75030, Pakistan
[3]School of Information and Communication Engineering, Hainan University, Haikou, 570228, China
*Corresponding Author: Ziaur Rahman. Email: ziaurrahman167@yahoo.com

Received: 30 November 2022; Accepted: 15 March 2023

Abstract: Object detection in images has been identified as a critical area of research in computer vision image processing. Research has developed several novel methods for determining an object's location and category from an image. However, there is still room for improvement in terms of detection efficiency. This study aims to develop a technique for detecting objects in images. To enhance overall detection performance, we considered object detection a two-fold problem, including localization and classification. The proposed method generates class-independent, high-quality, and precise proposals using an agglomerative clustering technique. We then combine these proposals with the relevant input image to train our network on convolutional features. Next, a network refinement module decreases the quantity of generated proposals to produce fewer high-quality candidate proposals. Finally, revised candidate proposals are sent into the network's detection process to determine the object type. The algorithm's performance is evaluated using publicly available the PASCAL Visual Object Classes Challenge 2007 (VOC2007), VOC2012, and Microsoft Common Objects in Context (MS-COCO) datasets. Using only 100 proposals per image at intersection over union (IoU) = 0.5 and 0.7), the proposed method attains Detection Recall (DR) rates of (93.17% and 79.35%) and (69.4% and 58.35%), and Mean Average Best Overlap (MABO) values of (79.25% and 62.65%), for the VOC2007 and MS-COCO datasets, respectively. Besides, it achieves a Mean Average Precision (mAP) of (84.7% and 81.5%) on both VOC datasets. The experiment findings reveal that our method exceeds previous approaches in terms of overall detection performance, proving its effectiveness.

Keywords: Deep neural network; deep learning features; agglomerative clustering; localizations; refinement; region of interest (ROI); object detection

1 Introduction

Object Detection (OD) is a critical issue in machine and computer vision. Numerous applications require OD to be binding and a significant component. Environmental reasoning applications and systems frequently use it in various reasoning tasks that rely on visual input [1,2]. As a result, a diligent study in this domain is perpetually active and significant. OD seeks to automatically locate and classify various objects in an image [3]. The detecting process can be harmed by multiple variables, including poor image quality, noise, and interference in the image's background [4]. Developing a system that offers hi-tech detection performance is a demanding task. OD has been extensively explored over the years, and several ways for efficiently detecting an object in images have been developed to address these issues [5,6].

Generic OD can be done via a region-based proposal approach or regression analysis [7]. Region-based OD systems consist of two stages, i.e., generating proposed object positions within an image and applying a classifier to these sampling locations to determine the object's category. Most of these systems begin by developing proposed object locations and then classifying them using object classifiers. Enhancing both localization and classification is necessary to attain the desired detection performance. Improving one of these stages will not achieve high detection accuracy [8]. The generation of object proposals is a technique that seeks to construct a bounding box around the parts of the image that include the objects of interest. The methods for generating object proposals are commonly classified into three categories: grouping-based, deep learning-based, and window scoring-based [9]. The major goal of the object proposals generation process is to reduce the number of computations required to determine the pixel-by-pixel similarity between the required object and the whole image while still focusing on the candidate object proposals believed to contain the desired object. The performance of the object proposal generators can be determined by the number of proposals generated and the recall of intended objects. An ideal object proposal generator obtains high-recall and produces lesser proposals [10].

Over the years, numerous approaches have been presented to obtain high-quality proposals with increased performance [11–26]. Nonetheless, some implausible limitations would be apparent when using these techniques for OD, including low localization accuracy, the absence of a scoring mechanism, a high computational cost, low precision, class-dependent, object-specific, redundant, and an excessive number of proposals. These constraints may significantly impair performance in the second stage, reducing overall detection efficiency. In the second step, regions are classed to identify the object's category; in recent years, detection systems' classification accuracy has been boosted via the development of deep learning networks. Researchers have spent considerable effort developing deep learning architectures capable of performing robust object classification. The Convolutional Neural Network (CNN) and its variants are robust deep learning architectures widely employed in OD's classification stage [27–32]. Numerous regularization strategies have been developed to enhance these networks' classification performance [33,34]. However, these networks have constraints, including model size, computational cost, and memory usage. They are redundant and problematic concerning specific proposal generation systems.

Current proposal-based object detectors leverage the promising properties of CNNs to determine the localization and objectiveness of each proposal [35–41]. While we analyze the ramifications and features of typical two-stage OD approaches, one cannot disregard single-stage methods. The practitioners have devoted considerable work to developing single-stage and proposal-free algorithms that are favorable and applicable to a wide variety of real-time applications [42,43]. Compared to the two-stage detectors, these approaches are reasonably quick but produce lower detection accuracy.

According to the authors' knowledge, based on these research findings, the issue of OD in natural images is still in its infancy in the experimental results and requires appropriate consideration. As a result, the authors present a two-stage method for generic OD. The proposed methodology represents a significant improvement over previous techniques.

The main contribution of this research is summed up as follows:

- ❖ Proposing a method for effectively generating a small number of high-quality object locations in imagery while avoiding repetitive locations.
- ❖ Proposing a refining network to reduce the volume of proposals significantly.
- ❖ Proposing a system that enhances the overall performance of detection and classification tasks.
- ❖ Comparing the proposed method to previously implemented generic OD methods. The proposed method demonstrated superior detection performance compared to existing methods.

The rest of the manuscript is structured as follows: Section 2 examines previous research that is relevant to the proposed study, Section 3 describes the proposed technique in detail, Section 4 presents the experimental results and their analysis, Section 5 contains the discussion, and Section 6 presents with the conclusion and future work.

2 Related Work

Detecting objects in images is a crucial challenging task in machine vision. It is developing, achieving enormous economic success, and establishing a foothold in virtually all industries. Contemporary research on this topic continues to yield fresh perspectives and prospects for future study, with several strategies being presented to reach state-of-the-art detection performance [44–46]. To learn more about how the brain processes visual information, it's helpful to know how objects are recognized. The human brain can efficiently detect, process, and interpret visual information. A sizable portion of the individual's brain is devoted to visual information processing. Compared to human intelligence, systems could not sense or process data adequately when confronted with the following demanding factors: varying viewpoints and perspectives, illumination, small object scale, deformation, occlusion, rotation, and a high intra-class variation. These complex elements affect the object, reducing overall effectiveness and complicating the detecting work [47]. To increase the generalizability of OD algorithms, researchers must commit significant effort to modify and improve detection systems in response to specific conditions. This section delves deeply into the conventional approaches for OD based on deep learning. These detectors attracted substantial attention due to their demonstrated ability to perform well in various detection tasks.

Region-based Convolutional Neural Network (R-CNN) is the first to demonstrate that a CNN-based model can outperform a classic Histogram of Oriented Gradients (HOG) based model. The network generates region proposals from the top to the bottom of the image using Selective Search (SS) and then classifies them to forecast their category. Using the VOC dataset [48], the network improved detection performance and achieved an mAP of 53.7%. It eliminates many simple negatives before training, which accelerates learning and minimizes the frequency of false positives. Regardless of how far this network has advanced, it remains limited due to redundant feature computations from highly overlapped regions. The training and testing take excessive time, resulting in an abnormally slow detection speed. Another primary risk of employing R-CNN is their fixed-size input. Spatial Pyramid Pooling Network (SPP-net) were then introduced to extend the capabilities of R-CNN, which already allows variable input sizes and share processing. The critical role of the SPP-net is the insertion of an SPP layer, which enables a CNN to generate a predetermined length feature vector representation regardless of the size of the image's an important region without rescaling. SPP-net is more efficient and faster than R-CNN and attained an mAP of 59.2% on VOC2007 imagery. Regardless of how effectively it increased detecting speed and accuracy, there are still specific weaknesses connected with them. The learning scenario is multiphase, consumes a significant amount of disk space, and relies mainly on its Fully Connected (FC) layers for OD by completely disregarding all prior layers.

Girshick et al. later enhanced detection efficiency by introducing the Fast-R-CNN, a multitask learning detector capable of generating preset length feature vectors from a feature map like the SPP-net. Compared to R-CNN, which processes each image region separately, this technique processes the entire image simultaneously. Fast R-CNN extracts region-specific information using the ROI pooling. It is a subset of SPP in that it partitions the underlying proposal using a single scale and propagates the error signals to the convolution kernels via backpropagation. The technique outperforms R-CNN and SPP-net in computational complexity and memory usage. The method required far less forward computation than R-CNN and significantly accelerated the training process. Even though Fast R-CNN successfully integrates R-CNN and SPP-net's advantages, the system's detection speed is still restricted. It employs a SS method for extracting proposals from images, which is computationally expensive. While substantial progress has been achieved in learning detectors, proposal generation has relied chiefly on conventional techniques such as SS and EdgeBox (EB), which depend on trivial visual signals and are difficult to train using a data-driven approach. To address these concerns, Ren et al. developed Faster R-CNN shortly after Fast R-CNN in 2015. Faster R-CNN consists of deep CNN and Fast R-CNN detector, responsible for proposal generation and classification, respectively.

The model solves the limitations of previous techniques, such as limited speed and high computing costs. The model substitutes SS with a powerful Region Proposal Network (RPN) to extract the regions in the image. The method offers fewer proposals; it takes a long to attain convergence. Since it achieved an mAP of 73.2% on the VOC2007 images and 70.4% on the VOC2012 dataset, Faster R-CNN can be deemed the first end-to-end and near-real-time deep learning detection model. The network outperforms Fast R-CNN in terms of detection speed. Nevertheless, subsequent detection phases experience some computational idleness. Deep layer features are significant semantically but insignificant regionally, whereas shallow layer characteristics are important semantically but not so much regionally.

Lin et al. investigated this phenomenon and proposed a Feature Pyramid Network (FPN), a classifier incorporating deep and shallow layer features for detection and recognition at many scales. The key idea is to reinforce spatially particular features with meaningful semantic features at deeper levels. Including this FPN into a standard Faster R-CNN model, significantly improved detection performance on the MS-COCO dataset, with an mAP of 59.1%, is achieved. The FPN has become a critical component of current detectors. He et al. developed the Mask R-CNN, a straightforward, intuitive, easy-to-use model based on pixel-level image segmentation. The approach achieved a high rate of speed and precision (high classification accuracy, high detection accuracy, high instance segmentation accuracy, etc.). The network is an extension of the Faster-R-CNN, which enhanced detection accuracy dramatically. Although the model is relatively stable, it requires considerable memory and is slow in detecting objects, making it unsuitable for real-time applications. The limitation of Faster R-CNN is that it computes a feature map for an input image and then constructs region-based feature vectors from it, dispersing feature extraction computation across the image's regions. These additional computations may be costly, as each input image may have hundreds of proposals. Dai et al. developed a Region-based Fully Convolutional Network (R-FCN) to eliminate this additional calculation expense. The model significantly enhances detection performance, and the model's shared convolution directly impacts the entire image. The model is highly accurate compared to previous approaches and beats them when speed is irrelevant. To improve detection performance, R-CNN is further extended into a Cascade R-CNN [49]. In reality, cascading can be helpful for any two-stage object detector. Several variants were also developed to ensure the detection process is as accurate as feasible, each with a somewhat different architecture [50–56].

To address the issue of slow detection speed, the one-stage algorithms You Only Look at Once (YOLO) and Single-Shot Detector (SSD) have been announced to predict the position and classification of objects. Numerous modifications have been proposed to improve the overall accuracy, including [57–62]. The YOLO method creates several grids on the images to perform localization and classification, which results in increased detection accuracy. But it cannot detect small objects due to a lack of low-level high-resolution information. For small objects, the approach yields unsatisfactory detection results. SSD is developed to address issues with YOLO systems. It identifies objects by combining low-level and high-level feature maps with high-resolution information. This approach is highly accurate and effective when dealing with small objects. Yet, the model is still subject to refinement. It was noted that current detectors heavily rely on anchors regardless of whether the detection is one-stage or two-stage. Typically, classic techniques generate anchors using RPN. These anchors are then immediately categorized and regressed. The quantity and shape of anchors significantly affect the effectiveness of OD algorithms. Nonetheless, two-stage algorithms are slower than one-stage algorithms but more accurate in their detection.

3 Proposed Method

This research aims to provide a helpful model for OD in images. This section discusses each significant stage in depth. Fig. 1 depicts the proposed model's architecture. Our proposed methodology is comprised of the crucial phases described below:

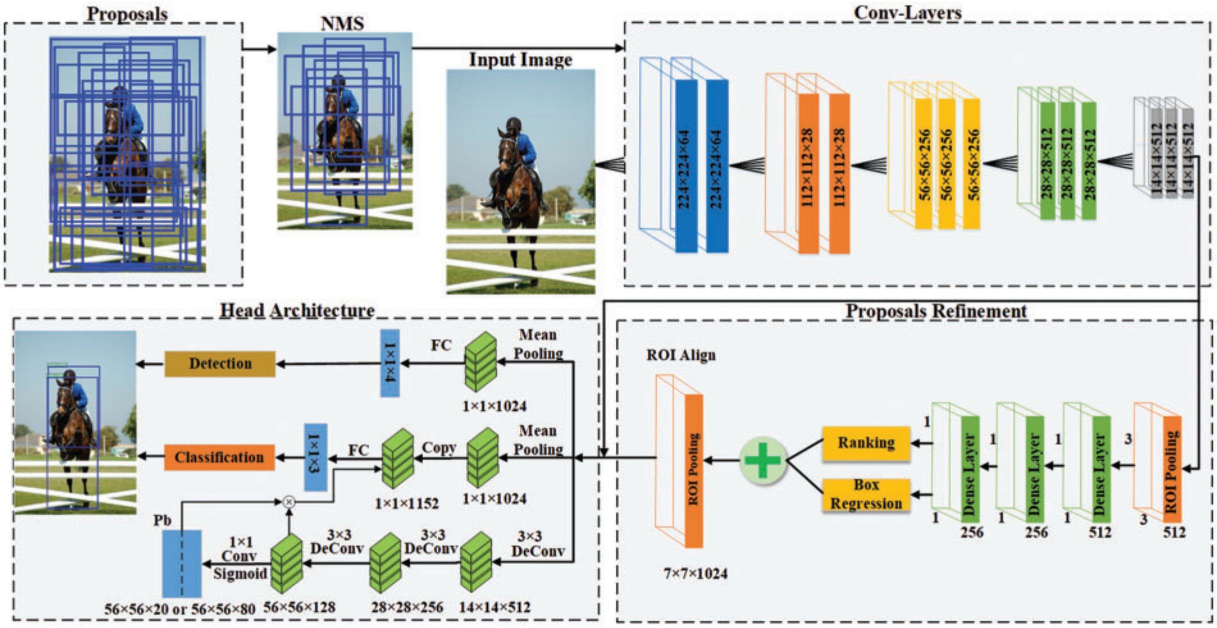

Figure 1: The proposed method's architecture

3.1 Object Proposals

The initial critical stage is to produce high-quality, class-independent, and sparse object proposals. Based on these earlier study findings, the small but high-quality collection of proposals can significantly improve OD performance. Previous strategies are insufficient to generate an adequate quantity of high-quality proposals. We started by segmenting the image to establish a set of initial regions. Segmentation enhances OD performance. Compared to pixels regions containing more information, obtaining object proposals using region-based characteristics is a solid approach. This work uses a technique developed in Ref. [63] to quickly and accurately generate a collection of initial regions. Each acquired region is referred to as a cluster. Then, surrounding regions are grouped according to their similarities using a hierarchical clustering-based bottom-up grouping approach. After determining the similarities between nearby regions, the most similar neighbors are joined to produce a single region. Then, similarity values between the two previously combined adjacent regions are assessed, and comparable regions are combined into a single region. Iteratively comparable regions are grouped until all similar regions are consolidated into a single region to make an image. Thus, the regions are clustered based on four complementary characteristics: color, size, gap, and texture similarities.

Our goal is to obtain as many proposals as feasible. The search for regions is varied, and more regions are acquired by using a clustering approach based on multiple color spaces, modifying the starting regions, and generating region similarities. The acquired regions have been sorted, and duplicates have been deleted. At this stage, the regions formed through grouping serve as proposals. After obtaining the proposals, the following step is to score and rank them. This is accomplished by employing the structural edge detector to extract the edges from the source image. Compared to other edge detectors, this one is reasonably quick and exact. Then, these edges are linked together depending on their similarity with adjacent edges. Eight adjacent edges with greater than pi/2 variance in orientation are connected to form edge groups. We then estimated the affinities between adjacent groups using their mean positions x_i and x_j and orientations θ_i and θ_j. To speed up the process, only affinities with a value larger than the 0.05 criterion are retained; all other affinities are removed.

$$a\left(S_i, S_j\right) = \left|\cos\left(\theta_i - \theta_{ij}\right)\cos\left(\theta_i - \theta_{ij}\right)\right|^{\gamma} \tag{1}$$

It is possible to tweak the affinity's sensitivity to orientation changes by changing the value γ; nevertheless, $\gamma = 2$ it is commonly used in practice.

We calculated the score for our proposals using these edge groups and their affinities. Each group's continuous value $w_b\left(S_i\right)$ is computed to determine whether the particular edges S_i are contained inside the candidate bounding box b. Suppose it is not entirely included inside the box $w_b\left(S_i\right) = 0$. In that case, the mathematical formula for determining if a collection of edges S_i is entirely encompassed within the candidate bounding box b is as follows:

$$w_b\left(S_i\right) = 1 - max_t \prod_{j}^{|T|-1} a\left(t_j - t_{j+1}\right) \tag{2}$$

where "a" is the affinity between two edge groups, and "t" represents the ordered path of edge groups, has a length of $|T|$, begins at $t_1 \in S_b$, and finishes at $t_{|T|} = S_i$, If no such path exists, then $w_b\left(S_i\right) = 1$. The scoring function may be represented as, based on the values derived using Eq. (2).

$$\frac{\sum_i w_b\left(S_i\right) \mathrm{m_i}}{2\left(b_w + b_h\right)^k} \tag{3}$$

Where “b_w” signifies the box’s width, “b_h” denotes the box’s height, “k” represents the bias value for larger boxes, which is 1.5, and where m_i is the total magnitude of all of the edges that belong to group S_i. Next, these resultant proposals are sorted by their Eq. (3) scoring system and fed into the backbone network, which optimizes them even more, enabling them to execute refinement tasks more effectively. The scored proposal has multiple duplicates (i.e., significantly overlapping proposals). We applied Non-Maximum Supersession (NMS) to improve proposals and select those with the highest objectivity score.

3.2 Proposals Refinement

As illustrated in Fig. 2, proposals should be refined to achieve high-quality detection performance. Numerous methods require that proposals to be of the highest quality and fewest in quantity. The refinement system fine-tuned the proposals generated in the previous stage. Following that, high-quality proposals are submitted for classification. Our technique is structured so that our detector’s refinement and classification components can exchange convolutional features to achieve resilient performance.

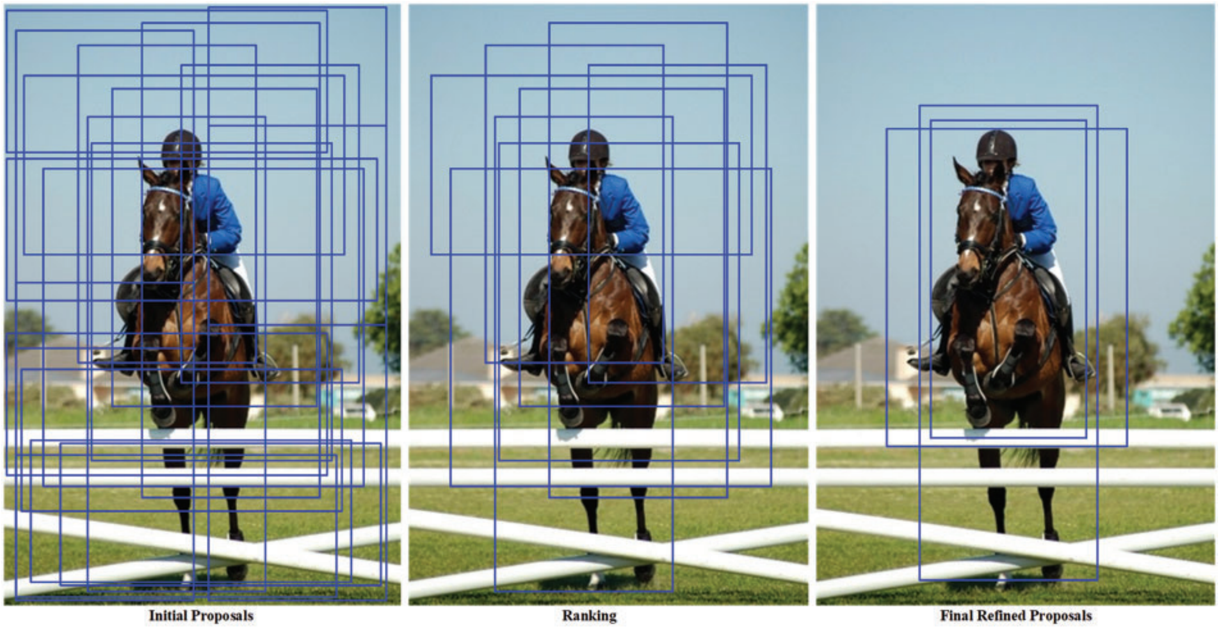

Figure 2: The initial proposals and their refinement

We extract the image’s deep features using a pre-trained Visual Geometry Group CNN (VGGNet) [32]. VGG considerably enhanced both precision and speed. This was mostly due to the improved depth of the model and the use of pre-trained models. Non-linearity is always an advantage in deep learning, and it increases as the number of layers with smaller kernels increases. VGG introduced a variety of structures based on the same concept. This expands our application’s architectural possibilities. It is one of the most prevalent approaches for image classification and is straightforward to implement with transfer learning. The network is updated after the 13th layer. Two new branches are added to refine and classify proposals. The model is supplied input from the first stage’s proposals and corresponding natural imagery. The input image and its corresponding proposals are traversed via the first to the thirteenth layers to extract features. We wish to minimize computing costs and time and thus use the proposal refinement concept presented in [37,64]. Following the thirteenth layer in our network, we apply ROI pooling to resize the feature map 512 × 3 × 3, which serves as the starting point for our proposal refinement. To generate the feature vector for the object proposal, the convolutional feature map is passed through one 512-d and two 256-d FC layers. Finally, a ranking branch is included as an FC layer for recalculating the proposal's score. This layer has two output neurons representing the likelihood of an object being present. Besides, another subfield of box regression incorporates an FC layer for determining the starting positions of proposals and forecasting the box regression values.

We may rank the revised proposals by their score and choose a few with the highest objectivity score to achieve final proposals. Initial proposals are allocated a binary class label during the network's training phase to signify whether they are objects. The loss function is defined as follows:

$$L_{obj}\left(p,u\right) = -\left[1_{\{u=1\}}\log p_1 + 1_{\{u\neq 1\}}\log p_0\right] \tag{4}$$

where "p" is the value computed using SoftMax from the two outputs of a FC layer, p_1 and p_0 are predicted probabilities per ROI, and "u" denote the current box's label. The box regression layer is used to learn the coordinate offsets. The coordinates are parameterized as follows:

$$\begin{aligned}
&t_x = (x - x_{in})/w_{in}, t_y = \left(y - y_{in}\right)/h_{in},\\
&t_w = \log\left(w/w_{in}\right), t_h = \log\left(h/h_{in}\right),\\
&v_x = (x^* - x_{in})/w_{in}, v_y = (y^* - y_{in})/h_{in},\\
&v_w = \log\left(w^*/w_{in}\right), v_h = \log\left(h^*/h_{in}\right)
\end{aligned} \tag{5}$$

where "h", "w", "x", and "y" represent the height, width, and candidate box's center coordinates. The quantities x^*, x_{in}, and "x" represent the ground truth box, input box, and predicted box. The same definitions apply to variables "y", "h", and "w". Variable "v" refers to the regression target and "t" denote the predicted tuple. The box regression loss is numerically described as follows:

$$\begin{aligned}
&L_{reg} = \sum_{i\in\{x,y,w,h\}} smooth_{L_1}\left(t_i - v_i\right),\\
&smooth_{L_1}\left(x\right) = \begin{Bmatrix} 0.5x^2\, if \left|x\right| < 1 \\ \left|x\right| - 0.5\, otherwise \end{Bmatrix}
\end{aligned} \tag{6}$$

where smooth $L_1\left(x\right)$ is the eponymous loss function for regression. As a result, the joint loss function has the following definition:

$$L\left(p,u,t,v\right) = L_{obj}\left(p,u\right) + \lambda \cdot 1_{\{u=1\}} L_{reg}\left(t,v\right) \tag{7}$$

where "λ" is a balancing parameter; in this case, it is set to one.

3.3 Classification

The network's FC layers required fixed-size input to accomplish subsequent operations, one of the critical obstacles to OD. Since the proposals produced would vary in size and shape. All generated proposals must be translated into a specific size or form. ROI pooling is applied after receiving the updated proposals to obtain a fixed-length feature vector 7 × 7. ROI pooling's output size does not depend on the number of proposals or the input feature map but rather on the number of sections into which proposals are partitioned. Implementing ROI pooling accelerates computation and allows all generated proposals to utilize the same input feature. This also significantly enhances the overall object detection performance. Convolutional features from the input image are incorporated into all produced proposals. This improves overall detection accuracy significantly [37]. These proposals are transmitted with the proposed network's final component to accomplish the needed objectives. Mean pooing is used to minimize the size of the proposal feature map to 1, and an FC layer of size 1 × 1 × 4 is added to display the object's bounding box and score. The feature map of the proposals from the previous stage is of low resolution. Three de-convolutional layers and one convolutional layer are added to increase the resolution. The sigmoid function is applied to the output of deconvolutional layers to derive object class probability. The classification branch receives the deconvolutional feature maps. The feature map of the size 1 × 1 × 1152 is obtained. This 1152 feature channel sum is obtained from the backbone's 1024 feature channels and the deconvolution output's 128 feature channels. Combining feature channels from two sources increased the classification performance significantly. The probability "p" of each output class "u" is determined using the SoftMax activation function, which is specified as,

$$L_{classificaiton}(p, u) = -\log(p_u) \tag{8}$$

4 Evaluation and Results

Based on prior research on generic OD, the effectiveness of the proposed technique is examined using the difficult VOC2007, VOC2012, and MS-COCO datasets [65–67]. These are the most frequently used and well-known OD datasets. The computer system used for the experiment is equipped with two RTX2080Ti Graphics Processing Units (GPUs) and has a RAM of 32 gigabytes. To enable the creation and reproduction of network models utilized in our research, PyTorch, a deep learning framework, is integrated with Python and Linux. The proposed network is trained using the Stochastic Gradient Descent (SGD) method. Each SGD mini-batch was developed from an image, yielding 512 boxes as a training sample. Each batch contains 512 training examples uniformly distributed between positive and negative data. Positive samples were defined as training boxes with an overlapping value of at least 0.7 with the ground truth boxes. A negative sample is defined as those boxes whose overlapping values with the ground truth are within the interval [0.1, 0.5]. The experiments are evaluated for a total of 32 epochs. The model layers are fine-tuned by setting a constant learning rate of 0.0001 for all epochs. We considered 512 object proposals of an image generated for each mini-batch to train our framework's detection module. In Fast R-CNN, 25% of proposals are positive if their overlap value with the ground truth boxes is 0.5. The maximum overlap value for negative samples is between [0.1, 0.5]. The detection model was trained using the top 2000 improved proposals, with the model learning rate set at a constant value of 0.0001 for all epochs. The proposed model has been verified against 100 high-quality proposals per image, whereas previous methodologies needed large numbers of proposals to demonstrate the model's usefulness.

To evaluate the overall detection performance in terms of localization and classification efficiency, the following generally known and widely used performance measures are adopted: mAP, DR, and MABO. Numerous state-of-the-art approaches have been chosen to compare the superiority and resilience of the proposed method [11–26,35–39,42–43,68–80]. The proposals generated through these methods were entered into the detection module to ensure the proposals' quality during the entire detection task. To evaluate the proposed model's mAP to earlier models, the main basic techniques used can be seen in Table 2.

Tables 1 and 3 compare the proposed model's DR performance to existing approaches utilizing VOC2007 and MS-COCO datasets. The DR trends are displayed in Figs. 3, 4, 7, and 8. The DR is determined based on 100, 300, 500, and 1000 proposals received under two separate IoU conditions, particularly 0.5 and 0.7. The proposed technique has achieved the highest DR. Variation in IoU values on both datasets and the number of proposals boosts performance significantly. Despite producing high-quality proposals, many cutting-edge algorithms failed to achieve high recall due to a small number of candidate proposals. At the same time, fewer methods are relatively competitive but unable to produce high recall due to poorly matched proposals. The proposed method yielded significant recall performance of 93.17%, 94.75%, 95.2%, and 97.3% on VOC2007 across various proposal settings at IoU = 0.5. With the MS-COCO dataset at IoU = 0.5, the proposed approach attained recall results of 69.4%, 73.75%, 77.57%, and 82.66% approximately. At IoU = 0.7, the proposed approach completed approximate recall values of 79.35%, 81.15%, 84.4%, and 88.1% on VOC2007 imagery, respectively. With an IoU of 0.7 and MS-COCO Imagery, the proposed methodology provided recall values of roughly 58.35%, 62.97%, 68.66%, and 73.81% for 100, 300, 500, and 1000 proposals, respectively. None of the strategies investigated thus far has shown exceptionally high performance for either IoU threshold value. Our technique makes use of the intrinsic convolutional features throughout the network. The proposed technique obtained a 93.17% recall value with around 100 proposals per image at IoU = 0.5. In contrast, the DeepBox (DB), RPN and Deep Mask Zoom (DMZ), methods require 300, 500, and 1000 proposals per image to reach approximately the same performance. Similarly, the model has reached 79.35% at IoU = 0.7. Compared to the proposed model, the other models had the highest number of proposals yet achieved the same performance.

Table 1: DR performance for various proposals settings at IoU 0.5 and 0.7 VOC2007

Ref.	Methods name	Proposals (IoU = 0.5)				Proposals (IoU = 0.7)			
		100	300	500	1000	100	300	500	1000
[11]	Rantalankia	16.05	18.85	22.95	27.4	9.37	12.1	15.55	19.8
[12]	Geodesic object proposals (GOP)	61.32	64.75	69.9	74.6	37.22	39.7	43.15	47.3
[13]	Rahtu	63.55	66.1	70.1	74.7	47.22	49.9	53.3	57.3
[14]	Objectness	66.65	69.05	72.9	77.3	31.15	34	37.55	41.7
[15]	Binarized normed gradients (BING)	71.42	73.4	76.1	81.8	26.35	29.25	32.6	36.5
[16]	Random prim (RP)	72.4	74.55	77.95	82.4	46.07	48.65	52.45	56.8
[17]	SS	73.5	75.55	78.9	83.7	51.25	54.2	57.55	60.5
[18]	Cascade support vector machines (CSVMs)	76.25	78.6	82.15	86.1	27.45	29.95	33.25	37.4
[19]	Learning to propose objects (LPO)	76.95	79.3	82.9	87.4	50.37	52.55	55.7	59.9
[20]	EB	77.3	79.95	84.05	89.2	62.57	65.35	68.85	74.1
[21]	Multiscale combinatorial grouping (MCG)	83.5	85.5	87	89.7	62.02	64.75	68.15	73.2
[22]	Endres	85.05	87.05	88	90.2	61.35	64.05	67.3	71.4
[23]	DB	86.32	88.25	90.35	95.1	72.15	74.75	78.4	83.6
[24]	RPN	90.25	92.4	93.4	96.2	66.32	69.25	72.85	77.1
[25]	DMZ	91.5	93.05	93.85	96.7	72.95	75.75	79.85	84.7
[26]	Sharp mask zoom (SMZ)	92.15	94.65	94.3	96.9	76.02	78.95	82.5	86.9
Ours		93.17	94.75	95.2	97.3	79.35	81.15	84.4	88.1

Table 2: MABO Performance of the proposed approach on different proposals and mAP performance on 100 proposals per image on VOC2007 images

Methods	Proposals (MABO)				mAP using 100 proposals per image
	100	300	500	1000	
Rantalankia	23.2	28.85	32.7	36.6	11.3
Objectness	56	60.9	64.15	68.3	45.05
BING	57.05	61.95	65.6	71.7	44.45
GOP	57.85	63.1	66.75	72.9	28.2
Rahtu	58.85	63.9	67.35	73.3	42.8
CSVMs	59.6	65	68.75	74.8	46.15
RP	63.75	68.6	72.95	79.5	42.1
SS	65.45	70.85	74.95	81.2	44.15
LPO	66.7	71.85	76.1	82.4	44.6
EB	67.35	72.45	77.3	84.8	51.6
RPN	71.2	76.55	81.4	87.7	65.4
Endres	71.5	77.25	81.85	88.2	55.7
MCG	72.4	78	82.85	88.6	54.2
DB	72.85	78.4	82.7	89.4	60.55
DMZ	77.2	82.55	86.85	90.5	63.55
SMZ	78.25	83.45	88.1	91.8	64.5
Ours	79.25	85.15	90.1	92.7	71.9

Table 3: DR performance for various proposals settings at IoU = 0.5 and 0.7 using the MS-COCO dataset

Methods	Proposals (IoU = 0.5)				Proposals (IoU = 0.7)			
	100	300	500	1000	100	300	500	1000
BING	30.2	34.5	38.35	44.42	5.15	9.67	15.41	20.54
Rahtu	31.5	35.8	39.65	45.72	21.95	26.27	32.11	37.19
Objectness	32.9	37.25	41.07	47.16	12.25	17.02	22.63	27.82
SS	33.1	37.35	41.22	47.28	17.75	21.87	27.81	32.84
RP	38.7	42.95	46.82	52.88	22.25	26.82	32.53	37.67
EB	39.5	43.6	47.55	45.57	28.65	33.07	38.86	43.96
GOP	42.3	46.7	50.5	56.60	26.6	30.85	36.72	41.78
LPO	43.5	47.8	51.65	57.72	25.55	30.12	35.83	40.97
DB	48.9	53.2	57.05	63.1	35.45	40.07	45.76	50.91
MCG	51.8	56.2	60.0	66.1	35.25	39.92	45.58	50.75
RPN	66.5	70.75	74.62	78.685	47.5	52.05	57.77	62.91
DMZ	67.6	72.0	75.8	80.9	53.9	58.05	63.97	69.01
SMZ	68.1	72.35	76.22	81.28	54.95	59.07	65.23	70.15

(Continued)

Table 3: Continued

Methods	Proposals (IoU = 0.5)				Proposals (IoU = 0.7)			
	100	300	500	1000	100	300	500	1000
Ours	69.4	73.75	77.57	82.66	58.35	62.97	68.66	73.81

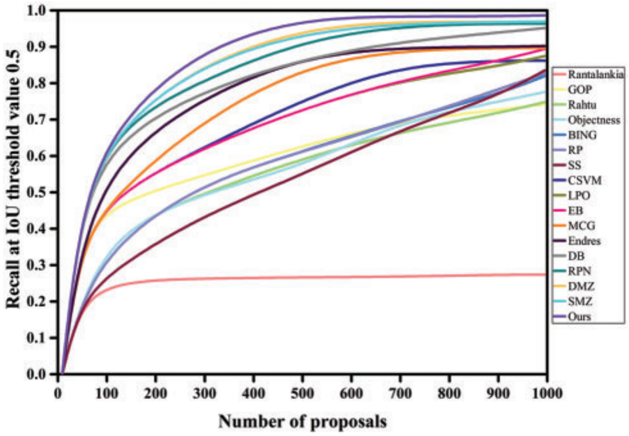

Figure 3: DR patterns for a variety of proposals at IoU 0.5 on VOC2007

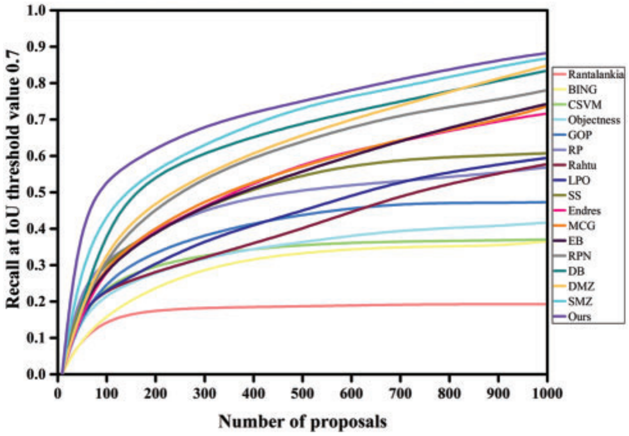

Figure 4: DR patterns for a variety of proposals at IoU 0.7 on VOC2007

Due to the greater diversity of our proposals, we can achieve a high recall with a smaller number of proposals. This enables the proposed model to be used for tasks requiring high recall with fewer proposals. Thus, the proposed strategy displayed significant robust performance on a smaller number of proposals while attaining comparatively acceptable results on a more substantial number of proposals, demonstrating the proposed method's efficacy. The results indicate that the proposed strategy is substantially more accurate than earlier methods, owing to the high quality and refinement of the presented object proposals. On the other hand, prior approaches are ineffective at achieving high recall with lower and higher proposals than the proposed strategy.

Tables 2 and 4 compare the proposed model's MABO performance to prior techniques on both imaginaries. The MABO relationship between the various methods is visualized in Figs. 5 and 9. The proposed model outperforms existing approaches regarding proposal accuracy. The proposed scheme achieves the highest MABO values of 79.25%, 85.15%, 90.1%, and 92.7%, respectively, for all proposal settings using VOC2007 images. Similarly, utilizing MS-COCO images acquired MABO performance of 62.65%, 66.27%, 71.46%, and 76.86%, indicating the proposed method's usefulness. The findings show that current techniques have a high recall but insufficiently yield high MABO. This is because of the larger volume of low-quality proposals. Due to the excellent quality of the proposals, the proposed method achieves acceptable performance using a small number of proposals. The fewer methods resulted in a MABO value that was greater than that predicted by our model. The heterogeneity in object classes and the lowered processing cost significantly enhance the proposed method's overall detection performance. The detection efficiency of the proposed method compared to competing methods is presented in Table 2. The mAP is estimated using only the top 100 proposals.

Table 4: MABO Performance of the proposed approach on different proposals using the MS-COCO dataset

Methods	Proposals (MABO)			
	100	300	500	1000
BING	32.25	36.62	41.43	47.02
Rahtu	33.35	37.62	42.48	48.05
Objectness	33.9	38.05	42.97	48.51
SS	35.9	39.75	44.82	50.28
EB	38.8	43.2	48	53.6
RP	40.75	45.07	49.91	55.49
GOP	43.1	47.7	52.4	58.05
LPO	44.2	48.65	53.42	59.03
DB	46.8	51.1	55.95	61.52
MCG	50.2	54.85	59.52	65.18
RPN	56.35	60.67	65.51	71.09
DMZ	60.75	65.17	69.96	75.56
SMZ	61.45	65.85	70.65	76.25
Ours	62.65	66.27	71.46	76.86

In contrast, previous approaches required many proposals to determine the model's efficacy. According to the evaluation, the proposed technique resulted in an mAP of around 71.9% for 100 high-quality and enhanced proposals. Fig. 6 illustrates the detection performance of several models contrary to the IoU overlap threshold, and our method achieved a maximum DR of 93.17% using only 100 proposals per image. The proposed technique outperformed EB, SS, BING, and RP in terms of mAP by the high margin of 20.3%, 27.75%, 27.45%, and 29.8%, respectively. They also achieve higher mAP values compared to other approaches.

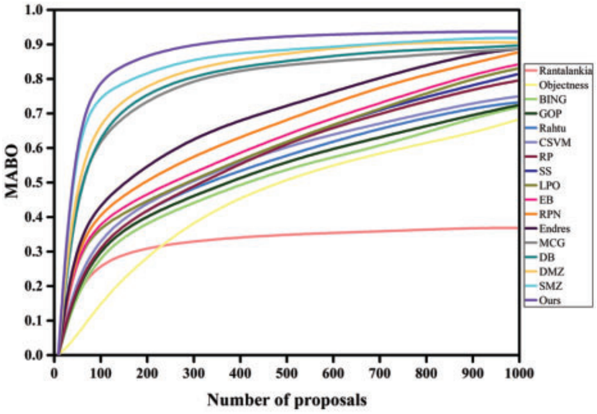

Figure 5: MABO performance of the proposed model on VOC2007

.

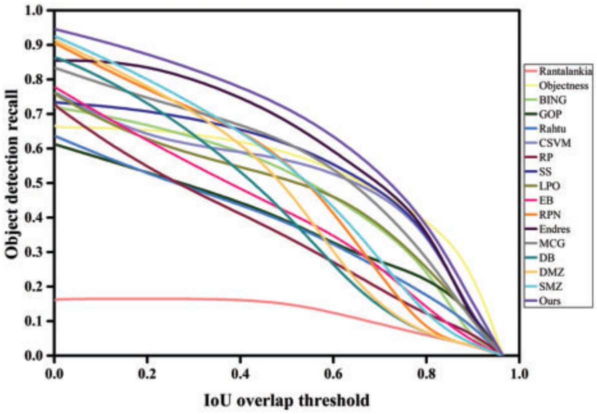

Figure 6: mAP performance of the proposed model for 100 proposals using VOC 2007

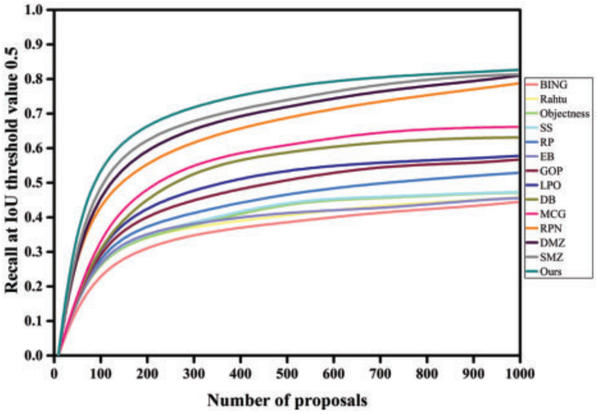

Figure 7: DR patterns for a variety of proposals at IoU 0.5 using the MS-COCO dataset

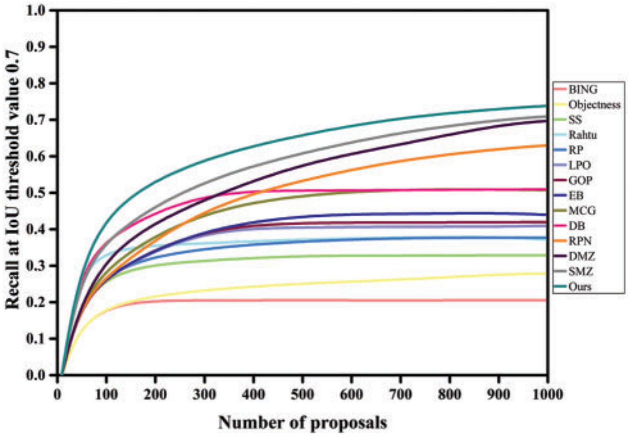

Figure 8: DR patterns for a variety of proposals at IoU 0.7 using the MS-COCO dataset

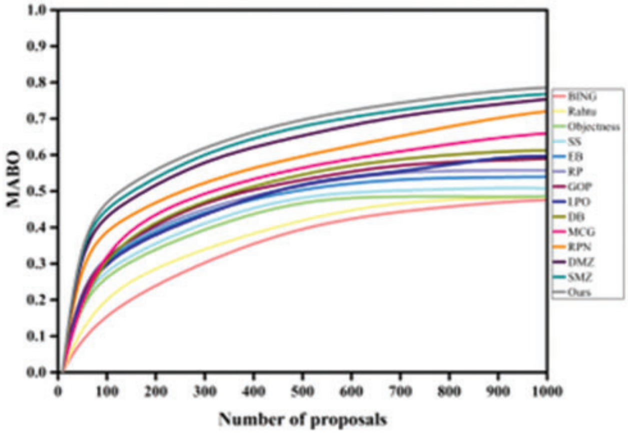

Figure 9: MABO performance of the proposed model using the MS-COCO dataset

To further validate the proposed model's efficiency, the overall efficiency in terms of overall mAP is tested using VOC2007 and VOC2012 datasets. As demonstrated in Table 5, compared to other current detectors on both datasets, the proposed network achieved mAP of 84.7% and 81.5%, respectively, which is greater than previous techniques. The proposed design generated high-quality, low-volume, and accurate proposals with high recall. Besides, it preserves feature information and achieves a high level of detection accuracy. Even though it is acknowledged as a critical restriction for many methods to enhance detection performance. To do this, convolutional features are shared across the whole network of the proposed model, which does not affect the proposal generation performance. The sharing of features enhances the overall detection performance greatly. The results indicate that past approaches to lower mAP aggravated the problem of low precision and may significantly hinder real-time performance. Due to the low detection performance, obtaining and meeting the real-time detection performance criterion was problematic. As a result, the results indicate that the proposed strategy is more efficient, performs better, and is more advantageous. The qualitative results of the proposed model are visualized in Fig. 10.

Table 5: Comparison of the proposed method's mAP performance with existing techniques to OD

Methods	Backbone architecture	Proposed year	Input size (Test)	mAP (%)	
				VOC2007	VOC2012
[35]	AlexNet	2014	600 × 1000	50.2	–
[35]	VGG-16	2014	Arbitrary	66.0	62.4
[36]	AlexNet	2014	224 × 224	63.1	–
[37]	VGG-16	2014	600 × 1000	70.0	68.4
[38]	VGG-16	2015	600 × 1000	73.2	70.4
[68]	VGG-16	2015	Multi-Scale	78.2	73.9
[39]	ResNet-101	2016	600 × 1000	80.5	77.6
[54]	VGG-16	2016	600 × 1000	79.2	76.4
[55]	VGG-16	2016	600 × 1000	76.3	
[56]	VGG-16	2016	600 × 1000	74.6	71.9
[69]	VGG-16	2016	600 × 1000	75.7	71.3
[70]	VGG-16	2016	600 × 1000	78.4	74.8
[71]	ResNet-101	2017	600 × 1000	82.6	–
[72]	ResNet-101	2017	600 × 1000	82.7	80.4
[73]	ResNet-101	2017	512 × 512	77.1	73.9
[74]	ResNet-101	2018	600 × 1000	82.4	81.1
[75]	ResNet-101	2018	600 × 1000	83.3	81.3
[76]	ResNet-101+	2018	Arbitrary	84.0	81.2
[42]	VGG-16	2016	448 × 448	63.4	57.9
[77]	DarkNet	2017	544 × 544	78.6	73.5
[43]	VGG-16	2016	300 × 300	77.2	–
[43]	VGG-16	2016	512 × 512	79.8	78.5
[59]	ResNet-101	2017	321 × 321	78.6	–
[59]	ResNet-101	2017	513 × 513	81.5	80.0
[78]	DenseNet	2017	300 × 300	77.7	76.3
[79]	VGG-16	2017	384 × 384	75.4	73.0
[80]	ResNet101	2019	512 × 512	78.7	–
Ours	VGG-16	2022	600 × 1000	84.7	81.5

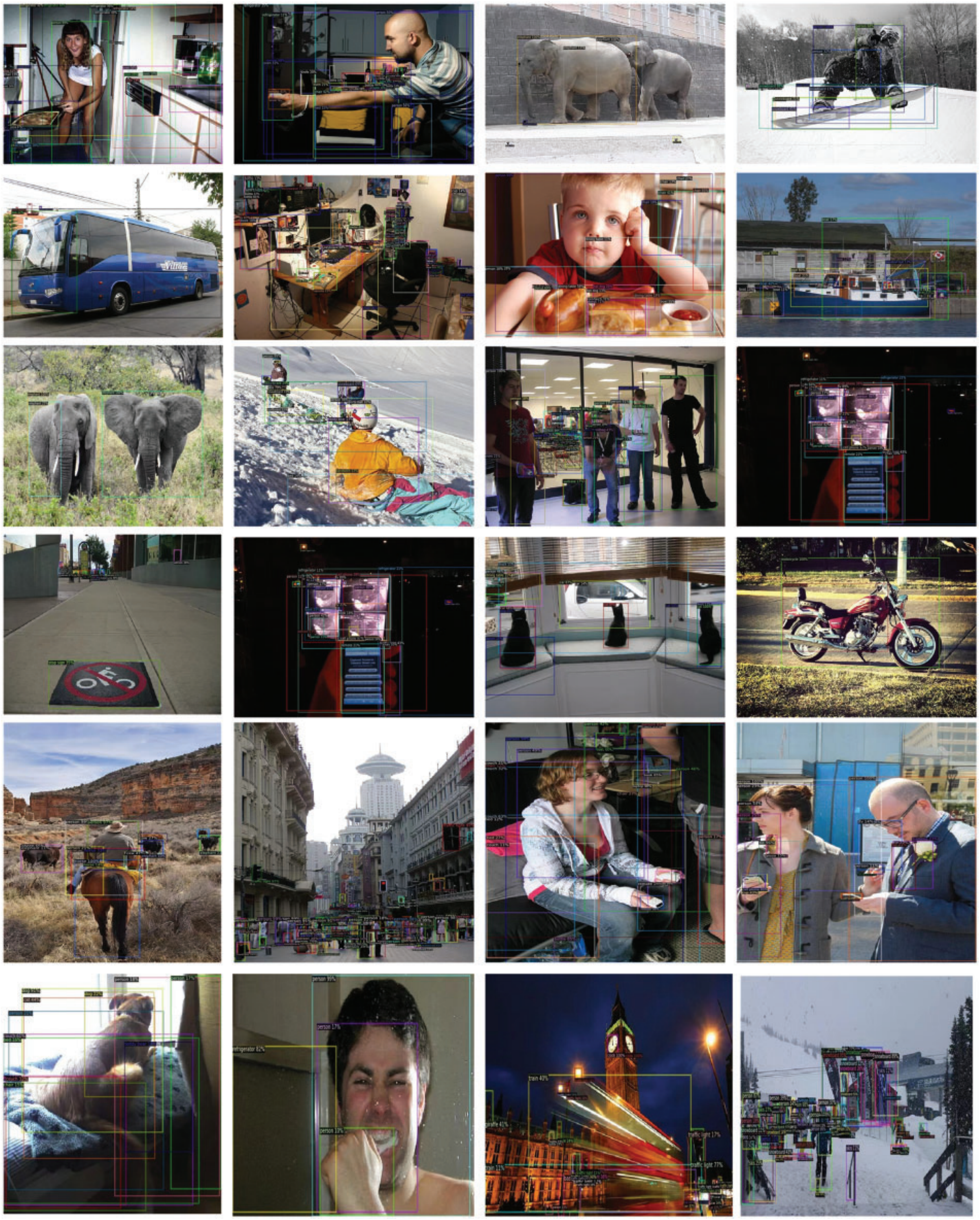

Figure 10: The qualitative results of the proposed model

5 Discussion

This study aims to present a better method for recognizing objects in images based on recent advances in OD. The proposed technique can significantly increase overall detection performance compared to existing approaches. While the authors assess and provide their experimental findings in Part 4 of this study, this section discusses the proposed method's significance and compares the results to previous studies (Tables 1–5). Compared to Rantanlankia and GOP models, the proposed technique produced a significant margin difference in all quality indicators over 100 proposals with varying IoU levels utilizing both datasets. The proposed method surpassed their competitors in terms of recall and detection efficiency despite employing a smaller and more diverse set of proposals, which is considered a drawback of such models. Direct control over proposal generation, scoring, and generating a higher proportion of high-quality and diversified proposals ensures that the proposed model performs optimally to achieve quality detection efficiency.

In contrast to Rahtu's study, their algorithm prioritizes initial top-rank proposals over top best proposals, lowering total detection performance. The proposed scheme performs significantly better in detection rate, MABO, and mAP. It generates a small number of high-quality proposals for performing high-quality detection. The author's technique outperforms the Objectness model at a high IoU threshold, resulting in robust detection performance. As the IoU threshold is increased, the efficiency of BING drops. This method underperforms for certain types of objects than the proposed technique. Compared to RP, their approach lacks a scoring system and a mechanism for controlled proposal generation. Due to the reduced proposal count and more concentrated proposal selection, the proposed technique beats these strategies significantly.

The study SS used no learning in the proposal generation process, yet the system outperforms other existing methods. Compared to their research, the proposed technique outperforms it by producing higher-quality and fewer proposals, significantly improving average detection performance. Our method scores, ranks, and refine proposals to increase detection efficiency and surpasses their model by a large margin in recall (20.4%), MABO (14.5%), and mAP (24.1%), which demonstrates their resilience when compared to their method. The proposed methodology outperforms CSVMs and significantly improves recall (17.7%), MABO (19.8%), and mAP (21.7%). The LPO generates insufficient high-quality proposals to improve overall detection performance. However, this strategy outperforms others in the recall, MABO, and mAP. The proposed scheme significantly outperforms theirs while maintaining a high level of performance across all quality indicators. The authors' analysis demonstrates that it is possible to achieve high-quality performance despite numerous proposals. The EB features a scoring approach to regulate proposal generation, and it does not achieve consistent detection performance due to the absence of additional proposal refining. The proposed method outperforms their strategy in the recall, MABO, and mAP.

Compared to MCG, the proposed model consistently beat its approach and superiority. The increased threshold value resulted in a significant improvement in their overall detection ability. The Endres model is slow and requires substantial processing power. This model is unable to detect every object across each image. Compared to the authors' method, their model fared poorly in recall, MABO, and mAP. Compared to DB, the authors' technique improved various performance metrics considerably. The proposed strategy generates a more significant number of superior proposals than they do. There is no way to reduce the number of proposals, making them more vulnerable to obtaining quality detection performance. Compared to RPN, DMZ, and SMZ, the proposed technique produces fewer proposals and generates fewer false positives, boosting final performance. These methods have similar results to the proposed solution; yet, the significant marginal difference in performance efficiency remains obvious. After acquiring high-quality proposals, these are incorporated into the OD process. A range of advanced object detectors is included to further evaluate the proposed model's detection efficiency. The comparison results are presented in Table 5. Notably, the proposed method achieved a greater detection precision than others due to generating fewer but higher-quality proposals. Due to improved recall and proposal accuracy, the proposed technique produced robust results for objects of various classes. Thus, the provided object detector's greatest mAP of 84.7% and 81.1% using VOC2007 and VOC2012 indicate its superiority over earlier detectors in the OD challenges.

Given that the fundamental objective of this research is to generate fewer, high-quality, and class-independent proposals, improve detection performance, and raise the usefulness of quality assessment criteria. The methodologies described above have many advantages and disadvantages. The proposed technique is helpful, notably in a recall, MABO, and mAP.

6 Conclusion

This article discussed an efficient framework for OD in images. The proposed methodology is a helpful addition to the family of object detectors. Our method initially generated less high-quality, class-independent, and accurate object proposals. Then, using these resulting proposals, efficiently determine the object's class. This strategy successfully decreases the number of affirmative proposals from 1000 to 100; it also prevents the need for trivial computations. By sharing convolutional features inside the network, the proposed scheme maintained a good recall and a high detection accuracy on fewer proposals. Previously, this was viewed as a disadvantage due to the overwhelming number of proposals, which dramatically reduced detection performance. We substantially increased our model's detection capabilities and improved OD performance by achieving a high recall of actual objects through rapid proposal generation and refinement. The experimental results advocate that the proposed model outperforms existing approaches in all quality measuring indicators. The improved performance demonstrates the proposed model's effectiveness and robustness, indicating that it can be used in various OD applications.

Given that the effectiveness of the proposed model is proportional to the number of initial proposals, it may be less productive for complicated images with a large number of initial proposals. Because our model is constructed from small feature maps created by down-sampling in the backbone network, images with a significant number of little objects would impair its performance. CNN deepening and scale interpolation always result in the loss of feature information in images. Therefore, enhancing feature extraction and interaction is essential for future detection performance improvements. Numerous high-level applications, such as multi-label image classification, disease diagnosis, pedestrian detection, etc., are anticipated to function more efficiently with fewer but more exact concepts. We intend to apply our method to a range of other high-level applications in the future.

Acknowledgement: Thanks to reviewers and editors for their suggestions during the review process.

Funding Statement: This study was funded by Huanggang Normal University, China, Self-type Project of 2021 (No. 30120210103) and 2022 (No. 2042021008).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] R. P. Esther and S. S. Jamiya, “ShortYOLO-CSP: A decisive incremental improvement for real-time vehicle detection,” *Journal of Real-Time Image Processing*, vol. 20, no. 1, pp. 3, 2023.

[2] H. Beomyeon, S. Lee and H. Han, “LNFCOS: Efficient object detection through deep learning based on LNblock,” *Electronics*, vol. 11, no. 17, pp. 2783, 2022.

[3] M. Aamir, P. Y. Fei, Z. Rahman, M. Tahir, H. Naeem *et al.,* “A framework for automatic building detection from low-contrast satellite images,” *Symmetry*, vol. 11, no. 1, pp. 3, 2019.

[4] G. Yurong, M. Aamir, Z. Hu, Z. A. Dayo, Z. Rahman *et al.,* “An object detection framework based on deep features and high-quality object locations,” *Traitement du Signal*, vol. 38, no. 3, pp. 719–730, 2021.

[5] G. Yurong, M. Aamir, Z. Hu, W. A. Abro, Z. Rahman *et al.,* “A region-based efficient network for accurate object detection,” *Traitement du Signal*, vol. 38, no. 2, pp. 481–494, 2021.

[6] M. Aamir, P. Y. Fei, Z. Rahman, W. A. Abro, H. Naeem *et al.,* “A hybrid proposed framework for object detection and classification,” *Journal of Information Processing Systems*, vol. 14, no. 5, pp. 1176–1194, 2018.

[7] C. Ying, D. Guo, Y. Shao, Z. Wang, C. Shen *et al.,* “Joint classification and regression for visual tracking with fully convolutional siamese networks,” *International Journal of Computer Vision*, vol. 130, no. 2, pp. 550–566, 2022.

[8] M. Aamir, P. Y. Fei, W. A. Abro, H. Naeem and Z. Rahman, “A hybrid approach for object proposal generation,” in *Proc. of the Int. Conf. on Sensing and Imaging*, Chengdu, China, pp. 251–259, 2017.

[9] W. Xiongwei, D. Sahoo and S. C. Hoi, “Recent advances in deep learning for object detection,” *Neurocomputing*, no. 396, pp. 39–64, 2020.

[10] D. Yao, H. Liang and Z. Y. Yi, “An improved approach for object proposals generation,” *Electronics*, vol. 10, no. 7, pp. 794, 2021.

[11] P. Rantalankila, J. Kannala and E. Rahtu, “Generating object segmentation proposals using global and local search,” in *Proc. of CVPR*, Columbus, Ohio, USA, pp. 2417–2424, 2014.

[12] P. Krähenbühl and V. Koltun, “Geodesic object proposals,” in *Proc. of European Conf. on Computer Vision*, Zurich, Switzerland, pp. 725–739, 2014.

[13] E. Rahtu, J. Kannala and M. Blaschko, “Learning a category independent object detection cascade,” in *Proc. of ICCV*, Barcelona, Spain, pp. 1052–1059, 2011.

[14] B. Alexe, T. Deselaers and V. Ferrari, “Measuring the objectness of image windows,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2189–2202, 2012.

[15] M. M. Cheng, Z. Zhang, W. Y. Lin and P. Torr, “BING: Binarized normed gradients for objectness estimation at 300fps,” in *Proc. of CVPR*, Columbus, Ohio, USA, pp. 3286–3293, 2014.

[16] S. Manen, M. Guillaumin and L. V. Gool, “Prime object proposals with randomized prim’s algorithm,” in *Proc. of Int. Conf. on Computer Vision*, Sydney, NSW, Australia, pp. 2536–2543, 2013.

[17] J. Uijlings, K. van de Sande, T. Gevers and A. Smeulders, “Selective search for object recognition,” *International Journal of Computer Vision*, vol. 104, no. 2, pp. 154–171, 2013.

[18] Z. Zhang and P. H. Torr, “Object proposal generation using two-stage cascade SVMs,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 1, pp. 102–115,2016, 2016.

[19] P. Krahenbuhl and V. Koltun, “Learning to propose objects,” in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, Boston, Massachusetts, USA, pp. 1574–1582, 2015.

[20] C. Zitnick and P. Dollár, “Edge boxes: Locating object proposals from edges,” in *Proc. of ECCV*, Zurich, Switzerland, pp. 391–405, 2014.

[21] P. Arbeláez, J. Pont-Tuset, J. T. Barron, F. Marques and J. Malik, “Multiscale combinatorial grouping,” in *Proc. of IEEE Conf. on CVPR*, Columbus, Ohio, USA, pp. 328–335, 2014.

[22] I. Endres and D. Hoiem, “Category-independent object proposals with diverse ranking,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 2, pp. 222–234, 2014.

[23] W. Kuo, B. Hariharan and J. Malik, “DeepBox: Learning objectness with convolutional networks,” in *Proc. of the ICCV*, Santiago, Chile, pp. 2479–2487, 2015.

[24] Q. Fan, W. Zhuo, C. K. Tang and Y. W. Tai, "Few-shot object detection with attention-RPN and multi-relation detector," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Virtual, pp. 4013–4022, 2020.

[25] P. O. Pinheiro, R. Collobert and P. Dollár, "Learning to segment object candidates," *Advances in Neural Information Processing Systems*, vol. 28, pp. 1990–1998, 2015.

[26] P. O. Pinheiro, T. Y. Lin, R. Collobert and P. Dollár, "Learning to refine object segments," in *Proc. of the European Conf. on Computer Vision*, Amsterdam, Netherlands, pp. 75–91, 2016.

[27] A. Krizhevsky, I. Sutskever and G. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[28] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in *Proc. of the CVPR*, Las Vegas, Nevada, USA, pp. 770–778, 2016.

[29] G. Huang, Z. Liu, K. Q. Weinberger and L. van der Maaten, "Densely connected convolutional networks," in *Proc. of the CVPR*, Honolulu, Hawaii, USA, pp. 4700–4708, 2017.

[30] M. Lin, Q. Chen and S. Yan, "Network in network," in arXiv: 1312.4400, 2013.

[31] C. Szegedy, W. Liu, Y. Jia and P. Sermanet, "Going deeper with convolutions," in *Proc. of the CVPR*, Boston, Massachusetts, USA, pp. 1–9, 2015.

[32] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in arXiv:1409.1556, 2014.

[33] N. Srivastava, G. E. Hinton and A. Krizhevsky, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[34] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. of Int. Conf. on Machine Learning*, Lille, France, pp. 448–456, 2015.

[35] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. of the CVPR*, Columbus, Ohio, USA, pp. 580–587, 2014.

[36] K. He, X. Zhang, S. Ren and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[37] R. Girshick, "Fast R-CNN," in *Proc. of the ICCV*, Santiago, Chile, pp. 1440–1448, 2015.

[38] S. Q. Ren, K. M. He, R. Girshick and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[39] J. F. Dai, Y. Li, K. M. He and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Advances in Neural Information Processing Systems*, Barcelona, Spain, vol. 1, pp. 379–387, 2016.

[40] K. He, G. Gkioxari, P. Doll'ar and R. Girshick, "Mask R-CNN," in *Proc. of the ICCV*, Venice, Italy, pp. 2961–2969, 2017.

[41] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan *et al.,* "Feature pyramid networks for object detection," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, Hawaii, USA, pp. 2117–2125, 2017.

[42] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. of the CVPR*, Las Vegas, Nevada, USA, pp. 779–788, 2016.

[43] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed *et al.,* "Ssd: Single shot multibox detector," in *Proc. of the ECCV*, Amsterdam, Netherlands, pp. 21–37, 2016.

[44] N. F. Soliman, E. A. Alabdulkreem, A. D. Algarni, G. M. El Banby, F. E. Abd El-Samie *et al.,* "Efficient deep learning modalities for object detection from infrared images," *Computers, Materials & Continua*, vol. 72, no. 2, pp. 2545–2563, 2022.

[45] K. Jot Singh, D. Singh Kapoor, K. Thakur, A. Sharma and X. Gao, "Computer-vision based object detection and recognition for service robot in indoor environment," *Computers, Materials & Continua*, vol. 72, no. 1, pp. 197–213, 2022.

[46] H. Beomyeon, S. Lee and H. Han, “DLMFCOS: Efficient dual-path lightweight module for fully convolutional object detection,” *Applied Sciences*, vol. 13, no. 3, pp. 1841, 2023.

[47] Z. Xin, Y. Liu, C. Huo, N. Xu, L. Wang *et al.,* “PSNet: Perspective-sensitive convolutional network for object detection,” *Neurocomputing*, vol. 468, pp. 384–395, 2022.

[48] M. Everingham, L. V. Gool, C. K. Williams, J. Winn and A. Zisserman, “The pascal visual object classes (voc) challenge,” *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[49] Z. Cai and N. Vasconcelos, “Cascade R-CNN: Delving into high quality object detection,” in *Proc. of the IEEE Conf. on CVPR*, Salt Lake City, Utah, USA, pp. 6154–6162, 2018.

[50] E. C. Kaya and A. A. Alatan, “Improving proposal-based object detection using convolutional context features,” in *Proc. of the IEEE Int. Conf. on Image Processing (ICIP)*, Athens, Greece, pp. 1308–1312, 2018.

[51] B. Liu, W. Zhao and Q. Sun, “Study of object detection based on Faster R-CNN,” in *Proc. of the Chinese Automation Congress (CAC)*, Jinan, China, pp. 6233–6236, 2017.

[52] Y. Zhang, Y. Chen, C. Huang and M. Gao, “Object detection network based on feature fusion and attention mechanism,” *Future Internet*, vol. 11, no. 1, pp. 9, 2019.

[53] Y. Xiao, X. Wang, P. Zhang, F. Meng and F. Shao, “Object detection based on faster R-CNN algorithm with skip pooling and fusion of contextual information,” *Sensors*, vol. 20, no. 19, pp. 5490, 2020.

[54] S. Bell, C. L. Zitnick, K. Bala and R. Girshick, “Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks,” in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, Nevada, USA, pp. 2874–2883, 2016.

[55] T. Kong, A. Yao, Y. Chen and F. Sun, “Hypernet: Towards accurate region proposal generation and joint object detection,” in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, Nevada, USA, pp. 845–853, 2016.

[56] A. Shrivastava, A. Gupta and R. Girshick, “Training region-based object detectors with online hard example mining,” in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, Nevada, USA, pp. 761–769, 2016.

[57] H. Zhao, Y. Zhou, L. Zhang, Y. Peng, X. Hu *et al.,* “Mixed YOLOv3-LITE: A lightweight real-time object detection method,” *Sensors*, vol. 20, no. 7, pp. 1861, 2020.

[58] D. Jiang, B. Sun, S. Su, Z. Zuo, P. Wu *et al.,* “FASSD: A feature fusion and spatial attention-based single shot detector for small object detection,” *Electronics*, vol. 9, no. 9, pp. 1536, 2020.

[59] C. Y. Fu, W. Liu, A. Ranga, A. Tyagi and A. C. Berg, “Dssd: Deconvolutional single shot detector,” in arXiv: 1701.06659, 2017.

[60] Z. Li and F. Zhou, “FSSD: Feature fusion single shot multibox detector,” in arXiv: 1712.00960, 2017.

[61] J. Jeong, H. Park and N. Kwak, “Enhancement of SSD by concatenating feature maps for object detection,” in arXiv: 1705.09587, 2017.

[62] J. Leng and Y. Liu, “An enhanced SSD with feature fusion and visual reasoning for object detection,” *Neural Computing and Applications*, vol. 31, no. 10, pp. 6549–6558, 2019.

[63] P. F. Felzenszwalb and D. P. Huttenocher, “Efficient graph-based image segmentation,” *International Journal of Computer Vision*, vol. 1, no. 59, pp. 167–181, 2004.

[64] Y. Liu, L. Shijie and C. M. Ming, “Refinedbox: Refining for fewer and high-quality object proposals,” *Neurocomputing*, vol. 406, no. 406, pp. 106–116, 2020.

[65] M. Everingham, L. V. Gool, C. K. Williams, J. Winn and A. Zisserman, “The PASCAL visual object classes challenge 2007 (VOC2007) results,” 2007. Available online: http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html

[66] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn and A. Zisserman, “The PASCAL visual object classes challenge 2012 (VOC2012) results,” 2012. Available online: http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html

[67] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona *et al.,* “Microsoft COCO: Common objects in context,” in *Proc. of the European Conf. on Computer Vision*, Zurich, Switzerland, pp. 740–755, 2014.

[68] S. Gidaris and N. Komodakis, "Object detection via a multi-region and semantic segmentation-aware CNN model," in *Proc. of the ICCV*, Santiago, Chile, pp. 1134–1142, 2015.

[69] B. Yang, J. Yan, Z. Lei and S. Z. Li, "Craft objects from images," in *Proc. of the CVPR*, Las Vegas, Nevada, USA, pp. 6043–6051, 2016.

[70] S. Gidaris and N. Komodakis, "Locnet: Improving localization accuracy for object detection," in *Proc. of the CVPR*, Las Vegas, Nevada, USA, pp. 789–798, 2016.

[71] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang *et al.,* "Deformable convolutional networks," in *Proc. of the ICCV*, Venice, Italy, pp. 764–773, 2017.

[72] Y. Zhu, C. Zhao, J. Wang, X. Zhao, Y. Wu *et al.,* "Couplenet: Coupling global structure with local parts for object detection," in *Proc. of the ICCV*, Venice, Italy, pp. 4126–4134, 2017.

[73] L. Tychsen-Smith and L. Petersson, "Denet: Scalable real-time object detection with directed sparse sampling," in *Proc. of the ICCV*, Venice, Italy, pp. 428–436, 2017.

[74] T. Kong, F. Sun, W. Huang and H. Liu, "Deep feature pyramid reconfiguration for object detection," in *Proc. of the ECCV*, Munich, Germany, pp. 169–185, 2018.

[75] H. Xu, X. Lv, X. Wang, Z. Ren and R. Chellappa, "Deep regionlets for object detection," in *Proc. of the ECCV*, Munich, Germany, pp. 798–814, 2018.

[76] B. Cheng, Y. Wei, H. Shi, R. Feris, J. Xiong *et al.,* "Revisiting RCNN: On awakening the classification power of faster RCNN," in *Proc. of the ECCV*, Munich, Germany, pp. 453–468, 2018.

[77] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in *Proc. of the CVPR*, Honolulu, Hawaii, USA, pp. 7263–7271, 2017.

[78] Z. Shen, Z. Liu, J. Li, Y. G. Jiang, Y. Chen *et al.,* "Dsod: Learning deeply supervised object detectors from scratch," in *Proc. of the IEEE ICCV*, Venice, Italy, pp. 1919–1927, 2017.

[79] T. Kong, F. Sun, A. Yao, H. Liu, M. Lu *et al.,* "RON: Reverse connection with objectness prior networks for object detection," in *Proc. of the CVPR*, Honolulu, Hawaii, USA, pp. 5936–5944, 2017.

[80] X. Zhou, D. Wang and P. Krähenbühl, "Objects as points," in arXiv:1904.07850, 2019.