electronics

MDPI

Article

An Improved Approach for Object Proposals Generation

Yao Deng [1], Huawei Liang [2,*] and Zhiyan Yi [3]

[1] Department of Automation, University of Science and Technology of China, Hefei 230026, China; baiyue@mail.ustc.edu.cn
[2] Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei 230031, China
[3] Department of Civil and Environmental Engineering, University of Utah, 110 Central Campus Dr. RM 1650, Salt Lake City, UT 84112, USA; zhiyan.yi@utah.edu
* Correspondence: hwliang@iim.ac.cn

Abstract: The objectness measure is a significant and effective method used for generic object detection. However, several object detection methods can achieve accurate results by using more than 1000 candidate object proposals. In addition, the weight of each proposal is weak and also cannot distinguish object proposals. These weak proposals have brought difficulties to the subsequent analysis. To significantly reduce the number of proposals, this paper presents an improved generic object detection approach, which predicts candidate object proposals from more than 10,000 proposals. All candidate proposals can be divided, rather than preclassified, into three categories: entire object, partial object, and nonobject. These partial object proposals also display fragmentary information of the objectness feature, which can be used to reconstruct the object boundary. By using partial objectness to enhance the weight of the entire object proposals, we removed a huge number of useless proposals and reduced the space occupied by the true positive object proposals. We designed a neural network with lightweight computation to cluster the most possible object proposals with rerank and box regression. Through joint training, the lightweight network can share the features with other subsequent tasks. The proposed method was validated using experiments with the PASCAL VOC2007 dataset. The results showed that the proposed approach was significantly improved compared with the existing methods and can accurately detect 92.3% of the objects by using less than 200 proposals.

Keywords: objectness detection; fixation prediction; salient object segmentation; detection rate (DR)

Citation: Deng, Y.; Liang, H.; Yi, Z. An Improved Approach for Object Proposals Generation. *Electronics* **2021**, *10*, 794. https://doi.org/10.3390/electronics10070794

Academic Editor: Youngbae Hwang

Received: 10 March 2021
Accepted: 24 March 2021
Published: 27 March 2021

1. Introduction

Studies in cognitive psychology and neurobiology showed that humans can perceive objects before identifying objects. However, humans' first glances of an object usually focus on a certain partial detail, while the rest of the details of the object remain unperceived. Considering this behavior, we utilized a mechanism similar to the visual perception of humans. When a camera captures a frame, the number of objects included in the image must be counted before categorizing the objects. Generally, image windows are used to represent regions where objects are detected. Object detection and recognition have been research hotspots in recent years. Great endeavors have been made to meet actual applications. The results of these endeavors were used as basis to introduce several standardized databases, such as ImageNet and the Pascal Visual Object Classes (VOC) [1]. Since the datasets contain a rich variety of objects, conferring additional challenges for the detection and recognition systems, the variability of the object categories is intractable and must be solved. In particular, the different object categories demonstrate varied degrees of deformation in the images. On the one hand, the natural structure of the objects varies—that is, living creatures, such as dogs, are more deformed than human-made machines, such as airplanes. On the other hand, different object locations lead to various viewing distances and angles in the same image; such that rigid objects can display deformations

in various view angles, whereas deformable objects appear rigid at a certain distance. Despite the significant advances made in object detection, many state-of-the-art detection techniques use category classifiers to estimate the image windows in a sliding window framework [2,3]. This sliding window technique entails high computing costs when a large number of object classes exist or when the applied classifiers evaluation costs ate high [4]. To address this limitation, scholars have developed numerous detection methods. For instance, techniques based on multistage cascade architectures have been proposed; in the early stages of detection, a larger portion of the bad bounding box hypotheses are filtered compared with the accurate ones [5]. Hence, a small set of windows can estimate costly classifiers. However, this technique is unsuitable for huge datasets in various object categories because a huge number of windows are generated in each image. Objectness is presented as a value to describe the likelihood of an image window overwriting an object [6]. This insensitivity can be regarded as a feature of objectness, because general objects contain a well-defined closed boundary and center when resizing into a small fixed size. As predicted, an objectness measure that deals with specific classifiers in each category must be used. The region-based convolutional neural networks (R-CNNs) and Fast R-CNN [7] achieve significant performance in object detection. When ignoring the computing time of proposals, the speed of object detection approaches real-time with Fast R-CNN deep networks. The time consumption of the proposals becomes a problem to be solved in detection systems. In addition, the properties of computational efficiency and accuracy of detection must also be considered. Therefore, high object detection rate (DR), high computational efficiency, and good generalization ability need be considered in a general object detection method.

Based on the region method, the object detection is divided into two substages. In the first stage, the regional proposal generation network and deep convolution neural network are combined to generate high-quality candidate proposals. In the second stage, these candidate proposals need to be classified and refined. The region proposal phase can eliminate most of the proposals with backgrounds, thus greatly reducing the search space of the object detection. Using very deep CNNS [8,9], the Fast R-CNN pipeline [7], has recently shown high accuracy on the mainstream target detection benchmark [10–13]. A multistage training process (e.g., [14,15]) is usually used for the joint operation of region proposal generation and postdetection. In Fast R-CNN [10], the region-based subnet repeatedly evaluates the positive and negative signs of thousands of regions to generate detection scores. Faster R-CNN and detection network share full image convolution features, which makes it possible to propose regions with almost no cost. Recently, R-FCN [14] attempted to make Faster R-CNN's per ROI unshared calculation shareable by position-sensitive scoring maps. However, R-FCN still requires a regional proposal generated by the region proposal network [10]. A popular technique for classification models is to use a bag-of-words model [16–18] that transforms documents into vectors where each word in the document is assigned a score. To ensure the accuracy of detection, the images are resized to a large enough size in all methods (usually the shortest edge is 600 pixels). When the images are feed into the deep network, both in the training time and the inference time, it is quite a waste of resources and time.

More than 1000 proposals containing true positive object windows are obtained by using the existing proposal methods. The weight of each proposal is weak and cannot distinguish object proposals. In order to obtain the advantages of the traditional proposal generation methods and the effective representation ability of CNNs, we present an improved proposal method to search for objects in an image. The existing object proposal methods divide all proposals into object and nonobject proposals. However, we observed that some nonobject proposals that covered or were located around object proposals display objectness features (Figure 1). To efficiently obtain accurate detection results, we divide all proposals into three categories: object, partial object, and nonobject proposals. We also demonstrate how partial object proposals can be used to increase the weight of the entire object proposals and reconstruct the boundary of the object. In contrast with the

state-of-the-art techniques, our method for model classification not only achieves objectness discrimination but also reduces the search space of the true positive object proposals.

The first thing to consider is that the network does not have too many computations. Secondly, the improved training process can be implemented in an end-to-end manner. Since the statistics of all the proposals are lightweight and easy to calculate, we can share the convolution features with other networks through joint training. In this paper, we combined our module with the famous Fast R-CNN [7] detection framework. The proposed module connects behind the last layer of the neural network, like VGG16 [9]. Therefore, by sharing the basic convolutional layer, our optimized network can achieve more efficient optimization effects with subsequent object detection network.

To evaluate the performance of the algorithm, we use the PASCAL VOC2007 dataset. The experiment results show that our technique exhibits significant improvements compared with previous methods. Our method uses only 200 proposals in each image and achieves a detection recall rate of 80.4% and 67.9% for the intersection over union (IoU) 0.5 and 0.7, respectively. The experimental results show that our method can generate high quality proposals when the number of proposals is limited. Additionally, the proposed method can accurately detect 92.3% of the objects by using fewer than 200 proposals.

The remaining part of this paper is organized as follows. Section 2 describes related works. Our method is presented in detail in Section 3. Then, Section 4 shows the experimental results. Finally, Section 5 reveals our conclusion.

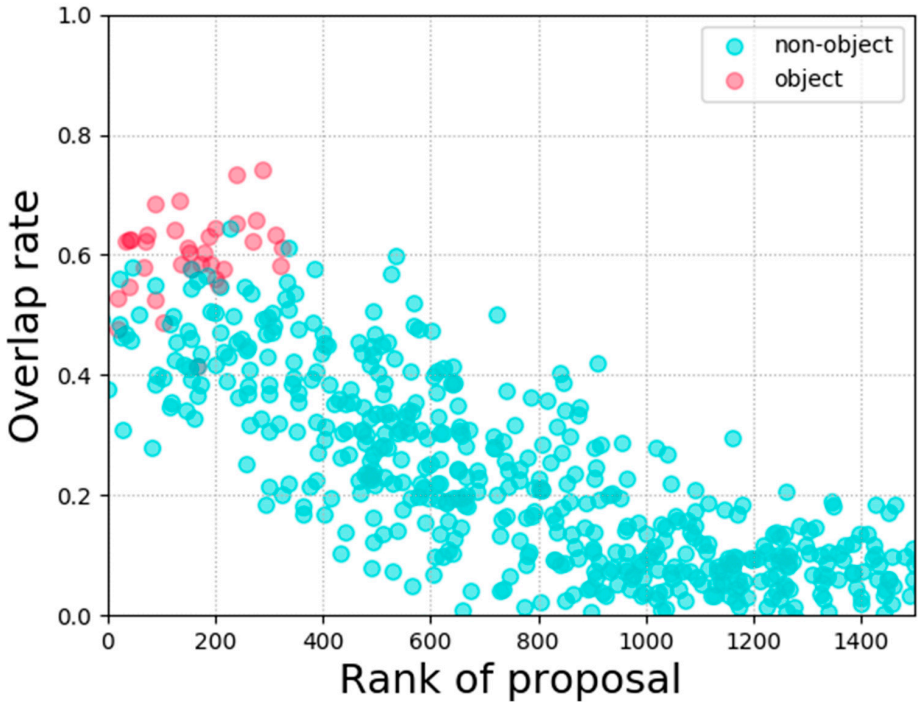

Figure 1. Proposals classified by existing methods: ranking proposals based on objectness detection scores (red and blue, respectively, represent the proposals with entire object, nonobject, partial object).

2. Related Works

Three categories of related models are presented considering the mechanism of human visual perception, in which the location of a potential object is first determined before identifying the object. In this section, we briefly discuss the three models, respectively, fixation prediction, salient object segmentation, and object proposal generation.

2.1. Fixation Prediction

Ko, J.; et al. [19] propose a fixation prediction method using the hybrid saliency-based attention model for calculating the saliency of the eye movement behavior of humans in

the computer vision community. Further studies with patch-based or pixel-based [20] features are utilized in fixation prediction models. Harel et al. [21] introduced the feature map surrounded by the normalized center, which is used for the highlighted part of the object. Fixation prediction has greatly contributed to object recognition in recent years. However, the models cannot accurately detect all general object proposals.

2.2. Salient Object Segmentation

Salient pixels form blobs, which are noticeable in an image. Inspired by studies of human eye movements [22], a selective visual attention is implemented from a bottom-up perspective. Typically, the bounding box is scored in combination with some low-level features, and then the selected bounding boxes have higher scores. To determine the noticeable regions in an image, Liu et al. [23] presented a mixed measure involving boundary and saliency information. Huang et al. [24] proved that existing fixation algorithms provide poor results when the F-measures of a PR curve are benchmarked. MCG [25] is an efficient image segmentation algorithm, which effectively utilizes multiscale information. By exploring the combination space, the multiscale regional hierarchy is combined into an object proposal. Palazzo et al. [26] introduced an overview of the features and performance of recent salient object algorithms. Manen et al. [27] constructed a connected graph of the image superpixel, then adopted the random version of the Prim algorithm to generate a spanning tree with a large number of the expected edge weights. DeepMask's [28] training has two targets: firstly, the designed system outputs a categorically uncertain split mask, and then the likelihood is obtained. To enhance the target segmentation of the feedforward network, a new top-down refinement method is proposed by Sharp mask [29]. The proposed architecture can effectively generate high-fidelity object masks. To form the segmentation hierarchy, Rantalankila et al. [30] proposed that the local search on superpixels and then the graph segmentation at the intermediate level can be obtained by global search. Although the methods [31–34] are successfully used in image scene analysis, a complicated image containing numerous objects remains difficult to solve.

2.3. Objectness and Object Proposals

According to previous studies [35–39], the production of rough segmentations as object proposals could lead to huge number of computations and requires 2–7 min per image. To obtain efficient and precise prediction, a measured approach of the "objectness" of a bounding box is presented by Alexe et al. [40]. The objecness score is calculated based on the number of outlines completely contained in each bounding box. Zhang et al. [37] introduced a cascading sorting support vector machine method to generate proposals using oriented gradient features. Cheng et al. [41] proposed the binarized normed gradient (BING) feature to quickly capture the objectness of an image proposal and generally achieved more than 1000 proposals in a short span of time. In order to accurately locate the objects in an image [6], it is necessary to refine the object proposals in postprocessing. RPN [10] predicts the objectness score and the image CNN feature each position of the object boundary. Zhang et al. [35] proposed an optimization method for object proposal generation, which used superpixels to optimize the proposal box. Therefore, the system accelerates the superpixel generation speed in MTSE with optimizational segmentation. He et al. [42] proposed that object-oriented schemes with different directions, rather than just the vertical boxes used in conventional methods.

Compared with other methods, our method builds a refinement network and has good performance in object proposal generation evaluation and object detection evaluation by reducing the true positive object proposal space to less than 200.

3. Materials and Methods

Human visual perception involves two cascades, wherein objects of the environment are first perceived in the basic cascade, followed by identification of the objects found in the succeeding cascade. Similar to this mechanism, computer vision relatively demonstrates

the ability to perceive a hierarchical system. Compared with the method presented in [41], we propose an approach to cluster proposals produced by other proposal generations and reduce the range of true positives. Using the feature information in the deep neural network [6], the objectness score of each box is recalculated.

3.1. Objectness Detection

Objectness always acts as a seal around the objects. When resizing windows to a fixed size of 8×8, the closed objectness boundaries slightly change. This characteristic is a feature particularly used to describe the structure of objects. Although one object demonstrates various visions in different illumination conditions and distances, the normed gradients eliminate the effects of the environment and maintain the uniformity of the entire object. We first resize the original image into 28 kinds of scales to make sure target windows contain different sizes of objects and then compute the normed gradients of each scale image. When we search the resized window pixel by pixel, each predefined 8×8 region contains 64 values of normed gradient that are defined as the 64D normed gradient feature. Changing object position and scale minimally influences objectness structure because the objectness normed gradient feature is compact and insensitive to changes in translation.

We also use a two-stage cascaded SVM as in [37] to train the normed gradient model. First, linear SVM is utilized for a single model w with truth object window and random background window; w has high weights along the boundary and will split objects from the background. Second, each window size possesses different correlation coefficient learning from a linear SVM. As such, we obtain the filter scores and re-sort the proposals by using Equation (1). The large weights along the w borders can help to separate an object from its background. We found that w captures a more complex and natural a priori than the hand-designed central surround pattern. Lower object areas tend to be obscured from above object. Therefore, the w places less confidence in the lower areas.

$$\boldsymbol{w} = \sum_{\boldsymbol{k=1}}^{\boldsymbol{N}} \boldsymbol{\beta_k x_k} \tag{1}$$

where N denote the total number of basis vectors, $x_k \in \{-1, 1\}^{64}$ is a basis vector, $\beta_k \in R$ presents the coefficient to each basis vector. To approximate w, we present Equation (2) in terms of a set of basis vectors $x_k = x_k^+ - \overline{x_k^+}$, where $x_k^+ \in \{0, 1\}^{64}$.

$$\langle \boldsymbol{w}, \boldsymbol{y} \rangle \approx \sum_{\boldsymbol{k=1}}^{\boldsymbol{N}} \boldsymbol{\beta_k} (\langle \boldsymbol{x_k^+}, \boldsymbol{y} \rangle - \langle \overline{\boldsymbol{x_k^+}}, \boldsymbol{y} \rangle) \tag{2}$$

To calculate and binarized the normed feature, we use Equation(3) to approximate H with top M binary bits of the BYTE values.

$$\boldsymbol{H} = \sum_{\boldsymbol{j=0}}^{\boldsymbol{M}} \boldsymbol{2^{8-j} y_j} \tag{3}$$

where H is the 64D normed feature.

The filter score of binary normed gradient features S can be estimated by Equation (4).

$$\boldsymbol{S} \approx \sum_{\boldsymbol{k=1}}^{\boldsymbol{N}} \boldsymbol{\beta_k} \sum_{\boldsymbol{j=1}}^{\boldsymbol{M}} \boldsymbol{2^{8-j}} (\boldsymbol{2} \langle \boldsymbol{x_k^+}, \boldsymbol{y} \rangle - \left| \boldsymbol{y_j} \right|) \tag{4}$$

3.2. Objectness Detection Constructing Object Boundary with Partial Object Proposals

When generating an object proposal, we need to be concerned about whether the boundaries are contained precisely, rather than the pixels inside the object. By observation, the good bounding boxes are those that tightly cover the boundaries of an object. When the boundaries of a proposal do not intersect with any the edge points, we can assume that

the proposal contains no objects. On the other hand, the proposal has not yet reached the boundaries of the object. In all of these cases, the proposal needs to be modified to push it toward the real object boundaries. Therefore, we suggest refining the proposal using the point closest to the proposed boundary as a fast and weak indicator of the object boundary.

In each window scale, the 8×8 binary normed gradient feature that shifts by one step represents a new proposal, and each proposal corresponds to a score of objectness detection (Equation (4)). The results show that more than 10,000 proposals are predicted. The key challenge is how to screen and reduce the range of true positive object windows in all proposals. We rank all proposals basing on objectness detection scores; meanwhile, red blue and yellow circles are used to represent the predicted proposals. To efficiently and rapidly estimate the proposals, we utilize the nonmaximum suppression for removing the useless proposals. Subsequently the scope of the ranking is decreased to 1500.

After filtering the scores of the proposals, the most possible proposals are distributed in the front of the ranking. When we display the 64D normed gradient feature region of 1500 proposals in the original gradient image, the proposals in the first 400 spots in the ranking mostly bear heavy weights and are surrounded by pixels; each of these proposals is represented by a red circle. The yellow points represent the proposals that partly display objectness and belong to the middle part of the ranking. The end part of the ranking shows low objectness in the original gradient image.

We presume that objectness detection can not only separate object proposals from nonobject proposals but can also separate partial object proposals. Therefore, to verify our presumption, we use three proposals: entire object inside, partial object inside, and nonobject inside. We determine the statistical distribution of each type of proposal in the ranking. Figure 2 shows that most of the entire objects, partial objects, and nonobjects are distributed in the rankings as follows: 0:400, 0:800, 700:1500.

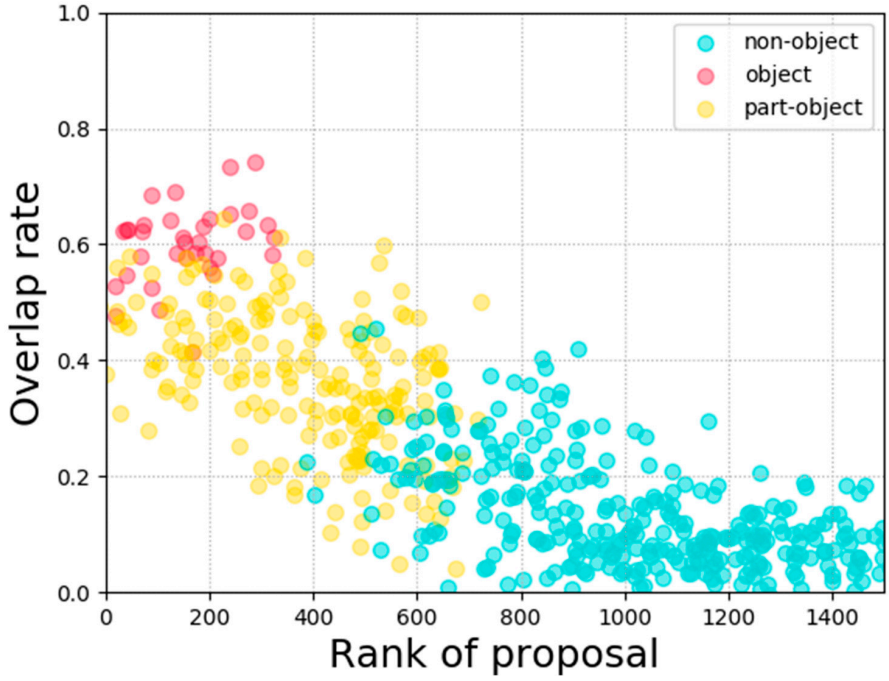

Figure 2. Proposals classified by our method: ranking proposals based on objectness detection scores (red, blue and yellow circles, respectively, represent the proposals with entire object, nonobject, partial object).

3.3. Clustering the Most Possible Proposals

As mentioned in Section 3.1, more than 1500 proposals are obtained in an image. These proposals contain entire objects, partial objects, and nonobjects. We compare the proposals with the true locations of objects by using testing image annotations. If the proportion of

the overlap of one proposal window and a truth object window is more than 80%, then the proposal is marked as the most possible proposal. When the proposals containing entire objects are all clustered in the first 200 spots in the ranking, this phenomenon becomes advantageous in finding regions where the most possible proposals are distributed, thereby reducing computing time. In fact, the most possible proposals are dispersed in the rankings. To cluster the most possible proposals, we provide a method to enhance the weight of objectness in the proposals and ensure that the most possible proposals are included in the ranking. When we resize each proposal to the corresponding windows in an original image, the corresponding object windows always cover or intersect with the small windows. The small windows also contribute to constructing the parts of an object. If a corresponding window contains or intersects with another window, then the score of the corresponding window increases. We choose the most possible proposals as classes and randomly initialize their respective central points. The intersecting proposals act as a data point vector. A central point is a location at the same distance from each data point vector. This requires us to predict the number of classes in advance. Then, we calculate the distance between each data point and the center point and divide the data points into the category closest to the center point. By recalculating the center point of each class, we achieve the new center point. We repeat the above steps until the center position of each category is fixed after each iteration; that is, the new most possible proposal position. Similarly, when we have a new proposal, we use its center coordinate as a reference and calculate the minimum hamming distance from the center point of each category we have obtained in the previous step, attributing it to the nearest category. After recalculating the weight of the proposal and re-sorting the score ranking, we obtain a new proposal ranking. All of the most possible proposals are clustered and distributed in the first 200 proposals in the ranking.

Windows corresponding to the 8×8 region of the resized normed gradient maps with high scores display the necessary objectness to construct an entire object in each scale, given that the filter eliminates most of the 64D normed gradient windows with low scores. Therefore, we use proposals that are distributed in 300 to 1000 in the ranking to determine the regions where the objects are located in the original image and then transform the image into a grey image. By projecting each grey images into a white image (value = 255), we reveal the contours corresponding to the objects (Figure 3). This result proves that the partial object proposals can be used to construct the outline of the entire object. Although the true position of an object is lacking in some situations, we can still reconstruct the object's contour by using partial object proposals.

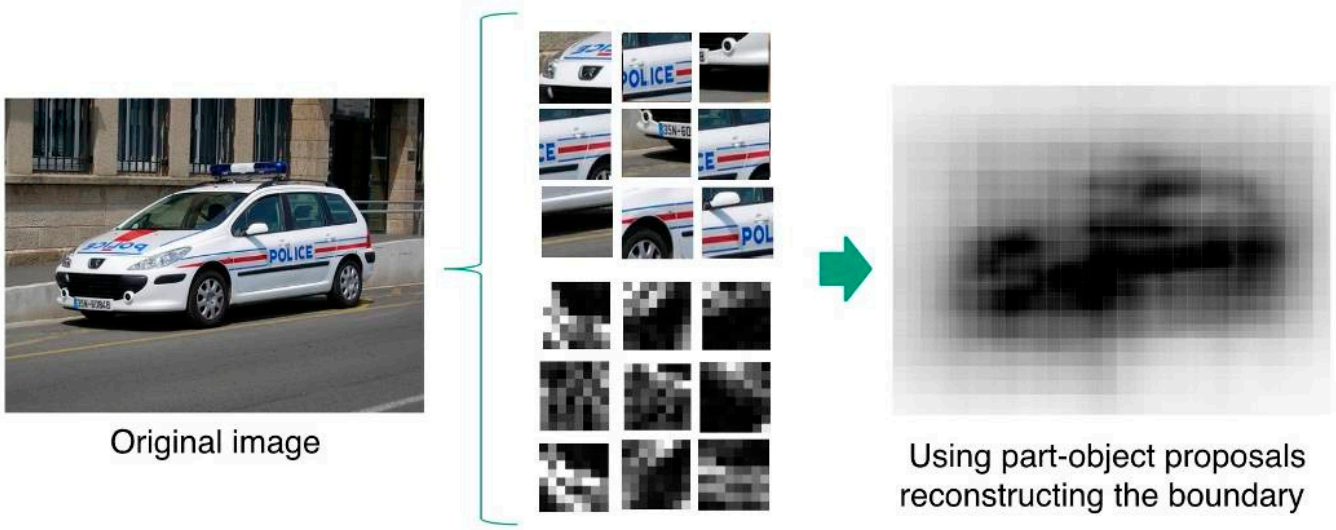

Figure 3. Proposals with partial objects are regarded as components of the entire object; clustering the objectness of a partial object proposal can reconstruct the boundary of the entire object.

3.4. Network Architecture

Our object detection system has two modules (Figure 4). We model the first module with a deep fully convolutional network. The values in the convolutional layers mean the number of channels. Ren [10] suggests that the proposal quality of RPN+VGG is better than that of RPN+ZF. Inspired by this result, we use VGG as the deep fully convolutional network. VGG is the basic network architecture which is widely used in deep learning. In this paper the first module has 13 convolutional layers and 3 fully connected layers. The second module is the region proposal network. Its input is an image with any size and its output is a set of rectangular object proposals, each with an objectness score. When we train an image with a car, we first need to perform multiscale scaling of the image to obtain proposals of different sizes and angles. Then, using Bing to speed up the calculation of the objectness of each proposal, we cluster the most possible proposals with the corresponding position of the object. Finally, the reranking proposals are obtained. These rectangular object proposals serve as inputs into the next network.

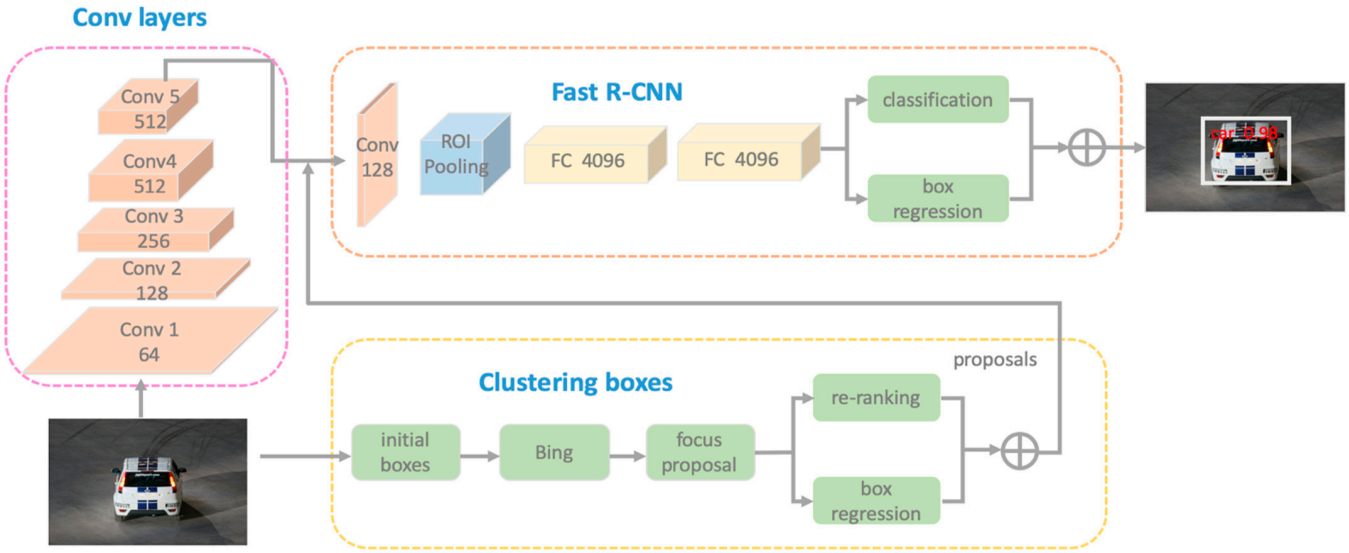

Figure 4. Overview of our network architecture. The object detection system is composed of two modules. The proposed network is designed to generate the initial boxes, and then input the clustered boxes into the branch of Fast region-based convolutional neural networks (R-CNN) for classification.

After the 13th convolutional layer, we firstly connect a kernel size of 3×3 convolutional layer, which reduces the number of channels from 512 to 128. Each initial box will be sampled downward and fixed to the size of 8×8 feature map through the ROI Pooling layer. The input feature grid map has the same width and height. ROI Pooling is performed for the maximum pooling of each grid. It is then connected to a full connection layer with 512 output neurons. Finally, by two branches of box regression and ranking the objectness score is recalculated, as well as the positional offset of each box. Additionally, the ranking branch then consists of two fully connected layers with two neurons, each of which outputs a probability value of whether it is an object or not.

Following the multitask loss in Fast R-CNN [7], we minimize an objective function. The loss function is:

$$L_{obj}(p, u) = -\left[\mathbf{1}_{\{u=1\}} \mathbf{log} p_1 + \mathbf{1}_{\{u \neq 1\}} \mathbf{log} p_0\right] \quad (5)$$

where p is calculated by a softmax on the two outputs in a fully connected layer. If the box is an object, the label u is 1. Otherwise, the label u is 0. Considering the learning of coordinate offsets, the box regression layer is designed as a fully connected layer.

The box regression values is predicted as described below:

$$\begin{aligned} t_x &= (x - x_k)/w_k, & t_y &= (y - y_k)/h_k, \\ t_w &= \log(w/w_k), & t_h &= \log(h/h_k), \\ t_x^* &= \frac{(x^* - x_k)}{w_k}, & t_y^* &= (y^* - y_k)/h_k, \\ t_w^* &= \log(w^*/w_k), & t_h^* &= \log(h^*/h_k), \end{aligned} \tag{6}$$

where x and y are the central coordinates of the box. Similarly w and h are the box's width and height. The variables x denotes the predicted box. The variables x_k and x^* have a similar definition; respectively, input box, and ground-truth box. The variable t^* and t denote the prediction target and the prediction tuple. When making location predictions for boxes, consider this as a regression from a box to the ground-truth box boundary. The box regression loss is defined as:

$$L_{reg}(t_i, t_i^*) = R(t_i - t_i^*) \tag{7}$$

where R denotes the robust loss function which is defined in [7]. Then, joining Equation (5) and Equation (6) the loss function is:

$$L(p, u, t, t^*) \; = \; L_{obj}(p, u) + \; \lambda \cdot \mathbf{1}_{\{u=1\}} L_{reg}(t, t^*) \tag{8}$$

where the balance parameter λ is set as 1 in this paper.

4. Experiment

Based on recent works on general objectness detection, we evaluated our target detection performance using the challenging Pascal VOC 2007 dataset. The dataset has 4952 images in 20 different categories. All the images are marked with boundary box annotations (Figure 5). The PASCAL VOC 2007 has become a famous benchmark dataset for general multiple-class object detection because it can be used for various objects, such as humans and airplanes. The detection performance of each category is measured by average precision. We need to detect all objects in the images regardless of the object categories; hence, various categories and sufficient number of objects make PASCAL VOC 2007 suitable for our verification.

During the training, each stochastic gradient descent (SGD) requires 256 proposals from an image. In each batch, half of the samples are positive and the other half are negative. The initial learning rate is set to 0.001. After running 12 epochs, the learning rate needs to be divided by 10. We run SGD for 16 epochs throughout the training process. Our experimental training and testing take place on the GTX 970 GPU. All parameters of these comparison methods are set to default values.

4.1. Proposal Quality Comparisons

The method proposed by Cheng [41] for general objectness detection can achieve more than 1000 proposals with 96.2% DR. Each proposal has a filter score obtained from the linear SVM as the 1D feature. We used testing image annotations to check the labels of the proposals and rank their filter score (Figure 5). When we mark the entire object proposals as red circles, the partial object proposals as yellow circles and the nonobject proposals as blue circles in the rank chart, all circles are clearly distributed in three different regions of the rank chart. Proposals with nonobject are useless and mostly found below the 1000th rank. Every image generates around 10,000 proposals. Then, we use the proposed method to exclude the proposals that are marked as partial objects and nonobjects. More than 8000 proposals are removed in each image. The standard recall-over-IoU performance is used to evaluate our performance. Each method proposes 1000 bounding boxes and we test each method's recall with different IoUs. In order to explore the statistics of the object proposals

of each class, we combined the aspect ratio and size of the object proposal, and used the weighted contour for scoring. Part-object proposals are chosen from a pool proposal with 80% to 35% overlap with some object and non-object proposals have no more than 35% overlap with any object.

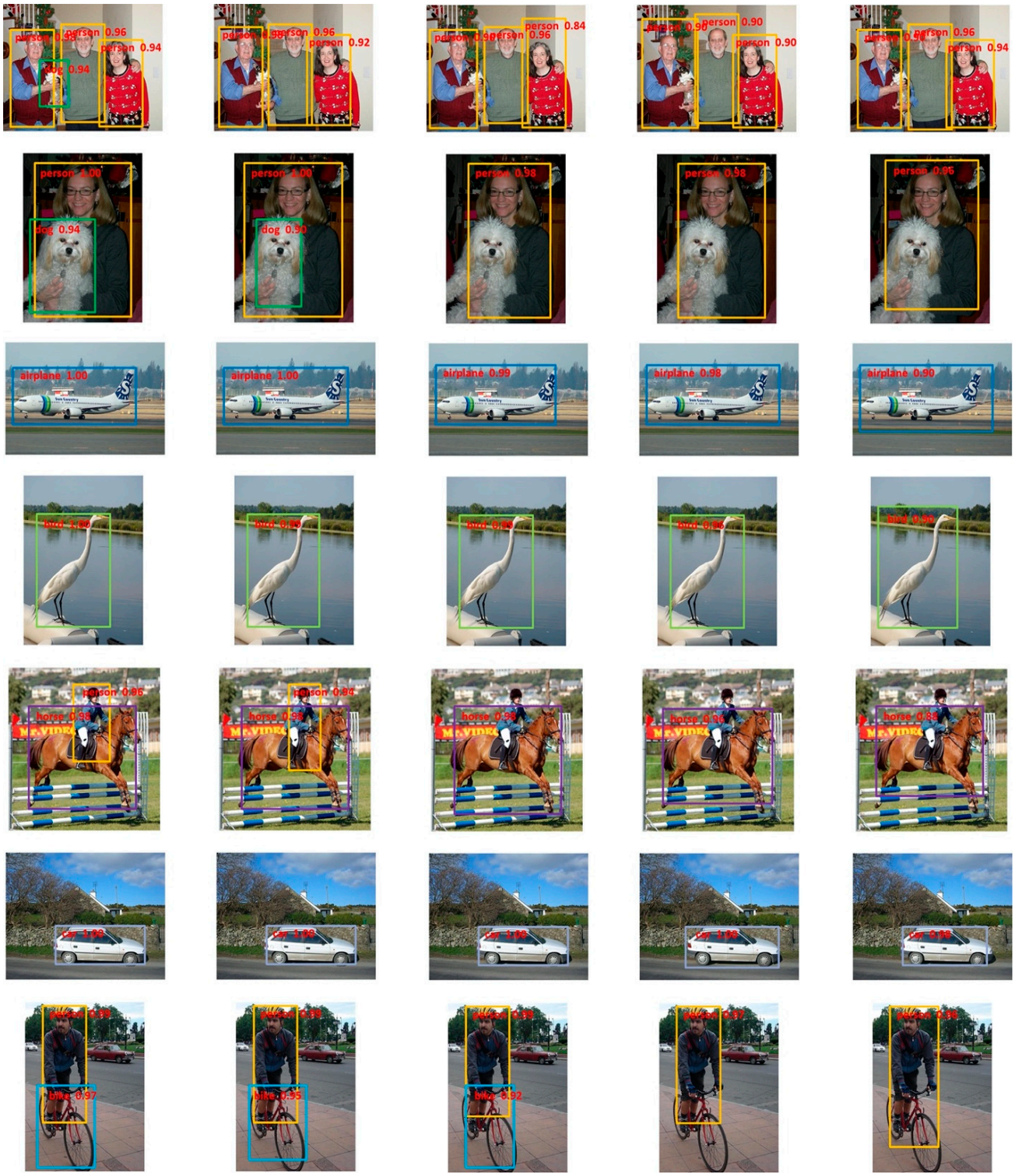

Figure 5. Under the same conditions, the object detection results obtained by different methods. Our data are from VOC2007 test set.

Inspired by [43], we tried to use high IoU (0.7) for object detection and compared the refinement methods of the different proposals. We selected edge boxes to generate the initial proposals, and then input them into these optimization algorithms. Changing the nonmaximum suppression parameter from 0.7 to 0.9, we also found that more edge boxes could be obtained. Note that our method uses contour score to select 1000 proposals from the selective search [44]. Additionally, they are significantly better than almost all of the IoU's 1000 proposals in the selective search report. Fewer proposals and higher detection recall rates would be of great benefit for subsequent advanced applications. With just 100 proposals, we can implement 91.3% and 83.5% detection recalls, with IoU of 0.5 and 0.7. We obtain a small amount of high quality object proposals. Selecting the good proposals from all the proposals generated by the previous method is the goal of our method. The goal of our method is to select the most possible proposal from all the proposals generated by the previous method.

4.2. Object Detection Performance

Based on image annotations, we marked the most possible entire object proposals with red circles. Note that not all of entire object proposals are distributed together in front of the ranking line; hence, we increase the weight of entire object proposals with objectness of the partial object proposals. Our approach obtains a small set of high quality proposals in dataset. Figure 6 shows that the most possible proposals are distributed in the first 200 spots. The proposed approach can be used to avoid most windows with no significance and thus save computation time. The Fast R-CNN network is retrained by the first 200 proposals for each image. We train the VOC2007 TrainVal set with all these methods and test them on the test set (Figure 7). The results are shown in Tables 1 and 2. To measure the computing cost of a network, we typically use floating point operations (FLOPs) as a metric. A floating point operation implies a multiplication and addition operation. For each proposal, the FAST R-CNN branch has 120 million FLOPs in the fully connected layers, while our branch only has 3 million FLOPs in the fully connected layers. Therefore, the improved proposal branch will only increase the extra computation a little bit. BING is simple, using advanced speed-up techniques to make the computational time tractable. The runtime of our method for each image is about 0.23 s, which is very fast when compared with the traditional proposal generation methods. Compared with the original methods, i.e., Rath, GOP, MCG, and EdgeBox, the improved proposals improve performance 12.8%, 33.6%, 25.1% and 27.2%, respectively. Under the same conditions, our results showed a 1.2% improvement in performance compared with the state-of-art Refinedbox [6]. Figure 6 shows the comparison of our method with other methods, namely, Refinedbox, MCG, and GOP [44], which use more than 1000 proposals to obtain 97.8% DR. By contrast, our approach can achieve 92.3% DR by using less than 200 proposals. If we use 1000 proposals, the result will be higher than 98.6%. The behavior of these curves demonstrates the excellent performance of the proposed approach over the other methods.

To ensure that the experimental results are meaningful and that statistical validation is possible, we repeat the experiment 30 times with the same dataset of 20 categories. Here, we randomly use different image detection initialization cores with a total of 30 kinds; that is, for each image, we will obtain 30 score values as a group. In order to better verify our results, we conduct paired t-tests on the results of each method, and eight groups of comparison results are obtained, assuming that there is no difference in the results of the object detection by different methods. Calculating the statistics, we obtain the score of each method and the p value with our method in Table 3. According to the standard, $p < 0.05$ is considered statistically significant. For quantitative analysis, the difference is statistically significant, and it can be considered that the identification results of the two methods are different.

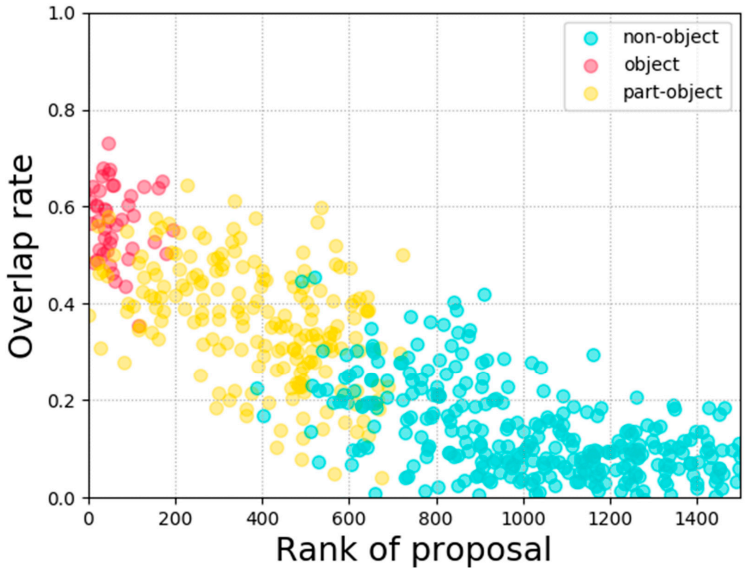

Figure 6. After clustering the most possible proposals with partial proposals.

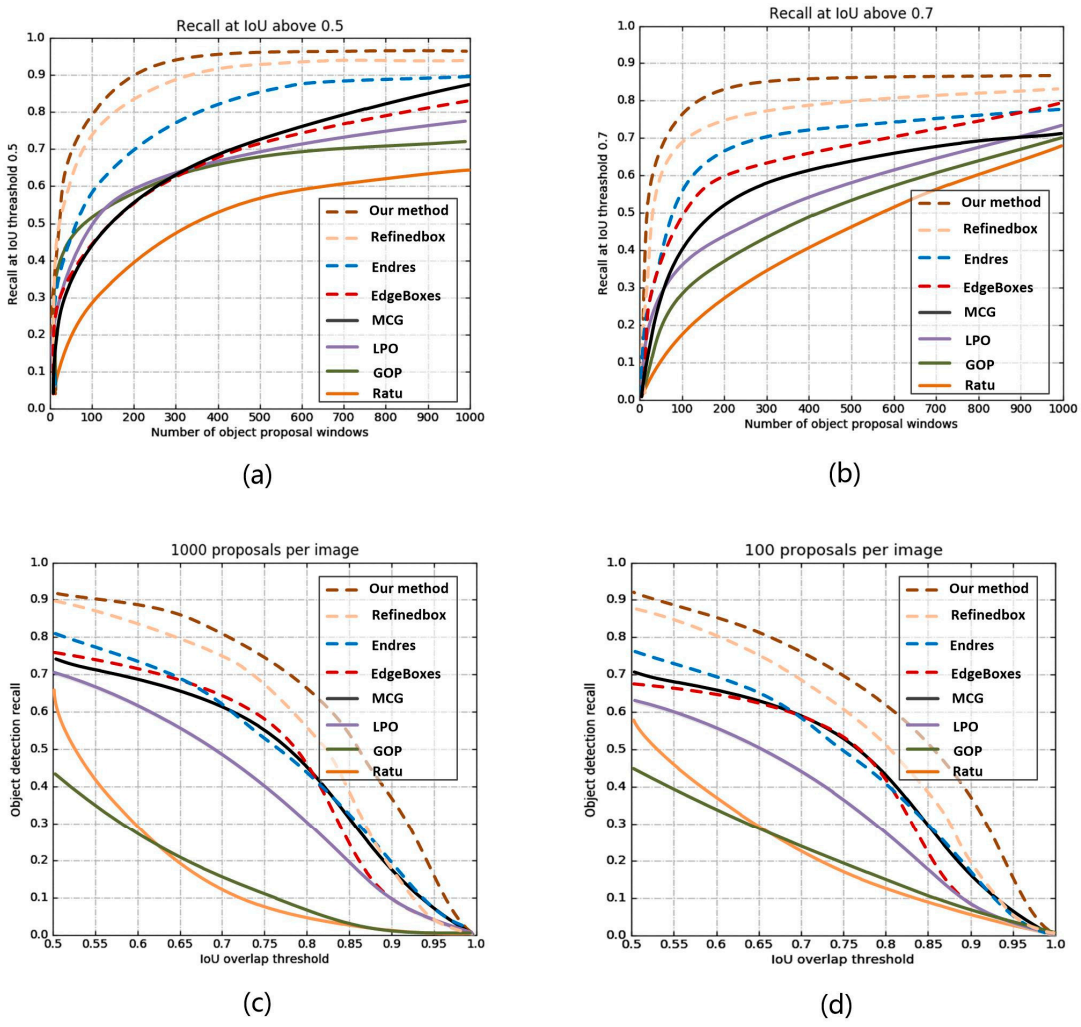

Figure 7. Experiment results on PASCAL VOC2007 test set: (**a**,**b**) show the recall of object detection versus the number of proposals at IoU threshold 0.5 and 0.7, respectively. (**c**,**d**) show the recall of object detection versus IoU overlap threshold using 100 proposals and 1000 proposals each image.

Table 1. Detection recall (%) using intersection over union (IoU) 0.5 thresholds on the VOC2007 test set.

Method \ Proposals	IoU = 0.5			Time(s)
	100	500	1000	
EdgeBoxes	75.3	90.2	94.3	0.24
Endres	83.6	93.2	94.0	18.77
GOP	62.1	93.8	96.2	0.28
LPO	73.6	95.5	95.9	0.47
MCG	83.3	96.0	97.1	18.02
Rahtu	61.7	87.3	89.2	0.69
Refinedbox	89.2	97.3	97.8	0.31
Our method	91.3	98.4	98.6	0.23

Table 2. Thresholds on the VOC2007 test set.

Method \ Proposals	IoU = 0.7			Time(s)
	100	500	1000	
EdgeBoxes	61.1	83.2	86.7	0.24
Endres	59.6	78.3	80.3	18.77
GOP	35.2	74.5	83.6	0.28
LPO	48.3	76.5	82.8	0.47
MCG	59.8	81.3	85.2	18.02
Rahtu	46.2	68.2	72.3	0.69
Refinedbox	78.2	84.5	85.4	0.31
Our method	83.5	85.2	86.8	0.23

Table 3. Paired t-test with our method on the VOC2007 test set.

Method \ Classes																				
	Paired t-Test with Our Method: Statistically Significant *p*-Value for the t-Statistic																			
EdgeBoxes	0.021	0.041	0.023	0.011	0.009	0.014	0.032	0.025	0.034	0.027	0.022	0.012	0.008	0.019	0.025	0.007	0.024	0.025	0.013	0.021
Endres	0.034	0.033	0.030	0.017	0.012	0.023	0.024	0.033	0.023	0.028	0.042	0.022	0.006	0.012	0.018	0.012	0.025	0.021	0.015	0.022
GOP	0.027	0.025	0.012	0.028	0.014	0.034	0.027	0.028	0.021	0.034	0.040	0.028	0.012	0.016	0.010	0.011	0.028	0.027	0.017	0.019
LPO	0.012	0.041	0.018	0.029	0.017	0.045	0.028	0.022	0.022	0.025	0.037	0.027	0.005	0.013	0.034	0.010	0.024	0.024	0.019	0.025
MCG	0.007	0.051	0.016	0.013	0.024	0.012	0.031	0.018	0.028	0.029	0.033	0.024	0.021	0.018	0.028	0.009	0.012	0.019	0.009	0.028
Rahtu	0.018	0.022	0.025	0.027	0.022	0.026	0.027	0.019	0.016	0.030	0.028	0.030	0.025	0.016	0.022	0.006	0.036	0.017	0.007	0.018
Refinedbox	0.007	0.027	0.028	0.022	0.032	0.008	0.027	0.012	0.014	0.026	0.022	0.028	0.002	0.011	0.030	0.012	0.028	0.010	0.004	0.015

5. Conclusions

In this paper, we presented an improved approach for object proposals generation. By classifying all of the candidate proposals into three categories, we enhance the weak weight of entire object proposals and enable them to distinguish object proposals. This approach can reduce the number of true positive object proposals from 1000 to less than 200 and avoid insignificant computation. Since the network is easy to optimize, it is possible to combine it with subsequent applications for training. The evaluation of the object detection proves the effectiveness of this method. The experiments using the PASCAL VOC2007 dataset show that our approach achieves 92.3% DR of generic objects by using a small set of candidate proposals. Since this method is processed for the small feature map generated by subsampling in the backbone, when there are many small targets in the image, the performance will be affected, such as the target detection method [45,46]. With a small number of high-quality object proposals, we can do more advanced work in the future.

Author Contributions: Validation, Z.Y.; writing—original draft preparation, Y.D.; writing—review and editing, H.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by National Key Research and Development Program of China (Nos. 2020AAA0108103, 2016YFD0701401, 2017YFD0700303 and 2018YFD0700602), Youth Innovation Promotion Association of the Chinese Academy of Sciences (Grant No. 2017488), Key Supported Project in the Thirteenth Five-year Plan of Hefei Institutes of Physical Science, Chinese Academy of Sciences (Grant No. KP-2019-16), Natural Science Foundation of Anhui Province (Grant No. 1508085MF133) and Technological Innovation Project for New Energy and Intelligent Networked Automobile Industry of Anhui Province.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Everingham, M.; van Gool, L.; Williams, C.; Winn, J.; Zisserman, A. The PASCAL visual object class challenge 2007 results. *Int. J. Comput. Vis.* **2007**, *88*, 303–338. [CrossRef]
2. Yao, H.; Xu, C. Joint person objectness and repulsion for person search. *IEEE Trans. Image Process.* **2020**, *30*, 685–696. [CrossRef] [PubMed]
3. Wang, W.; Lai, Q.; Fu, H.; Shen, J.; Ling, H.; Yang, R. Salient object detection in the deep learning era: An in-depth survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**. [CrossRef]
4. Verma, A.; Meenpal, T.; Acharya, B. Object Proposals Based on Variance Measure. In *Computational Intelligence in Pattern Recognition*; Springer: Singapore, 2020; pp. 307–320.
5. Van De Sande, K.E.A.; Uijlings, J.R.R.; Gevers, T.; Smeulders, A.W.M. Segmentation as selective search for object recognition. In Proceedings of the 2011 International Conference on Computer Vision (ICCV), Barcelona, Spain, 6–13 November 2011; pp. 1879–1886.
6. Liu, Y.; Li, S.; Cheng, M.M. Refinedbox: Refining for fewer and high-quality object proposals. *Neurocomputing* **2020**, *406*, 106–116. [CrossRef]
7. Girshick, R. Fast R-CNN. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
8. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016.
9. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015), San Diego, CA, USA, 7–9 May 2015.
10. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. In Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS 2015), Montreal, QC, Canada, 8–13 December 2014.
11. Everingham, M.; Eslami, S.M.A.; Van Gool, L.; Williams, C.K.I.; Winn, J.; Zisserman, A. The Pascal visual object classes challenge: A retrospective. *Int. J. Comput. Vis.* **2015**, *111*, 98–136. [CrossRef]
12. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
13. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollar, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In *European Conference on Computer Vision 2014*; Springer: Cham, Switzerland, 2014.
14. Dai, J.; Li, Y.; He, K.; Sun, J. R-fcn: Object detection via region-based fully convolutional networks. In Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, 5–10 December 2016.
15. Kong, T.; Yao, A.; Chen, Y.; Sun, F. Hypernet: Towards accurate region proposal generation and joint object detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016.
16. Polap, D.; Wlodarczyk-Sielicka, M. Classification of non-conventional ships using a neural bag-of-words mechanism. *Sensors* **2020**, *20*, 1608. [CrossRef] [PubMed]
17. Qiu, D.; Jiang, H.; Chen, S. Fuzzy information retrieval based on continuous bag-of-words model. *Symmetry* **2020**, *12*, 225. [CrossRef]
18. Liu, X.; Zhang, S.; Huang, T.; Tian, Q. E2BoWs: An end-to-end Bag-of-Words model via deep convolutional neural network for image retrieval. *Neurocomputing* **2020**, *395*, 188–198. [CrossRef]
19. Ko, J.; Cheoi, K.J. A novel distant target region detection method using hybrid saliency-based attention model under complex textures. *Comput. Sci. Inf. Syst.* **2021**. [CrossRef]
20. Hou, X.; Harel, J.; Koch, C. Image signature: Highlighting sparse salient regions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *34*, 194–201. [CrossRef]
21. Harel, J.; Koch, C.; Perona, P. Graph-based visual saliency. In Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, 5–10 December 2016; pp. 542–552.
22. Chen, Z.; Zhou, H.; Lai, J.; Yang, L.; Xie, X. Contour-aware loss: Boundary-aware learning for salient object segmentation. *IEEE Trans. Image Process.* **2021**, *30*, 431–443. [CrossRef] [PubMed]

23. Liu, T.; Sun, J.; Zheng, N.; Tang, X.; Shum, H. Object recognition as ranking holistic figure-ground hypotheses. In Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA, 13–18 June 2010; pp. 1712–1719.
24. Huang, Z.; Guo, B.; Wang, G.; Li, C.; Wei, Z. Robust and accelerated frequency-tuned salient object detection via color quantization. *J. Phys. Conf. Ser.* **2020**, *1575*, 012140. [CrossRef]
25. Arbelaez, P.A.; Pont-Tuset, J.; Barron, J.; Marques, F.; Malik, J. Multiscale combinatorial grouping. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014; pp. 328–335.
26. Palazzo, S.; Rundo, F.; Battiato, S.; Giordano, D.; Spampinato, C. Visual saliency detection guided by neural signals. In Proceedings of the 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), Buenos Aires, Argentina, 18–22 May 2020; pp. 525–531.
27. Manen, S.; Guillaumin, M.; Van Gool, L. Prime object proposals with randomized Prim's Algorithm. In Proceedings of the 2013 IEEE International Conference on Computer Vision (ICCV), Sydney, Australia, 1–8 December 2013; pp. 2536–2543.
28. Pinheiro, P.O.; Collobert, R.; Dollár, P. Learning to segment object candidates. *Adv. Neural Inf. Process. Syst.* **2015**, 1990–1998.
29. Lu, X.; Wang, W.; Shen, J.; Tai, Y.-W.; Crandall, D.J.; Hoi, S.C.H. Learning video object segmentation from unlabeled videos. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 8957–8967.
30. Rantalankila, P.; Kannala, J.; Rahtu, E. Generating object segmentation proposals using global and local search. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014; pp. 2417–2424.
31. Rahtu, E.; Kannala, J.; Blaschko, M. Learning a category independent object detection cascade. In Proceedings of the 2011 International Conference on Computer Vision (ICCV), Barcelona, Spain, 6–13 November 2011; pp. 1052–1059.
32. Jain, A.; Lioutikov, R.; Niekum, S. Screwnet: Category-independent articulation model estimation from depth images using screw theory. *arXiv* **2020**, arXiv:2008.10518.
33. Tang, P.; Ramaiah, C.; Wang, Y.; Xu, R.; Xiong, C. Proposal learning for semi-supervised object detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACF), Waikoloa, HI, USA, 5–9 January 2021; pp. 2291–2301.
34. Zitnick, C.L.; Dollár, P. Edge boxes: Locating object proposals from edges. In *European Conference on Computer Vision 2014*; Springer: Cham, Switzerland, 2014; pp. 391–405.
35. Zhang, Z.; Liu, Y.; Chen, X.; Zhu, Y.; Cheng, M.-M.; Saligrama, V.; Torr, P.H. Sequential optimization for efficient high-quality object proposal generation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 1209–1223. [CrossRef] [PubMed]
36. Chen, X.; Ma, H.; Wang, X.; Zhao, Z. Improving object proposals with multi-thresholding straddling expansion. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 2587–2595.
37. Zhang, Z.; Torr, P.H. Object proposal generation using two-stage cascade SVMs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 102–115. [CrossRef]
38. Endres, I.; Hoiem, D. Category independent object proposals. In *European Conference on Computer Vision 2010*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 575–588.
39. Wang, X.; Dong, Y.; Zhang, Q.; Wang, Q. Region-based depth feature descriptor for saliency detection on light field. *Multimed. Tools Appl.* **2020**, 1–18. [CrossRef]
40. Alexe, B.; Deselaers, T.; Ferrari, V. Measuring the objectness of image windows. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 2189–2202. [CrossRef]
41. Cheng, M.-M.; Zhang, Z.; Lin, W.-Y.; Torr, P. BING: Binarized normed gradients for objectness estimation at 300fps. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014; pp. 3286–3293. [CrossRef]
42. He, S.; Lau, R.W.H. Oriented Object Proposals. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Las Condes, Chile, 11–18 December 2015; pp. 280–288.
43. Du, Y.-C.; Muslikhin, M.; Hsieh, T.-H.; Wang, M.-S. Stereo vision-based object recognition and manipulation by regions with convolutional neural network. *Electronics* **2020**, *9*, 210. [CrossRef]
44. Uijlings, J.R.R.; Van De Sande, K.E.A.; Gevers, T.; Smeulders, A.W.M. Selective search for object recognition. *Int. J. Comput. Vis.* **2013**, *104*, 154–171. [CrossRef]
45. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
46. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.