This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2576759, IEEE Transactions on Circuits and Systems for Video Technology

Object Proposal Generation with Fully Convolutional Networks

Zequn Jie, Wen Feng Lu, Siavash Sakhavi, Yunchao Wei, Eng Hock Francis Tay, Shuicheng Yan

***Abstract*—Object proposal generation, as a pre-processing technique, has been widely used in current object detection pipelines to guide the search of objects and avoid exhaustive sliding window search across images. Current object proposals are mostly based on low-level image cues, such as edges and saliency. However, "objectness" is possibly a high-level semantic concept showing whether one region contains objects. This paper presents a framework utilizing fully convolutional networks (FCN) to produce object proposal positions and bounding box location refinement with SVM to further improve proposal localization. Experiments on the PASCAL VOC 2007 show that using high-level semantic object proposals obtained by FCN, the object recall can be improved. An improvement in detection mean average precision (mAP) is also seen when using our proposals in the Fast R-CNN framework. Additionally, we also demonstrate that our method shows stronger robustness when introduced to image perturbations, e.g., blurring, JPEG compression and "salt and pepper" noise. Finally, the generalization capability of our model (trained on the PASCAL VOC 2007) is evaluated and validated by testing on PASCAL VOC 2012 validation set, ILSVRC 2013 validation set and MS COCO 2014 validation set.**

***Index Terms*—Object proposals, fully convolutional networks, box location refinement, deep learning.**

I. Introduction

Object proposal generation has become crucial for object-based vision tasks, like class-specific object detection and semantic segmentation. Instead of dealing with 10^6 to 10^7 bounding boxes across all possible scales in a sliding window manner [1], object proposal generation aims to find all candidate regions that may contain objects in an image [2]. Compared with the sliding window scheme, object proposals benefit the object detection in two aspects: saving computation time spent on the tremendous number of sliding windows and improving the detection accuracy by enabling the use of more sophisticated detectors [3], [4], [5] due to the smaller number of inputs passed to the detector.

A generic object proposal generator should normally satisfy the following requirements: it should be able to capture objects of all scales, have small biases towards object class, achieve high recall with a manageable number of proposals (from several hundred to a few thousand per image) and be computationally efficient.

Current object proposal generators primarily rely on low-level image cues, such as saliency, gradient and edge information [6], [7]. The main rationale behind these methods is that all objects of interest share common visual properties that can easily distinguish them from the background. However, sometimes visual appearance variation of objects makes it difficult for low-level cues to distinguish them from background (e.g., a girl wearing green dress running on the grassland, or in the forest with messy background and strong texture). Therefore, "objectness" is more of a high-level semantic concept showing semantic information of a region, which implies the presence of objects better than the low-level cues. In addition, when faced with image perturbations (e.g., blurring, JPEG compression and "salt and pepper" noise) which may cause big low-level appearance variation, such a semantic definition of objectness also provides stronger robustness and stability.

In this paper, we present a data-driven learning pipeline to produce a high-level semantic objectness score, which shows to what extent a specific region may contain an object. Briefly, we train an object/non-object binary classifier using a fully convolutional network (FCN) [8] on patches from images with annotated objects. The fully convolutional network can take an input image of arbitrary size and output a dense "objectness map" showing the probability of containing an object for each corresponding box region in the original image. An example is shown in Fig. 1. To predict the objectness for boxes of different scales, we rescale the original image into multiscales and feed them to the network to obtain the objectness maps of different scales correspondingly. Then, non-maximal suppression (NMS) is performed to remove redundant low-quality proposals. Finally, we train an SVM on the image gradients to refine the proposals by finding the proposal with the highest objectness score among the neighboring boxes of each rough proposal obtained by FCN.

Extensive experiments on the PASCAL VOC 2007 [9] demonstrate the superiority of our approach both on the object recall rate and class-specific object detection mAP. The robustness is investigated by testing perturbed images from PASCAL VOC 2007 and the generalization ability of the approach is validated using ILSVRC 2013 [10]. The remainder of the paper is organized as follows. First, we survey the related works on object proposal generation in Section II. Then, we elaborate the multiscale fully

Zequn Jie is with Keio-NUS CUTE center of Interactive and Digital Media Institute, National University of Singapore, e-mail: jiezequn@u.nus.edu

Wen Feng Lu and Eng Hock Francis Tay are with Department of Mechanical Engineering, National University of Singapore.

Siavash Sakhavi is with Department of Electrical and Computer Engineering, National University of Singapore and Institute for Infocomm Research (I2R), A*STAR.

Yunchao Wei is with the Institute of Information Science, Beijing Jiaotong University and also with Department of Electrical and Computer Engineering, National University of Singapore.

Shuicheng Yan is with Department of Electrical and Computer Engineering, National University of Singapore.

1051-8215 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2576759, IEEE Transactions on Circuits and Systems for Video Technology

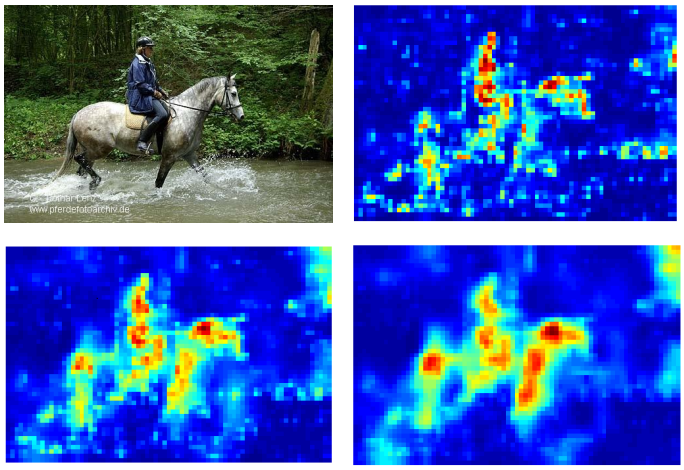


Fig. 1: Illustration of original image (left top), objectness map of scale 32*32 (right top), objectness map of scale 64*64 (left bottom) and objectness map of scale 128*128 (right bottom). All the objectness maps have been scaled up to the same size as the original image. Each pixel in the objectness map shows the probability of a corresponding box region containing an object.

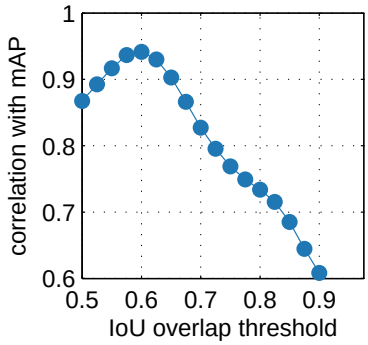


Fig. 2: Correlation between detection mAP of Fast R-CNN on the PASCAL VOC 2007 and recall at different IoU thresholds [11].

convolutional networks for object proposal generation in Section III. Subsequently, the proposal refinement with SVM is introduced in Section IV. After showing the experimental results and analysis in Section V, we make some discussions and the conclusion in Section VI.

The main contributions of the paper can be summarized as follows.

1) We propose a novel way to predict the objectness of densely distributed image patches simultaneously in a single pass of Fully Convolutional Networks (FCN). Based on the high-level semantic objectness from CNN, object/non-object prediction becomes more accurate. Benefited from FCN, time cost is significantly reduced compared to the one-by-one CNN pass strategy. Moreover, the positions of proposals can be directly decided by mapping the output neuron in the objectness map back to its receptive field in the image. No position regression is needed like other CNN-based methods.

2) We train a new small-scaled (much shallower than VGG and GoogleNet) object/non-object classification CNN model which can be adapted to the dense objectness prediction for images with arbitrary sizes. Combined with the multiscale sampling strategy, the model shows strong discrimination power which is validated by extensive experiments.

II. Related Work

The existing approaches for generating object proposals can be classified into two types: *Segment grouping methods* and *Window scoring methods* [11]. Apart from these, we also list the related approaches for object proposals/detection which are based on Convolutional Neural Networks (CNN).

Segment grouping methods aim to generate multiple segments that may contain objects. This type of methods typically depends on an initial oversegmentation (e.g., superpixels [12]). Then different merging strategies are adopted to group the similar segments into object proposals. Similarity measures usually rely on diverse low-level cues, e.g., shape, color and texture. For example, Selective Search [13] greedily merges superpixels to generate proposals in a hierarchical scheme without learning. Randomized Prim [14] learns a randomized merging strategy based on the superpixel connectivity graph. Rantalankila et al. [15] used superpixel merging combined with graph cuts to generate proposals. Multiscale Combinatorial Grouping (MCG) [16] utilizes multi-scale hierarchical segmentation and merges them based on edge strength to obtain proposals. Geodesic Object Proposal (GOP) [17] starts from over-segmentation, and then computes a geodesic distance transform and selects certain level sets of the distance transform as the object proposals.

Usually this type of methods achieves high recall when the intersection over union (IoU) threshold criterion is relatively large (>0.7), indicating the precise localization ability. However, when choosing a relatively loose IoU threshold criterion (<0.7), the recall may not be as good as *Window scoring methods*. In addition, high quality proposals of these methods are often obtained by multiple segmentations in different scales and colorspaces, thus they are computationally expensive and more time-consuming.

Window scoring methods are designed to show how likely a candidate window is to contain an object of interest. Generally, this type of methods first initializes a set of candidate bounding boxes across scales and positions in the image, and then sorts them with a scoring model and selects the top ranked boxes as object proposals. Objectness [18] selects some salient locations from an image, and then sorts them according to multiple low-level cues, e.g., color, edge, location and size. Zhang et al. [19] proposed a cascade of SVMs trained on gradient features to estimate the objectness. The SVMs are trained for different scales and the method outputs a pool of boxes at each scale, followed by another SVM to rank all these obtained boxes. BING [6] trains a simple linear SVM classifier over the gradient map and applies it in a sliding window manner when testing. Using binary approximation enables it to be finished within 10ms

1051-8215 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2576759, IEEE Transactions on Circuits and Systems for Video Technology

per image. Edge Boxes [7] is also performed in a sliding window manner and scores the windows based on the edge maps obtained by some edge detection techniques [20]. Then, box refinement is used to improve localization precision.

Compared to *segment grouping methods*, *window scoring methods* are usually computationally efficient as they do not output a segmentation mask. Another advantage of them is the high recall when setting a relatively low IoU threshold criterion (<0.7). The main drawback of this type is the poor localization accuracy due to the discrete sampling of the sliding windows, leading to a low recall given a high IoU threshold criterion. However, the recent findings [11] showed that object detection mean average precision (mAP) has the strongest correlation with the recall at IoU threshold around 0.6, and the correlation decreases with the increasing of the IoU threshold, as shown in Fig. 2. This suggests that high recall at a relatively low IoU threshold is more important than precise localization of the proposals for achieving a good detection mAP.

CNN in object proposal/detection. CNN, as a popular deep learning model, is also utilized for object proposal/detection tasks. Overfeat [21] trains a deep CNN to simultaneously predict the box coordinates and category confidence for each object in a sliding window manner to solve the class-specific object detection problem. MultiBox [22], [23] trains a CNN to directly regress a fixed number of proposals without sliding the network over the image and then ranks the proposals by their CNN confidences of being the bounding box of an object. They achieve top results on the ImageNet detection task. Karianakis [24] extracted the convolutional responses of an image from first layers of the CNN, and then fed them to a boosting model which differentiates object proposals from background. Pinheiro [25] trained a CNN to output a class-agnostic segmentation mask and the likelihood of the patch being centered at a full object for each patch in an image. Trained on the expensive pixel-level labeled images, they reported top recall on both PASCAL and Microsoft COCO benchmarks [26].

Another type of approaches does not output the proposals by themselves, and instead they re-rank the proposals generated by other methods. DeepBox [27] re-ranks the proposals of other methods based on their CNN output values which reflect the high-level objectness and improve the object recall. Each proposal is fed into the network to obtain an objectness value and thus a high time cost is required to pass all the proposals (usually several thousand) separately to the CNN network. Salient Object Subitizing [28] trains a CNN to identify the number of salient objects in an image and selectively reduces the number of retrieved proposals according to the predicted number of salient objects. The recall of other object proposal methods can be improved by allocating a proper number of proposals in this way.

Our method can also be categorized as a *Window scoring method*. The difference between our approach and the existing *Window scoring methods* is the window scoring scheme. We use fully convolutional networks to output high-level semantic objectness maps instead of judging from low-level cues. The most similar method to ours may be Region Proposal Networks (RPN), which is used in the Faster R-CNN detection pipeline for class-agnostic proposal generation. RPN predicts proposals for each image region in a sliding window manner based on a set of pre-defined anchors in the region. Compared to other CNN-based object proposal methods, our FCN neither generates only a fixed number of proposals nor needs expensive pixel-level labeled training samples, and it can be end-to-end both in training and testing stages. Another difference is that we do not regress the box coordinates like OverFeat [21], MultiBox [22], [23] and RPN [29], and instead decide the window which the pixel in the output map corresponds to as a proposal. Combining such a mapping localization method with the multi-scale scheme obtains better precision than the box coordinates regression. To improve the localization precision, a learning based refinement method is utilized to iteratively search for a window with a higher objectness score.

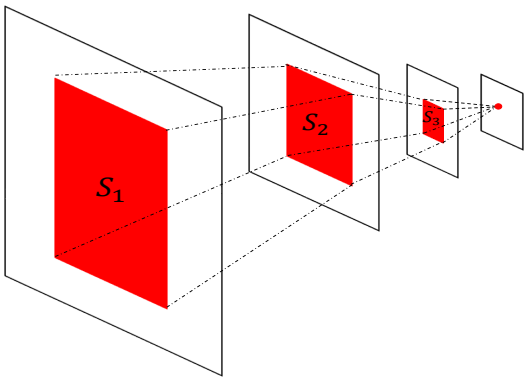


Fig. 3: Illustration of fully convolutional network. Red pixels in the output map show the classification confidence of the red window region S_1 in the input image and will not be affected by other regions in the input image.

III. Multiscale Fully Convolutional Networks

A. Fully Convolutional Networks for Dense Objectness Prediction

Convolutional Neural Network (CNN) can be seen as an automatic hierarchical feature extractor combined with a single classifier. Such a learning-based deep feature extraction pipeline avoids hand-crafted feature designing which may not be suitable for a particular task, and meanwhile strengthens the discrimination power of the feature. Recently, as an extension of the classic CNN for classification problems [30], [31], [32], fully convolutional networks can take an input of arbitrary size and output a map whose size corresponds to the input, which can be used for dense prediction problems (e.g., semantic segmentation [8], [33], image restoration [34] and depth estimation [35]).

1051-8215 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2576759, IEEE Transactions on Circuits and Systems for Video Technology

TABLE I: **Fully convolutional network architecture.** The spatial size of the feature map depends on the input image size, which varies during our inference step. Here we show training spatial sizes.

Layer	1	2	3	4	5	6	7	8	9	10	11
Type	conv	conv	conv+max pool	conv	conv	conv+max pool	conv	conv	conv+max pool	conv	conv
#channels	64	64	64	128	128	128	256	256	256	512	2
Conv. kernel size	3×3	3×3	3×3	3×3	3×3	3×3	3×3	1×1	1×1	3×3	1×1
Conv. stride	1×1	1×1	1×1	1×1	1×1	1×1	1×1	1×1	1×1	1×1	1×1
Pooling size	-	-	2×2	-	-	2×2	-	-	2×2	-	-
Pooing stride	-	-	2×2	-	-	2×2	-	-	2×2	-	-
Zero-padding size	1×1	1×1	1×1	1×1	-	-	-	-	-	-	-
Spatial input size	40×40	40×40	40×40	20×20	20×20	18×18	8×8	6×6	6×6	3×3	1×1

We feed the whole image into the fully convolutional network to obtain a dense objectness map. This feed-forward process can be seen as object/non-object binary classification for the densely sampled sliding windows in the input image. Each output pixel in the objectness map shows the classification confidence of one specific sliding window in the input image, as illustrated in Fig. 3. To map back to the input object proposal boxes from the output objectness map, we have to decide how big the area is that the output pixel can correspond to in the input image (receptive field size). Assume the receptive field size of each layer is S_i (i=1, 2, ..., n) and S_1 is the receptive field size in the input image. The receptive field size of each layer can be computed using the recursive formula below:

$$S_{i-1} = up(S_i) = s_i(S_i - 1) + k_i \tag{1}$$

where s_i and k_i represent the stride and the convolution kernel size of the i^{th} convolutional or pooling layer. S_{i-1} and S_i denote the receptive field size of the $(i-1)^{th}$ and the i^{th} layer respectively. To accurately map an output pixel back to the window region it covers in the input image, apart from knowing the receptive field size in the input image, the sliding window sampling stride *Str* is also indispensable. A fully convolutional network has its inherent sampling stride *Str*, which is the product of the strides of all the layers, i.e.,

$$Str = \prod_{i=1}^{n} s_i \tag{2}$$

where s_i indicates the stride of the i^{th} convolutional or pooling layer. With the known S_1 and *Str*, the window region which corresponds to the output pixel (x_o, y_o) can be decided as below:

$$\begin{aligned} x_{min} &= x_o Str \\ x_{max} &= x_o Str + S_1 \\ y_{min} &= y_o Str \\ y_{max} &= y_o Str + S_1. \end{aligned} \tag{3}$$

In contrast to other sliding window approaches that compute the entire pipeline for each window, fully convolutional networks are inherently efficient since they naturally share computation common to different overlapping regions. When applying a fully convolutional network to the input of an arbitrary large size in testing, convolutions are applied in a bottom-up manner so that the computation common to neighboring windows only needs to be done once.

B. Network Architecture and Patch-wise Training

For the implementation of our idea, a new fully convolutional network architecture for objectness prediction is designed and trained from scratch. The detailed architecture of the network is shown in Table I. The network architecture is similar to VGG [31]: the first two pooling layers follow three convolutional layers with kernel size 3. The last two 1×1 convolutional layers follow the idea of Network in Network (NIN) [36], and they can be seen as the cascaded cross channel pooling structure allowing complex and learnable interactions of cross channel information. All the convolutional layers are followed by a ReLU non-linear activation layer. On the top of the network, a softmax normalization layer is used to ensure the output confidence within the range (0, 1). The loss function to be optimized is the cross-entropy loss, i.e.,

$$E = t_k \ln(y_k) + (1 - t_k) \ln(1 - y_k) \tag{4}$$

where t_k denotes the k^{th} target value and y_k represents the k^{th} prediction value. According to Eqn. (1) and Eqn. (2), the receptive field size of the input S_1 for this network is 40 and the sampling stride for the network is 8. It is worth mentioning that the sampling stride is 0.2 timing the receptive field size (also the window size), which is close to the empirically optimal sliding window sampling stride ratio recommended by [7]. This is the main factor we consider in designing the network.

In terms of training, we treat the network as an object/non-object binary classification network and use a patch-wise training strategy instead of using the whole images to train a dense structured prediction network. To this end, we crop the patches from the images with annotated objects and resize them to 40×40, the same as S_1. Among the cropped patches, those with IoU $\geq$ 0.5 with a ground-truth box are treated as positive samples and the rest as negatives. To balance the number between the positives and the negatives, we crop multiple patches around each ground-truth box while sparsely sampling the patches in the background regions.

1051-8215 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2576759, IEEE Transactions on Circuits and Systems for Video Technology

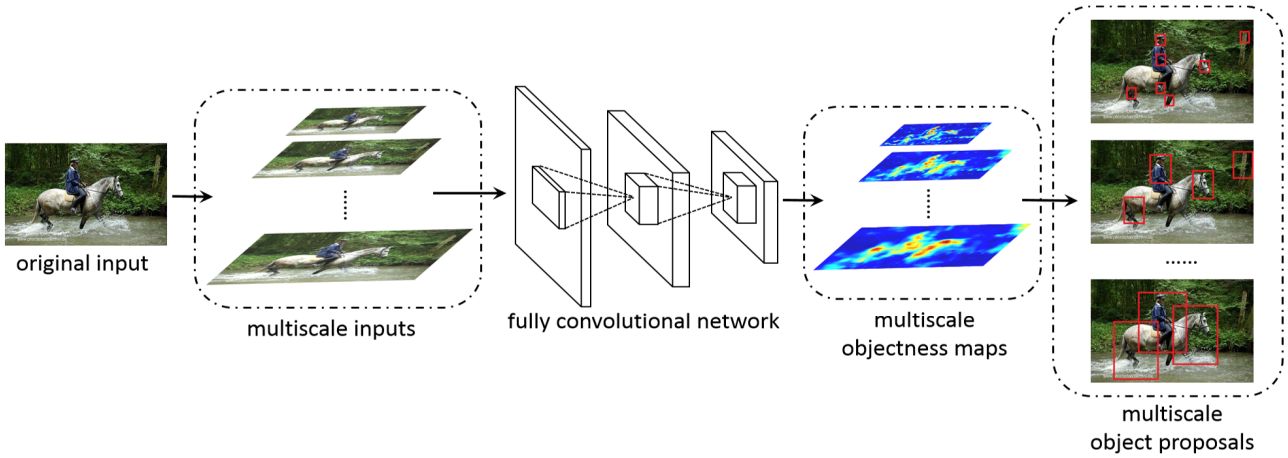


Fig. 4: Pipeline of multiscale object proposal generation by a single fully convolutional network.

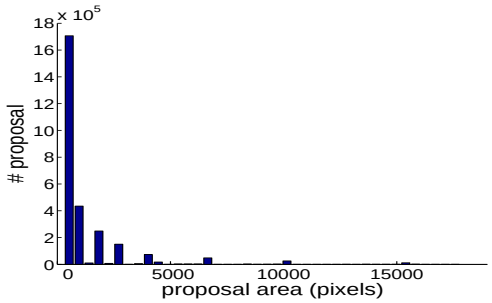


Fig. 5: The distribution of the proposal area.

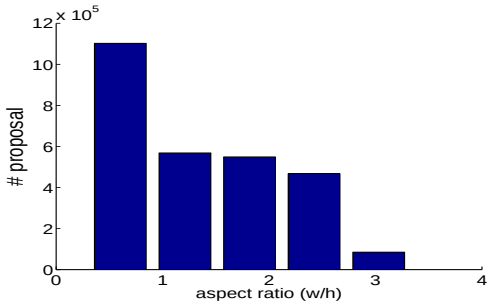


Fig. 6: The distribution of the aspect ratio of the proposals.

For the stochastic gradient descent (SGD) training process, the weights of all the layers are initialized with a zero-mean Gaussian distribution with the standard deviation 0.01 and the biases are initialized with 0. The learning rate is 0.01, which will be reduced by a scale of 10 after every 20 epochs. The minibatch size is set as 256.

C. Multiscale Inputs Inference

Using the above mentioned fully convolutional network, each pixel in the output map only covers a window region with a fixed size 40. To enable the network to predict the object proposals with different sizes and aspect-ratios, we rescale the original image to different scales. By doing this, a window with the size equal to the receptive field size 40 in the rescaled inputs will correspond to windows of different scales in the original image. The rescaled input size S_r can be computed according to the original input size S_o and the size of a window region which is needed to be detected using Eqn. (5) denoted as S_w.

$$\frac{S_r}{40} = \frac{S_o}{S_w}. \tag{5}$$

Subsequently, the multiscale inputs after rescaling are fed into the network individually to obtain the multiscale objectness maps (see Fig. 1). It can be seen that the map corresponding to a small scale (32×32) characterizes the boundaries better but can hardly capture the internal regions of the objects. In contrast, the map corresponding to a large scale (128×128) focuses more on the localization of the whole big objects but is unable to depict the boundary details well. Therefore, we utilize the multiscale strategy to generate the object proposals in all kinds of scales. The pipeline of our method is illustrated in Fig. 4.

Here we present the multiscale setting in detail to specify the scales needed in our approach. Generally, the more and the denser the scales are, the more a concentrated set of bounding boxes near the areas is likely to contain an object. However, the downside is that noisy bounding boxes which may lower the recall of the top candidate boxes will be produced as well. This issue introduces a tradeoff in parameter selection for the multiscale setting.

Specifically, we define α as the stepsize indicating the IoU for neighboring boxes. In other words, the step sizes in scale and aspect ratio are determined such that one step results in neighboring boxes having an IoU of α. The scale values range from a minimum box area of 1000 pixels to the full image. The aspect ratio changes from 1/3 to 3. The exact values of the scale and the aspect ratio can be computed with Eqn. (6) and Eqn. (7).

$$\text{scale} = \sqrt{1000}(\sqrt{1/\alpha})^{\text{s}}, \tag{6}$$

$$\text{aspect ratio} = (\frac{1+\alpha}{2\alpha})^{\text{r}}. \tag{7}$$

Here the index s can be any integer from 0 to $\lfloor \log(\text{image size}/\sqrt{1000})/\log(\sqrt{1/\alpha}) \rfloor$, and the index r can

1051-8215 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2576759, IEEE Transactions on Circuits and Systems for Video Technology

be any integer from $-\lfloor \log(3)/\log(\frac{1+\alpha}{2\alpha})^2 \rfloor$ to $\lfloor \log(3)/\log(\frac{1+\alpha}{2\alpha})^2 \rfloor$. A value of α = 0.65 is ideal for most of the cases [7] so we fix α as 0.65 in the experiments. The distribution of the proposals in terms of their areas and aspect ratios are shown in Fig. 5 and Fig. 6 respectively, from 100 images which are randomly selected from PASCAL VOC 07 test set when setting α as 0.65. For the multiscale proposals, we first remove those with objectness lower than 0.2, reducing the total proposal number from several tens of thousands to less than 10000. Next, we sort all the remained proposals based on their objectness in a descending order. Finally, non-maximal suppression (NMS) is performed on the sorted proposals. Specifically, we find the proposal with the maximum objectness score and remove all the proposals with an IoU larger than an overlap threshold (we use 0.8 in all our experiments).

Algorithm 1 Refine the proposals $[P_1, P_2, ..., P_n]$

Require: : A set of raw proposals $[P_1, P_2, ..., P_n]$
  for $P = [P_1, ...P_n]$ **do**
    $Obj \Leftarrow svm(P)$
    $S_c \Leftarrow 0.2P_w$ (P_w is the proposal box width)
    $S_r \Leftarrow 0.2P_h$ (P_h is the proposal box height)
    while $S_c > 2$ *and* $S_r > 2$ **do**
      $[P_{c1}, P_{c2}, ..., P_{cn}] \Leftarrow ColomnNeighbors(P, S_c)$
      $P_{cmax} \Leftarrow argmax(svm(P_{ci})),\ i = [1, 2, ..., n]$
      if $svm(P_{cmax}) > Obj$ **then**
        $Obj \Leftarrow svm(P_{cmax})$
        $P \Leftarrow P_{cmax}$
      end if
      $[P_{r1}, P_{r2}, ..., P_{rn}] \Leftarrow RowNeighbors(P, S_r)$
      $P_{rmax} \Leftarrow argmax(svm(P_{ri})),\ i = [1, 2, ..., n]$
      if $svm(P_{rmax}) > Obj$ **then**
        $Obj \Leftarrow svm(P_{rmax})$
        $P \Leftarrow P_{rmax}$
      end if
      $S_c \Leftarrow S_c/2$
      $S_r \Leftarrow S_r/2$
    end while
  end for

IV. Box Refinement with Gradients Cues

Due to the fixed multiscale setting and box sampling strategy, the above obtained raw proposals have their inherent weakness of being pre-defined both in scales and locations which may cause misdetection of ground-truth boxes. To overcome this, we adopt a greedy iterative search method to refine each raw proposal.

Previous works show that objects are stand-alone things with well-defined closed boundaries and centers [18], [37], [38]. Based on this observation, gradient and edge information are widely used for implying the presence of objects in early works, e.g., BING [6] and Edge Boxes [7]. Considering this, we rely on the low-level gradients cues instead of 3-channel RGB information for the efficient implementation of our method. Specifically, we train a linear SVM object/non-object classifier on the gradient maps of the patches from the images with annotated objects. We use the ground-truth boxes of the annotated objects as positive samples, and crop the patches in the images and treat those with IoU $<$ 0.3 for all the ground-truth boxes as negative samples. For all the chosen samples, we resize them to 16×16 before training the SVM. Having trained the 256-d SVM, to refine the proposals, we maximize the SVM score of each box over the neighboring positions, scales and aspect ratios. After each iteration, the search step is reduced in half. The search is stopped once the translational step size is less than 2 pixels. The procedure is summarized in pseudo-code in Algorithm 1.

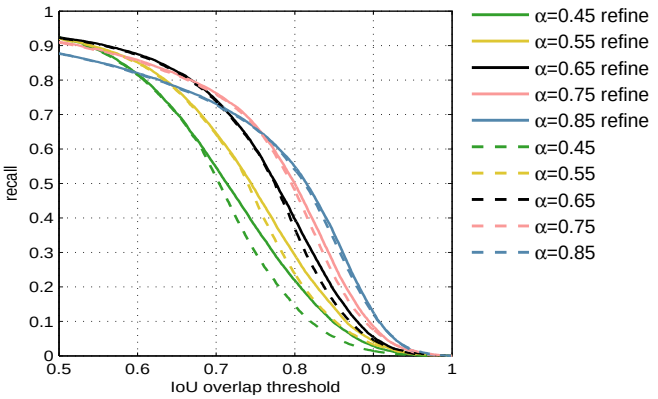


Fig. 7: Recall versus IoU threshold for various search stepsizes α (1000 proposals per image) on the PASCAL VOC 2007 test set.

TABLE II: Runing speed of FCN with different α and the refinement step.

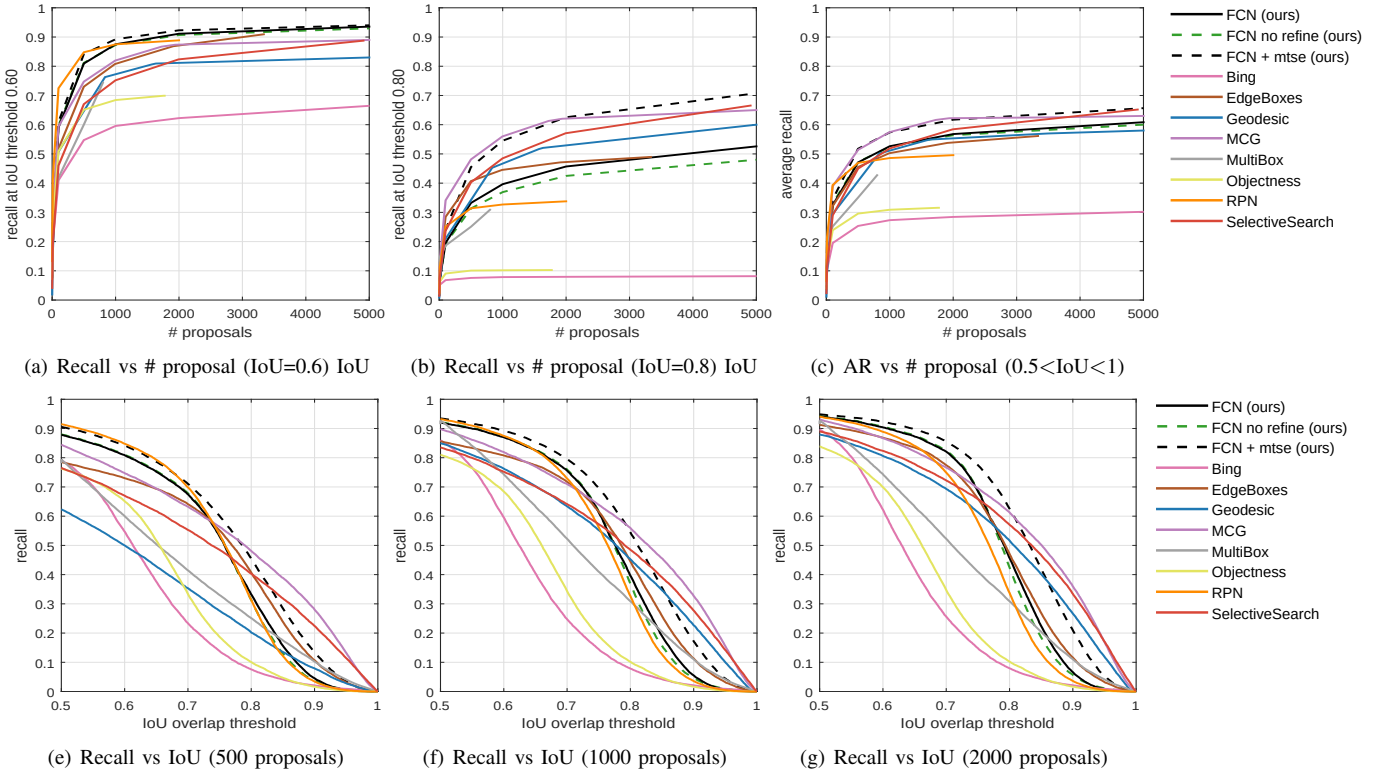
	Runing time per image
α = 0.45 no refine	0.53s
α = 0.45 refine	0.62s
α = 0.55 no refine	0.66s
α = 0.55 refine	0.77s
α = 0.65 no refine	0.95s
α = 0.65 refine	1.10s
α = 0.75 no refine	2.12s
α = 0.75 refine	2.39s
α = 0.85 no refine	5.23s
α = 0.85 refine	5.83s

V. Experiments and Discussion

In this section, we evaluate the performance of our method on PASCAL VOC 2007 test set, PASCAL VOC 2012 validation set, ILSVRC 2013 validation set and MS COCO 2014 validation set. To be fair, similar to other supervised learning based methods, we train the fully convolutional network on the PASCAL VOC 2007 trainval set, which contains 5011 images and around 15000 annotated objects. Our method will be compared with the state-of-the-art in terms of the following four parts: object

1051-8215 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

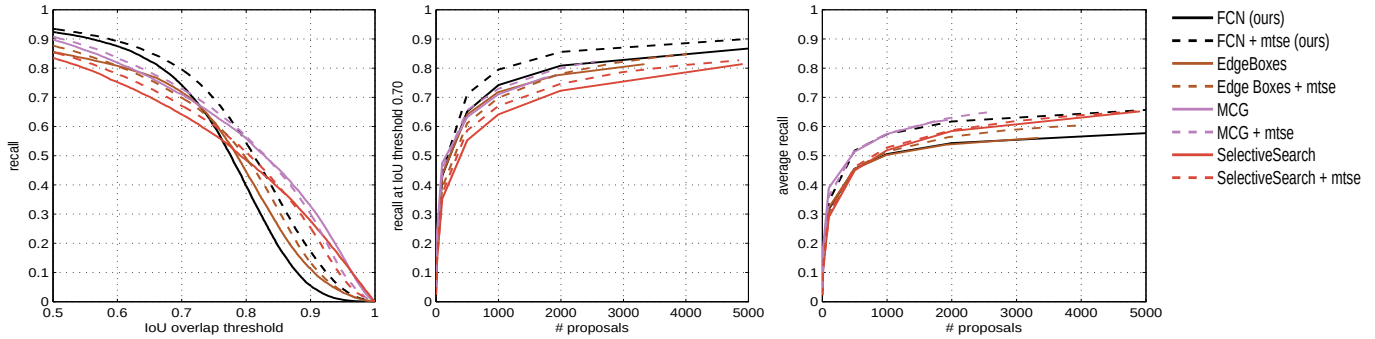
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2576759, IEEE Transactions on Circuits and Systems for Video Technology



(a) Recall vs # proposal (IoU=0.6) IoU (b) Recall vs # proposal (IoU=0.8) IoU (c) AR vs # proposal (0.5<IoU<1)

(e) Recall vs IoU (500 proposals) (f) Recall vs IoU (1000 proposals) (g) Recall vs IoU (2000 proposals)

Fig. 8: Recall comparison between the FCN method and other state-of-the-art methods on PASCAL VOC 2007 test set.



(a) 1000 proposals per image (b) Recall at 0.7 IoU (c) Average recall (IoU between (0.5, 1))

Fig. 9: Recall comparison between methods with MTSE refinement w.r.t different IoU thresholds on PASCAL VOC 2007 test set.

recall, detection mAP, robustness to image perturbation, generalization to unseen categories.

A. Approach Variants

We begin the experiments by testing different variants of the approach with various parameter settings. First, we analyze the effects of the granularity of the multi-scale search as well as the box refinement step. Fig. 7 shows the algorithm's behavior based on the search stepsize parameter α and the refinement step, when generating 1000 proposals per image.

As the stepsize α increases, the scales to be computed are increased, leading to more CNN feed-forward passing times. From Fig. 7, when α is between 0.45 to 0.65, as α increases, recall increases for all the IoU thresholds between 0.5 to 1. This is natural as more scales provide more chances to have a proposal close to the groundtruth bounding box. However, when α exceeds 0.65, as α increases, recall at high IoU thresholds ($>$0.7) increases while recall at low IoU thresholds ($<$0.7) decreases. The reason probably lies in that too many boxes concentrated on a small area are introduced, resulting in a loss of the recall for top-selected candidate proposals. From Fig. 7, α should be set as 0.65 or 0.75.

Another critical component to be evaluated is the box refinement step. Fig. 7 also shows the effect of the refinement step for different search stepsizes α. As can be seen from Fig. 7, the refinement step indeed improves the recall for all the stepsizes α. However, the smaller the stepsize α is, the more recall improvement is brought by the refinement step. Another finding is that the refinement step only improves the recall at high IoU thresholds and has little effect on the recall at low IoU thresholds. This suggests that

1051-8215 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2576759, IEEE Transactions on Circuits and Systems for Video Technology

the refinement step mainly refines the coarsely localized proposals to fine localized ones, which means improving the IoU of the coarsely localized proposals from > 0.5 to even higher values (e.g., > 0.7).

We also conduct the running time comparison experiment for each search stepsize α and the refinement step on the PASCAL VOC 2007 test set. Table II presents the detailed running time for various values of α and the refinement step. It is found that for a certain value of α, the time spent on the refinement step is relatively much less than the multi-scale FCN feed-forward computation, e.g, 0.09s for α=0.45, 0.11s for α=0.55 and 0.15s for α=0.65. The major time cost is on the multi-scale FCN computation and when α=0.75, the running time can reach a rather 2.39s with the refinement step. Although setting α as 0.75 achieves a higher recall at high IoU thresholds than 0.65 according to Fig. 7, we still fix α as 0.65 in all the later experiments for the tradeoff between the recall and the running speed.

B. Object Recall

When using object proposals for detection, it is crucial to have a good coverage of all the objects of interest in the testing image, because the missed objects can never be recovered in the subsequent classification stage. Therefore it is a common practice to evaluate the proposal quality based on the object recall. We compare our method with many state-of-the-art methods, including BING [6], CPMC [39], Edge Boxes [7], Geodesic Object Proposal [17], MCG [16], Objectness [18], and Selective Search [13].

*1) **Metrics**:* In class-independent object proposals, one of the primary metrics is the object recall, for a fixed IoU threshold, as the number of proposals is changed. Another widely used metric is, for a fixed number of proposals, the object recall as the IoU threshold is varied.

*2) **Results**:* We first evaluate recall on the PASCAL VOC 2007 test set, which contains 4952 images with about 15000 annotated objects (including the objects labeled as "difficult") in 20 categories. For the recall computation, the same as [11], we compute the matching between the proposals and the ground-truths so that one proposal cannot cover two ground-truth objects. Fig. 8(a), 8(b) and 8(c) present the recall when varying the number of proposals for different IoU thresholds. We choose two commonly used IoU thresholds, i.e., 0.6 and 0.8 for evaluation, around which the recall shows the strongest correlation with detection mAP (see Fig. 2). In addition, we plot the average recall (AR) versus the number of proposals curve for the methods. This is because AR summarizes proposal performance across IoU thresholds and correlates well with detection performance. It can be seen that our approach performs better than most of the existing methods at IoU threshold 0.6 for both small and large numbers of proposals. The advantage of our approach reaches the maximum for a small number of proposals (e.g., < 1000), suggesting that our approach can roughly localize the positions of objects with a small number of proposals. For IoU threshold 0.8, our method does not work well, even though the box refinement step boosts the recall by about 5%. This implies that our method does not perform well in localizing objects with very high precision (with IoU < 0.8). As for average recall, our method is only slightly lower than MCG which is the best one in terms of AR. Fig. 8(e), 8(f) and 8(g) demonstrate the recall when the IoU threshold changes within the range [0.5, 1]. It can be seen that no single method can take the dominant place across all IoU thresholds. However, our approach takes the lead by a wide margin when IoU ranges from 0.5 to about 0.75. Please note that we directly employ the publicly available MultiBox model trained on ILSVRC benchmark to extract the proposals. It is surprising that MultiBox does not work well compared to other state-of-the-art methods. We attribute its inferior performance to the poor generalization from ILSVRC benchmark to PASCAL VOC benchmark. For RPN, we directly use the publicly released model which is trained on PASCAL VOC 2012 dataset. It is found that RPN performs slightly better than ours for low IoU thresholds (e.g., 0.6) with a small number of proposals (e.g., $<$1000), but suffers from poor localization accuracy for high IoU thresholds. This is probably because that RPN does not utilize the multi-scale prediction strategy since multi-scale inference generates much more proposals which brings better results for a large number of proposals and high IoU thresholds but worse recall for a small number of proposals and low IoU thresholds.

Another finding is that the recall of our approach decreases sharply with the increasing of IoU threshold when it is above 0.75. The phenomenon that *window scoring methods* usually outperform *segment grouping methods* for low IoU thresholds while fall behind for high IoU thresholds is also found for other *window scoring methods*, e.g., BING and Edge Boxes. A possible explanation lies in the inherent drawback of *window scoring methods* that discretely sample windows over pre-defined positions and scales.

To improve the poor recall of our method for high IoU thresholds ($>$0.8), Multi-Thresholding Straddling Expansion (MTSE) [40] can be introduced to adjust our proposals to be better aligned with the boundaries of the superpixels. From Fig. 8, it can be seen that FCN+MTSE almost takes the first place in all the evaluation cases. To be fair, we also conduct experiments to compare the recall of other state-of-the-art methods with MTSE refinement with ours. Fig. 9 demonstrates that compared with other methods with MTSE refinement, the FCN method with MTSE achieves better recall for low IoU thresholds (i.e., < 0.8). When looking at high IoU thresholds and average recall between IoU 0.5 to 1, MCG with MTSE refinement performs the best.

For better visualization of our proposals, we show the distribution of the proposals of our method as well as Edge Boxes and MCG for comparison in Fig. 10. The distribution figures are obtained by assigning red color to the proposal regions according to the density of the proposals in that region. It is clear that the proposals of our method are more tightly concentrated on the objects. In contrast, the proposals of Edge Boxes and MCG often spread evenly across a much larger region rather than the objects of interest.

1051-8215 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2576759, IEEE Transactions on Circuits and Systems for Video Technology

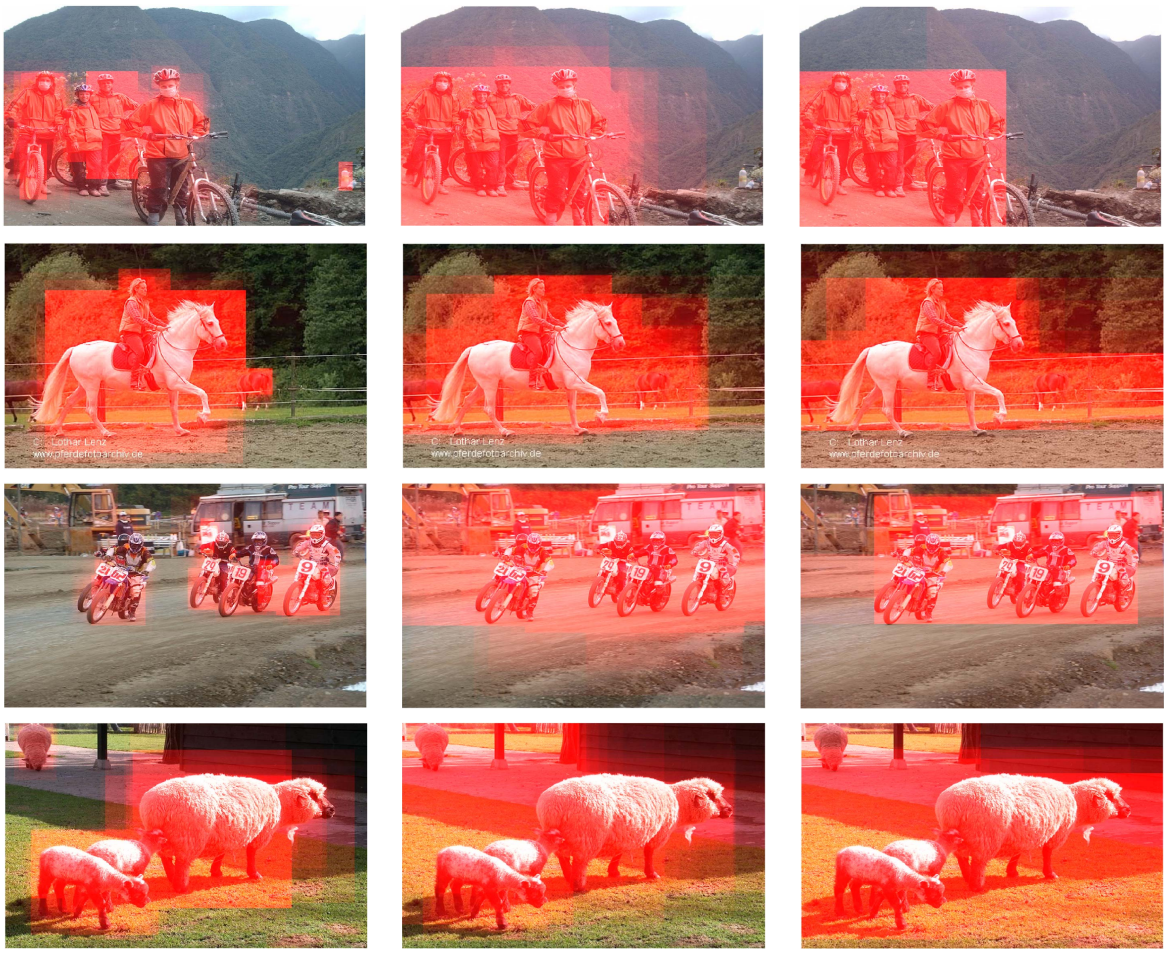


Fig. 10: Examples of the proposal distribution of Edge Boxes, MCG and our method. Top 2000 proposals are illustrated for each image. For each row, our FCN is on the left, Edge Boxes is in the middle and MCG is on the right.

3) ***Speed****:* The detailed running speed of our FCN method as well as other state-of-the-art methods is presented in Table III. The detailed setting of parameters for each method is as follows. We choose the single color space (i.e., RGB) proposal computation for BING, and the "Fast" version for selective search. For the rest methods, we directly run their default codes. Inference for an image of PASCAL VOC size takes 1.1s for our FCN method. Although it is not one of the fastest object proposal methods (compared to BING and Edge Boxes), our approach is still competitive in speed among the proposal generators. We do, however, require use of the library Caffe [41] which is based on GPU computation for efficient inference like all CNN based methods. To further reduce the running time, some CNN speedup methods such as FFT, batch parallelization, or truncated SVD could be used in the future.

TABLE III: Runing speed of the state-of-the-arts and our method.

	Runing time per image
BING	0.01s
CPMC	250s
Edge Boxes	0.3s
Geodesic	1s
MCG	30s
Objectness	3s
Selective Search	10s
Our method (no refine)	0.95s
Our method	1.1s

C. Object Detection Performance

In this subsection we analyze object proposals for use with object detectors to evaluate the effect of proposals on the detection quality. We utilize the recently released Fast R-CNN [42] detector as the benchmark. For fast evaluation, we adopt the AlexNet [30] instead of the VGG net [31] as the model. The proposals obtained by our approach and another three state-of-the-art object proposal generators, i.e., Edge Boxes, Selective Search and MCG are used as training samples for fine-tuning the Fast R-CNN detector. Object proposals having IoU $\geq$ 0.5 with a ground-truth bounding box are positive samples and the rest are negatives. For each method, only the top 2000 proposals are chosen to fine-tune

1051-8215 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2576759, IEEE Transactions on Circuits and Systems for Video Technology

the Fast R-CNN detector.

The detection mean average precision (mAP) and the average precisions of all the 20 categories are presented in Table IV. It can be seen that our approach wins on 8 categories among the 20 categories of PASCAL VOC 2007 in terms of detection average precision and also achieves the best mAP 57.3%. Considering that our approach cannot obtain as good recall as the rest three methods when IoU threshold is greater than 0.8, the good detection performance of our approach supports the finding that recall at a very high IoU threshold is not a good predictor for detection mAP compared with the recall at around 0.6 [11], which is shown in Fig. 2.

D. Robustness

The distribution of the object proposals is quite different from that of sliding windows both on the positive and negative samples used for training a class-specific detector. This requires the proposal generators to be able to consistently propose stable object proposals on the slightly different images with the same image content. This property is associated with the object proposal robustness (called "repeatability" in [11]) when faced with image perturbation. To investigate the proposal robustness, we generate the perturbed versions of the images in the PASCAL VOC 2007 test set and evaluate the robustness faced with three kinds of perturbation, i.e., JPEG artifacts, blurring and "salt and pepper" noise (see Fig. 11).

1) ***Metrics****:* For each pair of the original image and the perturbed image we generate the proposals (top 1000 proposals) for each method. The proposals of the perturbed image are mapped back to the original image and matched to the proposals of the original image. Matching is performed at different IoU thresholds. Next, we plot the recall for every IoU threshold and define the robustness as the area under this "recall versus IoU threshold" curve between IoU 0 and 1. By doing this, the methods which generate proposals at similar locations for the original image and the perturbed image will obtain higher robustness.

2) ***Results****:* Fig. 12(a) shows the robustness of the methods faced with JPEG artifacts. The perturbed images are obtained by writing the images with Matlab "imwrite" function with different compression quality settings from 5% to 100% (see Fig. 11). Because even 100% quality setting is still lossy in the image quality, we include a lossless setting. It can be seen that except for 5% quality, our methods (both the refined one or the non-refined one) lead across all the compression qualities by a wide margin. Fig. 12(b) demonstrates the robustness after blurring with different degrees. The blurred images are obtained by smoothing the original images using a Gaussian kernel with standard deviations $0 \leq \sigma \leq 8$ (see Fig. 11). Similarly, our methods outperform the others significantly in all the cases. It is worth mentioning that the non-refined version of FCN outperforms the refined version, mainly because the refined version relies on the image gradients which are heavily affected by the blurring. Fig. 12(c) presents the robustness

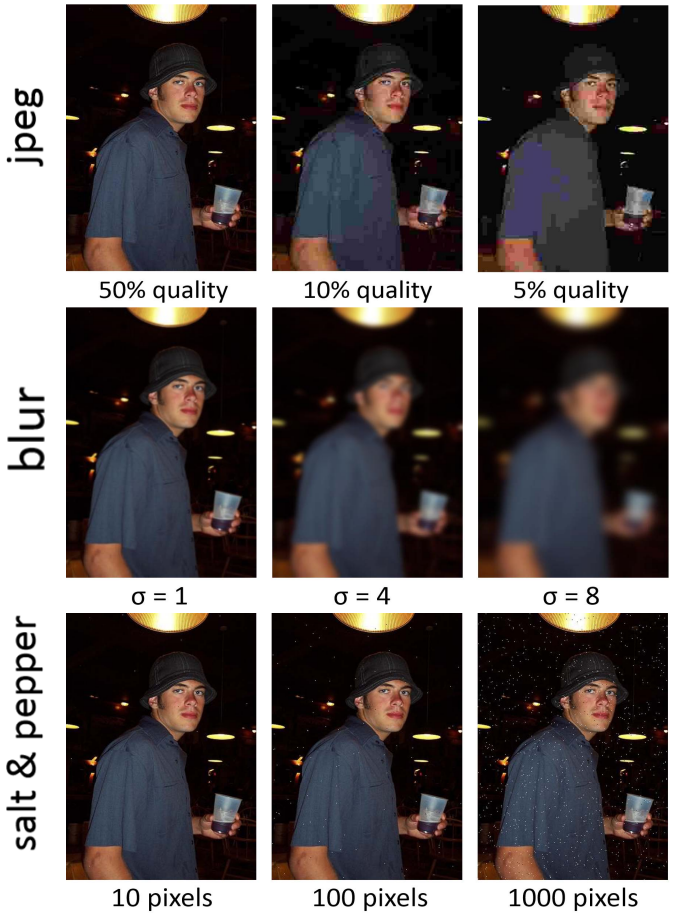


Fig. 11: Illustration of the perturbation images for the robustness experiments.

faced with salt and pepper noise. The noise is produced by adding the noise to the image in between 1 to 1000 random locations. Our methods (both the refined one and the non-refined one) almost achieve the same robustness with BING, which is the best among the state-of-the-art ones.

In general, we find that the *segment grouping methods* which are based on superpixels are more prone to small perturbation and have worse robustness compared to the *window scoring methods* (e.g., our method, BING and Edge Boxes). This may be due to the fact that superpixels strongly depend on the low-level cues which are more sensitive to small image perturbation. In contrast, our method keeps the best robustness in most of the perturbation cases. The superiority reflects that the high-level semantic learning based objectness not only helps to achieve good recall but also provides more stable proposals in the perturbed images.

E. Generalization to Unseen Categories

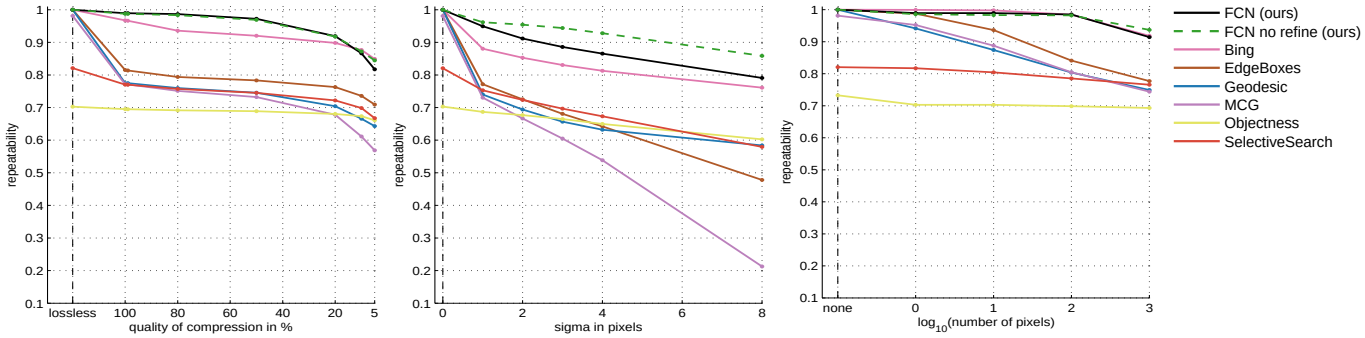
The good recall our approach achieves on the PASCAL VOC 2007 test set does not guarantee it to have learned the generic objectness notion or be able to predict the object proposals for the images containing novel objects in unseen categories. Because it is possible that the model is highly tuned to the 20 categories of PASCAL VOC. To investigate

1051-8215 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2576759, IEEE Transactions on Circuits and Systems for Video Technology

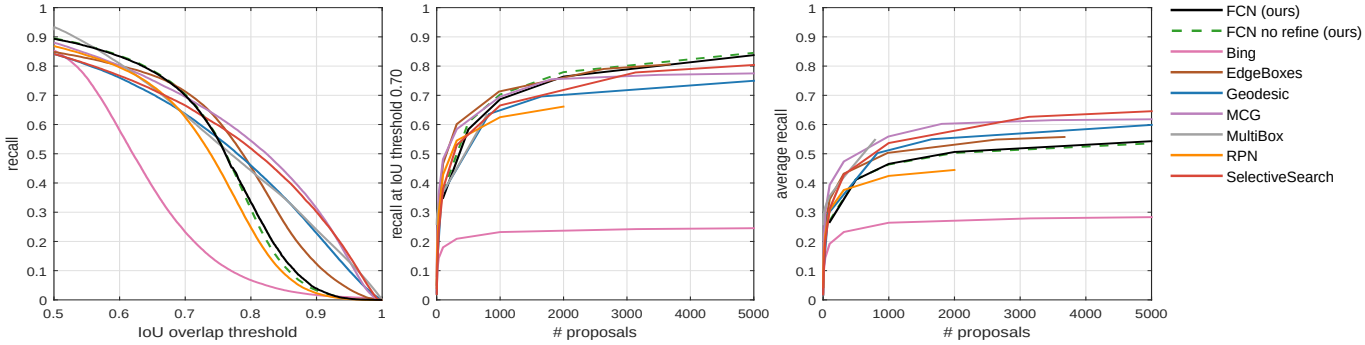
TABLE IV: Object detection average precision for all the 20 categories as well as the mean average precision (mAP) on the PASCAL VOC 2007 test set using Fast-RCNN trained on several different proposals.

	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
Sel search	**64.8**	**70.5**	**55.8**	40.2	22.7	67.0	69.2	70.9	30.1	62.1	60.7	62.7	72.1	67.3	56.3	26.0	49.2	57.5	**69.2**	56.2	56.5
Edge Box	62.2	65.0	50.9	41.8	29.2	**70.5**	**71.4**	70.1	**30.2**	**63.7**	56.2	61.2	72.8	66.6	60.9	28.5	**53.0**	54.3	68.2	56.1	56.6
MCG	61.8	64.1	49.9	38.2	21.4	63.4	61.1	67.6	27.1	53.0	**63.0**	58.7	67.9	59.4	49.4	22.4	46.0	**59.8**	64.9	57.7	52.8
Ours (no refine)	63.1	67.4	54.2	42.0	**33.1**	68.9	71.2	69.7	29.1	58.7	51.6	**63.7**	74.4	66.4	**62.8**	**30.8**	51.4	58.1	65.3	**59.3**	57.0
Ours	63.5	69.9	52.9	**44.6**	31.9	68.5	71.2	**71.1**	29.9	62.4	50.8	62.8	**75.0**	**68.0**	62.1	29.2	52.0	56.6	65.7	57.5	**57.3**



(a) JPEG artifacts (b) Blur (c) Salt and pepper noise

Fig. 12: Robustness results under various perturbation.



(a) 1000 proposals per image (b) Recall at 0.7 IoU (c) Average recall (IoU between (0.5, 1))

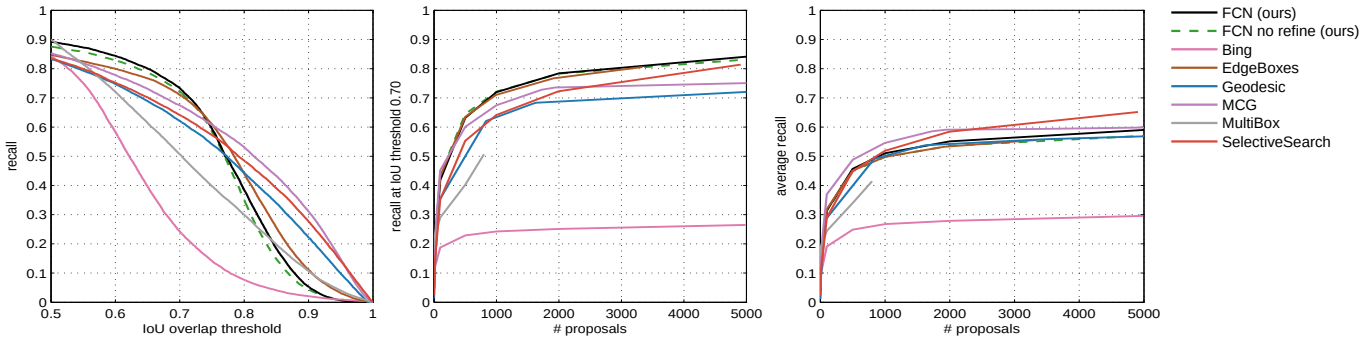
Fig. 13: Recall versus IoU threshold on ImageNet ILSVRC 2013 validation set.

whether it is capable of predicting the proposals for the unseen categories beyond training, we evaluate our approach on the ImageNet ILSVRC 2013 validation set which contains more than 20k images with around 50k annotated objects in 200 categories. Note that the 200 categories are not fine grained versions of the 20 categories of PASCAL VOC. Many of them are totally different from the PASCAL VOC categories, such as food (e.g., bananas) or sports equipment (e.g., rackets). We also conduct the generalization test on PASCAL VOC 2012 validation set which is more difficult to overfit on. In addition, MS COCO validation set which contains lots of small challenging annotated objects is also used for this evaluation.

For ILSVRC 2013 evaluation, we plot several recall curves in Fig. 13. Here we include the MultiBox method from google to compare our FCN method with other CNN-based object proposal methods. Since MultiBox only produces 800 proposals for each image, we set the number of proposals for MultiBox as 800 in Fig. 13(a). From Fig. 13(a), we find that MultiBox achieves high recall at low IoU thresholds (i.e., 0.5<IoU<0.55) but also decreases fast with the increasing of IoU threshold. From Fig. 13(b), it is seen that MultiBox almost keeps performance as good as the state-of-the-art at 0.7 IoU threshold with a very limited number of proposals. In terms of average recall, MultiBox also shows its superiority when generating very few proposals (less than 800 per image). However, due to the limitation of its maximum number (i.e., 800 per image) of proposals, MultiBox cannot boost its recall further by generating more proposals. To summarize, MultiBox is able to roughly localize the objects with a small number of proposals. As for our FCN method, the overall trend of the

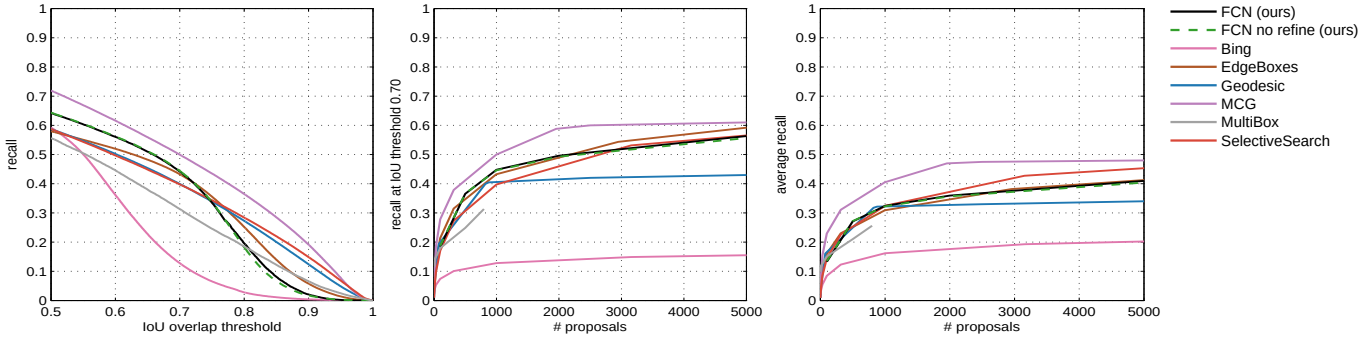
1051-8215 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2576759, IEEE Transactions on Circuits and Systems for Video Technology



(a) 1000 proposals per image (b) Recall at 0.7 IoU (c) Average recall (IoU between (0.5, 1))

Fig. 14: Recall versus IoU threshold on PASCAL VOC 2012 validation set.



(a) 1000 proposals per image (b) Recall at 0.7 IoU (c) Average recall (IoU between (0.5, 1))

Fig. 15: Recall versus IoU threshold on MS COCO 2014 validation set.

recall remains consistent with that on the PASCAL VOC 2007. Specifically, our approach almost keeps the same recall as the best method, i.e., Edge Boxes across a broad range of proposal numbers (see Fig. 13(b)). Fig. 13(a) demonstrates that the recall of our method is still competitive across a wide range of IoU thresholds (from 0.5 to 0.7). In terms of AR, our approach is slightly worse than selective search, which achieves the highest AR. Please note that we also directly use the publicly released RPN model trained on PASCAL VOC in the generalization evaluation here. It is observed that RPN does not perform as well as on PASCAL VOC 2007. This may be because object class information is utilized when training the layers of RPN which are shared with class-specific detectors on PASCAL VOC. Therefore, the generalization to ILSVRC may be influenced by the class-aware training of RPN on PASCAL VOC.

Fig. 14 shows the results of all the methods on PASCAL VOC 2012 validation set. The overall trends of all the methods are consistent with those on PASCAL VOC 2007 test set. Benefited from similar visual appearance and the same categories of PASCAL VOC 2007 and PASCAL VOC 2012, the proposed FCN method keeps similar good performance with that on PASCAL VOC 2007 test set, which is better than on ILSVRC 2013 validation set. The poor generalization ability from ILSVRC to PASCAL VOC of MultiBox results in similar inferior results to those on PASCAL VOC 2007 test set.

As for MS COCO 2014, a similar set of recall figures are shown in Fig. 15. Different from the PASCAL VOC and ILSVRC 2013, it is found that MCG is the best one among all the evaluation cases. Our method shows a similar trend with the previous two benchmarks while all the recalls are lower than the best method. We attribute the difference to different statistics of the datasets, especially the different size distributions of objects (see Fig. 16). As can be seen, MS COCO 2014 contains a large proportion of small objects. This is challenging for *Window scoring methods* as they need add more small scales to avoid missing the small groundtruth objects, which will lead to a higher chance of false positives and much higher computation cost.

Considering the significantly different statistics of MS COCO 2014, based on the above results on ILSVRC 2013, PASCAL VOC 2012 and MS COCO 2014, no significant overfitting towards the PASCAL VOC categories is found in our approach. In other words, the proposed approach has learned a generic notion of objectness and can generalize well to the unseen categories on the whole.

VI. Summary and Conclusion

In this paper, we utilize fully convolutional network (FCN) to generate object proposals in images. The novel high-level semantic objectness concept produced by FCN enables more accurate judgement on whether a patch contains an object or not. Moreover, proposals produced according to their high-level objectness scores are more stable when faced with image perturbation compared to low-level based methods. Both advantages of our proposals benefit the object recall and detection mean avearge

1051-8215 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2576759, IEEE Transactions on Circuits and Systems for Video Technology

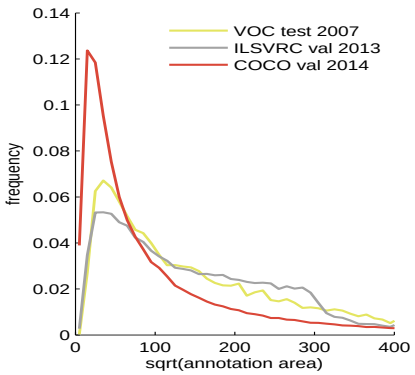


Fig. 16: Comparison of the distribution of the sizes of the groundtruth objects among all considered datasets: PASCAL VOC 2007 test set, ILSVRC 2013 validation set and MS COCO 2014 validation set.

precision. In addition, the novel localization way which directly maps the output neuron in the objectness map to its receptive field in the image does not involve any coordinates regression and shows to be more effective. Apart from this, a proper setting of the multiscale scheme is also critical. Although crossing many scales means a higher chance to localize the objects precisely, it may also bring more false positive objects and higher computation cost. We finalize the setting by fixing α as 0.65 after the tradeoff between recall and speed. Finally, the generalization of our model to unseen categories is also evaluated and validated to ensure that the network can be used to locate generic objects in images, which should be meaningful in real-world applications.

It should be mentioned that the proposed FCN method does not perform well in finding very small objects (i.e., containing less than 500 pixels). This is because of the inherent weakness of *Window scoring methods* that they need smaller scales to find small groundtruth objects, but have a higher chance to have false positives and higher computation cost at the same time.

Although our FCN method focuses on dealing with the static image object proposal problem, it can also be extended to dynamic video sequences. As is well known, optical flow is the most widely used feature to describe the motion information in videos. It is possible to combine the optical flow map with the 3-channel RGB information together as the FCN input for each frame in the videos. Trained on the samples with 4-channel input, the FCN is expected to gain stronger power in finding moving objects in the videos. In the future, we will do research considering the motion information as input for video sequences.

In future work, we will also continue using the semantic objectness obtained by CNN to solve segment proposal problems as the segment proposal provides more precise information about the locations and the shape of the objects of interest.

Acknowledgment

This research is supported by the National Research Foundation, Prime Minister's Office, Singapore under its International Research Centres in Singapore Funding Initiative.

References

[1] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 9, pp. 1627–1645, 2010.

[2] B. Alexe, T. Deselaers, and V. Ferrari, "What is an object?" in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 73–80.

[3] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 580–587.

[4] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from rgb-d images for object detection and segmentation," in *Computer Vision–ECCV 2014*. Springer, 2014, pp. 345–360.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *Computer Vision–ECCV 2014*. Springer, 2014, pp. 346–361.

[6] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr, "Bing: Binarized normed gradients for objectness estimation at 300fps," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 3286–3293.

[7] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Computer Vision–ECCV 2014*. Springer, 2014, pp. 391–405.

[8] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *arXiv preprint arXiv:1411.4038*, 2014.

[9] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, 2014.

[10] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, pp. 1–42, 2014.

[11] J. Hosang, R. Benenson, P. Dollár, and B. Schiele, "What makes for effective detection proposals?" *arXiv:1502.05082v3*, 2015.

[12] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 11, pp. 2274–2282, 2012.

[13] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *International journal of computer vision*, vol. 104, no. 2, pp. 154–171, 2013.

[14] S. Manen, M. Guillaumin, and L. Van Gool, "Prime object proposals with randomized prim's algorithm," in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 2536–2543.

[15] P. Rantalankila, J. Kannala, and E. Rahtu, "Generating object segmentation proposals using global and local search," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 2417–2424.

[16] P. Arbelaez, J. Pont-Tuset, J. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 328–335.

[17] P. Krähenbühl and V. Koltun, "Geodesic object proposals," in *Computer Vision–ECCV 2014*. Springer, 2014, pp. 725–739.

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

[19] Z. Zhang, J. Warrell, and P. H. Torr, "Proposal generation for object detection using cascaded ranking svms," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 1497–1504.

[20] P. Dollár and C. L. Zitnick, "Structured forests for fast edge detection," in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 1841–1848.

[21] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," in *International Conference on Learning Representations*, 2014.

1051-8215 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2576759, IEEE Transactions on Circuits and Systems for Video Technology

[22] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov, "Scalable object detection using deep neural networks," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*.

[23] C. Szegedy, S. Reed, D. Erhan, and G. Anguelov, "Scalable, high-quality object detection," *arXiv preprint arXiv:1412.1441*, 2014.

[24] N. Karianakis, T. Fuchs, and S. Soatto, "Boosting convolutional features for robust object proposals," *arXiv preprint arXiv:1503.06350*, 2015.

[25] P. O. Pinheiro, R. Collobert, and P. Dollr, "Learning to segment object candidates," *arXiv preprint arXiv:1506.06204*, 2015.

[26] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollr, "Microsoft coco: Common objects in context," *arXiv preprint arXiv:1506.06204*, 2015.

[27] W. Kuo, B. Hariharan, and J. Malik, "Deepbox:learning objectness with convolutional networks," in *Computer Vision (ICCV), 2015 IEEE International Conference on*. IEEE, 2015.

[28] J. Zhang, S. Ma, M. Sameki, S. Sclaroff, M. Betke, Z. Lin, X. Shen, B. Price, and R. Mech, "Salient object subitizing," in *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*.

[29] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 91–99.

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

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

[32] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *arXiv preprint arXiv:1409.4842*, 2014.

[33] P. H. Pinheiro and R. Collobert, "Recurrent convolutional neural networks for scene parsing," *arXiv preprint arXiv:1306.2795*, 2013.

[34] D. Eigen, D. Krishnan, and R. Fergus, "Restoring an image taken through a window covered with dirt or rain," in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 633–640.

[35] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Advances in Neural Information Processing Systems*, 2014, pp. 2366–2374.

[36] M. Lin, Q. Chen, and S. Yan, "Network in network," in *International Conference on Learning Representations*, 2014.

[37] D. A. Forsyth, J. Malik, M. M. Fleck, H. Greenspan, T. Leung, S. Belongie, C. Carson, and C. Bregler, *Finding pictures of objects in large collections of images*. Springer, 1996.

[38] G. Heitz and D. Koller, "Learning spatial context: Using stuff to find things," in *Computer Vision–ECCV 2008*. Springer, 2008, pp. 30–43.

[39] J. Carreira and C. Sminchisescu, "Cpmc: Automatic object segmentation using constrained parametric min-cuts," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 7, pp. 1312–1328, 2012.

[40] X. Chen, H. Ma, X. Wang, and Z. Zhao, "Improving object proposals with multi-thresholding straddling expansion," in *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*. IEEE, 2015.

[41] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.

[42] R. Girshick, "Fast r-cnn," *arXiv preprint arXiv:1504.08083*, 2015.

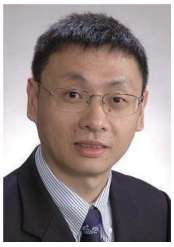
Zequn Jie received his B.E. degree in Mechanical Engineering from University of Science and Technology of China. He is currently a Ph.D. student from Vision and Machine Learning Group, directed by Professor Shuicheng Yan and Jiashi Feng of National University of Singapore. His current research interests mainly include object localization related topics in computer vision, such as object proposal, object detection.

Wen Feng Lu is currently the Associate Professor of Department of Mechanical Engineering at National University of Singapore (NUS). He received his PhD in Mechanical Engineering from the University of Minnesota, USAand had been a faculty at the University of Missouri, USA for ten years after receiving his PhD degree. He later worked as the group manager and senior scientist in Singapore Institute of Manufacturing Technology for six years before joining NUS in 2005. His research interests include IT in Product Design, Sustainable Design and Manufacturing, 3D printing, and Intelligent Manufacturing. He is the recipient of 1997 Ralph R. Teetor Educational Award from Society of Automotive Engineers of USAand 2011 ASME Virtual Environments and Systems Technical Committee Best Paper Award.

Siavash Sakhavi is currently working towards his Ph.D. degree in Department of Electrical and Computer Engineering, National University of Singapore. His current research is learning representations and designing machine learning architectures for time-series data, more specifically, for Electroencephalography (EEG) data. In addition, he is also working on deep learning for object detection in computer vision.

Yunchao Wei is a Ph.D. student from the Institute of Information Science, Beijing Jiaotong University, China. He is currently working at National University of Singapore as a Research Intern. His research interests mainly include semantic segmentation, object detection and classification in computer vision and multi-modal analysis in multimedia.

Eng Hock Francis Tay is currently an Associate Professor with the Department of Mechanical Engineering, Faculty of Engineering, National University of Singapore. Dr. Tay is the Deputy Director (Industry) for the Centre of Intelligent Products and Manufacturing Systems, where he takes charge of research projects involving industry and the Centre. Dr. Tay was also the founding director of the Microsystems Technology Initiative (MSTI), and had established the Microsystems Technology Specialization.



Shuicheng Yan is currently an Associate Professor at the Department National University of Singapore, and the founding lead of the Learning and Vision Research Group (http://www.lv-nus.org). Dr. Yan's research areas include machine learning, computer vision and multimedia, and he has authored/co-authored nearly 400 technical papers over a wide range of research topics, with Google Scholar citation>12,000 times. He is ISI highly-cited researcher 2014, and IAPR Fellow 2014. He has been serving as an associate editor of IEEE TKDE, CVIU and TCSVT. He received the Best Paper Awards from ACM MM'13 (Best Paper and Best Student Paper), ACM MM'12 (Best Demo), PCM'11, ACM MM'10, ICME'10 and ICIMCS'09, the runnerup prize of ILSVRC'13, the winner prizes of the classification task in PASCAL VOC 2010-2012, the winner prize of the segmentation task in PASCAL VOC 2012, the honorable mention prize of the detection task in PASCAL VOC'10, 2010 TCSVT Best Associate Editor (BAE) Award, 2010 Young Faculty Research Award, 2011 Singapore Young Scientist Award, and 2012 NUS Young Researcher Award.

1051-8215 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.