Neurocomputing 405 (2020) 138–148

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Brief papers

# RFP-Net: Receptive field-based proposal generation network for object detection

Lin Jiao [a,b,*], Shengyu Zhang [a,c], Shifeng Dong [a,b], Hongqiang Wang [a,b]

[a] Institute of Intelligent Machines, Hefei Institutes of Physical Science, Chinese Academy of Science, Hefei, China
[b] University of Science and Technology of China, Hefei, China
[c] Institutes of Physical Science and Information Technology, Anhui University, Hefei, China

ARTICLE INFO

*Article history:*
Received 26 November 2019
Revised 19 April 2020
Accepted 26 April 2020
Available online 8 May 2020
Communicated by **Shen Jianbing Shen**

*Keywords:*
Object detection
Convolutional neural network (CNN)
Proposals generation
Receptive field
Effective receptive field

ABSTRACT

Recently, object detection has achieved great improvements due to deep CNNs. In this paper, we propose a novel proposal generation network named RFP-Net by mimicking human visual system for high-quality proposals generation. Specifically, RFP-Net takes receptive fields (RFs) as reference boxes to remove many hyper-parameters of anchor boxes that have large sensibility to object detection results. During network training, we select positive samples using an effective RF (eRF) rule instead of the Intersection-over-Union (IoU) rule, which only requires the centroid of a ground truth box to be within the eRF region. This renders RFP-Net learn the representation of region proposals not limited to be of a fixed range of scales and accurately localize the bounding boxes of region proposals around the eRF. RFP-Net also solves the imbalance problem between negative and positive samples with less computational cost. The proposed RFP-Net significantly improves multiply state-of-the-art two-stage and multi-stage detectors. For example, it achieves 43.1% AP by combined it with Cascade RCNN on MS COCO dataset, outperforming previous approaches.



## 1. Introduction

Object detection is one of fundamental machine vision challenges [1,2], aiming to localize and recognize object instances from images and videos with bounding boxes. It is the crucial basis of many computer vision tasks, for example, video object tracking [3–6], edge detection [7], instance segmentation [8] and so forth. Over past years, many deep learning-based object detectors have been developed. Among them, R-CNN [9] and its variants have led to a revolutionary success of object detection in many real-world applications. The idea behind them is to produce region proposals as pre-processing for the final detection of objects in the second stage. Due to scale variation and poor localization, proposal generation, however, is still a bottleneck in most of state-of-the-art R-CNN detection systems. Traditionally, proposal generation methods include selective search (SS) [10], MCG [11], objectness measuring[12] and EdgeBoxes [13]. Most of them perform unsatisfactorily and are time-consuming so that they are impractical in many applications. Recently, Ren et al. [14] presented a deep CNN-based object detection network, Faster R-CNN, which employs a region proposal network (RPN) to generate proposals by sharing features extracted from deep CNN. The RPN introduces multiple anchors at each sliding window of the top feature map as reference boxes for proposal regression, largely improving the accuracy of proposal generation. The anchor mechanism has been thought of to be a corner-stone paradigm for generating region proposals in the two stage object detection frameworks [15–18]. At present, most state-of-the-art one-stage object detection methods also employ such an anchor mechanism for the bounding boxes research, e.g., SSD [19], YOLOv3 [20], and RetinaNet [21], where multiple anchor boxes are preset with different scales and different aspect ratios for regressing the bounding boxes of objects. Nevertheless, such anchor mechanism has the following limitations: (1) Many hyper-parameters need to be preset in advance, e.g., base size, scales and aspect ratios of the anchor boxes, and there exists no guideline for choosing these parameters for a particular data scenario. For example, different base sizes could lead to very different mean Average Precisions (mAPs), as shown in Fig. 1. Moreover, it will become more complex when multi-scale feature architectures, for example, pyramids of image and features, are adopted [17,19]. (2) Positive and negative samples are often seriously imbalanced. The anchor mechanisms determine positive and negative samples based on the IoU overlap between anchor

* Corresponding author at:Institute of Intelligent Machines, Hefei Institutes of Physical Science, Chinese Academy of Science, Hefei, China
*E-mail address:* linj93@mail.ustc.edu.cn (L. Jiao).

https://doi.org/10.1016/j.neucom.2020.04.106
0925-2312/

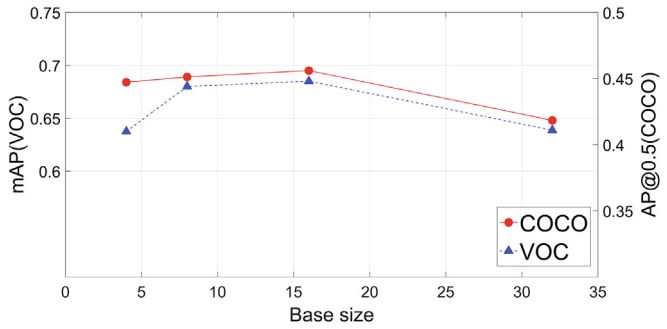


**Fig. 1.** Performances of Faster R-CNN with different base sizes on Pascal VOC2007 (VOC) and MS COCO (COCO) datasets.

boxes and ground-truth (GT) boxes for network training [14–17]. By the IoU rule, a large number of anchors are needed to ensure large overlaps with any GT box, which, however, induces too many negative samples with small or no overlaps with GT boxes. (3) It is deficient to detect small objects. Small objects tend to have little overlap with the anchors preset. For example, the average max IoU for small objects is only 0.29 on COCO dataset [22]. Therefore, there is little chance for small objects to be positive samples, thus reducing the efficiency of learning small objects. Some methods have been proposed to improve the anchor mechanism for proposal generation. For example, Wang et al. [23] proposed to pre-learn anchors for each location in the feature map. The pre-learned anchors reflect possible shapes of objects around the corresponding position and make it efficient to generate proper proposals. Yang et al. [24] presented a flexible anchor mechanism, MetaAnchor, which dynamically generates anchor functions instead of modeling anchors in a predefined manner. Specifically, anchor functions are learned from the customized prior boxes. The resulting anchors are robust to the anchor setting and bounding box distribution. Li et al. [25] devised a map attention decision unit to weight the feature channel input for better proposal generation. Recently, some researches went away with anchors for proposal generation. For example, DeNet [26] established corners-based region-of-interest estimator for proposal generation. The estimator is scene adaptive and does not require predefined reference bounding boxes as usual. Similarly, Zhu et al. [27] designed an anchor-free branch into each level of the feature pyramid of FPN, which allows box encoding and decoding in an anchor-free manner. The discoveries in neuroscience revealed that the human visual system relies on a receptive field (RF) when paying attention to something [28]. Biologically, the RF mechanism has a characteristic of homocentric opponent in response to stimuli of the retina. Inspired by this phenomenon, we here present a new proposal generation method for object detection named RFP-Net. Specifically, we employ the receptive field in place of the dense anchors in the anchor mechanism, and utilize an effective receptive field (eRF) to refine proposal generation. In network training, a new eRF-based rule is applied to determine positive and negative samples, which allows including positive samples of different scales for better learning the representation of region proposals. The main contributions of this paper are listed as follows: (1) The proposed RFP-Net adopts the RFs, instead of random dense anchor boxes, as reference boxes to predict object proposals, and then refines proposals according to the eRFs. Our method effectively avoids designing hyper-parameters in RPN. The adoption of RFs and eRFs is more reasonable and more robust to estimate the distribution of region proposals. (2) We design a novel eRF-based matching rule to select positive and negative samples for network training. It means that only objects paid more attention in the RFs are detected, which conforms to the homocentric opponent mechanism of human visual system. The eRF rule improves the quality of training samples and addresses the imbalance problem between negative and positive samples. (3) Our proposed approach can obtain significant improvements over many state-of-the-art two-stage and multi-stage detectors. For example, it achieves 43.1 AP based on the baseline of Cascade RCNN on MS COCO dataset, outperforming previous approaches. Besides, it can be easily embedded into multiply detectors with few changes.

## 2. Related work

In this section, we briefly review the applications of human visual system in object detection (Section 2.1), then in Section 2.2 and Section 2.3, we introduce two-stage object detection approaches and object proposal generation methods, respectively.

### 2.1. Applications of human visual system on object detection

The cognitive studies of human visual attention behavior confirmed that the HVS can quickly turn attention to the most informative area in the visual scene [29]. Many recent computer vision tasks are solved by mimicking human visual system (HVS), achieving excellent performance. As we know, the most popular field is salient object detection, aiming at extracting salient object regions in the static image or dynamic video [30]. Shen et al. developed a novel object tracking approach by introducing the attention mechanism into Siamese network, outperforming most state-of-the-art trackers on popular tracking benchmarks [3]. Wang et al. designed an effective and efficient deep model for salient detection in videos. Similarly, visual attention prediction task also takes advantage of HVS, and it aims to recognize the fixation locations that human observers would fixate at first glance [31,32]. In [32], a new detection framework, Attentive Saliency Network (ASNet), was proposed, and it could detect the salient object with the help of fixation maps. Wang et al. addressed the problem of photo cropping by building a neural network including attention box prediction (ABP) network and aesthetics assessment (AA) network. By leveraging attention information, much important information can avoid discarding [33]. The comparison experiments show excellent performance of the proposed method. Inspired by the successful application of attention mechanism in vision problems, Hu et al. [34] firstly proposed an adapted attention module for object detection, achieving the first fully end-to-end object detection. In [35], a Receptive Fields module was proposed by mimicking the relationship between the size and eccentricity of receptive field in HVS, which improves deep features.

### 2.2. Two-stage object detectors

Two-stage detection approaches are the mainstream of modern object detections. They implement object detection in two stages: The first stage generates a spare set of region proposals, and the second stage refines the detection based on the resulting proposals. For example, Faster R-CNN [14] uses a region proposal network (RPN) in the first stage to generate proposals from a set of predefined anchor boxes, and then these proposals are fed as regions-of-interest into Fast R-CNN detector for final multi-classes detection. Similarly, Dai et al. [16] developed a R-FCN framework by constructing a set of positive-sensitive score maps in a fully convolutional way, leading to decrease computation time and improve detection accuracy. In FPN [18], Lin et al. constructed feature pyramids using a top-down architecture with lateral connections, accurately addressing the issue of multi-scale object detection.

### 2.3. Proposal generation methods

An object can be located at any position and any scale in one image, therefore, it is natural and cost-efficient to generate proposals in advance [36,37]. Generally, there are three ways for proposal generation: sliding window-based, segmentation-based and grouping-based methods. Among sliding window-based methods, Edgeboxes [13] uses structural edges and a contour detector to compute proposals scores in a sliding window fashion without learning any parameter. Alexi et al. [38] proposed an objectness measure based on image saliency and other cues to score all siding windows and sampled desired number of windows according to their scores. An alternative approach to sliding window methods is the segmentation-based algorithm. For example, Carreira et al. [39] proposed to segment the object of interest based on Graph Cuts algorithm. It produces segments from randomly generated seeds, and each segment denotes a proposal bounding box. Following the grouping-based strategy, Uijlings et al. [10] designed a data-driven grouping-based strategy, Selective Search, to obtain a small set of high-quality object locations. Recently, deep convolutional neural networks have been introduced to proposal generation. In Deepbox [40], Kuo et al. presented a convolutional network model that learns to re-rank proposals generated by EdgeBox [13], a bottom-up method for bounding box proposals. Ghodrati et al. [41] proposed a Deep-Proposal object proposals framework that uses deep convolutional layer features in a coarse-to-fine inverse cascading to obtain possible object proposals in an image. Recently, RPN [14] uses a deep convolutional network to predict object bounding boxes and confidence scores at each position of feature map and obtain high-quality region proposals. Instead of using RPN for proposal generation, we proposed a novel RFP-Net to produce high-quality object proposals by exploiting the homocentric opponent phenomenon in human visual system, and the relationship between receptive field and effective receptive field.

## 3. Proposed method

Our method can combine with two-stage or multi-stage detection frameworks for replacing RPN and achieve multi-classes object detection via an end-to-end way. In this section, we first introduce the theoretical details of the RFP-Net (Section 3.1) and definitions of RF and eRF in CNNs (Section 3.2). Then, we present more details of RFP-Net (Section 3.3), including its network architectures, eRF-based matching strategy, loss function, and the filter module. Finally, the proposed RFP-Net can take place of RPN and be merged into any two-stage object detection framework for object detection.

### 3.1. The homocentric opponent phenomenon (HOP) in human visual system

In real visual system, each neuron responds to a specific area of stimuli that fall on the retina, named receptive field (RF). When eyes are paying attention to something, there occurs a homocentric opponent phenomenon in the RF: Exciting central area but inhibiting peripheral area [42,28], as shown in Fig. 2(a). Rodieck et al. [43] established a Gaussian distribution model for characterizing such homocentric opponent phenomenon: The closer it is to the center of receptive field, the stronger the human eye senses, and the center of the RF are most sensitive visually, as shown in Fig. 2(b).

### 3.2. Definitions of RF and eRF in deep CNNs

Similar to human visual system, we introduce RF for a pixel or a sliding window over the top feature maps in CNNs. As shown in Fig. 2(c), the RF for a pixel $p$ can be defined as a rectangle region in an input image $I$ that the pixel sees:

$$\mathscr{R}_i = \mathscr{G}(I; p; \vartheta) \tag{1}$$

where $\vartheta$ denotes the scaling coefficient($\vartheta = 16$ for VGG16). For a $k \times k$ sliding window with a centroid $(x, y)$, we can obtain its RF with a centroid $(x_r, y_r)$, where $x_r = \vartheta x$ and $y_r = \vartheta y$, and of size $w_r \times h_r$, where $w_r = h_r = k\vartheta$. Given a convolutional feature map of size $w \times h$, we can obtain $w \times h$ RFs totally.

According to the HOP, the magnitude of signals that the pixel can perceive is non-uniform within the RF and follows a Gaussian distribution centered at the centroid of the RF. Then an effective Receptive Field (eRF) can be defined for each RF to be the area that effectively perceives targets within the RF. We argue that only objects whose centroids lie in the eRF can be precisely recognized and localized by the sliding window associated with the RF. Specifically, we take the eRF as a circle region centered at the centroid of the RF with a radius $r$. In CNN, radius $r$ of eRF can be defined as follows:

$$r = kS\sqrt{2}/2 \tag{2}$$

where $k$ is a regulatory factor, and $S$ is the up-sampling factor from the top feature map to the input image. For VGG16 network, $S$ is 16, and $r = k8\sqrt{2}/2$. Note that $S$ can be adjusted according to its up-sampling factor of each pyramid in FPN. The regulatory factor essentially decides the size of eRF and has an influence on localizing and recognizing objects in a CNN-based object detection framework. In this paper, $k$ is set to 1 by experiments as mentioned in Section 4.5.

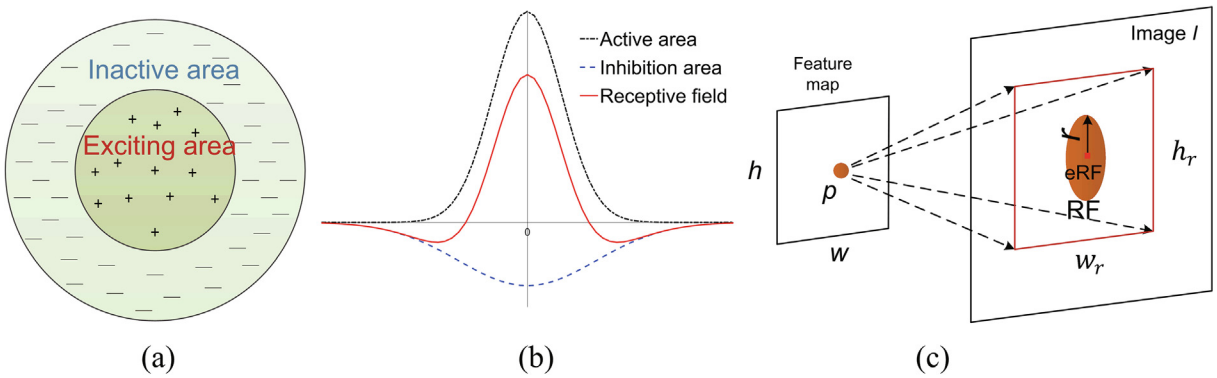


**Fig. 2.** Homocentric opponent phenomenon (a) and its Gaussian model (b) in human visual system and the definitions of RF and eRF in CNNs (c).

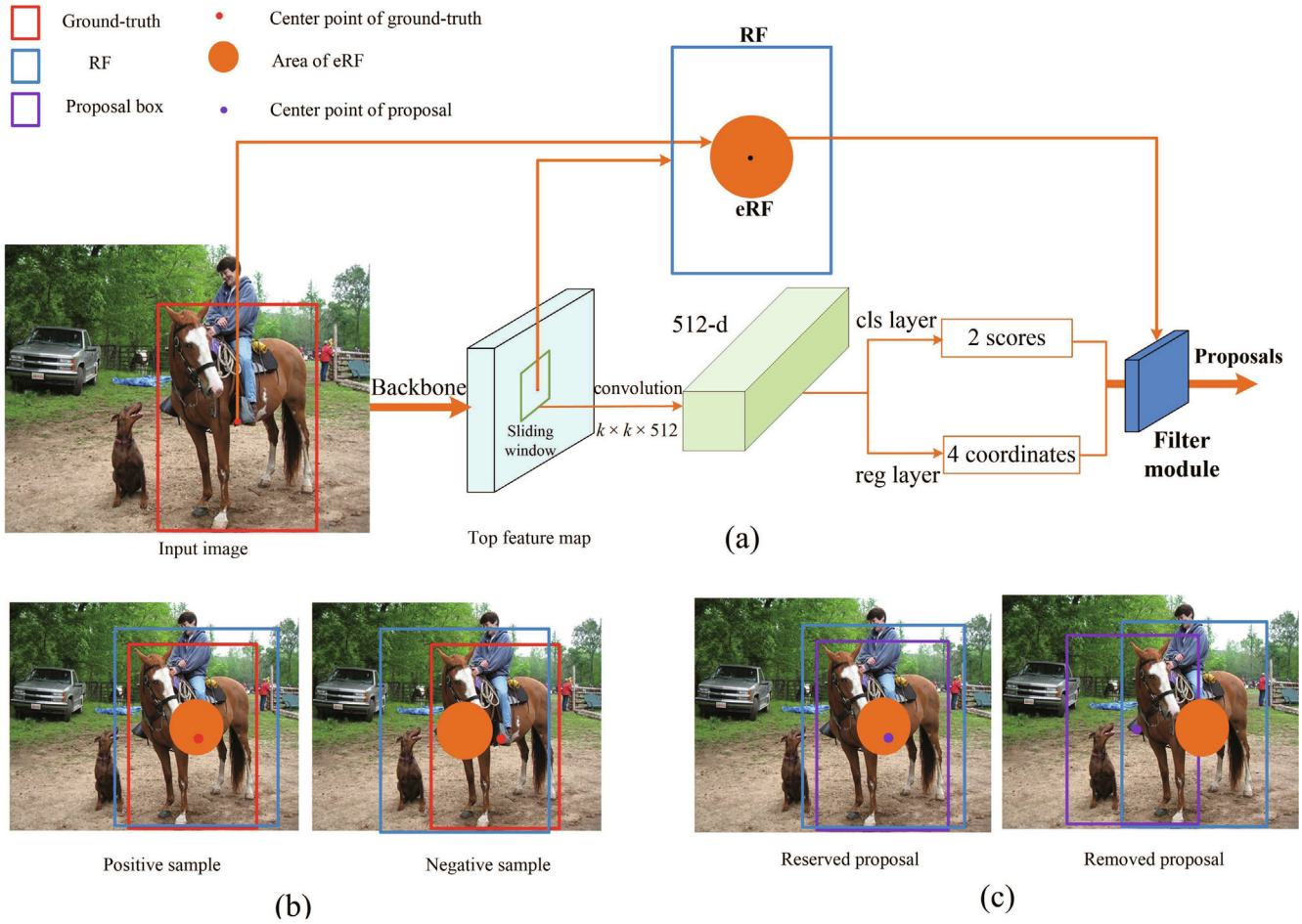


**Fig. 3.** (a) The architecture of our RFP-Net. It includes backbone network (VGG and ResNet) for extraction features, a $k \times k$ convolutional layer for feature transformation, two-sibling branches for refining object proposals, the positive sample selection module(b) and the filter module(c). In (b), the negative sample whose centroid is outside in the associated eRF. In (c), we will reserve the proposal if its center point locates in the area of eRF. See Section 3.3.2 and 3.3.4 for more details about (b) and (c).

### 3.3. RFP-Net

#### 3.3.1. Network structure of the RFP-Net

Fig. 3(a) presents the network structure of our method. Briefly speaking, RFP-Net inputs an image and outputs a set of candidate region proposals with objectness scores and locations. We use VGG16 [44] and ResNet [45] as the backbone for extracting deep features. Over the top convolutional feature map, we slide a small network that takes as input a $k \times k$ window ($k = 3$ in this paper) and transforms the window into a $d$-dimensional feature vectors (512-$d$ for VGG16). Specifically, the small network can be a $k \times k$ convolutional layer with 512 channels followed by ReLu transformation. The 512-$d$ feature is then fed into two siblings $1 \times 1$ convolutional layers: one is with 2 output neurons encoding the probabilities of being objects or not for classification, and another is with 4 output neurons encoding the coordinates of a box for localization. Note that the localization is parameterized as offsets relative to the associated RF. During training, we assign a label (the positive or negative) for each RF sample according to its eRF, as shown in Fig. 3(b). It shows that the RF is positive samples if there exists an object whose centroid lying in the associated eRF. Otherwise, it is a negative sample. Finally, in order to remove low-quality proposals, all output proposals are passed to a filter module, as shown in Fig. 3(c). Following the characteristic of eRF, we design this filter module. Specifically, we screen out the proposals whose centroid is beyond the associated eRF. Our method has an important property of translation invariance in terms of the RFs and the functions that compute proposals relative to the RFs. When an object in an image is translated, the proposal ought to translate and the same function ought to be able to predict the proposals. The MultiBox method [46], in contrast, uses the k-means algorithm to generate 800 anchors without translation invariance. Therefore, MultiBox does not generate the same proposal when an object is translated. Our RFP-Net significantly reduces the numbers of network parameters compared with traditional proposal generation methods. As we know, MultiBox [46] has a $(4+1) \times 800$-dimensional fully-connected output layer, and RPN [12] has a $(4+2) \times 9$-dimensional convolutional output layer, whereas our RFP-Net has a (4 + 2)-dimensional output layer. Totally, our RFP-Net has $3 \times 10^3$ parameters in the output layer, one order of magnitude fewer than RPN and three orders of magnitude fewer than MultiBox.

#### 3.3.2. An eRF-based matching rule for sample selection

Considering the eRF has a higher visual sensitivity than the rest in RF and objects whose centroids lie in the eRF are more likely recognized and localized, we introduce the eRF into the training sample selection. Specifically, we assign a binary class label (an object or not) to each RF (sample) by testing if there is at least one ground truth which is visible by the corresponding eRF or not. To implement the strategy, we design the following eRF matching rule:

Given an RF sample in an image, firstly, we calculate the Euclidean distance between the RF and all GT boxes:

$$d_i = \sqrt{(x_r - x_i^*)^2 + (y_r - y_i^*)^2} \tag{3}$$

where $i$ denotes the order of GT boxes in the image. $(x_r, y_r)$ denotes the center coordinate of the RF and $(x_i^*, y_i^*)$ denotes the center coordinate of the $i-th$ GT boxes. Then, we select the minimum value $d_{min}$ of all distances to compare with the radius $r$ of eRF. If $d_{min} < r$, this RF sample is assigned the positive label and participates in training; otherwise it is the negative label and will be ignored during training. As we know, in the anchor mechanism, there are 9 or more anchors at each sampling location. By the IoU rule for sample selection, there are more chances for these anchors to be negative samples, which constitutes the basic reason for the imbalance between positive and negative samples. In contrast, our proposed approach only samples one RF at each sampling location to be the training sample, and by the eRF-based rule, the chances for being positive and negative samples are equal. This eventually leads to more likely balanced negative and positive samples for network training. Specifically, by RPN, the ratio of foreground and background classes is 1:1500 on PASCAL VOC dataset, while by the eRF-based rule, the ratio reduces to 1:300. Therefore, the RFP-Net can greatly relieve the sample imbalance problem for proposal generation. On the other hand, the use of the eRF-based rule also relieves the scale variation problem. For object detection, the most challenging problem is to recognize objects with different size, especially small objects. RPN designs multi-anchors to solve scale variation problem, however, it needs to adjust scale and aspect ratio for detecting different objects, and when training the model, most small objects will be ignored by its IoU-based matching rule. Therefore, it cannot address scale variation, especially small objects. In contrast, by going away with the IoU-based matching rule, our eRF-based matching strategy can select objects of any size to be training samples, making it fair to learn different sizes of objects.

#### 3.3.3. Loss function

For training RFP-Net, we minimize a multi-task loss function based on the positive and negative samples described above. For each sample, the definition of the loss function is described:

$$L(l, l^*, t, t^*) = L_{cls}(l, l^*) + \gamma[l^* \geqslant 1]L_{loc}(t, t^*) \tag{4}$$

where the classification loss $L_{cls}$ is a cross-entropy loss function for two classes, and the localization loss $L_{loc}$ takes the $smooth - L_1$ bounding box regression loss [39]. $l$ and $l^*$ denote the predicted and ground-truth labels, respectively. $t^*$ denotes the parameterized coordinates of predicted $(t^* = \{t_x^*, t_y^*, t_w^*, t_h^*\})$ and $t$ denotes the parameterized coordinates of the GT bounding box $(t = \{t_x, t_y, t_w, t_h\})$. We parameterize the coordinates of the GT box and the predicted bounding box as follows:

$$\begin{aligned} t_x^* &= (x^* - x_r)/w_r, t_y^* = (y^* - y_r)/h_r \\ t_w^* &= log(w^*/w_r), t_w^* = log(h^*/h_r) \end{aligned} \tag{5}$$

$$\begin{aligned} t_x &= ((x - x_r))/w_r, t_y = ((y - y_r))/h_r \\ t_w &= log(w/w_r), t + w = log(h/h_r) \end{aligned} \tag{6}$$

where $(x^*, y^*)$ is the center coordinate and $w^*, h^*$ are the width and height of GT box, $(x, y, w, h)$ are the counterparts for the predicted box, and $(x_r, y_r, w_r, h_r)$ are for the receptive field. This can be viewed as a regression from a receptive field to its nearby GT box. The term $[l^* \geqslant 1]L_{loc}(t, t^*)$ indicates that the regression loss is activated only for a positive receptive field $(l^* = 1)$. The two terms $L_{cls}$ and $L_{loc}$ are weighted by a balance parameter $\gamma$. We set $\gamma$ to 1 in this paper, which means that the scoring and bounding box regression losses are optimized without biases.

#### 3.3.4. Filter module for duplicate removal

Some proposals may highly overlap with each other. We design a filter module for reducing the redundancy. Like the eRF-based matching strategy, our filter module also depends on eRF. Assume an object proposal centered at a location $(X_p, Y_p)$ and its associated eRF centered at a location $(X_e, Y_e)$ with a radius $r$. The distance between them can be calculated as follows:

$$D = \sqrt{(X_P - X_e)^2 + (Y_p - Y_e)^2} \tag{7}$$

We only filter those proposals with $D \geqslant r$, otherwise, we will reserve them as final object proposals.

### 3.4. Combination of RFP-Net and two-stage detectors

For two-stage detectors, in the first stage, an image is fed into the backbone for extracting deep features and passed to the RFP-Net for generating region proposal candidates. In the second stage, top-300 proposals ranked by scores are recommended to the second stage for fine-tuning detection, predicting the objects categories and a class-specific bounding box. Non-maximum suppression (NMS) algorithm is finally employed for removing redundant boxes for the same objects [47] Because separate training will lead to different convolutional layers. We, therefore, joint-train RFP-Net and the fine-tuning network in the second stage via an end-to-end way, allowing for shared convolutional layers. In each SGD iteration, the forward pass generates proposals which are then fed into the fine-tuning network for training. The backward propagation happens as usual, and for the sharing convolutional layer, the backward propagated signals come from the combination of RFP-Net loss and fine-tuning loss in the second stage.

## 4. Experiment results

### 4.1. Experiment details

**Baseline network:** In order to evaluate the performance of our proposed RFP-Net, we use it to replace RPN and combine it with four popular baseline detectors: Faster RCNN [14], R-FCN [16], FPN [18]and Cascade RCNN [48]. These baselines are mainstream two-stage and multi-stage object detection frameworks, achieving state-of-the-art detection results. Note that we use their default settings (except where noted). **Training settings:** Considering RFP-Net is a fully-convolutional network, we train it via an end-to-end way. Each SGD mini-batch is constructed from a single image that contains 256 positive and negative samples. For each mini-batch, positive and negative samples are randomly selected such that the ratio between positive and positive samples is 1:1. When the number of positive samples is fewer than 128 in an image, we will fill the SGD mini-batch with negative ones. All new layers are initialized from a zero-mean Gaussian distribution with standard deviations 0.01, and the shared convolutional layers are initialized by pre-training a model for ImageNet classification. **Datasets:** Three object detection datasets were used: (1) MS COCO 2017 [1], which involves 80 object categories and contains ~118 k training images, 5 k validation and 20 k testing images (test-dev). Compared with the PASCAL VOC dataset, MS COCO poses challenges in terms of more object classes and smaller objects. We report final results on test-dev set without annotation labels. (2) PASCAL VOC2007 + 2012 [2], which consists of 10 k trainval images for training and the VOC2007 test set for testing. **Evaluation metrics:** For MS COCO dataset, results are verified by the Average Precision (AP) (mean AP for IoU @ [0.5:0.95]), $AP_{50}$ (AP for IoU 50%) and $AP_{75}$ (AP for IoU 75%). We also evaluate the results using $AP_S$, $AP_M$, and $AP_L$, which represent the mAP for small, medium

and large objects, respectively. Besides, we also use mAP and recall to evaluate detection results on PASCAL VOC dataset.

### 4.2. Comparision with state-of-the-art detectors

To evaluate the generalization ability of our proposed method, we conduct experiments on MS COCO test-dev dataset. Table 1 reports the detection results, demonstrating that our method outperforms previous detectors. Compared with baselines, the RFP-Net brings consistent improvement. For example, it improves FPN and Cascade RCNN by 1.3% and 0.7% on ResNet-101 backbone, which is significant for the MS COCO dataset. Note that we obtain the improvement by just taking place of region proposal network of two-stage or multi-stage detectors, therefore, it can be applied conveniently to other detectors with few changes.

### 4.3. Detection results on PASCAL VOC

We also report the performance of the RFP-Net on PASCAL VOC dataset. The networks were trained on VOC 07 + 12 trainval set and tested on VOC07 test set. We use Faster RCNN (with VGG16) and R-FCN (with ResNet) as baselines. The COCO evaluation metrics were used for exploring the detection performance. Table 2 reports the resulting APs on PASCAL VOC data set. It shows that we obtain an improvement of 4.7 and 3.3 points respectively for Faster RCNN and R-FCN detectors. This may result from the robustness of the proposed RFP-Net.

We further examine the mAPs and recalls under different numbers of proposals recommended, as shown in Fig. 4(a). From this figure, it can be seen that our method outperforms Faster R-CNN in both mAPs and recalls, regardless of the numbers of proposals, suggesting that our network is more cost-effective. Specifically, with 20 proposals, our approach has an mAP of 68.0%, 9.6% higher than Faster R-CNN, and a recall of 81.5%, 16.9% higher than Faster R-CNN. Fig. 4(b) shows the changes of recalls and mAPs with the IoU cutoffs by our method and previous methods. We can clearly see that our method always achieved the highest mAPs and highest recalls among the three methods regardless of IoU cutoffs. For example, with an *ad hoc* IoU cutoff of 0.7, our method achieves a 56.0% mAP and 80.8% recall, which are 10%and 23.3% higher than Faster R-CNN, respectively.

### 4.4. Proposals quality of RFP-Net

The role of RFP-Net is to generate region proposals for fine detection in the second stage. Fig. 5 examines the changes of recalls and APs of region proposal generation with the IoU cutoffs in top 50, 100 and 200 proposals. For comparison, Fig. 5 also shows the results by two previous methods, RPN and selective search. From this figure, we can see that RFP-Net always achieves the highest recall and AP rates at all IoU thresholds, regardless of the number of proposals. Specifically, when considering 50 top region proposals, RFP-Net obtained a recall of up to 91% and 32% at an IoU cutoff of 0.5, surpassing RPN by 9% and 11%, respectively, and when considering more region proposals, e.g., 100 and 200, RFP-Net still outperforms RPN by 5% and 2% on recalls and 11% and 11% on APs, respectively.

By fixing the IoU threshold at 0.5, we further investigated the changes of Recalls and APs with the number of proposals, as shown in Fig. 6. From this figure, we can see that compared with the previous methods, RPN and SS, RFP-Net obtained higher Recalls and APs with fewer proposals, suggesting the higher quality of proposals by RFP-Net. Specifically, with top 20 proposals, the recall by RFP-Net is up to 79.2%, 11.6% higher than that by RPN, and the AP is 31.8%, 15.1% higher than that by RPN. Finally, to look into the proposals, we taken three images as examples and laid the top 20 proposals onto the original images, as shown in Fig. 7. For comparison, the results by previous methods, RPN and SS are also illustrated in Fig. 7. We can clearly see that RFP-Net largely reduces redundant and duplicate recommendations and generates high-quality proposals compared with previous methods.

Fig.8(a) shows the relative scale distribution of proposals generated by RFP-Net, RPN and SS, as well as that of GT boxes. Note that the relative scale for a region proposal or a GT box is calculated as the size ratio to the whole image. We can see that RFP-Net results in a similar distribution of scales to that of GT boxes with a wide scale coverage, showing no scale bias. In contrast, RPN is obviously biased toward large objects, while SS is biased toward small

**Table 1**
Comparison results with state-of-the-art methods on COCO dataset. The best results are shown in bold text.

| Methods | Backbone | AP | $AP_{50}$ | $AP_{75}$ | $AP_S$ | $AP_M$ | $AP_L$ |
|---|---|---|---|---|---|---|---|
| YOLO [20] | DarkNet-19 | 21.6 | 44.0 | 19.2 | 5.0 | 22.4 | 35.5 |
| SSD513 [19] | ResNet-101 | 31.2 | 50.4 | 33.3 | 10.2 | 34.5 | 49.8 |
| RetinaNet [21] | ResNet-101 | 39.1 | 59.1 | 42.3 | 21.8 | 42.7 | 50.2 |
| ION [49] | VGG16 | 23.0 | 42.0 | 23.0 | 6.0 | 23.8 | 37.3 |
| DeNet-101 [26] | ResNet-101 | 33.8 | 53.4 | 36.1 | 12.3 | 36.1 | 50.8 |
| Faster RCNN [14] | VGG16 | 23.5 | 43.9 | 22.6 | 8.1 | 25.1 | 34.7 |
| R-FCN [16] | ResNet-101 | 30.3 | 52.2 | 30.8 | 12.0 | 34.7 | 44.3 |
| FPN [18] | ResNet-101 | 36.2 | 59.1 | 39.0 | 18.2 | 39.0 | 50.9 |
| Cascade RCNN [48] | ResNet-101 | 42.4 | 61.1 | 46.1 | 23.6 | 45.4 | 54.1 |
| Faster RCNN + RFP-Net | VGG16 | 24.7 | 45.8 | 24.4 | 8.5 | 27.2 | 38.1 |
| R-FCN + RFP-Net | ResNet-101 | 31.4 | 52.0 | 32.8 | 12.4 | 35.2 | 48.5 |
| FPN + RFP-Net | ResNet-101 | 37.5 | 58.0 | 43.2 | 18.7 | 40.1 | 52.5 |
| Cascade RCNN + RFP-Net | ResNet-101 | **43.1** | **63.8** | **47.3** | **24.4** | **48.3** | **55.3** |

**Table 2**
Object detection results on PASCAL VOC dataset. The best results are shown in bold.

| Methods | Backbone | AP | $AP_{50}$ | $AP_{75}$ |
|---|---|---|---|---|
| Faster RCNN [14] | VGG16 | 41.8 | 73.2 | 43.1 |
| R-FCN [16] | ResNet-50 | 43.8 | 77.1 | 46.8 |
| Faster RCNN + RFP-Net | VGG16 | 46.5 | 76.5 | 50.1 |
| R-FCN + RFP-Net | ResNet-50 | **47.1** | **79.5** | **53.5** |

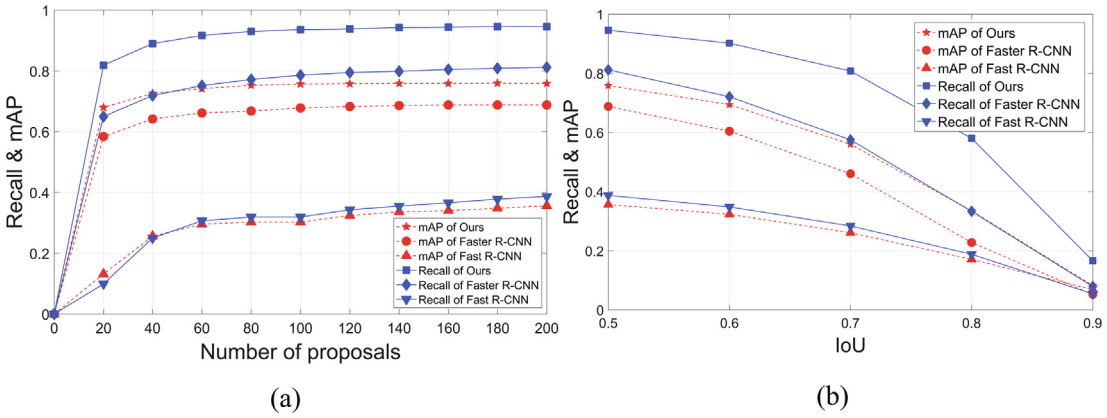


**Fig. 4.** Changes of mAP and Recall with the number of proposals (a) and with IoU cutoffs (b) by our method, Faster R-CNN, and Fast R-CNN. Note that the experiments were performed on the Faster RCNN (baseline).

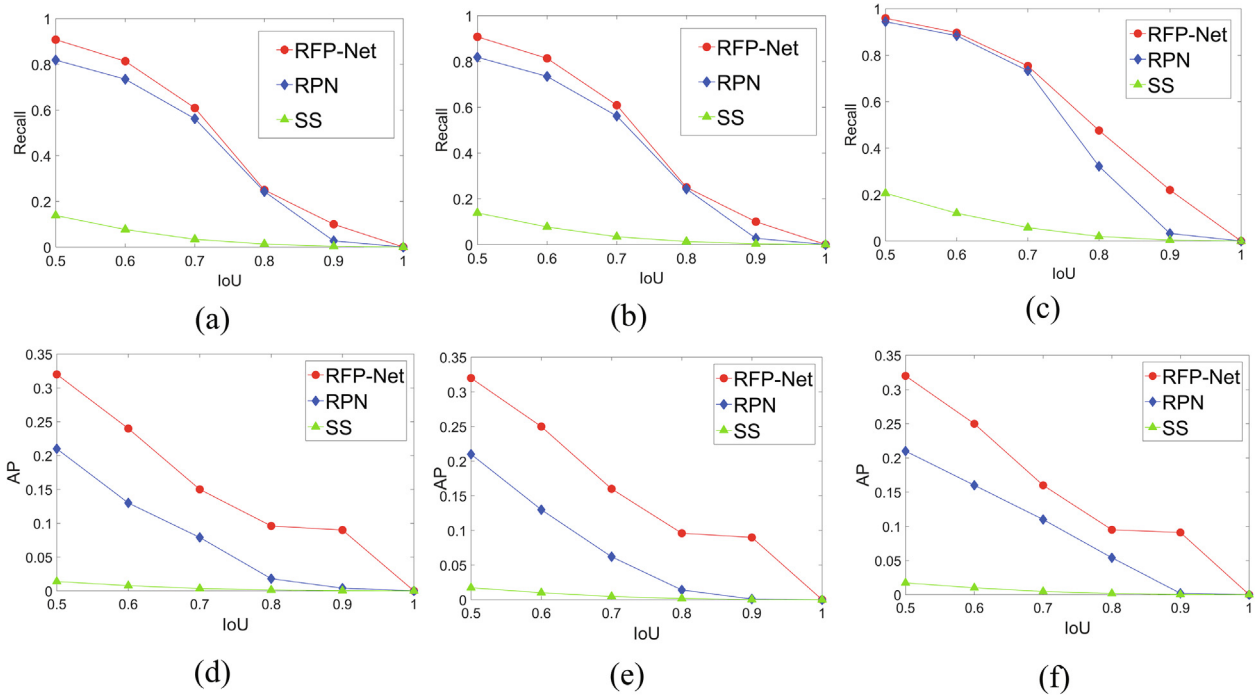


**Fig. 5.** Recalls and APs versus IoU thresholds. (a), (d): 50 proposals. (b), (e): 100 proposals. (c), (f): 200 proposals.

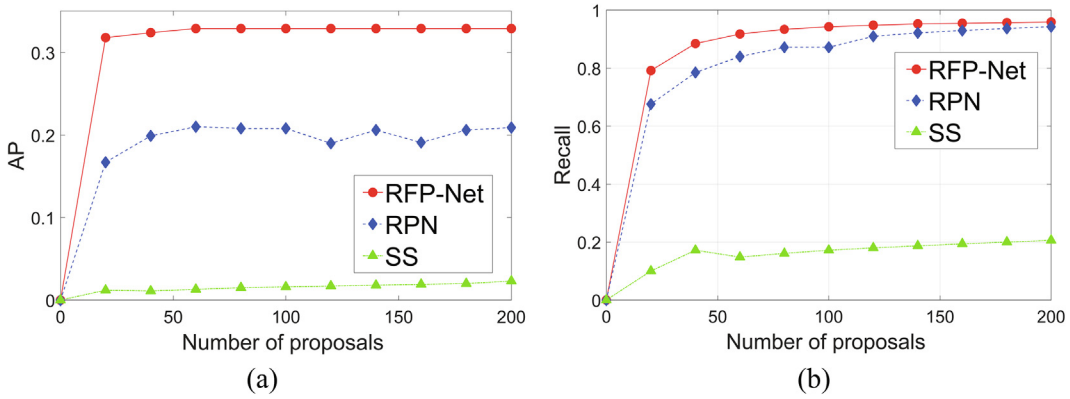


**Fig. 6.** Recall and AP versus Number of proposals with IOU = 0.5 on PASCAL VOC dataset.


**Fig. 7.** Examples of top 20 SS proposals (top row), RPN proposals (middle row) and RFP-Net proposals (bottom row). The red boxes are GT boxes, and the blue boxes are predicted proposals.

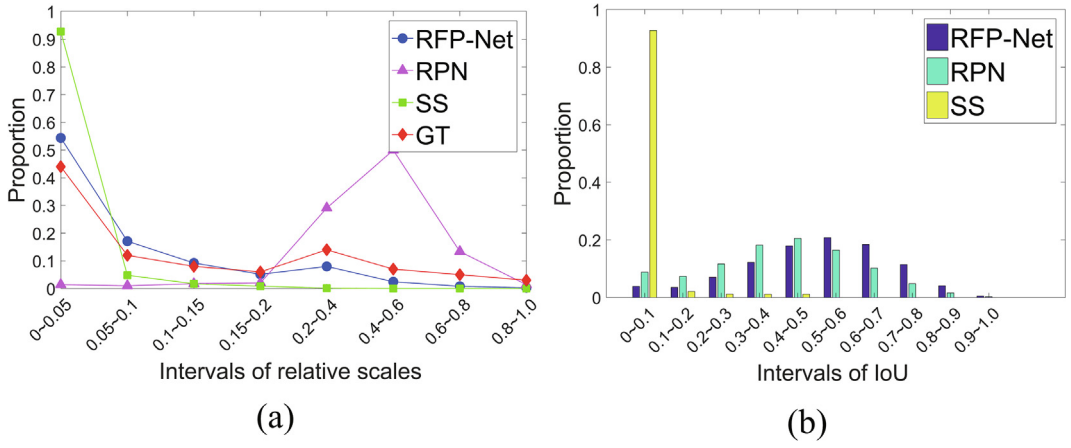


**Fig. 8.** Proportions of GT boxes and proposals by RFP-Net, RPN and SS at different relative scales (a) and IoU intervals (b).

objects. We further examined the IoU of the predicted proposals by different methods with GT boxes. Note that for each proposal, the maximum IoU with any GT boxes were considered. Fig. 8(b) shows the proportions of proposals at different IoU intervals for RFP-Net, RPN and SS. It reveals that compared with RPN and SS, RFP-Net generates a larger proportions of proposals with larger IoUs, suggesting that RFP-Net can more accurately localize the bounding boxes of region proposals. This should be attributed to the inclusion of RF and eRF information for determining positive training samples. Taken altogether, the results suggest that RFP-Net can recommend all possible proposals for objects of any scales.

### 4.5. Influence of the hyper-parameter k

RFP-Net has only one hyper-parameter, the regulatory factor $k$ of the eRF, which plays an important role by specifying the area of eRF. We conduct several experiments to select optimal parameter $k$ on PASCAL VOC and MS COCO datasets, respectively. Fig. 9 shows the changes of mAP with $k$. From this figure, we observe that both mAP and AP first increase sharply and then drop gradually as $k$ increases, suggesting that there exists a best eRF area for an RF. The poor results at small $k$s may be associated with a few positive samples caused by small eRF, while the reduced performance at

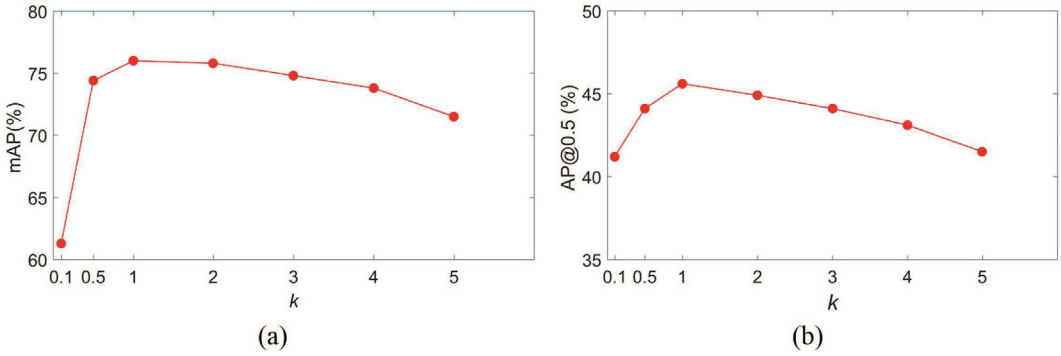


**Fig. 9.** Changes of detection mAPs with the parameter $k$ on PASCAL VOC dataset (a) and MS COCO dataset (b), respectively. Experiments were conducted on Faster RCNN with RFP-Net by using VGG16 backbone.

**Table 3**
Training time (ms/iter) of our method and Faster R-CNN on two benchmarks.

| Methods | PASCAL VOC | MS COCO |
|---|---|---|
| Faster R-CNN | 290 | 315 |
| Faster R-CNN with RFP-Net | 240 | 304 |

**Table 4**
Testing time (ms/img) of our method and Faster R-CNN on two benchmarks.

| | | PASCAL VOC | | MS COCO | |
|---|---|---|---|---|---|
| Methods | Number of proposals | *ms*/img | mAP | *ms*/img | $AP_{50}$ |
| Faster R-CNN | 50 | 62 | 68.2 | 64 | 37.3 |
| | 300 | 71 | 73.2 | 81 | 43.9 |
| Faster R-CNN with RFP-Net | 50 | 58 | 73.8 | 63 | 39.9 |
| | 300 | 67 | 76.5 | 74 | 45.8 |

large *ks* should be due to the degradation of RFP-Net to RPN. Thus, in this paper, the parameter $k$ is set to 1 for VOC dataset and MS COCO dataset.

### 4.6. Computation efficiency

We finally examined the computation efficiency of our method on a single NVIDIA TITAN X (PASCAL) GPU. Tables 3,4 compare the training and testing time of our method with those of Faster R-CNN on PASCAL VOC and MS COCO data sets, respectively. From Table 3, we can see that training our network takes 240 ms/iteration, 40 ms faster than Faster R-CNN on PASCAL VOC, and 304 ms/iteration, 11 ms faster than Faster R-CNN on MS COCO. This is consistent with the case of test time regardless of the number of proposals, as shown in Table 4. These results suggest the higher computation efficiency of our method over previous methods.

## 5. Conclusion

In this paper, we have proposed a new proposals generation network named RFP-Net for object detection. The RFP-Net introduces the concepts of RF and eRF for accurate generation of region proposals. Specifically, the method takes the RF of each sliding window as reference boxes and exploits the eRF area for filtering out low-quality proposals. Additionally, we designed an eRF-based matching strategy to determine positive and negative samples for training RFP-Net, which effectively relieves the imbalance problem of positive and negative training samples and the scale variation problem in object detection. Numerous comparison experiments on PASCAL VOC and MS COCO benchmarks demonstrate the superior region proposals performance of RFP-Net over existing proposal generators. However, as we described in this paper, we imitate the HOP in HVS for selecting positive samples and designing filter module, but do not refer to the network, therefore, the radius of eRF could not learn from CNN. Inspired by applications of attention mechanism in [31,32,50,51], we may try to introduce the attention mechanism into the RFP-Net for predicting the radius of eRF and helping localize objects with a little search space. In the future, we hope these ideas provide improvements to our work on object detection.

## CRediT authorship contribution statement

**Lin Jiao:** Conceptualization, Methodology, Software, Investigation, Writing - original draft. **Shengyu Zhang:** Validation, Formal analysis, Visualization, Software. **Shifeng Dong:** Validation, Formal analysis, Visualization. **Hongqiang Wang:** Resources.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Nos. 61773360, 61973295) and the Anhui Provinces Key Research and Development Project (No. 201904a07020092).

## References

[1] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: common objects in context, in: European conference on computer vision, Springer, 2014, pp. 740–755.
[2] M. Everingham, S.A. Eslami, L. Van Gool, C.K. Williams, J. Winn, A. Zisserman, The pascal visual object classes challenge: a retrospective, Int. J. Comput. Vision 111 (1) (2015) 98–136.
[3] J. Shen, X. Tang, X. Dong, L. Shao, Visual object tracking by hierarchical attention siamese network, IEEE Trans. Cybern. (2019) 1–13, https://doi.org/10.1109/TCYB.2019.2936503.
[4] X. Dong, J. Shen, F. Porikli, Quadruplet network with one-shot learning for visual tracking, arXiv preprint arXiv:1705.07222..
[5] Z. Liang, J. Shen, Local semantic siamese networks for fast tracking, IEEE Trans. Image Process. 29 (2020) 3351–3364, https://doi.org/10.1109/TIP.2019.2959256.

[6] X. Dong, J. Shen, Triplet loss in siamese network for object tracking, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 459–474.
[7] W. Wang, S. Zhao, J. Shen, S.C. Hoi, A. Borji, Salient object detection with pyramid attention and salient edges, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 1448–1457.
[8] W. Wang, J. Shen, R. Yang, F. Porikli, Saliency-aware video object segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 40 (1) (2017) 20–33, https://doi.org/10.1109/TPAMI.2017.2662005.
[9] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580–587..
[10] J.R. Uijlings, K.E. Van De Sande, T. Gevers, A.W. Smeulders, Selective search for object recognition, Int. J. Comput. Vision 104 (2) (2013) 154–171.
[11] P. Arbeláez, J. Pont-Tuset, J.T. Barron, F. Marques, J. Malik, Multiscale combinatorial grouping, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 328–335.
[12] B. Alexe, T. Deselaers, V. Ferrari, Measuring the objectness of image windows, IEEE Trans. Pattern Anal. Mach. Intell. 34 (11) (2012) 2189–2202.
[13] C.L. Zitnick, P. Dollár, Edge boxes: locating object proposals from edges, in: European Conference on Computer Vision, Springer, 2014, pp. 391–405.
[14] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: towards real-time object detection with region proposal networks, Adv. Neural Inf. Process. Syst. (2015) 91–99.
[15] T. Kong, A. Yao, Y. Chen, F. Sun, Hypernet: towards accurate region proposal generation and joint object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 845–853.
[16] J. Dai, Y. Li, K. He, J. Sun, R-fcn: object detection via region-based fully convolutional networks, Adv. Neural Inf. Processing Syst. (2016) 379–387.
[17] T. Kong, F. Sun, A. Yao, H. Liu, M. Lu, Y. Chen, Ron: Reverse connection with objectness prior networks for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5936–5944..
[18] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2117–2125.
[19] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.C. Berg, Ssd: Single shot multibox detector, in: European conference on computer vision, Springer, 2016, pp. 21–37.
[20] J. Redmon, A. Farhadi, Yolov3: An incremental improvement, arXiv preprint arXiv:1804.02767..
[21] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988.
[22] M. Kisantal, Z. Wojna, J. Murawski, J. Naruniec, K. Cho, Augmentation for small object detection, CoRR abs/1902.07296. arXiv:1902.07296.http://arxiv.org/abs/1902.07296..
[23] J. Wang, K. Chen, S. Yang, C.C. Loy, D. Lin, Region proposal by guided anchoring, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2965–2974.
[24] T. Yang, X. Zhang, Z. Li, W. Zhang, J. Sun, Metaanchor: Learning to detect objects with customized anchors, in: Advances in Neural Information Processing Systems, 2018, pp. 320–330..
[25] H. Li, Y. Liu, W. Ouyang, X. Wang, Zoom out-and-in network with map attention decision for region proposal and object detection, Int. J. Comput. Vision 127 (3) (2019) 225–238.
[26] L. Tychsen-Smith, L. Petersson, Denet: Scalable real-time object detection with directed sparse sampling, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 428–436..
[27] C. Zhu, Y. He, M. Savvides, Feature selective anchor-free module for single-shot object detection, arXiv preprint arXiv:1903.00621..
[28] D.H. Hubel, T.N. Wiesel, Receptive fields, binocular interaction and functional architecture in the cat's visual cortex, J. Physiol. 160 (1) (1962) 106–154.
[29] D.-P. Fan, W. Wang, M.-M. Cheng, J. Shen, Shifting more attention to video salient object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2019, pp. 8554–8564.
[30] W. Wang, Q. Lai, H. Fu, J. Shen, H. Ling, Salient object detection in the deep learning era: An in-depth survey, arXiv preprint arXiv:1904.09146..
[31] W. Wang, J. Shen, X. Dong, A. Borji, R. Yang, Inferring salient objects from human fixations, IEEE transactions on pattern analysis and machine intelligence..
[32] W. Wang, J. Shen, X. Dong, A. Borji, Salient object detection driven by fixation prediction, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1711–1720.
[33] W. Wang, J. Shen, H. Ling, A deep network solution for attention and aesthetics aware photo cropping, IEEE Trans. Pattern Anal. Mach. Intell. 41 (7) (2018) 1531–1544.
[34] H. Hu, J. Gu, Z. Zhang, J. Dai, Y. Wei, Relation networks for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3588–3597.
[35] S. Liu, D. Huang, et al., Receptive field block net for accurate and fast object detection, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 385–400.
[36] H. Harzallah, F. Jurie, C. Schmid, Combining efficient object localization and image classification, in: 2009 IEEE 12th international conference on computer vision, IEEE, 2009, pp. 237–244.
[37] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, IEEE Trans. Pattern Anal. Mach. Intell. 32 (9) (2009) 1627–1645.
[38] B. Alexe, T. Deselaers, V. Ferrari, What is an object?, in: 2010 IEEE computer society conference on computer vision and pattern recognition, IEEE, 2010, pp 73–80.
[39] J. Carreira, C. Sminchisescu, Constrained parametric min-cuts for automatic object segmentation, 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE (2010) 3241–3248.
[40] W. Kuo, B. Hariharan, J. Malik, Deepbox: learning objectness with convolutional networks, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 2479–2487.
[41] A. Ghodrati, A. Diba, M. Pedersoli, T. Tuytelaars, L. Van Gool, Deepproposal: Hunting objects by cascading deep convolutional layers, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 2578–2586.
[42] D.H. Hubel, T.N. Wiesel, Receptive fields of single neurones in the cat's striate cortex, J. Physiol. 148 (3) (1959) 574–591.
[43] R.W. Rodieck, Quantitative analysis of cat retinal ganglion cell response to visual stimuli, Vision Res. 5 (12) (1965) 583–601.
[44] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556..
[45] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[46] C. Szegedy, S. Reed, D. Erhan, D. Anguelov, S. Ioffe, Scalable, high-quality object detection, arXiv preprint arXiv:1412.1441..
[47] A. Rosenfeld, M. Thurston, Edge and curve detection for visual scene analysis, IEEE Transactions on computers (5) (1971) 562–569..
[48] Z. Cai, N. Vasconcelos, Cascade r-cnn: delving into high quality object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 6154–6162.
[49] S. Bell, C. Lawrence Zitnick, K. Bala, R. Girshick, Inside-outside net: detecting objects in context with skip pooling and recurrent neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2874–2883.
[50] W. Wang, J. Shen, Deep visual attention prediction, IEEE Trans. Image Process. 27 (5) (2017) 2368–2378.
[51] Q. Lai, W. Wang, H. Sun, J. Shen, Video saliency prediction using spatiotemporal residual attentive networks, IEEE Trans. Image Process. 29 (2019) 1113–1126.

**Lin Jiao** received the M.S. degree at the College of Mechanical and Electronic Engineering, Northwest A&F University in 2018, shaanxi, China. She is currently pursuing the Ph.D. degree in computer science and application with the University of Science and Technology of China, Hefei. His current research interests include image processing and computer vision.

**Shengyu Zhang** is currently pursuing the M.S. degree at Anhui University Hefei, China. His rearch interests include pattern recognition and object detection.

**Shifeng Dong** received the B.E. degree in automation from the Hefei University, Hefei, China, in 2017. He is currently pursuing the M.S. degree in control engineering with the University of Science and Technology of China, Hefei. His current research interests include computer vision and deep learning.

**Hongqiang Wang** received the B.E. degree in mechanical automation from the Hefei University of Technology, Hefei, China, in 2000, and studied for M.S. degree and the Ph.D from 2000 to 2005 in pattern recognition and intelligent system from the University of Science and Technology of China, Hefei, China, in 2005. He has been a Research Associate and Research Fellow with the Hong Kong Polytechnic University and City University of Hong Kong from 2006 to 2008. He has been a Postdoctoral Reacher Fellow in the University of Georgia, U.S.A. He is currently a Researcher with the Institute of Intelligent Machinery, Chinese Academy of Sciences. His main research interests include Intelligent computing and bioinformatics.