Deep contour and symmetry scored object proposal

Wei Ke[a,b], Jie Chen[b], Qixiang Ye[a,**]

[a] *School of Electronic, Electrical and Communication Engineering, University of Chinese of Academy of Sciences, Beijing, 101408, China*
[b] *Center for Machine Vision and Signal Analysis, University of Oulu, 90570, Finland*

ABSTRACT

Object proposal has been successfully applied in recent supervised and weakly supervised visual object detection tasks to improve the computational efficiency. The classical grouping-based object proposal approach can produce region proposals with high localization accuracy, but incorporates significant redundancy for the lack of object confidence to evaluate the proposals. In this paper, we propose leveraging the essential properties of images, i.e., contour and symmetry, to score the redundant region proposals. Specifically, the contour and symmetry are extracted by a Simultaneous Contour and Symmetry Detection Network (SCSDN) and used to score the bounding box with a Bayesian framework, which guarantees that the scoring procedure is adaptive to general objects. A subset of high-scored proposals reserves the recall rate, while can also significantly decrease the redundancy. Experimental results show that the proposed approach improves the baseline by increasing the recall rate from 0.87 to 0.89 on the PASCAL VOC 2007 dataset. It also outperforms the state-of-the-art on AUC and uses much fewer object proposals to achieve comparable recall rate.

1. Introduction

Object localization is the first important step for visual object detection. Conventional detection approaches (Dalal and Triggs, 2005; Dollár et al., 2014; Felzenszwalb et al., 2010) which use a sliding-window strategy to localize objects tend to generate up to millions of candidate windows. The classification of such a big number of windows in the subsequent steps is computationally expensive, particularly when complex features and/or classification methods are used. Recently, an alternative way, i.e., object proposal, has been investigated to improve the efficiency of object localization. Object proposal tends to produce much fewer (up to two orders of magnitude) windows than the sliding-window strategy, which by no doubts significantly improves the computational efficiency (Girshick et al., 2014). Besides, object proposal is necessary for weakly supervised object detection (Wan et al., 2016; Ye et al., 2017) and instance-based semantic segmentation (Dai et al., 2016). As there is only image level annotation without object bounding boxes in these tasks, it is a good choice to discovery object candidates in an unsupervised way. Higher recall rate with the fewer number of object proposals can increase the robust of models and reduce the object-level search space.

Object proposal approaches in literatures can be coarsely categorized into two: grouping-based and objectness-based ones. In the first category, bounding boxes are produced via hierarchically merging super-pixels. With various super-pixel/region grouping strategies, the generated proposals contribute to high recall rate and localization accuracy, but the generated proposals are redundant and lack object confidences. In the second category, a multi-scale sliding window strategy is used to produce object bounding boxes. Delicate objectness measurement is then designed to measure how likely a bounding box is an interesting object. Nevertheless, such objectness approaches, without precise segmentation procedure, reports lower recall rate and localization accuracy than the grouping-based ones. Considering that missed objects cannot be recovered in the subsequent stages, objectness-based approaches are not competent for tasks where recall rate is the first concern.

In this paper, we propose a new object proposal approach that fully utilizes the advantages of both super-pixel grouping and objectness strategies. Our approach first produces redundant bounding boxes using super-pixel grouping approaches to achieve a high recall rate. It then scores each bounding box us-

[**] Corresponding author

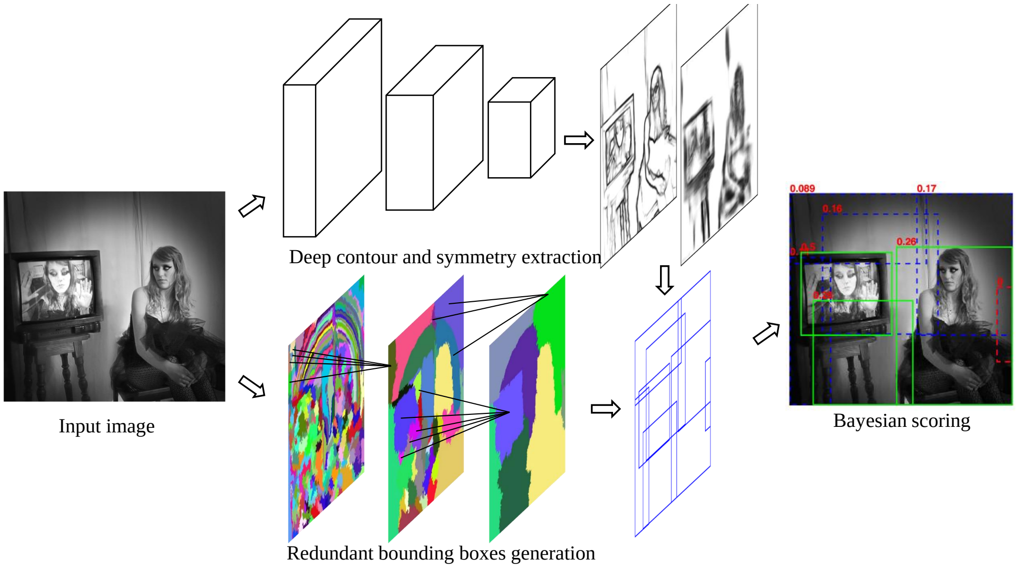

Deep contour and symmetry extraction
Input image
0.089
0.17
0.16
0.26
Bayesian scoring
Redundant bounding boxes generation

Fig. 1: Flowchart of the proposed approach. For an input image, a uniform deep network, i.e, Simultaneous Contour and Symmetry Detection Network (SCSDN) is designed to extract deep contour and symmetry maps. On the other hand, hierarchical super-pixel grouping is used to generate redundant bounding boxes, each of which is scored with Bayesian scoring using deep contour and symmetry. High-scored bounding boxes are outputted as object proposals (green boxes). As we use the pre-trained SCSDN to extract contour and symmetry, our approach is still unsupervised manner.

ing multiple objectness properties, i.e., deep contour and symmetry, as well as choosing a subset of high-scored proposals as solution. Our motivation is that a true object region should have one or more distinct properties contrasting with its surroundings, i.e., some objects could have clear contour, while the others have distinct symmetry parts, or both. Such properties are beneficial to reduce effectively the redundant regions produced by the super-pixel grouping approaches. The contributions of our approach are summarized as follows:

- We propose the Simultaneous Contour and Symmetry Detection Network (SCSDN), which can uniformly extract image properties including contour and symmetry.
- We propose Similarity Adaptive Search (SAS) to generate object bounding boxes, which improves Selective Search by adaptively calculating the similarity between super-pixel subsets.
- We propose using the Bayesian framework to score redundant bounding boxes and choose a subset of high-scored proposal as solution. Experiments demonstrate that we can use significant fewer high-scored object proposals to achieve comparable recall rates with the baseline Selective Search.

The remaining parts of this paper are organized as follows: the related works are represented in Section 2. The proposed approach is detailed in Section 3. Experimental results are given in Section 4, and we finally conclude the approach in Section 5.

2. Related works

Object proposal approaches are coarsely categorized into grouping-based and objectness-based ones. Grouping-based approaches usually root in image segmentation and region grouping strategies, in a bottom-up manner. Objectness-based approaches, in contrast, adopts a sliding window strategy to generate object candidates, in a top-down manner.

The extensively investigated image segmentation algorithms, e.g., Constrained Parametric Min-Cuts (CPMC) (Carreira and Sminchisescu, 2012), can be directly used to generate object proposals. Considering that segmented regions are insufficient to cover objects with a high recall rate, Uijlings et al. propose a hierarchical strategy to merge color homogeneous regions and generate object proposals. They leverage multiple low-level features and multiple merging functions, named Selective Search, to generate redundant bounding boxes so that as many objects as possible are covered (Uijlings et al., 2013). Manen et al. further improve the merging strategy in Selective Searcing by using learned weights as measurement to merge super-pixels (Manen et al., 2013). Different with Selective Search that uses single-scale segmentation, MCG (Arbeláez et al., 2014), explores multi-scale hierarchical segmentation regions for merging, achieving higher recall rate, at the cost of computational efficiency. By taking advantages of both CPMC and Selective Search, Rantalankila et al. propose using a grouping process with a large pool of features and generate segmentations using a CPMC-like process (Rantalankila et al., 2014). Xiao et al. propose a complexity-adaptive metric distance for super-pixel merging, which improves region grouping in different levels of complexity (Xiao et al., 2015). Chen et al. focus on the object proposal localization bias and propose multi-thresholding straddling expansion (MTSE) to reduce localization bias using super-pixel tightness (Chen et al., 2015).

Although grouping-based approaches produce region proposals with a high recall rate, they tend to produce many redundant proposals. Furthermore, their involved fundamental image segmentation procedure is usually time-consuming. Recently, more and more attempts are made to generate object

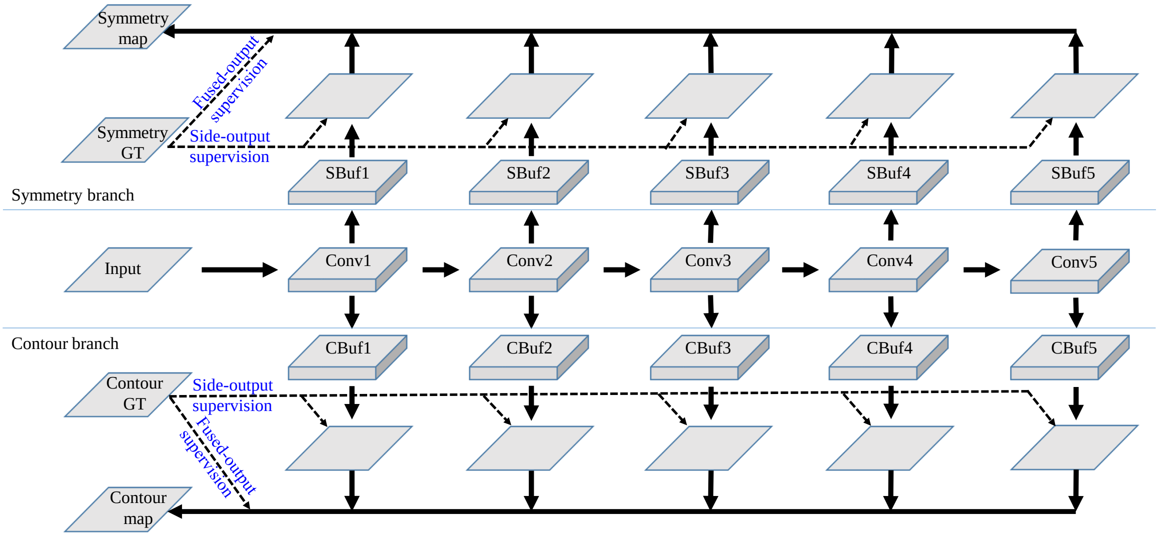


Fig. 2: The convolutional neural network architecture of Simultaneous Contour and Symmetry Detection Network (SCSDN), which outputs contour and symmetry maps Simultaneous. We add buffer convolutional layers (CBuf and SBuf) on the trunk network to form contour branch network and symmetry branch network and the two branches are trained orderly.

proposal based on the object confidence of sliding windows, which avoid the image segmentation procedure, increasing the computational efficiency, and decreasing the proposal number.

Objectness (Alexe et al., 2012), scoring how likely a detection window contains an object, is the pioneer exploring sliding window based object proposal. The score is estimated based on a combination of multi-cues including saliency, color contrast, edge density, location and size statistics. Feng et al. score sliding windows with saliency cues and Randomized-Seeds score it with super-pixel straddling (Feng et al., 2011). Cheng et al. propose Binarized Normed Gradients (BING) using sliding window for object proposal, based on an efficient weak classifier trained using binarized gradients. With a delicate design of binary computations, a low computation cost of BING can be guaranteed, which, as reported, reaches 300 FPS on a PC platform (Cheng et al., 2014). EdgeBoxes is conducted in a sliding window manner of multiple scales and multiple aspect-ratios (Zitnick and Dollár, 2014). The scores of objects are estimated by the number of complete contours detected with the structure forest method.

To take the implemented advantages of grouping-based and objectness-based approaches to reduce the number of object proposals, Krhenbh and Koltun train a regressor model to compute the object confidence with the object ground-truth and the binary segmentation masks. However, it is in an supervised way and limited by the trained regressor as most object datasets are without segmentation masks. (Krähenbühl and Koltun, 2015, 2014).

The power of Convolutional Neural Networks (CNN) has been explored to compute objectness recently. In (Karianakis et al., 2015) the shallow CNN layers are fed into fast decision forests to produce robust object proposals. In (Kuo et al., 2015), a deep score is learned by CNN and is used for updating the confidence of object proposal of EdgeBoxes. Region Proposal Network (RPN) (Ren et al., 2016) utlizes deep learning features to score the sliding window. Benefit from the powerfull representation of convolutional features, the RPN needs only hundreds of bounding boxes to achieve a similar recall rate with other objectness-based approaches that usually output thousands of proposals. Pinheiro et al., train a CNN model to output segmentation masks as well as object confidence (Pinheiro et al., 2015). They also refine the segmentation results by a bottom-up/top-down CNN architecture (Pinheiro et al., 2016). Nevertheless, they are all data-driven and requires preciese annotated training samples to achieve high accuracy, which limits its applications in many tasks, e.g., semantic segmentation and weakly supervised object detection, where no precise object annotation is available.

3. Methodology

The proposed approach, as shown in Fig. 1, first produces redundant bounding boxes by grouping super-pixels hierarchically and extracts general object properties, i.e., deep contour and symmetry, with a uniform fully convolutional neural network. With the extracted object properties, Bayesian scoring is then proposed to score each bounding box. A subset of high-scored proposals is selected to guarantee a high recall rate, while significantly decreases the proposal number.

3.1. Deep contour and symmetry extraction

Contour is pixel-based low-level feature representation with good generalization ability. Usually, a trained contour model can be directly used on other images (Zitnick and Dollár, 2014), that is to say, contour is used with unsupervised manner. The symmetry factor has similar property. Instead of using traditional counter and symmetry approaches, we use Convolutional Neural Network (CNN) to extract the low level features. The recent developed Fully Convolutional Networks (FCN) (Long et al., 2015) makes it possible to output a pixel-based response map. Holistically-Nested Edge Detection (HED) (Xie and Tu, 2015) integrates FCN with deeply supervision on the side-outputs of a trunk network for contour detection. Considering symmetry has the similar property with contour, we use the similar architecture of HED to extract symmetry map. In order to reduce the computation time and the network size, we use only one trunk network for both contour and symmetry detection, as shown in Fig. 2. Using the 16-layer VGG (Chatfield et al., 2014) as the trunk network, the contour buffer convolutional layer (CBuf) and symmetry buffer convolutional layer (Ebuf) are added to each stage of VGG to form contour branch and symmetry branch. The uniform network is named Simultaneous Contour and Symmetry Detection Network (SCSDN). As the contour and symmetry detection are performed in an end-to-end manner with fully convolution, it takes about a hundred of millisecond to process one image.

In SCSDN, both contour branch and symmetry branch affects the parameters of the trunk network, which leads to some instability during the multi-task learning. Compared to HED, the buffer convolutional layers are added to prevent the loss of each branch from being back-propagated directly to the trunk network. With buffer layers, the two branches of SCSDN can be learned orderly.

In the training phase, supervision is taken on both side-outputs and the fused-output, i.e., weighted side-outputs. The loss functions for the contour branch the symmetry branch are

$$\mathcal{L} = \mathcal{L}_{side}(\mathbf{W}_t, \mathbf{W}_c, \Phi_c) + \mathcal{L}_{fused}(\mathbf{W}_t, \mathbf{W}_c, \Phi_c, \mathbf{h}_c), \qquad (1)$$

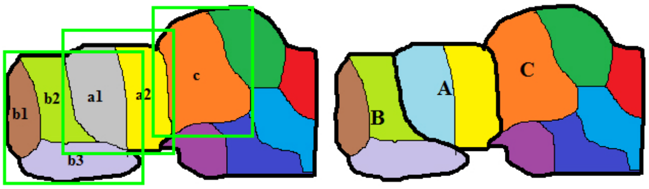
$$\mathcal{L} = \mathcal{L}_{side}(\mathbf{W}_t, \mathbf{W}_s, \Phi_s) + \mathcal{L}_{fused}(\mathbf{W}_t, \mathbf{W}_s, \Phi_s, \mathbf{h}_s). \qquad (2)$$

where $\mathbf{W}_t, \mathbf{W}_c, \mathbf{W}_s$ are the parameters of the trunk network, the buffer layers for the contour and symmetry branch; $\mathbf{\Phi}_c$ and $\mathbf{\Phi}_s$ are the classifiers of each side outputs; $\mathbf{h}_c$ and $\mathbf{h}_s$ are the fusing weights of the side-outputs; $\mathcal{L}_{side}$ and $\mathcal{L}_{fused}$ are the loss of side-outputs and fused-output, respectively. In the testing phase, the contour and symmetry maps are extracted by taking sigmoid processing on the fused-outputs.

Discussion: In the experiments of (Xie and Tu, 2015), one HED model is effective and efficient for edge detection comparing with the traditional contour detection methods. If we use one HED model for contour and another for symmetry, it takes 2× parameters and computational time of one HED model. In SCSDN, the parameters and computational time is similar with only one HED as parameter sharing in trunk network. It keeps the advantages of detection performance as well as saving computational resources.

3.2. Redundant bounding boxes generation

Given an image, we follow the idea of Selective Search (Uijlings et al., 2013) to generate redundant bounding boxes. In this paradigm, the image is partitioned into hundreds of super-pixels, and then group the super-pixels hierarchically with different similarity metrics of the adjacent super-pixel pairs. The initial super-pixel regions are generated by the fast segmentation method (Felzenszwalb and Huttenlocher, 2004). Similarity of all adjacent super-pixel pairs are then calculated and the two most similar regions are grouped together. By iteratively grouping, Selective Search ends until all the regions are merged together. The minimized bounding box is outputted which contains the merged super-pixel pairs, as shown in Fig. 3(a). The similarity for the region pair (r_i, r_j) is measured as:



(a) Bounding boxes (b) Super-pixel subsets

Fig. 3: Illustration of region grouping. (a) each super-pixel group is bounded with a solid line box. The super-pixel groups are $\{c\}$, $\{a1, a2\}$, $\{b1, b2\}$, and $\{b1, b2, b3\}$. (b) the new super-pixel subsets after grouping.

$$\begin{aligned} d(r_i, r_j) = {} & a_1 \cdot d_c(r_i, r_j) + a_2 \cdot d_t(r_i, r_j) \\ & + a_3 \cdot d_s(r_i, r_j) + a_4 \cdot d_f(r_i, r_j), \end{aligned} \qquad (3)$$

where d_c, d_t, d_s, and d_f are four base similarity measures, indicating to preferentially grouping the similar color, similar texture, small size and the region pair that fits into each other, respectively; $a_i \in \{0, 1\}$ is a indicator about using or disabling the base similarity measure. That is to say, it takes the similarity measurement when $a_i = 1$ or otherwise. The final redundant region pool is constituted with different combination of a_i.

3.2.1. Similarity Adaptive Search

To further improve the recall rate, we propose to use more powerful similarity measurements to update the classical Selective Search method. In its hierarchical grouping procedure, super-pixel subsets are generated, shown as the region A, B, and C in Fig. 3(b). The color and texture similarity of subset pair in Selective Search is calculated as

$$D_{mean}(R_m, R_n) = a_1 \cdot d_c(R_m, R_n) + a_2 \cdot d_t(R_m, R_n), \qquad (4)$$

where R_m and R_n are two super-pixel subsets that are seemed as adjacent super pixels. That is to say, merging A with B or merging A with C is only determined by the mean color and/or texture similarity of the whole regions in Fig. 3(b).

Usually, super-pixel subsets are complex that the mean color and texture of the two subsets are significantly different but the connected super-pixels in the subsets are similar. Taking Fig. 3 and color similarity as an example, merging A with B or merging A with C is not only determined by $d_c(A, B)$ but also by $d_c(a1, b2)$ and $d_c(a2, c)$. Considering the connected super-pixels, the low and high complexity similarities in (Xiao et al., 2015) are respectively defined as:

$$D_L(R_m, R_n) = \min\{a_1 \cdot d_c(r_i, r_j) + a_2 \cdot d_t(r_i, r_j) | r_i \in R_m, r_j \in R_n\}, (5)$$

$$D_H(R_m, R_n) = \max\{a_1 \cdot d_c(r_i, r_j) + a_2 \cdot d_t(r_i, r_j) | r_i \in R_m, r_j \in R_n\}. (6)$$

In our Similarity Adaptive Search (SAS), the distance between two super-pixels are calculated as

$$\begin{aligned} D(R_m, R_n) = {} & b_1 \cdot D_{mean}(R_m, R_n) \\ & + b_2 \cdot (\rho_{m,n} D_L(R_m, R_n) + (1 - \rho_{m,n}) \cdot D_H(R_m, R_n)) \quad (7) \\ & + b_3 \cdot D_s(R_m, R_n) + b_4 \cdot D_f(R_m, R_n), \end{aligned}$$

where $b_i \in \{0, 1\}$ denotes whether the similarity measure is used or not; $\rho_{m,n}$ indicates the complexity level of super-pixel subsets R_m and R_n; D_s and D_f are defined in Eq. (1). When R_m and R_n are individual super-pixels, Eq. (5) is equivalent with Eq. (1).

With defined similarity measures, we use the hierarchical merging procedure to obtain redundant bounding boxes. We also follow (Chen et al., 2015)) using tightness to rectify the bounding boxes so that the location of bounding boxes is more accuracy.

Discussion: Comparing with Selective Search (Uijlings et al., 2013), the proposed SAS considering not only the similarity of single super-pixel, but also the similarity of subsets of super-pixels. Selective Search is a part of SAS when complexity similarity is not considered, i.e., $b_2 = 0$. SAS keeps the diversity of super-pixel merging as well as increases the adaptivity.

3.3. Bayesian Scoring

Contour score: Following the hypotheses that the number of contours contained in a bounding box is indicative of the likelihood of the box containing an object (Zitnick and Dollár, 2014), we use the complete contour number as the objectness measurement, which is an orientation consistency edge group.

The contour response map is produced by the SCSDN is shown in Fig. 4(b). Every point is seemed as an edge point. Supposing the set of edge groups in an image is $S = \{s_i\}$, the edge group in a bounding box b is $S_b \subset S$, and $T = \{t_{ij}\}$ is the pixels on s_i. The score based on complete contour is computed as

$$w_e = \frac{\sum_i w_b(s_i) m_i}{2(b_w + b_h)^{\kappa}}, \tag{8}$$

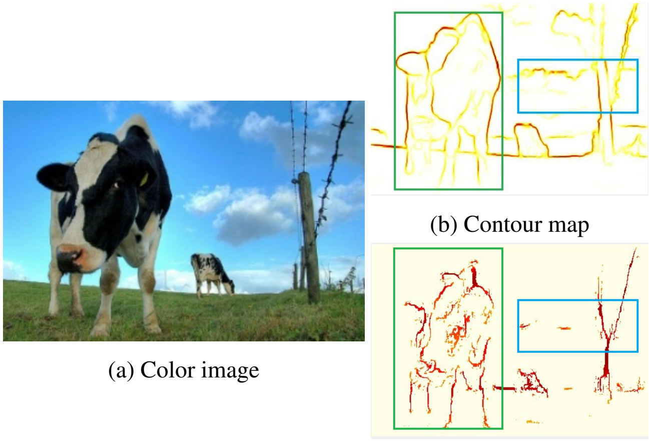
where m_i is the magnitude of s_i; b_w and b_h are the width and height of the bounding boxes. The perimeter of the bounding box is used for normalization, and $\kappa > 1$ is a parameter to offset the bias of larger windows having more edges on average and we set $\kappa = 2$ following EdgeBoxes (Zitnick and Dollár, 2014) . The score $w_b(s_i)$ of every edge group s_i is computed by

$$w_b(s_i) = \begin{cases} 1 & \text{if } s_i \text{ is in } b \\ 1 - \max\limits_{P} \prod\limits_{j=1}^{|P|-1} a(t_j, t_{j+1}) & \text{if } s_i \text{ overlap } b \\ 0 & \text{if } s_i \text{ is out of } b \end{cases}, \tag{9}$$

where $a\left(t_j, t_{j+1}\right)$ is the affinity of the orientations of t_j and t_{j+1}, and P is an ordered path of s_i with length $|P|$, which is from $t_1 \in b$ to the end of s_i.

On the pre-computed contour map, the confidence score for each region is calculated with Eq. (8). The larger the score is, the more compete contours exist in the box, which accounts for a higher confidence that the box is an object.

Symmetry score: Symmetry is an important object property, which has been explored in visual tasks including object detection (Bai et al., 2009) and object recognition (Zhang et al., 2015). If a bounding box has a considerable number of symmetry axes, it is more likely to contain an object. With the symmetry map, the objectness based on symmetry is calculated same with Eq. (8). A symmetry map is shown in Fig. 4(c).



(a) Color image

(b) Contour map

(c) Symmetry map

Fig. 4: Illustration of contour map and symmetry map. The green bounding box is more potential to contain an object than the blue one as the green one contains more complete contours and symmetry parts.

Bayesian Scoring: Let $B = \{b_1, b_2, \cdots, b_N\}$ denotes a set of bounding boxes, and $H = \{(w_e^j, w_s^j)\}_{j=1}^N$ denotes the objectness corresponding to contour and symmetry, respectively. In the Bayesian framework, the score $y = f(H, B)$ is drawn from a probabilistic model:

$$p(y|\mathcal{D}_N) \propto p(\mathcal{D}_N|y), \tag{10}$$

where $\mathcal{D}_N = \{(H_j, b_j)\}_{j=1}^N$. w_e and w_s are assumed conditional independent and we have

$$p(\mathcal{D}_N|y) = \prod_{i=1}^{2} P(\mathcal{D}_N^i|y), \tag{11}$$

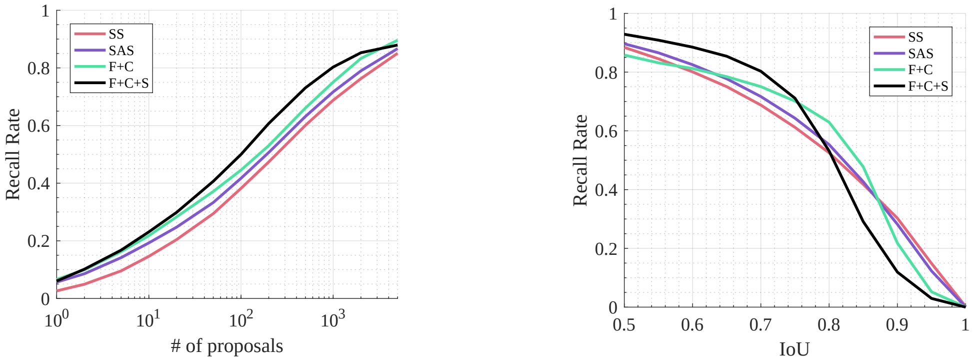
where $\mathcal{D}_N^1 = \{(w_e^j, b_j)\}_{j=1}^N$ and $\mathcal{D}_N^2 = \{(w_s^j, b_j)\}_{j=1}^N$. Sigmoid function is applied on w_e and w_s to transform them to probability to measure how likely the bounding box might be an object, namely, $P((w_e, b)|y) = sigmoid(w_e)$ and $P((w_s, b)|y) = sigmoid(w_s)$. As contour and symmetry are two independent low-level factors, we assume that they make equal contribution to the Bayesian score. Given a new bounding box b and $h = (w_e, w_s)$, the probability of the box to be an object is calculated with

$$p(y|(b, h)) \propto P((w_e, b)|y) \cdot P((w_s, b)|y). \tag{12}$$

4. Experimental results

4.1. Metrics

Dataset. Following (van de Sande et al., 2011; Alexe et al., 2012; Zitnick and Dollár, 2014; Hosang et al., 2014), we evaluate the proposed approach on the PASCAL VOC 2007 dataset (Everingham et al., 2015). The dataset consists of training (4501 images), validation (2510 images) and test datasets (4510 images). We compared our approach with baseline on validation dataset and with the state-of-the-art on test datasets. We also verify the performance of the proposed approach on the challenging MS COCO validation dataset, which contains 40504 images with 80 object categories. (Lin et al., 2014).



(a) Recall rate vs. number of proposals.

(b) Recall rate vs. IoU.

Fig. 5: Comparison with the baseline. SS is Selective Search (Uijlings et al., 2013), SAS is our proposed Similarity Adaptive Search, F+C denotes re-ranked results with contour score, and F+C+S with contour and symmetry scores, respectively.

Evaluation procedures. We follow the same evaluation procedure as (Hosang et al., 2014), using recall rate, proposal number, and proposal-object overlap (Intersection over Union, IoU).

- Recall rate: With higher recall rate, the following classifier is more potential to get high detection accuracy. Once some object is lost in the object proposal stage, the classifier can no longer detect the object.
- Proposal number: Less proposal number is the efficiency guarantee of the following classifier.
- IoU: Larger IoU means more accuracy localization, so that the following feature extraction approachs can extract more efficiency features.

Better approaches are recognized by smaller proposal numbers and larger IoU, while keeping high recall rate. There are three commonly used experimental setups: recall rate vs. window number with given IoU, recall rate vs. IoU with given proposal number, and the minimum proposal number with given recall rate and IoU.

Contour and symmetry detector. We use the dataset BSDS (Arbelaez et al., 2011) to train contour branch and the dataset SYMMAX (Tsogkas and Kokkinos, 2012) to train symmetry branch, which are not fine-tuned any more with the dataset for evaluation the performance of object proposal.

The hyper-parameters of SCSDN include: mini-batch size (1), learning rate (1e-6 for contour branch and 1e-8 for symmetry branch), momentum (0.9), weight decay (0.002). The number of training iterations (10,000, 8,000, 6,000, 4000, 2000, 1000 and 1000 for contour branch and symmetry branch orderly).

With this setting, the proposed SCSDN can output contour and symmetry response map at about 10 fps using single core GPU with better performance than the Structure Edge detector (Dollár and Zitnick, 2013) in EdgeBoxes.

4.2. Comparison with baseline

We evaluate the efficiency of the combination of grouping and objectness on PASCAL-VOC validation datasets. Fig. 5 indicates the comparison between our approach and the baseline approach, i.e., Selective Search (Uijlings et al., 2013). With the region generated by Similarity Adaptive Search (SAS) in Section 3.2, the contour and symmetry are cooperated orderly, as F+C and F+C+S shown in Fig. 5.

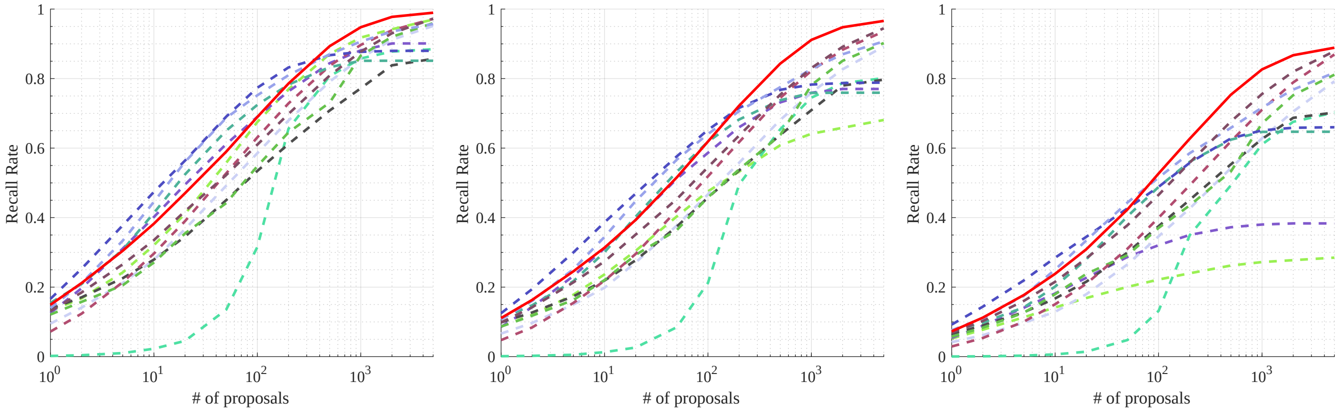
Recall rate vs. the number of proposals is shown in Fig. 5(a) with IoU=0.7. It can be seen that our approach has improved the recall rate by more than 10% when using 100 or 1000 proposals. Recall rate is improved to 0.89 while Selective Search is 0.85, respectively. In addition, our approach needs only 601 detection proposals to get 75% recall rate, while Selective Search needs 1777.

Recall rate vs. IoU is shown in Fig. 5(b) when using 1000 detection proposals. In Fig. 5(b), it can be seen that our approach is better than the baseline when IoU locates between 0.5 and 0.80. When IoU is larger than 0.80, our approach reports a lower recall rate than Selective Search. The reason lies in that our approach further employs a Nox-Maximum Suppression (NMS) procedure. Considering that the IoU=0.5 or IoU=0.7 are the two typical setting in the supervised / weakly-superwised object detection tasks, one can conclude from Fig. 5(b) that our approach outperforms the baseline on recall vs. IoU to get object candidates.

4.3. Comparisons with state-of-the-art

We compare the proposed approach with recent unsupervised approaches including Rahtu (Rahtu et al., 2011), Objectness (Alexe et al., 2012), CPMC (Carreira and Sminchisescu, 2012), Selective Search (Uijlings et al., 2013), RandomizedPrims (Manen et al., 2013), Rantalankila (Rantalankila et al., 2014), MCG (Arbeláez et al., 2014), BING (Cheng et al., 2014), EdgeBoxes (Zitnick and Dollár, 2014), Endres (Endres and Hoiem, 2014), Rigor (Humayun et al., 2014), MTSE (Chen et al., 2015), and CA(Xiao et al., 2015). The results of the compared approaches are provided by (Hosang et al., 2014), and curves are generated using the Structured Edge Detection Toolbox V3.0 (Zitnick and Dollár, 2014).

Recall rate versus number of proposal is illustrated in Fig. 6, and we compare recent approaches using IoU thresholds of 0.5, 0.6, and 0.7. The red curves show the recall performance of our approach. It can be seen in Fig. 6 that the maximum recall rate of our approach slightly outperforms the state-of-the-art, in particular when IoU = 0.7. Recall rate vs. IoU is shown in Fig. 7. The number of proposals is set to 100, 500, and 1000, respectively. Varying IoU from 0.5 to 0.7, Endres, CPMC and MCG perform slightly better than our approach with 100 proposals. But our approach achieves the highest recall rate given 500 and 1000 proposals.

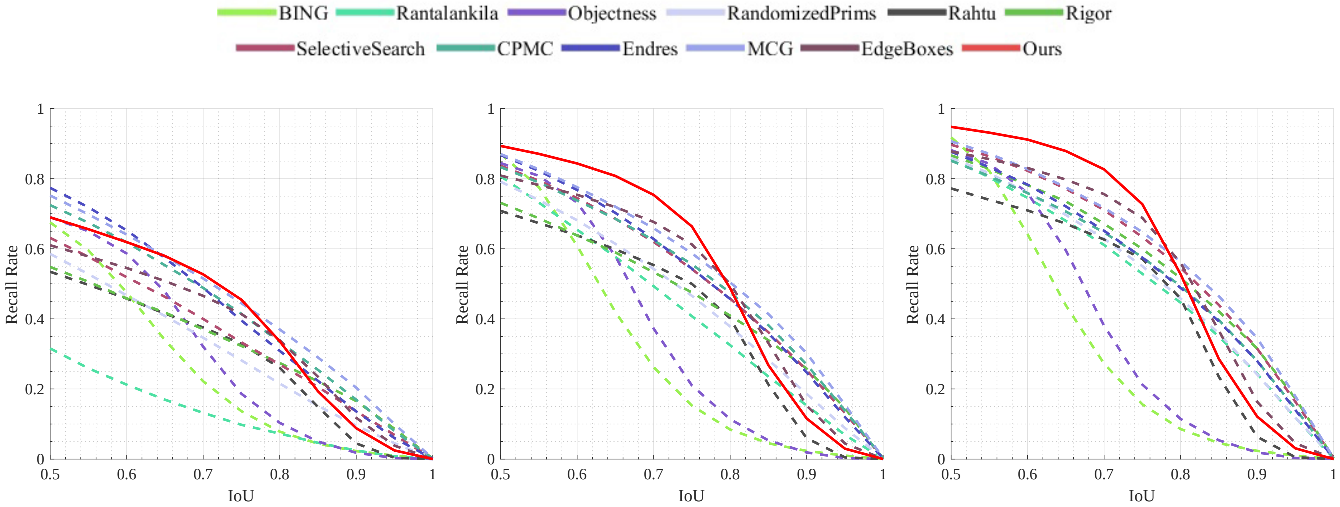


(a) IoU = 0.5

(b) IoU = 0.6

(c) IoU = 0.7

Fig. 6: Comparisons with the state-of-the-art using recall rate versus number of proposals (Best viewed in color).



(a) 100 proposals per image

(b) 500 proposals per image

(c) 1000 proposals per image

Fig. 7: Comparisons with state-of-the-art using recall rate versus IoU (Best viewed in color).

Table 1: of proposal numbers at 25%, 50% and 75% recall rate.

Method	AUC	N@25%	N@50%	N@75%	Recall
BING (Cheng et al., 2014)	0.20	302	-	-	0.28
Rantalankila (Rantalankila et al., 2014)	0.25	146	520	-	0.70
Objectness (Alexe et al., 2012)	0.27	28	-	-	0.38
RandomizedPrims (Manen et al., 2013)	0.35	42	358	3204	0.79
Rahtu(Rahtu et al., 2011)	0.36	29	310	-	0.70
Rigor (Humayun et al., 2014)	0.38	25	367	1961	0.81
Selective Search (Uijlings et al., 2013)	0.39	29	210	1416	0.87
CPMC (Carreira and Sminchisescu, 2012)	0.41	17	112	-	0.65
MTSE (Chen et al., 2015)	0.41	18	175	1112	**0.89**
CA (Xiao et al., 2015)	0.42	27	167	1418	0.88
Endres (Endres and Hoiem, 2014)	0.44	**7**	112	-	0.66
MCG (Arbeláez et al., 2014)	0.46	10	86	1562	0.82
EdgeBoxes (Zitnick and Dollár, 2014)	0.47	12	96	655	0.88
Our approach (C)	0.48	12	91	535	**0.89**
Our approach (C+S)	**0.49**	10	**71**	**476**	**0.89**

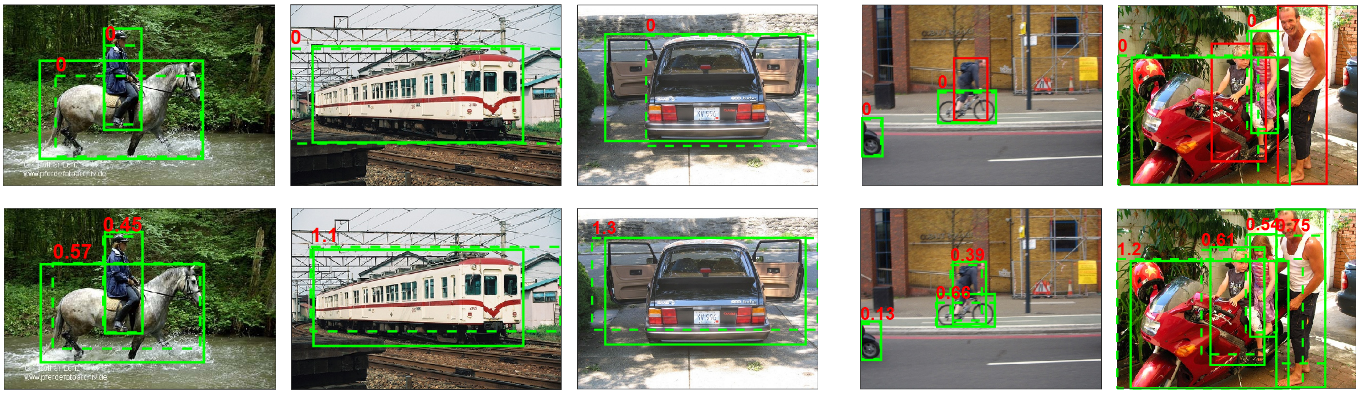
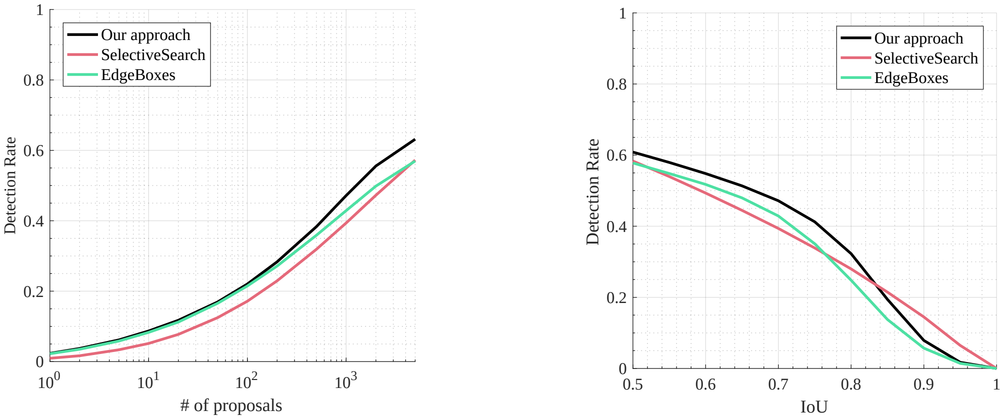


Fig. 8: Localization accuracy and recall rate comparison of the proposed approach in the first row and Selective Search in the second row. From the fist three columns, it is illustrated that the proposed approach has higher IoU, i.e., overlap between the object proposals (dashed line boxes) and the ground-truth (solid line boxes). From the last two columns, it can be seen that the proposals by Selective Search miss three true positives (red boxes), while the proposals by our proposed approach contain all true positives.



(a) Recall rate vs. number of proposals.

(b) Recall rate vs. IoU.

Fig. 9: Comparison results on MS COCO.

In Table 1, we compare the numbers of object proposal required by each approach with 25%, 50% and 75% recall rates and IoU 0.7. It can be seen that our approach keeps the highest recall rate of 0.89 compared to other methods. The AUC (Area Under the Curve) is increased to 0.48 when the countour is used, and to 0.49 when contour and symmetry are used. It was a trade-off between the recall rate and the number of object proposal and the recall rate about 75% is a good choice. To achieve 75% recall rate, our approach needs 476 detection proposals, which is the best among all compared approaches, showing that it can effectively reduce the redundancy. Fig. 8 shows examples about how our approach increases the localization accuracy and increase the recall rate.

4.4. *Evaluation on MS COCO*

In order to verify the performance of our proposed approach, we evaluate it on MS COCO (Lin et al., 2014), as shown in Fig. 9. We just compare the typical Selective Search based on superpixel merging and EdgeBoxes based on objectness. As the images and objects in COCO are more challenge than PASCAL VOC, the recall rate is much lower. However, our approach still gets performance gain and the recall rate is improved from 57% to 63%.

5. Conclusion

Object proposal methods can reduce object candidate windows from millions to thousands. Our motivation is that the proper integration of general object properties, i.e., color, contour, and symmetry contributes to fewer but better object proposal. To fully integrate advantages of grouping approaches with objectness based approaches, we propose a deep contour-symmetry scoring strategy in a Bayesian framework to further improve the performance of object proposal. Furthermore, the contour and symmetry properties are extracted with Simultaneous Contour and Symmetry Detection Network (SCSDN). Experiments demonstrate that a subset of high-scored proposal can guarantee the recall rate while decreasing the object proposal number, significantly. Our approach is easy to extend to some other proposal generation approaches and reduce the object proposal number.

Acknowledgement

This work was supported in part by the National Science Foundation of China under Grant 61671427, and Beijing Municipal Science and Technology Commission under Grant Z161100001616005.

References

Alexe, B., Deselaers, T., Ferrari, V., 2012. Measuring the objectness of image windows. IEEE Transactions on Pattern Analysis and Machine Intelligence 34, 2189–2202. doi:10.1109/TPAMI.2012.28.

Arbelaez, P., Maire, M., Fowlkes, C.C., Malik, J., 2011. Contour detection and hierarchical image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 33, 898–916.

Arbeláez, P.A., Pont-Tuset, J., Barron, J.T., Marqués, F., Malik, J., 2014. Multi-scale combinatorial grouping, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 328–335. doi:10.1109/CVPR.2014.49.

Bai, X., Wang, X., Latecki, L.J., Liu, W., Tu, Z., 2009. Active skeleton for non-rigid object detection, in: Proceedings of International Conference on Computer Vision, pp. 575–582. doi:10.1109/ICCV.2009.5459188.

Carreira, J., Sminchisescu, C., 2012. CPMC: automatic object segmentation using constrained parametric min-cuts. IEEE Transactions on Pattern Analysis and Machine Intelligence 34, 1312–1328. doi:10.1109/TPAMI.2011.231.

Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A., 2014. Return of the devil in the details: Delving deep into convolutional nets, in: Proceedings of the British Machine Vision Conference.

Chen, X., Ma, H., Wang, X., Zhao, Z., 2015. Improving object proposals with multi-thresholding straddling expansion, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 2587–2595. doi:10.1109/CVPR.2015.7298874.

Cheng, M., Zhang, Z., Lin, W., Torr, P.H.S., 2014. BING: binarized normed gradients for objectness estimation at 300fps, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 3286–3293. doi:10.1109/CVPR.2014.414.

Dai, J., He, K., Sun, J., 2016. Instance-aware semantic segmentation via multi-task network cascades, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 3150–3158.

Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 886–893. doi:10.1109/CVPR.2005.177.

Dollár, P., Appel, R., Belongie, S.J., Perona, P., 2014. Fast feature pyramids for object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 36, 1532–1545. doi:10.1109/TPAMI.2014.2300479.

Dollár, P., Zitnick, C.L., 2013. Structured forests for fast edge detection, in: Proceedings of International Conference on Computer Vision, pp. 1841–1848. doi:10.1109/ICCV.2013.231.

Endres, I., Hoiem, D., 2014. Category-independent object proposals with diverse ranking. IEEE Transactions on Pattern Analysis and Machine Intelligence 36, 222–234. doi:10.1109/TPAMI.2013.122.

Everingham, M., Eslami, S.M.A., Gool, L.J.V., Williams, C.K.I., Winn, J.M., Zisserman, A., 2015. The pascal visual object classes challenge: A retrospective. International Journal of Computer Vision 111, 98–136. doi:10.1007/s11263-014-0733-5.

Felzenszwalb, P.F., Girshick, R.B., McAllester, D.A., Ramanan, D., 2010. Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence 32, 1627–1645. doi:10.1109/TPAMI.2009.167.

Felzenszwalb, P.F., Huttenlocher, D.P., 2004. Efficient graph-based image segmentation. International Journal of Computer Vision 59, 167–181. doi:10.1023/B:VISI.0000022288.19776.77.

Feng, J., Wei, Y., Tao, L., Zhang, C., Sun, J., 2011. Salient object detection by composition, in: Proceedings of International Conference on Computer Vision, pp. 1028–1035. doi:10.1109/ICCV.2011.6126348.

Girshick, R.B., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587. doi:10.1109/CVPR.2014.81.

Hosang, J.H., Benenson, R., Schiele, B., 2014. How good are detection proposals, really?, in: Proceedings of the British Machine Vision Conference.

Humayun, A., Li, F., Rehg, J.M., 2014. RIGOR: reusing inference in graph cuts for generating object regions, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 336–343. doi:10.1109/CVPR.2014.50.

Karianakis, N., Fuchs, T.J., Soatto, S., 2015. Boosting convolutional features for robust object proposals. CoRR abs/1503.06350.

Krähenbühl, P., Koltun, V., 2014. Geodesic object proposals, in: In Proceedings of European Conference on Computer Vision, pp. 725–739.

Krähenbühl, P., Koltun, V., 2015. Learning to propose objects, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 1574–1582.

Kuo, W., Hariharan, B., Malik, J., 2015. Deepbox: Learning objectness with convolutional networks, in: Proceedings of International Conference on Computer Vision, pp. 2479–2487. doi:10.1109/ICCV.2015.285.

Lin, T., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft COCO: common objects in context, in: In Proceedings of European Conference on Computer Vision, pp. 740–755.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440. doi:10.1109/CVPR.2015.7298965.

Manen, S., Guillaumin, M., Van Gool, L., 2013. Prime object proposals with randomized prim's algorithm, in: In Proceedings of the International Conference on Computer Vision, pp. 2536–2543. doi:10.1109/ICCV.2013.315.

Pinheiro, P.H.O., Collobert, R., Dollár, P., 2015. Learning to segment object candidates, in: Advances in Neural Information Processing Systems, pp. 1990–1998.

Pinheiro, P.O., Lin, T., Collobert, R., Dollár, P., 2016. Learning to refine object segments, in: In Proceedings of European Conference on Computer Vision, pp. 75–91.

Rahtu, E., Kannala, J., Blaschko, M.B., 2011. Learning a category independent object detection cascade, in: Proceedings of International Conference on Computer Vision, pp. 1052–1059. doi:10.1109/ICCV.2011.6126351.

Rantalankila, P., Kannala, J., Rahtu, E., 2014. Generating object segmentation proposals using global and local search, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 2417–2424. doi:10.1109/CVPR.2014.310.

Ren, S., He, K., Girshick, R., Sun, J., 2016. Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence 9. doi:10.1109/TPAMI.2016.2577031.

van de Sande, K.E.A., Uijlings, J.R.R., Gevers, T., Smeulders, A.W.M., 2011. Segmentation as selective search for object recognition, in: Proceedings of the International Conference on Computer Vision, pp. 1879–1886. doi:10.1109/ICCV.2011.6126456.

Tsogkas, S., Kokkinos, I., 2012. Learning-based symmetry detection in natural images, in: Proceedings of the European Conference on Computer, pp. 41–54. doi:10.1007/978-3-642-33786-4_4.

Uijlings, J.R.R., van de Sande, K.E.A., Gevers, T., Smeulders, A.W.M., 2013. Selective search for object recognition. International Journal of Computer Vision 104, 154–171. doi:10.1007/s11263-013-0620-5.

Wan, F., Wei, P., Han, Z., Fu, K., Ye, Q., 2016. Weakly supervised object detection with correlation and part suppression, in: Proceedings of IEEE International Conference on Image Processing, pp. 3638–3642. doi:10.1109/ICIP.2016.7533038.

Xiao, Y., Lu, C., Tsougenis, E., Lu, Y., Tang, C., 2015. Complexity-adaptive distance metric for object proposals generation, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 778–786. doi:10.1109/CVPR.2015.7298678.

Xie, S., Tu, Z., 2015. Holistically-nested edge detection, in: Proceedings of International Conference on Computer Vision, pp. 1395–1403. doi:10.1109/ICCV.2015.164.

Ye, Q., Zhang, T., Ke, W., Qiu, Q., Chen, J., Sapiro, G., Zhang, B., 2017. Self-learning scene-specific pedestrian detectors using a progressive latent model, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 509–518.

Zhang, Z., Shen, W., Yao, C., Bai, X., 2015. Symmetry-based text line detection in natural scenes, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 2558–2567. doi:10.1109/CVPR.2015.7298871.

Zitnick, C.L., Dollár, P., 2014. Edge boxes: Locating object proposals from edges, in: Proceedings of the European Conference on Computer, pp. 391–405. doi:10.1007/978-3-319-10602-1_26.