Object-level Proposals

Jianxiang Ma[1] Anlong Ming[1] Zilong Huang[2] Xinggang Wang[2] Yu Zhou [1,*]
[1]Beijing University of Posts and Telecommunications
[2]Huazhong University of Science and Technology
{jianxiangma,minganlong,yuzhou}@bupt.edu.cn {hzl,xgwang}@hust.edu.cn

Abstract

Edge and surface are two fundamental visual elements of an object. The majority of existing object proposal approaches utilize edge or edge-like cues to rank candidates, while we consider that the surface cue containing the 3D characteristic of objects should be captured effectively for proposals, which has been rarely discussed before. In this paper, an object-level proposal model is presented, which constructs an occlusion-based objectness taking the surface cue into account. Specifically, the better detection of occlusion edges is focused on to enrich the surface cue into proposals, namely, the occlusion-dominated fusion and normalization criterion are designed to obtain the approximately overall contour information, to enhance the occlusion edge map at utmost and thus boost proposals. Experimental results on the PASCAL VOC 2007 and MS COCO 2014 dataset demonstrate the effectiveness of our approach, which achieves around 6% improvement on the average recall than Edge Boxes at 1000 proposals and also leads to a modest gain on the performance of object detection.

1. Introduction

Object proposal aims to generate a certain amount of candidate bounding boxes to determine the potential objects and their locations in an image, which is widely applied to many visual tasks for pre-processing, e.g., object detection [12], [29], segmentation [8], [28], object discovery [15], and 3D match [3]. Due to the great practicability, it has been a significant research recently.

As *Perception of the Visual World* writes, "*The elementary impressions of a visual world are those of surface and edge.*" Indeed, edge and surface are fundamental to perceive everything in vision, including objects. Most of existing approaches utilize the edge or edge-like cues to generate proposals, but the surface cue has been rarely discussed. The main reason is that achieving the high-level surface cue in an unsupervised manner is a challenging task.

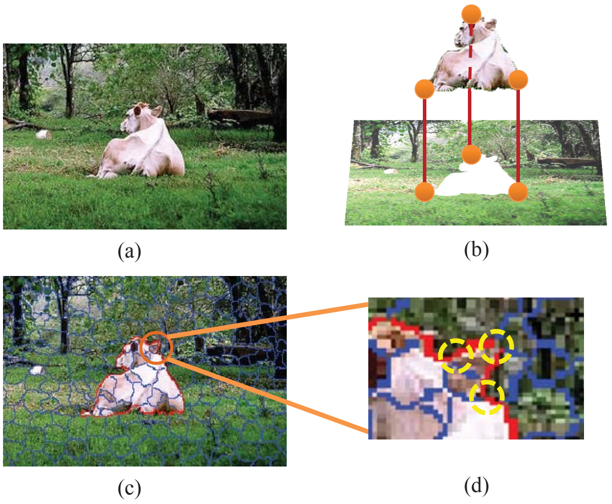

(a)
(b)
(c)
(d)

Figure 1. The formation of occlusion and contour. (a) a natural image, (b) the surface of the object cow and its projection, (c) the contour of the cow formed by occlusion edges in red, (b) the detailed occlusion edges.

From the perspective of optics [25], the smooth surface of an object presented on the 2D image forms the complete and continuous contour, which is produced by the occlusion events in the 3D space [23]. Fig.1 illustrates the formation of occlusion and contour. In (b), the optical rays (orange lines) project the object cow in the 3D space onto the background, and then its surface is delineated by contour in the 2D image. It is observed that the contour just occurs at the boundary where the surface of the cow occludes the background. Formally, *the contour is composed of a set of occlusion edges.* In this paper, an occlusion edge is an edge signalling depth discontinuity between regions, and the edge is called the basic edge for clarity. As shown in Fig.1 (c) and (d), the red occlusion edge is essentially a blue basic edge between two junctions, circled in yellow. Moreover, discontinuous occlusion edges form the complete contour, corresponding to the object surface. Consequently, occlu-

[*] Corresponding Author.

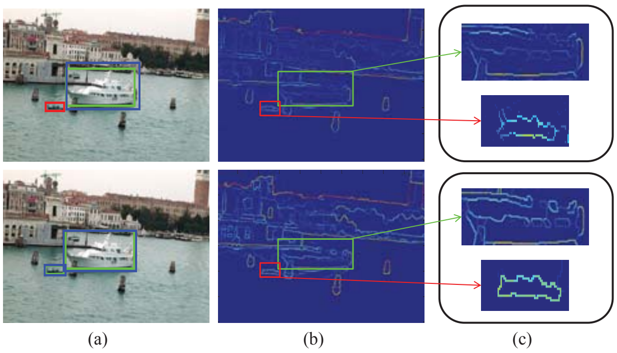


Figure 2. The comparative results of proposals, where the top corresponds to *Edge Boxes* [30] while the bottom is our approach with the surface cue added. (a) the proposals, where blue boxes are the best candidates for the found ground truth boxes in green, and the red are missing ones, (b) the edge maps, (c) the detailed edges contained in the green and red boxes in (b).

sion edges are employed to capture the surface cue. Based on the discussion above, the contour produced by connecting occlusion edges reflects the boundary of object surface in the 3D space, which is similar to [22]. Thus, a novel model based on the occlusion edges is presented to obtain the surface cue effectively to boost the performance of proposals.

The comparative performance is illustrated in Fig.2, where the top is the result of *Edge Boxes* [30], and the bottom corresponds to our method with the surface cue introduced. It is observed that the edge map of *Edge Boxes* in (b) produces weak and discontinuous response, which poorly delineates objects, e.g., the big boat in the green box and the small boat in the red box are not coherently complete, leading to the loss of proposals and their localization accuracy in (a). However, our informative occlusion edge map generated from the surface cue appears more consistent, and strengthens the objects' contours corresponding to their surface in the 3D world, which depicts objects more saliently for proposals. As shown in (b) and (c), the occlusion edges capturing the surface cue obviously contribute to the comprehensive and accurate discovery of objects, e.g., the small boat is found with our edge map, and the big boat is localized more precisely than *Edge Boxes*.

In this paper, a novel object-level approach of proposal generation is presented, where the surface cue is considered in the form of occlusion edge. To this end, occlusion edge detection is firstly demanded and formulated as a supervised learning task. Based on the work in [19], the edge cues are extracted to form feature samples and the kernel ridge regression is applied to acquire the occlusion edge map. Moreover, a novel sparsity induced optimization objective with Huber loss [14] is proposed to dynamically select a set of proper training samples, i.e., the basis. To further enrich the surface cue for proposals, an occlusion-dominated fusion is designed to obtain the overall contour information, namely, a more reliable occlusion edge map. In addition, it is observed that normalization is beneficial to most of proposals for small objects. Hence, a specific normalization criterion is proposed to measure its effect and determine whether the normalization should be done or not, which improves the occlusion edge map at utmost.

In summary, our contributions lie in:

1. An object-level proposal approach is presented with the surface cue considered. To the best of our knowledge, this is the first paper to introduce the surface cue into proposals.

2. Occlusion edges are novelly utilized for the capture of object surface cue to enhance proposals, and a whole occlusion-based framework is constructed for the better occlusion edge map and corresponding objectness.

3. In contrast with *Edge Boxes*, our approach achieves 6% improvement on the average recall at 1000 proposals, which also leads to a modest gain on the performance of object detection.

2. Related work

In general, two main categories may be distinguished for object proposals: window scoring approaches and grouping approaches. The former utilize a set of sampled windows to score and sort them based on the likelihood of containing an object to remove proposals with low rankings, e.g., *Objectness* [1] combines several image cues measuring characteristics of objects in a Bayesian framework, *Bing* [6] proposes a simple and powerful feature called binarized normed gradients to improve the search for objects using objectness scores. The latter usually partition an image into multiple patches and merge them with specific criteria to generate candidate region proposals, e.g., *Selective Search* [24] combines the strength of both an exhaustive search and segmentation, *CPMC* [4] exploits multiple graph-cut based segmentations with multiple foreground seeds and biases to propose objects, and *MCG* [2] develops the multiple hierarchical work by combinatorially grouping regions. However, these state-of-the-art methods rarely consider the 3D cues of object surface, while our object-level method takes the surface cue into account. In addition, recent deep learning based works achieve excellent performance for proposals, e.g., *Deep Mask* [20] and *Sharp Mask* [21], but they may be at the cost of efficiency.

2.1. Reviewing edge boxes

Since our object-level approach closely depends on occlusion edges, we deeply review *Edge Boxes* [30], which defines the specific objectness score based on an edge map to model the observation. Here, we thoroughly probe into the deficiency of the basic edge map [9] utilized in it:

- The basic edge map involves many tiny edges with weak response, so some relatively large candidate boxes containing them are likely to obtain higher scores than real yet small ones, which brings the great difficulty to small object proposals. In Fig.3 (b), the boat contained in the green dashed box is composed of weak edges, leading to the lower score, while other weak edges contained in a relatively large box score higher than the boat, so it is more likely to be an object, e.g., the red dashed box, which is a false judgement.
- Most of the basic edges are incomplete and weakly continuous, making the objects with large aspect ratio hard to find. The reason is that the candidate boxes intersecting the weak edges achieve the higher score ranking, e.g., for the train in the green dashed box in Fig.3 (b), the response gets so weak in the half that the red box truncating its weak edges acquires a higher score than the whole train.

Therefore, we attempt to address the issues above and improve the edge response to promote proposals, namely, enhance the response and consistency of object contours and weaken or remove false contours to obtain a more reasonable edge map, i.e., the occlusion edge map.

3. Object-level proposals

Since occlusion effectively captures the surface cue, we focus on occlusion estimation and occlusion-based objectness to propose objects. However, occlusion edges in the complex scenarios are hard to detect completely, which needs further improvement. Considering that basic edges provide overall yet weak response, an occlusion-dominated fusion is elaborately for a more reliable occlusion edge map to compensate for the lost surface cue.

3.1. Occlusion edge response

With the edge representation in [19], $\mathbf{F} \in R^{U\times N}$ denotes the sample matrix with N training edges, each of which has U dimensional features. However, such immense and miscellaneous samples greatly increase the complexity when training the occlusion edge detector, so a set of basic samples are necessary to accelerate learning and boost accuracy. Specifically, the basis matrix $\mathbf{B} \in R^{U\times M}$ is learnt to represent the original samples approximately and as exactly as possible, namely, $\mathbf{F} \approx \mathbf{BS}$, where $\mathbf{S} \in R^{M\times N}$ is the coefficient matrix for $\mathbf{B}$, M is the number of basis and $M \ll N$. [19] employs the Mean Shift Clustering [7] to obtain the cluster centers as representative samples, which are fixed and may include some noises. To avoid the adverse effects of them, dynamic basis learning is novelly introduced.

Motivated by the sparse coding [27], we present a sparsity induced optimization objective with the Huber loss [14], which is formulated as:

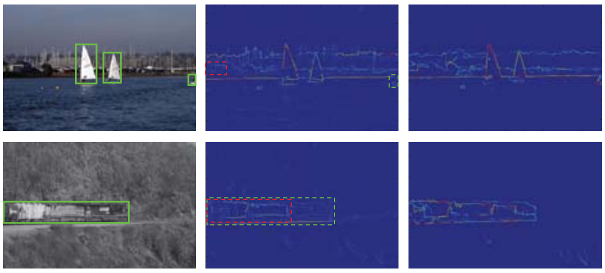


Figure 3. (a) the natural images, with ground truth boxes shown in green, (b) the basic edge maps, with ground truth boxes shown in green and false proposal boxes shown in red, (c) the occlusion edge maps with our approach.

$$\begin{aligned}\min_{\mathbf{B},\mathbf{S}} \quad & \sum_i \sum_j H_\alpha(r_{ij}) + \mu\|\mathbf{S}\|_1 \\ \text{s.t.} \quad & \mathbf{B} \geq 0, \mathbf{S} \geq 0,\ \|\mathbf{b}_i\|_2^2 \leq d,\ \forall\, i = 1, ..., M\end{aligned} \tag{1}$$

where d is a constant and $||\mathbf{S}||_1 = \sum_{i=1}^{M}\sum_{j=1}^{N}|s_{i,j}|$ denotes the ℓ_1-norm of the matrix. The residue $r_{i,j} = f_{i,j} - \mathbf{b}_{i\cdot}\mathbf{s}_{\cdot j}$ indicates the reconstruction error of each dimension. $H_\alpha(\cdot)$ denotes the Huber loss function with a parameter α. which is defined as:

$$H_\alpha(r) = \begin{cases} r^2/2 & |r| < \alpha \\ \alpha|r| - \alpha^2/2 & |r| \geq \alpha \end{cases} \tag{2}$$

According to Eq.(2), if the residue $|r| < \alpha$, representing the normal samples, the objective is the ℓ_2-regularized loss. Otherwise, it means that there may exist noises, i.e., the edges are useless for reconstruction, and hence the objective is the ℓ_1-regularized loss, which is insensitive to large errors. Therefore, the Huber loss is robust to accommodate noises caused by the arbitrary of the edges, while $||\mathbf{S}||_1$ encourages each edge to be approximated by a sparse combination of the basis.

Taking Eq.(2) into consideration, Eq.(1) can be approximately converted to the weighted least square problem with sparsity and non-negativity constraints:

$$\begin{aligned}\min_{\mathbf{B},\mathbf{S}} \quad & \frac{1}{2}\mathbf{W} \odot\ \|\mathbf{F} - \mathbf{BS}\|_F^2 + \mu\|\mathbf{S}\|_1 \\ \text{s.t.} \quad & \mathbf{B} \geq 0, \mathbf{S} \geq 0,\ \|\mathbf{b}_i\|_2^2 \leq d,\ \forall\, i = 1, ..., M\end{aligned} \tag{3}$$

where $\odot$ is the Hadamard product of matrices, and $\mathbf{W}$ can be interpreted as the weight matrix of the residue r. Given the pth iteration of the optimization procedure, each element of $\mathbf{W}$ is defined as:

$$w_{ij}^p = \begin{cases} 1 & |r_{ij}^p| < \alpha \\ \dfrac{\alpha}{|r_{ij}^p|} & |r_{ij}^p| \geq \alpha \end{cases} \tag{4}$$

Aiming to solve the optimization problem in Eq.(3), we alternate between updating $\mathbf{B}$ and $\mathbf{S}$. Fixing $\mathbf{B}$, with the non-negativity constraint to $\mathbf{S}$, the objective is similar to the sparse coding, and thus the update rule in [26] is employed to optimize $\mathbf{S}$. In turn, fixing $\mathbf{S}$, the formula is reduced to a conventional weighted least square problem with non-negativity constraint, which can be done efficiently by the Lagrange dual in [17].

With the optimized training samples $\mathbf{B}^*$, the kernel ridge regression in [19] is utilized to learn the occlusion classifier, which has a simple closed form solution, i.e., $\mathbf{v} = (\mathbf{K} + \gamma\mathbf{I})^{-1}\mathbf{L}$, where $\mathbf{K} = \kappa(\mathbf{B}^*, \mathbf{B}^*)$ is a kernel matrix calculated by the kernel function κ, γ is the regularization parameter, $\mathbf{I}$ is an identity matrix and $\mathbf{L}$ is the occlusion label vector. When testing, we calculate regression values for each edge with the trained classifier and only positive ones are retained, which is stated as:

$$c_e = \max(0, \mathbf{v}\kappa(\mathbf{B}^*, \mathbf{f}_e)) \tag{5}$$

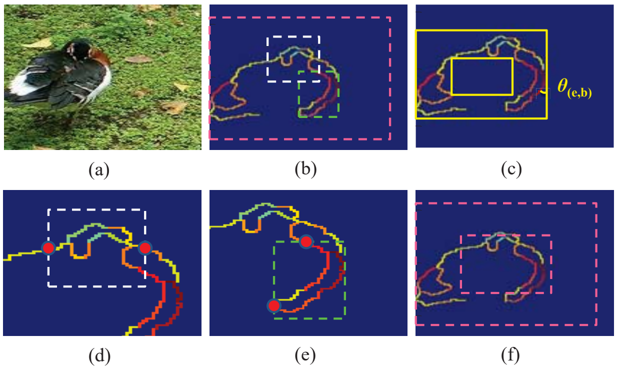
where $\mathbf{f}_e$ is the feature vector of the edge e. Consequently, the occlusion confidence c_p of each edge pixel $p \in e$ is c_e. Then, the occlusion edge map is constructed by assigning corresponding confidence c_p to each edge pixel p, denoted by the matrix $\mathbf{E}^c$. Note that the following edge maps with response known are obtained in the same way. As seen in Fig.3 (c), the occlusion edge maps strengthen the edge response of small objects, e.g., the boat, and remove a certain amount of irrelevant edges. Moreover, the edges of objects are more continuous and complete in comparison with edge response in Fig.3 (b), which can delineate objects more saliently and contribute to finding proper proposals, e.g., the boundary of long train.

3.2. Occlusion-based objectness

Since the contour produced by connecting occlusion edges reflects the boundaries of object surface in the 3D space, the discovery of objects directly from them seems so simple. Unfortunately, due to the complexity of natural scenes, e.g., the similar appearance to the background or the heavy shading, occlusion edges cannot be correctly and completely detected. However, our object-level model is no need of strictly closed and continuous occlusion edges to propose objects. With the rough outline of objects, the objectness of each box b is evaluated directly based on the degree of overlapping between occlusion edges and box boundaries, which is formulated as:

$$\Gamma(b) = \sum_{e \in b - b_o} C_e - \sum_{e \in O_b} \sin\theta_{(e,b)} C_e \tag{6}$$

where b_o is the inner box with half size centered in b, and C_e is the sum of occlusion edge confidence c_e for all pixels in the edge e. O_b is the set of occlusion edges overlapping the box b's boundary, which can be obtained efficiently with the two data structures in [30]. $\theta_{(e,b)} \in [0°, 90°]$ is the angle between e and its intersecting box boundary, and the weight $\sin\theta_{(e,b)}$ can be interpreted as the dissimilarity between them. Note that the score should be accordingly scaled like *Edge Boxes*.



(a) (b) (c)

(d) (e) (f)

Figure 4. (a) an object in the image, (b) the occlusion edge map with improper proposals in the three boxes, where the warmer color corresponds to the higher occlusion confidence, (c) the occlusion edge map with the best proposal in the large yellow box, which obtains a relatively high score ranking, (d), (e) and (f) respectively correspond to the three proposal boxes in (b).

According to Eq.(6), the objectness is mainly related to several factors: the number and confidence of occlusion edges straddling the box's boundary, the occlusion edges included in the inner box, and the orientation disparity between box's boundary and occlusion edges at the neighbourhood of the box. As shown in Fig.4, the large yellow proposal box in (c) obtains a higher score than the three boxes in (b). Firstly, the box's boundary is almost tangent to the detected occlusion edges of the bird, meaning that the proposal covers the object along its boundary approximately, i.e., $\theta(e, b)$ is small as marked in (c). Secondly, the inner yellow box hardly contains occlusion edges, which indicates its interior appearance is coherent and it is more likely to be an object. Hence, despite the contour of the bird is not complete, we still propose it correctly with our occlusion-based model. In contrast, (d), (e) and (f) illustrate several typical improper proposals, corresponding to the white, green and pink boxes in (b) respectively. The red points in (d) shows all intersection angles between occlusion edges and the box's boundary are large, which is much likely that the white box in (b) truncates the object and thus the score is degraded. The green box in (e) is better than (d) because some of the intersection angles are relatively small. For the large pink box in (f), although the boundary of the box is consistent with occlusion edges like (c), the inner pink box contains many occlusion edges with large confidence, which means there may exist a more suitable proposal box to represent an object.

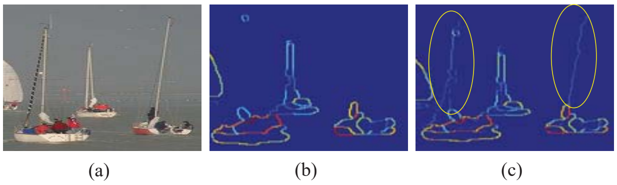


Figure 5. (a) the image whose occlusion edges are hard to detect completely, (b) the occlusion edge map $\mathbf{E}^c$, (c) the improved occlusion edge map $\tilde{\mathbf{E}}^c$ with the lost response circled in yellow .

3.3. Occlusion-dominated fusion

Considering that the basic edges can compensate for the detailed response lost in occlusion map, an occlusion-dominated fusion is introduced to further promote occlusion edge map, which can make our occlusion-based objectness more reliable. Specifically, due to the informative 3D characteristic of occlusion, we novelly regard occlusion confidence as the weighting term to dominate the fusion, which is formulated as:

$$\tilde{\mathbf{E}}^c = \mathbf{E}^c \odot \mathbf{E}^c + (\mathbf{E}^c_m - \mathbf{E}^c) \odot \mathbf{E}^g \tag{7}$$

where $\mathbf{E}^c_m$ is the matrix filled with the maximum confidence of all occlusion edges in the image. $\mathbf{E}^g$, $\mathbf{E}^c$ and $\tilde{\mathbf{E}}^c$ respectively represent the basic edge, occlusion edge and improved occlusion edge maps constructed like Section 3.1.

According to Eq.(7), for a certain edge pixel, we obtain the corresponding response in the edge map matrix with its location i and j, i.e., $\tilde{E}^c_{ij} = E^c_{ij}E^c_{ij} + (E^c_{m(ij)} - E^c_{ij})E^g_{ij}$. Thus, the occlusion response E^c_{ij} is the weight between E^c_{ij} and E^g_{ij}, which can adjust the response based on both edge and surface cue. For instance, a large E^c_{ij} makes $\tilde{E}^c_{ij}$ prefer E^c_{ij} itself, while a small one places more weight on E^g_{ij} to compensate for the lost response. As shown in Fig.5, the improved occlusion edge map $\tilde{\mathbf{E}}^c$ not only further enhances the occlusion edges of real objects, but also recovers weak yet necessary response of ambiguous boundaries lost in $\mathbf{E}^c$, e.g., the masts of ships in (a) are too narrow to be detected in the occlusion edge map (b), but with the supplement of basic edges, the occlusion edge map in (c) provides the weak response for them, circled in yellow, and thus can obviously contribute to precise discovery of ships.

However, some small objects with low score rankings are still difficult to find. To tackle this issue, we normalize the improved occlusion edge map into $[0, 1]$, namely, $\tilde{\mathbf{E}}^c_n = \tilde{\mathbf{E}}^c \oslash \tilde{\mathbf{E}}^c_m$, where $\oslash$ is the element-wise division and $\tilde{\mathbf{E}}^c_m$ is similar to $\mathbf{E}^c_m$. For these small objects, the normalization can diminish their score distance to obvious objects, which promotes their rankings and makes them easier to find, but it may risk losing some informative response. Thus a specific normalization criterion is designed to measure its effect. Given an image, a score ratio g is defined as:

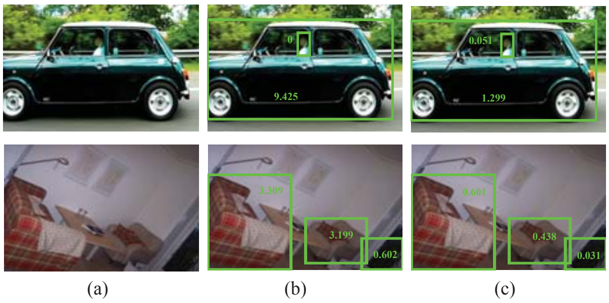


Figure 6. (a) the natural images, the top image is in need of normalization, while the bottom one is not, (b) the ground truth boxes with maximum and minimum scores obtained by $\tilde{\mathbf{E}}^c$, (c) the ground truth boxes with maximum and minimum scores obtained by $\tilde{\mathbf{E}}^c_n$.

$$g = \frac{\max_{b \in \Omega} \Gamma(b)}{\min_{b \in \Omega} \Gamma(b) + \varepsilon} \tag{8}$$

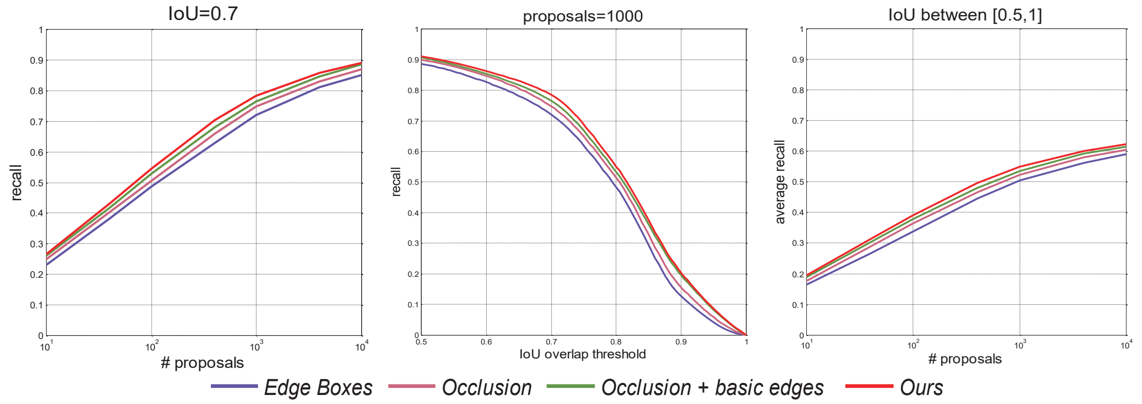
where Ω is the set containing all ground truth boxes, $\Gamma(b)$ is the objectness in Eq.(6) and ε is a sufficiently small number for smoothing. Eq.(8) is the ratio of maximum objectness to minimum objectness among ground truth boxes, which partly reflects the score distance between small and obvious objects. Based on $\tilde{\mathbf{E}}^c$ and $\tilde{\mathbf{E}}^c_n$, we can calculate the ratios g^c and g^c_n respectively with Eq.(8). Then the normalization criterion is stated as:

$$\mathbb{E} = \begin{cases} \tilde{\mathbf{E}}^c_n & g^c > g^c_n \\ \tilde{\mathbf{E}}^c & g^c \leq g^c_n \end{cases} \tag{9}$$

where $g^c > g^c_n$ means normalization works, otherwise it fails to reduce the distance so that $\tilde{\mathbf{E}}^c$ is unchanged. Fig.6 illustrates the two situations, where the top image is in need of normalization, while the bottom one is not. Scores of all ground truth boxes are marked near the green boxes in (b) and (c) respectively, and the corresponding score ratios are obtained based on Eq.(8). For the top image, $g^c \gg g^c_n$ indicates normalization is much significant to the image. But for the bottom one, $g^c < g^c_n$ means $\tilde{\mathbf{E}}^c$ is more effective.

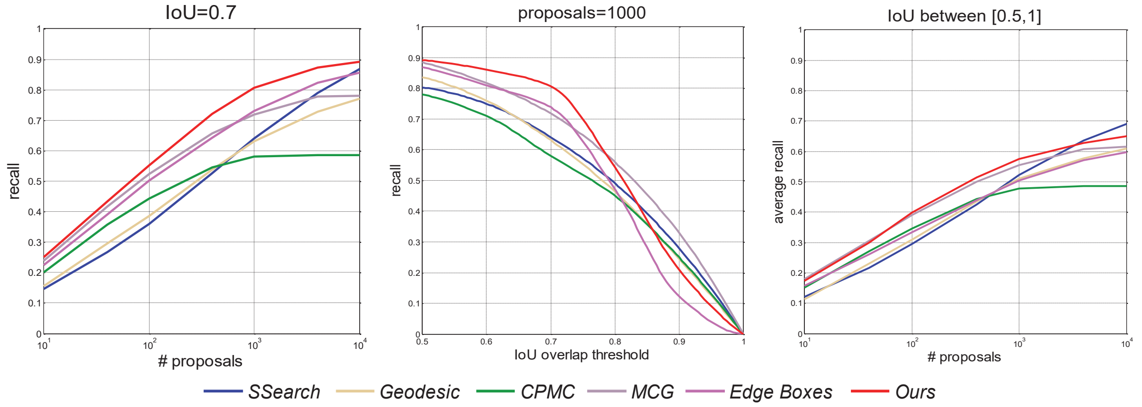
4. Experiments

In this section, we mainly evaluate the performance of our approach on the PASCAL VOC 2007 dataset [10]. Referring to the experimental settings of *Edge Boxes* [30], we employ the training and validation sets to report the results on variants of our algorithm, while the test set is used for contrast with state-of-the-art approaches. For each dataset, we measure the results with three proposal metrics: Firstly, we set the Intersection over Union (IoU) threshold to 0.7 and vary the number of object proposals from 10 to 10000.



(a) (b) (c)

Figure 7. Comparison on variants of our approach on the PASCAL VOC 2007 dataset. (a) recall versus number of proposals given IoU = 0.7, (b) recall versus IoU overlap threshold given 1000 proposals, (c) average recall versus number of proposals between IoU 0.5 to 1.



(a) (b) (c)

Figure 8. Comparison of our approach with state-of-the-art hand-crafted methods on the PASCAL VOC 2007 dataset. (a) recall versus number of proposals given IoU = 0.7, (b) recall versus IoU overlap threshold given 1000 proposals, (c) average recall versus number of proposals between IoU 0.5 to 1.

Secondly, given 1000 proposals, the IoU threshold ranges from 0.5 to 1. Thirdly, the average recall (AR) between IoU 0.5 to 1 is introduced in [13] to summarize proposal performance across IoU thresholds, varying from 10 to 10000 proposals too. In addition, we make comparison on the capability of variants of our algorithm to propose specific objects, and finally explore the effects of different existing methods on object detection. Recently, MS COCO has become the mainstream dataset for object proposal, especially for deep learning based works, thus some additional experiments are done on the MS COCO 2014 dataset [18].

4.1. Comparison on variants of the approach

First, we make comparison on variants of our approach, and the results are shown in Fig.7. *Edge Boxes* is regarded as the baseline, the second variant utilizes our occlusion-based objectness with primary occlusion edge map and the third employs the occlusion-dominated fusion. The final is our whole method including normalization criterion. It is observed that our occlusion-based objectness is better than *Edge Boxes*. The reason is that occlusion edges indicate the 3D discontinuity of object surface and provide the informative surface cue not involved in edges, which increases the accuracy of proposals and makes localization more precise. Then, when we properly add basic edges to improve occlusion edge map, the recall further rises. Even though for less proposals or higher IoU threshold, the performance is better than the methods with single edge map. Finally, with our normalization criterion introduced, there is still a modest gain in accuracy, which demonstrates that the refinement of occlusion edge response with our approach is effective to promote object proposals.

	Edge Boxes	(I)	(II)	*Ours*
Small size	0.142	0.191	0.204	0.212
Aspect ratio	0.381	0.453	0.477	0.483

Table 1. Comparison results of recall on the specific objects for 1000 proposals and IoU 0.7 on the PASCAL VOC 2007 dataset. (I) is the occlusion-based objectness only with occlusion edges. (II) utilizes the occlusion-dominated fusion. Ours is the whole framework with the normalization criterion added based on (II).

Additionally, as mentioned in Section 2.1, *Edge Boxes* based on the basic edges perform poorly for some specific objects, e.g., the objects with large aspect ratio or extremely small size. Thus, we compare the ability of different edge maps when proposing these difficult objects. In the experiment, a ground truth box b in the image I is defined as a small object if $area(b) < 0.01 * area(I)$, while the box with its aspect ratio more than 3 is also considered. Table 1 illustrates the results of recall for 1000 proposals and IoU 0.7. Small objects are too difficult to discover, but the occlusion-based objectness with primary occlusion edges (I) obtains 5% gain in contrast with *Edge Boxes*. When we further improve the occlusion edge map with occlusion-dominated fusion (II), the performance gets better. Finally, with the normalization criterion added, our whole approach achieves 7% improvement than *Edge Boxes*, which demonstrates the superior ability of our method to propose small objects. For the objects with large aspect ratio, the promotion is more significant. The recall of occlusion exceeds 7% than *Edge Boxes*, which indicates that occlusion effectively preserve the integrality of the objects. Similarly, when we refine the occlusion edge map with occlusion-dominated fusion and normalization criterion, the results also obtain the corresponding rise. Above all, our final occlusion edge map achieves 10% improvement than *Edge Boxes* for the objects with large aspect ratio.

Note that the pipeline of the presented method is the same as *Edge Boxes*, which only contains two parts: occlusion edge detection and object proposal generation with the occlusion edge map, namely, both of us train the specific edge detectors for supervised edge detection, and then use similar window scoring mechanism for unsupervised object proposal. For the PASCAL VOC 2007 dataset, the runtime of *Edge Boxes* is 0.55s, while ours is 0.7s, which is nearly as efficient as *Edge Boxes*, but achieves much higher recall when fixing the number of proposals.

4.2. Comparison with hand-crafted approaches

In Fig.8, several hand-crafted methods are selected from [13] to evaluate our approach, where their competing results are provided. *Selective Search* (*SSearch*) [24] and *Geodesic* [16] achieve promising accuracy and perform similarly for the three metrics. Both of them fall behind at a small number of proposals but rise rapidly with candidates increasing, especially at the larger number of proposals and higher IoU values, and *Selective Search* gets much powerful for proposals when the amount is very large (about $>$ 5000). *CPMC* [4] obtains relatively few yet high-quality proposals. In contrast, *Edge Boxes* [30], *MCG* [2] and our object-level approach achieve superior results of recall as a whole, and ours is the best. Both *Edge Boxes* and our approach perform well at the small or large number of proposals. However, due to the low localization accuracy of window scoring mechanism, their results get worse than grouping methods when IoU > 0.8, which adversely affects the average performance. *MCG* has a comparatively strong ability to propose objects and localize them precisely, leading to the competitive average accuracy across all proposals, as shown in Fig.8 (c). Nevertheless, when we introduce the informative surface cue into proposals, both the quality and localization precision are enhanced, and thus the average recall exceeds around 6% than *Edge Boxes* at 1000 proposals. Moreover, our approach outperforms *MCG* on overall performance, which demonstrates the effectiveness of our refined occlusion edge map.

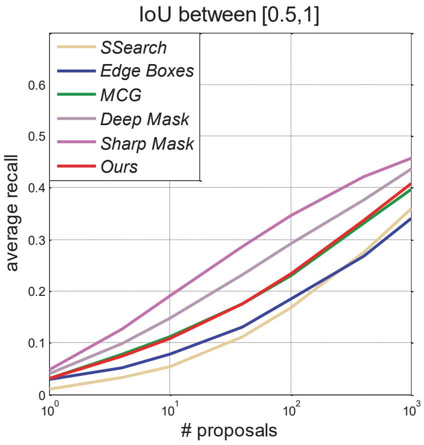


Figure 9. Comparison results of average recall for 1, 10, 100, 1000 proposals between typical hand-crafted and deep learning based methods on the MS COCO 2014 validation set.

4.3. Comparison with deep learning based approaches

Due to the powerful capability of feature extraction and well-designed structure of convolutional networks, recent deep learning based works like *Deep Mask* [20] and *Sharp Mask* [21] achieve excellent accuracy, and thus outperform hand-crafted methods. As shown in Fig.9, we compare the average recall (AR) between IoU 0.5 and 1 for 1, 10, 100, 1000 proposals on the MS COCO 2014 validation set, which is larger and more diverse. It is observed that our method still outperforms other hand-crafted ones, but *Deep*

	aero	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mean
CPMC	65.3	61.7	58.3	37.5	17.9	71.4	67.3	76.7	22.9	61.2	64.6	**70.3**	77.0	69.0	54.8	18.5	52.6	63.4	71.7	61.5	57.1
Geodesic	64.2	67.0	55.9	39.2	19.9	71.8	70.4	74.4	24.8	65.9	63.5	65.6	78.7	69.2	58.0	20.4	54.5	57.8	70.2	60.9	57.5
SSearch	**70.1**	67.1	61.5	42.5	21.4	68.3	68.7	76.4	27.6	65.7	**66.8**	70.0	75.5	68.9	57.9	25.6	53.6	63.7	**76.0**	**62.5**	59.6
MCG	66.4	69.3	60.3	42.3	28.5	71.3	72.3	**77.3**	30.1	61.4	62.4	69.8	77.4	68.2	62.2	27.4	**57.6**	**66.1**	75.8	59.4	60.3
Edge Boxes	67.1	69.8	59.7	**46.2**	28.3	72.9	72.3	73.9	28.7	68.1	62.4	67.6	**79.1**	**73.6**	62.4	28.3	55.8	61.2	70.4	59.7	60.4
Ours	68.1	**71.4**	**62.1**	45.7	**32.9**	**73.1**	**72.9**	76.1	**31.2**	**68.9**	62.8	67.9	79.0	72.9	**63.6**	**31.6**	56.9	61.7	70.7	59.8	**61.5**

Table 2. Fast R-CNN (model M) detection results (AP) on the PASCAL VOC 2007 dataset, where mean average precision is listed at the end. Bold numbers indicate the best proposal method per class. Our approach is better than other state-of-the-art methods for the majority of objects, and achieve the best mean performance of detection.

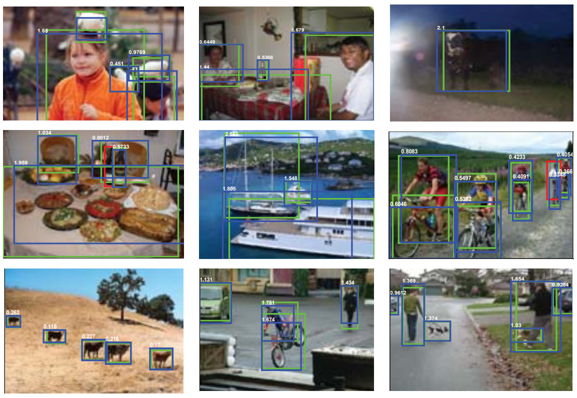


Figure 10. The results of our approach with 1000 proposals and IoU threshold of 0.7. Ground truth bounding boxes are shown in green and red, where red boxes indicate the objects are not found. Blue bounding boxes with their obtained scores nearby are the generated object proposals close to green ground truth boxes.

Mask and *Sharp Mask* perform better than hand-crafted methods, including ours. However, our method is entirely based on the unsupervised hand-crafted features and has comparable strengths. Firstly, it reflects a good tradeoff between recall and speed for object proposal, which takes less time than most deep learning based methods per image. Secondly, it does not require fully-labeled training images, and is easier to be generalized to work on other unlabeled data, compared with supervised deep learning methods.

4.4. Proposals for object detection

Proposals are commonly applied to object detection, whose precision is related to the average recall and localization accuracy of candidate boxes. Hence, we consider the well-known object detector, the Fast R-CNN [11]. After obtaining 2000 proposals with our approach, the Fast R-CNN is trained on the training and validation sets of the PASCAL VOC 2007 dataset, and then detect objects on the test set. For efficiency, proposals generated by different methods start from the same pre-trained VGG-M network [5]. Table 2 shows the per-class Fast R-CNN detection results of diverse approaches, as well as mean average precision (mAP). *Selective Search*, *MCG* and *Edge Boxes* achieve comparable accuracy results because their proposals are relatively high-quality. *Geodesic* and *CPMC* perform a little bit worse. In contrast, due to the improvements of the average recall and localization accuracy, our approach obtains the best results for the majority of objects and thus the highest mAP among these methods, which demonstrates that the high-quality proposals generated with our approach can be further utilized for effective object detection.

Finally, qualitative results of our proposal method are shown in Fig.10. Due to the introduced surface cue with improved occlusion edges, our approach almost discovers diverse objects effectively in various scenes, including the objects with large aspect ratio or small size, which are difficult to find only with the basic edge map.

5. Conclusion

This paper presents a novel object-level proposal model, where the occlusion-based objectness captures the surface cue reflecting abundant 3D characteristic of objects with occlusion edges. Specifically, to obtain the high-quality occlusion edge map for proposal, an optimization objective with Huber loss is first constructed to select proper samples for occlusion detection. Then an occlusion-dominated fusion is elaborately designed, with specific normalization criterion added, to further promote the occlusion edges and proposals. Experiments on the PASCAL VOC 2007 and MS COCO 2014 dataset demonstrate the superiority of our method over state-of-the-art methods, especially 6% improvement on the average recall at 1000 proposals than *Edge Boxes*.

6. Acknowledgement

This work was supported by the National Natural Science Foundation of China (NSFC) under grants U1435220 and 61503145, Young Elite Scientists Sponsorship Program by CAST (No. YESS20150077), the Fundamental Research Funds for the Central Universities under grants 2017RC53 and 2017RC52, and the National Key Technology Research and Development Program of the Ministry of Science and Technology of China under grant 2014BAK14B03.

References

[1] B. Alexe, T. Deselares, and V. Ferrari. Measuring the objectness of image windows. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2189–2202, 2012.

[2] P. Arbelaez, J. Pont-Tuset, J. Barron, F. Marqus, and J. Malik. Multiscale combinatorial grouping. In *CVPR*, pages 328–335, 2014.

[3] X. Bai, S. Bai, Z. Zhu, and L. Latecki. 3d shape matching via two layer coding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(12):2361–2373, 2015.

[4] J. Carreira and C. Sminchisescu. Cpmc: Automatic object segmentation using constrained parametric min-cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(7):1312–1328, 2012.

[5] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *BMVC*, 2014.

[6] M. Cheng, Z. Zhang, W. Lin, and P. Torr. Bing: Binarized normed gradients for objectness estimation at 300fps. In *CVPR*, pages 3286–3293, 2014.

[7] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002.

[8] J. Dai, K. He, and J. Sun. Convolutional feature masking for joint object and stuff segmentation. In *CVPR*, 2015.

[9] P. Dollar and C. L. Zitnick. Fast edge detection using structured forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(8):1558–1570, 2014.

[10] M. Everingham. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[11] R. Girshick. Fast r-cnn. In *ICCV*, pages 1440–1448, 2015.

[12] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1904–1916, 2015.

[13] J. Hosang, R. Benenson, P. Dollar, and B. Schiele. What makes for effective detection proposals? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(4):814–830, 2016.

[14] P. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

[15] C. Kading, A. Freytag, E. Rodner, P. Bodesheim, and J. Denzler. Active learning and discovery of object categories in the presence of unnameable instances. In *CVPR*, 2015.

[16] P. Krahenbuhl and V. Koltun. Geodesic object proposals. In *ECCV*, pages 725–739, 2014.

[17] H. Lee, A. Battle, R. Raina, and A. Y. Ng. Efficient sparse coding algorithms. In *NIPS*, 2006.

[18] T. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollar. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014.

[19] A. Ming, T. Wu, J. Ma, F. Sun, and Y. Zhou. Monocular depth-ordering reasoning with occlusion edge detection and couple layers inference. *IEEE Intelligent Systems*, 31(2):54–65, 2016.

[20] P. Pinheiro, R. Collobert, and P. Dollar. Learning to segment object candidates. In *NIPS*, 2015.

[21] P. Pinheiro, T. Lin, R. Collobert, and P. Dollar. Learning to refine object segments. In *ECCV*, 2016.

[22] W. Shen, X. Wang, Y. Wang, X. Bai, and Z. Zhang. Deepcontour: A deep convolutional feature learned by positive-sharing loss for contour detection. In *CVPR*, pages 3982–3991, 2015.

[23] R. Szeliski. *Computer Vision: Algorithms and Applications*. Springer-Verlag New York, 2011.

[24] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 104(2):154–171, 2013.

[25] R. Vaillant and O. D. Faugeras. Using extremal boundaries for 3-d object modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):157–173, 1992.

[26] N. Wang, J. Wang, and D. Yeung. Online robust non-negative dictionary learning for visual tracking. In *ICCV*, pages 657–664, 2013.

[27] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):210–227, 2009.

[28] Y. Zhang, X. Chen, J. Li, C. Wang, and C. Xia. Semantic object segmentation via detection in weakly labeled video. In *CVPR*, pages 3641–3649, 2015.

[29] Y. Zhu, R. Urtasun, R. Salakhutdinov, and S. Fidler. segdeepm: Exploiting segmentation and context in deep neural networks for object detection. In *CVPR*, pages 4703–4711, 2015.

[30] C. L. Zitnick and P. Dollar. Edge boxes: Locating object proposals from edges. In *ECCV*, pages 391–405, 2014.